Psychonomic Bulletin & Review
1999, 6 (3), 328-337

Asymmetries of comparison

CYNTHIA M. AGUILAR and DOUGLAS L. MEDIN
Northwestern University, Evanston, Illinois

Tversky's (1977) seminal work on directionality in judgments of similarity demonstrated that people may not judge the similarity of A to B to be equal to the similarity of B to A. In a series of studies, we investigated comparison asymmetries. In Experiment 1, our aim was to extend Tversky's findings to a heterogeneous stimulus set, but no reliable asymmetries were observed. Experiment 2 employed a variety of comparison judgments, and, although some of these measures showed asymmetries, we still failed to observe asymmetries in rated similarity. A final attempt to obtain asymmetries used direction as a within-subjects factor, and for the first time, rating asymmetries were observed. Our data reinforce the idea of comparison asymmetries but suggest that similarity rating asymmetries are only observed under quite circumscribed conditions.

Much of modern theorizing on similarity has focused on similarity *structure* and has used geometric scaling models that represent similarity relations as distances in some psychological space (see Schönemann, 1990; Shepard, 1987). The shorter the distance between two points, the more similar are the two items being compared. Multidimensional scaling (MDS) approaches have not, however, paid much attention to the processing side of similarity comparisons (see Krumhansl, 1978; Nosofsky, 1992, for notable exceptions). Tversky (1977) demonstrated the flexibility of comparison processes within the framework of a featural-theoretical approach to the analysis of similarity, known as the *contrast model.*

Tversky (1977) argued that metric models are not sufficient for representing similarity data and showed that similarity judgments violate axioms that must be satisfied by all distance models—most notably, for present purposes, symmetry. Symmetry is the assumption that the distance between two items is the same regardless of the direction of the comparison. However, people's judgments seem to show asymmetries. For example, Tversky reported that people rate the similarity of North Korea to Red China to be greater than the similarity of Red China to North Korea. In general, asymmetries of judgment raise serious problems for geometric models (but see Nosofsky, 1991, for a defense), but they are quite compatible with Tversky's model.

According to the contrast model, similarity is a weighted combination of (1) the number of features common to two objects (call them A and B), (2) the number of features distinctive to object A, and (3) the number of features distinctive to object B.

The contrast model can be expressed in the following equation:

$$S(A,B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

where the parameters θ, α, and β are weighting coefficients.[1] These weights may vary with the context and the judgment task. The f function measures the salience or prominence of the items. The contrast model distinguishes between the two terms A and B. According to the contrast model, the A term is the subject of the comparison, and the B term is the referent of the comparison. Such a directional comparison would be stated like this: How similar is (subject = A) to (referent = B)?

In the contrast model, judgment asymmetries are accounted for by the differential weighting of the distinctive features of stimuli being compared. Similarity is reduced more by the distinctive features of the subject (A term) than by the distinctive features of the referent (B term)—that is, $\alpha > \beta$ in the above equation. Therefore, if there are two stimuli and one (Y) is more prominent than the other (X), the similarity of X to Y will be greater than the similarity from Y to X. Consider again the comparison involving Red China and North Korea. Red China is more prominent (distinctive) than North Korea *and* its distinctive features receive more weight when it is the subject, rather than the referent, of a comparison. The combination of this difference in prominence and the weighting of the distinctive features of the subject and the referent implies that overall similarity will be reduced more when Red China is the subject and North Korea the referent than when their roles are reversed. In general, the contrast model

Preparation for this article was supported by the National Science Foundation Grant 92-11277. We thank Evan Heit for his continuous input and valuable advice throughout this project. We also thank Alexander Aminoff, Lance Rips, Peter Schönemann, Karen Solomon, Trish Van Zandt, Ed Wisniewski, and three anonymous reviewers for their comments and suggestions on earlier drafts of this paper. The members of our entire lab group (at both Northwestern University and the University of Michigan) deserve special thanks for their valuable advice, suggestions, and comments. Correspondence concerning this article should be addressed to C. M. Aguilar, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208 (e-mail: caguilar@merle.acns.nwu.edu).

implies that symmetry will only hold if the objects being compared are equally salient or if the judgment task is nondirectional.

Tversky and Gati (1978) provided converging evidence for their account of asymmetries. Twenty-one pairs of countries were used, and, in each pair, one element was more prominent than the other. Tversky and Gati employed two measures of prominence. In one task, the subjects were asked directly which one of the items in each pair was more prominent, and in the other task, the subjects were presented with each pair in two different orders and asked which order they preferred. These two measures of prominence had nearly perfect agreement and showed strong asymmetries. The data showed that the prominent item was preferred in the referent position.

In another task, a between-subjects design was used to assess the similarity of the same 21 pairs of countries described above. One group of subjects rated the similarity of the 21 pairs of countries in one direction (where the prominent item was in the subject position), whereas a different group of subjects rated similarity in the opposite direction. Tversky and Gati (1978) found that ratings for the pairs with the prominent item as the referent were significantly higher than those for pairs with the prominent item as the subject (e.g., North Korea [nonprominent] was judged to be more similar to China [prominent] than China was to North Korea).

In the present experiments, we sought to explore similarity asymmetries from the perspective of the contrast model (Tversky, 1977) as well as some more recent ideas concerning comparison asymmetries. To our surprise, we found ratings asymmetries to be considerably less robust than we had thought. This lack of asymmetries prompted us to aim to clarify the conditions under which comparison asymmetries are observed.

EXPERIMENT 1

Our goal in the first experiment was to replicate and extend Tversky's (1977) work on asymmetries of similarity by investigating the role of homogeneity of stimulus sets. Tversky's original research used stimuli that sampled a single domain (e.g., countries). This may have allowed subjects to focus on a small set of features (e.g., location, size) and to use these features for all of the judgments. These features may also have mediated the differences in prominence (preferred comparison order). On the other hand, if asymmetries are robust, they should appear whenever there are differences in prominence, regardless of the overall set of items being judged and regardless of whether the same small set of features can be used for all judgments.

In order to evaluate the role of homogeneity, both a stimulus set in which all the word pairs sampled a single domain (homogenous) and a set that sampled a variety of domains (heterogeneous) were used. The heterogeneous stimulus set included the same pairs as in the homogeneous condition, but these were intermixed with word pairs that sampled a variety of other domains (e.g., furniture, tools, toys). The homogeneous condition closely followed Tversky's (1977) original study. We employed many of the same countries that were used in the Tversky's study; however, some of the countries are no longer in existence, and pilot work suggested that Northwestern University undergraduates are not as knowledgeable of world geography as 1970's Hebrew University undergraduates. Hence, our list was slightly different from Tversky's.

In order to establish the expected direction of asymmetries, measures of the relative values of stimuli on the dimensions of interest were taken. Specifically, subjects made judgments about which of the countries was larger and more populous and which of the animals was bigger and more ferocious. Size and ferocity should be major determinants of prominence for the animal stimuli, given that these are the two dimensions that typically turn up in MDS solutions (e.g., Henley, 1969). The prediction is that, if an animal is judged to be bigger or more ferocious, it is the more prominent animal of the pair. For the countries, we expected that prominence would be based on size and population (or at least highly correlated with them), although we have no independent evidence supporting this conjecture. According to Tversky's (1977) contrast model, similarity should be greater when the prominent item of a pair is the referent, rather than the subject, of a comparison.

Measures of Prominence

Method

Subjects. The subjects from this study were 40 individuals on the Northwestern University campus who volunteered their time. The experimental session lasted approximately 5 min.

Materials and Design. The stimuli were 12 pairs of animals and 12 pairs of countries, where one item in each pair was more prominent than the other. Some of the word pairs were taken from the Tversky (1977) study, and others were chosen by the authors. The animal pairs were selected from an MDS representation of animals, where the dimensions were ferocity and size (Henley, 1969). The animal stimuli were chosen so that one of the members of the pair was more ferocious or bigger than the other. The country stimuli were selected in the same manner as the animal stimuli; however, an MDS scaling solution was not used.

To validate these expectations, the experimental condition had the subjects make a total of 24 choices. Domain was a within-subjects factor, so that all the subjects made choices about animals and countries. Within each domain, the subjects made judgments about some feature of each pair (for the animals, size or ferocity, and for the countries, population or size). The subjects made 12 choices about either the relative ferocity or the relative size of 12 pairs of animals and about either the relative size or the relative population of 12 pairs of countries. Overall, the subjects made 24 choices concerning two dimensions of two different domains (one dimension per domain).

The stimuli were presented in the context of a sentence. For example, for a stimulus from the domain of animals, where the frame of reference was ferocity, the sentence read: "Which is more ferocious, a dog or a cat?" All the stimuli and orders were randomized and counterbalanced.

Procedure. The subjects were given a three-sheet booklet. The first page listed instructions, which the subjects read and then proceeded with the task. Each of 12 pairs from a single domain was

listed on a single page. The pairs were counterbalanced so that each item of a pair appeared first in the sentence an equal number of times. The order of the pages were counterbalanced so that half the subjects rated the countries first and the other half rated the countries second.

Results

The measures of prominence showed strong differences for the animal and the country pairs. For the animal stimuli, where subjects judged which animal is more ferocious, a t test showed a significant difference from chance [.5; $t(11) = 4.88, p < .01$]. The size dimension of the animal domain also showed strong preferences and agreement across subjects. For all of the items, the subjects agreed on which animal was larger 75% or more of the time. A t test showed a significant difference from chance [.5; $t(11) = 15.42, p < .01$].

The country stimuli also showed strong agreement for both population and size. For population, a t test showed a significant difference from chance [.5; $t(11) = 6.97$, $p < .01$]. The size dimension showed even stronger consensus, where all but one pair showed a strong (75% or more) preference [the associated $t(11) = 9.94, p < .01$]. These observations set the stage for expected asymmetries in similarity ratings.

Ratings

As mentioned before, the goal of this experiment was to extend Tversky's (1977) work on similarity asymmetries by investigating the role of homogeneity in similarity comparisons. The predicted direction of asymmetries was derived from the prominence judgments. On the basis of the contrast model, the rated similarity when the prominent item is in the referent position should be greater than the similarity when the prominent item is in the subject position.

Method

Subjects. The subjects were 84 Northwestern University undergraduates, half of whom were from an introductory psychology class, who participated in this experiment in partial fulfillment of a course requirement, and the other half were paid for their participation. The subjects were run in groups of 1 to 6 students. The experimental session lasted approximately 10 min.

Materials and Design. The stimuli were the word pairs used in the prominence measure experiment above. In addition to the two domains used during pretesting, there were 12 other word pairs used in the heterogeneous condition. These pairs were chosen to sample a variety of domains and to be as heterogeneous as possible (e.g., toys, beverages, tools, etc.). These pairs were taken from the Battig and Montague (1969) category norms.

The experiment was a 2 × 2 between-subjects design, where the variables were condition (homogeneous or heterogeneous) and direction (prominent:nonprominent vs. nonprominent:prominent). The list of pairs was randomized with respect to order of comparison, and order was counterbalanced across subjects. In the homogeneous condition, the subjects rated the similarity of 12 pairs from the country domain, then rated 12 pairs from the animal domain. The domains were counterbalanced so that half of the subjects rated the animal domain first and the other half rated the animal domain second. In the heterogeneous condition, there were an additional 12 pairs that sampled a variety of domains. These pairs were intermixed with the 24 pairs from the homogeneous condition. All of these stimuli were completely randomized and presented on a computer screen.

Procedure. The subjects were run on Macintosh computers. Half of the subjects were presented with sentences of the form "How similar is X to Y?" whereas the other half were presented with the reverse form, "How similar is Y to X?" Each sentence was presented for 8 sec, and then the rating scale appeared. Once the rating scale appeared, the subjects clicked on the number that best assessed the similarity of the items presented. The rating scale ranged from 1 to 9, where 9 represented *maximal similarity*.

Results and Discussion

Our initial interest was in the role that homogeneity of the stimulus set might play in similarity asymmetries. But to our surprise, we did not replicate Tversky's (1977) judgmental asymmetries, even in the homogeneous condition. This occurred despite the fact that there were clear differences in measures of prominence for the animal and country stimuli. In addition, the pairs involving countries included a subset of those used earlier by Tversky, and we observed no asymmetries for these 5 pairs. Only those pairs that showed agreement on the two dimensions (7 for the animal pairs, 11 for the countries) were included in the analyses. (The analysis using the entire set also did not show any significant differences.)

The average similarity ratings across all the pairs for the animal condition when the stimuli were homogeneous were virtually identical. The mean similarity rating when the prominent animal was in the referent position agreed with the rating when the prominent animal was in the subject position, within rounding error (4.88). The means for the animal condition when the stimuli were heterogeneous (completely randomized) are also nearly the same (4.56 vs. 4.49).[2] The country stimuli also failed to yield the predicted pattern of asymmetries in either the homogeneous (5.48 vs. 5.39) or the heterogeneous condition (5.48 vs. 5.52). In short, no asymmetries were evident.

As a follow-up study, we tried what we thought was a stronger manipulation. Specifically, we ran a condition in which the dimension of interest was specified. For example, the subjects were asked "How similar is North Korea to China with respect to size?" The idea was that the feature set would be fixed and that differences in prominence along these dimensions would produce asymmetries. Specifically, comparing the alternative with the smaller value to the alternative with the larger value on the dimensions should have yielded higher ratings than reversed comparisons. But again, we did not observe asymmetries—the ratings were virtually identical across both directions of comparison.

Overall, we failed to find any asymmetries of judgment and failed to replicate Tversky's (1977) results. However, the measures of prominence used in this study, although similar, were not identical to those used by Tversky. In another attempt to produce asymmetries, the next experiment used the same measures of prominence taken by Tversky, as well as some additional ones.

EXPERIMENT 2

Experiment 2 used multiple measures of prominence. Many of the stimuli were taken from Medin, Goldstone, and Gentner's (1993) second experiment, because feature-listing data showed some properties consistent with asymmetries. A second aspect of this experiment was based on the contrast model's ideas concerning the flexible processing of similarity versus difference judgments. In the contrast model, more weight is placed on the common features when judging similarity, and more weight on the distinctive features when judging difference. This differential weighting makes the contrast model more flexible in handling nonequivalences between similarity and difference.

To test this idea concerning differential weighting, Tversky (1977) asked subjects to assess the similarity or difference of 20 sets of four countries; each set included a prominent pair and a nonprominent pair (e.g., East Germany–West Germany and Sri Lanka–Nepal, respectively). If the prominent pair has a greater number of both common and distinctive features than the other pair, it may be judged to be more similar and more different than a less prominent pair (i.e., one with fewer common and distinctive features). In support of this prediction, people judged prominent pairs, such as East Germany and West Germany, to be both more similar to and more different from each other than nonprominent pairs, such as Sri Lanka and Nepal.

In Experiment 2, it was also examined whether this presumed differential weighting is reflected in feature listing. A prediction consistent with the contrast model (but not demanded by it) is that a greater number of common features may be listed for similarity comparisons than for difference comparisons. In addition, a greater number of distinctive features might be listed for the first term of the comparison than for the second term of the comparison, because distinctive features are assumed to receive greater weight. (Again, this prediction would be consistent with the contrast model but is not required by it.)

Another goal of this experiment was to further explore the processing side of similarity, not only from the perspective of the contrast model, but also from that of other recent ideas concerning the comparison processes. In particular, Ortony (1979) has proposed that the salience or importance of a common feature may vary across concepts. Furthermore, Glucksberg and Keysar (1990) have argued that understanding comparisons may entail a focus on the referent, where properties of it become candidate properties for the subject (see, also, Clement & Gentner, 1991). In support of these ideas, Medin et al. (1993, Experiment 2) found that common properties of an A,B comparison were rated as being more closely associated with the B stimulus when A was compared with B and more closely associated with the A stimulus when B was compared with A. These observations raise the possibility that asymmetries arise from differential salience of common features. That is, the concept with the more prototypical common feature may be judged to be the more prominent item. For example, people may rate the similarity of North Korea to Red China to be greater than that of Red China to North Korea because the common feature, *communist country*, may be more salient or prototypical of China.

In short, Experiment 2 had three goals: One was another attempted replication of Tversky's (1977) rating asymmetries, on the basis of an assessment of converging measures of prominence and their relations to similarity and difference ratings. A second was to see whether differential weighting is reflected in feature listings. A final goal was to examine the questions of whether the focus of attention is on the referent or the subject term of the comparison and whether asymmetries arise out of distinctive features of the subject term or differential salience of common features.

Some groups of subjects were asked to make prominence judgments, other groups to make similarity and difference ratings, and, finally, a third group of subjects made both ratings and listed features to justify their ratings. The prominence judgments were used to make predictions about the direction of asymmetries in rated similarity and difference. The feature listings were used to investigate the role of differential salience.

Measures of Prominence

Method

Subjects. The subjects were 44 undergraduates from the University of Michigan, who were paid for their participation. The subjects were assigned to either the judgment condition ($n = 21$) or the preference condition ($n = 23$). The experimental sessions were conducted with groups of 2–5 subjects and lasted approximately 10 min.

Materials, Design, and Procedure. For all conditions, the stimuli used were 30 pairs of words. The majority of the word pairs (17 out of 30) were borrowed from Medin et al.'s (1993) second experiment. These pairs were chosen because they showed strong asymmetries in feature listings. The remaining word pairs were also chosen to sample a variety of domains and to show a difference in prominence.

In the judgment condition, the 30 word pairs were presented on a single page in a random order. The subjects read the instructions and chose the more prominent item in each of the 30 pairs.

For the preference condition, the stimuli were presented on a single page in two columns. One column displayed one order, and the other column displayed the reversed order. For example, in one column, the subjects would see the comparison "A brain is similar to a corporation," and in the other column, the subjects would see "A corporation is similar to a brain" (some of the comparisons were clearly metaphoric). The order of the comparisons was counterbalanced across subjects. The subjects indicated which direction of comparison seemed more natural to them.

Results

The results from the judged prominence measure showed reliable differences in prominence ratings for a strong majority of the pairs. On average, one item of a pair was selected as more prominent 70% of the time.[3] By a binomial test, a proportion of .71 or greater is reliable, and 18 of the 30 pairs met this criterion (compared with a

Table 1
Results From the Three Measures of Prominence-Direct Ratings, Preferred Similarity Order (s), and Preferred Difference Order (d)

p Prominent	q Nonprominent	Order Preference Π	Π_s	Π_d
kangaroo	rabbit	.86	.35	.61
surgeon	butcher	.86	.70	.56
US	England	.86	.30	.35
cherry	grape	.82	.48	.52
zebra	skunk	.82	.83	.39
Einstein	Franklin	.77	.39	.43
brain	computer	.77	.78	.48
elephant	gorilla	.77	.65	.61
chocolate bar	popcorn	.77	.35	.48
apple	prune	.77	.74	.30
brain	corporation	.72	.65	.48
stomach	box	.72	.39	.44
campfire	lantern	.72	.83	.56
wallet	purse	.72	.35	.30
doctor	engineer	.72	.65	.56
dog	cow	.72	.30	.26
house	tent	.72	.83	.83
brain	stomach	.72	.70	.78
ghost	shadow	.68	.43	.65
bicycle	skateboard	.68	.96	.70
orange	lemon	.64	.65	.43
shark	pitbull	.64	.83	.74
hats	earmuffs	.59	.91	.61
watermelon	football	.59	.65	.39
pencil	crayon	.59	.70	.61
English	Spanish	.59	.57	.61
porcupine	coconut	.59	.30	.30
car	blimp	.59	.70	.83
frisbee	boomerang	.50	.70	.83
onion	garlic	.50	.65	.56
Average		.70	.61	.54

chance expectation of 3 pairs). Our mean proportion is, however, considerably lower than the .92 proportion observed by Tversky and Gati (1978). The prominent member (p), based on this measure, of each pair is listed in the leftmost column of Table 1, and the second column contains the nonprominent item (q). The first numerical column gives the prominence proportions for each pair. These proportions are based on the proportion of people who chose p as the more prominent item in the pair. For example, in the zebra/skunk pair, .82 proportion chose *zebra* as the more prominent item.

The second measure of prominence, preferred comparison order in a similarity frame, also yielded reliable asymmetries. However, they were a bit weaker than those observed for direct ratings of prominence. The average deviation from .5 across the pairs was .20. A binomial test indicated statistically reliable differences ($p < .05$) for 8 of the 30 pairs. By chance, only 3 pairs should produce reliable differences.

The final measure of prominence, preferred comparison order in the frame of difference, yielded still weaker differences in prominence. The average deviation from .5 across the pairs was .14. A binomial test indicated statistically reliable differences for only 5 of the 30 pairs (again, chance expectation was that 3 pairs would be reliably different). The preference proportions for similarity and difference can be found in the last two columns, respectively, of Table 1. The proportions are represented in the direction favoring the prominent item (p term) in the referent position.

Using the first measure of prominence to fix the comparison order (see Table 1), the direction favoring the p term in the referent position was selected as more natural 61% of the time. For example, in the comparison zebra: skunk, *zebra* was the more prominent item of the pair; hence, the proportion .83 represents the proportion of people who chose the comparison "A skunk is similar to a zebra" as the more natural comparison.

Our data show far less agreement on measures of prominence than was observed by Tversky and Gati (1978). The correlation of preferred comparisons in the similarity frame with judged prominence was only $-.25$ (n.s.). These two measures were the ones used by Tversky and Gati. Thus, it is surprising that they did not show significant agreement. The correlation between preferred comparisons in the difference frame and judged prominence was marginally significant ($-.31$, $p = .09$). Note that the negative correlation means that people preferred to have the more prominent item as the referent rather than as the subject. The strongest correlation was found between

the two preferred comparison measures ($r = .52, p < .01$). This means that the same term tends to be preferred as the subject (or referent), regardless of whether the comparison involves similarity or difference.

We turn now to the question of whether these measures of prominence are accurate predictors of rating asymmetries in similarity and/or difference judgments. Equally important, we would like to know if these same prominence effects are present in other measures of similarity and difference (e.g., ratings and, possibly, feature listings).

Similarity and Difference Ratings

Method

Subjects. The subjects were 261 undergraduates from the University of Michigan, the majority of whom participated in this experiment in partial fulfillment of a course requirement, and some of whom were paid for their participation. The sessions were conducted with groups of 2–5 subjects. The sessions lasted approximately 35 min for the rating and feature-listing condition and approximately 10 min for the rating only condition.

Materials and Design. The stimuli were the same 30 pairs of words used in the prominence conditions. The word pairs were displayed in the following form: for similarity judgments, "How similar is A to B?" or, for difference judgments, "How different is A from B?"

The experiment was a 2 × 2 × 2 between-subjects design, where the factors were two conditions of a judgment task (subjects made either similarity judgments or difference judgments), two conditions of directionality (subjects received either a *pq* comparison or a *qp* comparison), and two conditions of the rating task (with or without feature listings). There were two orders of presentation for the stimuli—a random order and its reverse. If a word appeared in more than one pair (e.g., brain, stomach), at least one other item intervened between the pairs containing that word.

Procedure. There were two conditions of this experiment: the rating plus feature-listing condition and the rating only condition. In the rating plus feature-listing condition, each subject was presented with a booklet. The subjects read the instructions to themselves, and then the experimenter read the instructions aloud. The subjects did two things: (1) made a rating based on the similarity (or difference) between the pairs of words, using a 9-point scale (where 9 represented *maximal similarity* or *maximal difference*), and (2) listed the features/properties that came to mind as they made the ratings.

For each pair, the subjects always made their rating first and then listed features. The subjects were instructed to list both common and distinctive features and to specify which of the words in the pair they were describing. The groups were run in either a similarity condition or a difference condition; type of judgment was a between-subjects factor. All the subjects were given 1 min per item.

For the rating only condition, each subject was presented with the sheet of paper containing the 30 comparisons. The rating scale was the same as the one described above. All the subjects began at the same time and worked at their own pace to complete the task.

Results

Between-groups analyses. For purposes of an overall statistical test, difference ratings were converted to similarity ratings by subtracting each difference rating from 10. The analysis of variance revealed a marginal main effect of judgment [$F(1,248) = 3.26, MS_e = 53.33, p = .07$]. That is, there was a slight asymmetry between the ratings for judgments of similarity and those for judgments of difference, where similarity derived from similarity ratings was higher than that derived from difference ratings (mean similarity rating = 4.72, mean similarity score for difference ratings = 4.57). Table 2 presents the mean ratings for each of the 30 comparisons across judgment tasks (similarity and difference), directions (*p*/*q* and *q*/*p*), and conditions (with feature listings and without feature listings) as a function of prominence based on direct ratings of salience (the first prominence measure).[4]

Prominence and asymmetries. According to the contrast model, the main contributing factors to asymmetry are prominence differences and comparison direction. The less prominent item should be more similar to (and less different from) the prominent item than vice versa. The pairs for which both items were judged as being equally prominent are not included in the means. As noted before, some of the measures of prominence did not converge, and, therefore, the assessment of asymmetries was based on a combination of these measures (linear regression).

A regression analysis, with the three measures of prominence as predictors, was performed, with the dependent variable being the difference in rating between one order (*pq*) and the other (*qp*) for each of the four main comparisons. In all, four multiple regressions were performed. The results from these analyses show that the measures of prominence were not reliable predictors of rating asymmetries. The regression equation that used the measures of prominence to predict similarity ratings (with feature listing) produced an R^2 of only .03 [$F(3,26) = 0.31, p > .10$]. The same held true when these measures of prominence were used to predict similarity without feature listings. This regression equation produced an R^2 of .04 [$F(3,26) = 0.33, p > .10$]. The same lack of reliability held for asymmetries of difference judgments. Difference ratings with feature listings produced an R^2 of only .07 [$F(3,26) = 0.60, p > .10$], and for difference judgments without feature listings, the R^2 was .14 [$F(3,26) = 1.36, p > .10$]. In short, the three measures of prominence whether in isolation or in combination, failed to reliably predict asymmetries of either similarity or difference ratings. Rating asymmetries tended to be small and, at least with respect to our measures of prominence, unsystematic. We turn now to analyses based on the idea that differences in the salience of common features may mediate asymmetries of judgment.

Feature Listings Results

Common and distinctive features. The focus of the feature-listing analysis is to search for effects of direction, type of comparison, and differences in the salience of common and distinctive features within pairs of concepts (asymmetries). First of all, consider the idea, consistent with the contrast model, that similarity versus difference judgments lead to differential attention to common and distinctive features, respectively. One might expect more common features listed for judgments of similarity and more distinctive features listed for judgments of difference. Among the 22,940 features that were listed,

Table 2
Prominence Proportion by Direct Rating Measures and Average Similarities (s) and Differences (d) for Both Feature Listings and Rating Only Conditions

p prominent	q nonprominent	Ratings Plus Feature Listings				Ratings Only			
		$s(pq)$	$s(qp)$	$d(pq)$	$d(qp)$	$s(pq)$	$s(qp)$	$d(pq)$	$d(qp)$
brain	corporation	3.92	4.54	6.39	5.92	4.58	3.85	4.45	6.06
stomach	box	3.33	3.36	6.59	6.97	3.85	3.24	6.34	6.27
hats	earmuffs	6.03	6.05	4.08	3.82	5.33	5.39	4.19	4.30
watermelon	football	3.03	3.28	6.97	6.44	4.15	3.39	6.47	6.52
Einstein	Franklin	6.00	5.90	4.69	4.41	5.76	4.88	4.00	5.22
ghost	shadow	3.85	3.51	6.64	6.36	4.30	3.70	6.00	6.06
kangaroo	rabbit	5.97	4.87	5.00	4.87	4.39	4.76	4.59	5.45
surgeon	butcher	4.18	4.59	6.03	6.33	3.88	3.58	6.28	6.45
cherry	grape	6.72	6.54	3.90	3.54	5.21	5.45	3.66	4.12
campfire	lantern	5.95	5.31	4.74	4.56	5.06	5.42	6.52	5.03
brain	computer	6.00	6.28	4.85	4.77	6.76	5.52	3.91	4.56
elephant	gorilla	4.36	4.46	5.87	6.15	3.64	3.88	6.19	6.52
pencil	crayon	6.64	6.90	3.54	3.26	6.48	6.09	4.86	3.53
English	Spanish	5.36	5.28	4.64	5.49	4.36	4.73	5.16	5.12
orange	lemon	7.21	6.97	3.72	3.51	5.85	6.48	3.34	3.12
bicycle	skateboard	5.49	5.64	4.62	4.87	4.88	4.79	5.06	4.76
wallet	purse	7.08	7.26	3.54	3.44	6.42	6.61	3.70	3.34
zebra	skunk	4.49	4.05	5.59	5.10	3.85	3.97	6.15	6.03
chocolate bar	popcorn	3.79	4.26	5.64	6.10	3.76	3.15	6.48	6.22
doctor	engineer	5.72	5.21	5.64	5.26	4.61	4.97	5.45	5.66
porcupine	coconut	1.97	1.54	8.44	8.36	1.76	1.73	8.09	8.25
dog	cow	4.44	4.64	6.15	5.26	3.21	3.61	6.50	6.61
car	blimp	3.05	3.26	6.92	7.18	2.82	2.91	7.03	6.48
apple	prune	4.92	4.59	4.97	5.26	3.85	4.21	5.03	5.12
house	tent	5.51	5.46	5.67	5.79	4.64	5.09	5.19	5.30
US	England	5.95	6.13	5.28	5.10	4.88	5.58	4.78	4.94
shark	pitbull	4.56	4.72	5.85	6.05	4.18	3.88	5.70	6.25
brain	stomach	4.05	4.69	5.90	5.72	3.67	3.48	6.88	6.88
frisbee	boomerang	6.02	6.26	3.92	3.74	6.27	5.97	4.03	4.00
onion	garlic	6.72	6.38	3.38	3.44	5.70	5.85	3.66	3.58
Average		5.08	5.06	5.31	5.24	4.60	4.54	5.32	5.39

10,210 (45%) features were listed for similarity judgments, and 12,730 (55%) features were listed for difference judgments. For the features listed for similarity judgments, 3,301 (or 32%) were common features (68% were distinctive). For the difference judgments, only 2,431 (or 19%) were common. In brief, on this gross level, the relative proportions of common and distinctive features were consistent with the contrast model.

Distinctive features. As suggested before, it would be consistent with the focusing hypothesis of the contrast model if more distinctive features were listed for the subject term of the comparison than for the referent term. For similarity judgments, more distinctive features were listed for the subject term ($X_{subj} = 57.1$) than for the referent term ($X_{ref} = 51.1$). This held for 23 out of 29 pairs ($\chi^2 = 8.33, p < .01$). For difference judgments, the same pattern of results was observed: More distinctive features were listed for the subject term ($X_{subj} = 118.4$) than for the referent term ($X_{ref} = 108.3$). This held for 23 out of 28 pairs ($\chi^2 = 12.46, p < .01$). Thus far, it seems that the feature-listing results fit nicely with expectations generated from the contrast model.

There was also consistent agreement between the number of distinctive features listed and the preferred comparison order. For similarity judgments, there were 20 cases in which more distinctive features were listed for the preferred referent, 9 cases in which more distinctive features were listed for the preferred subject, and one tie ($Z = 1.92, p < .05$). For difference judgments, the opposite pattern appeared. There were 21 cases in which more distinctive features were listed for the preferred subject and only 9 cases in which more distinctive features were listed for the preferred referent ($Z = 2.48, p < .01$). However, these reliable comparison differences were not coupled with rating asymmetries.

Further predictions follow from a more detailed analysis of feature listings. More distinctive features should be listed for the more prominent member of pairs. For each of the four conditions (two directions × two judgments), the number of distinctive features was tallied. For example, in the comparison of zebra with skunk, the number of times that a feature distinctive to zebra (the prominent item of the pair) was mentioned was totaled across each of the four conditions (s(zebra:skunk), s(skunk:zebra), d(zebra:skunk), d(skunk:zebra)) and then compared to the number of times a feature distinctive to skunk was listed for each of the four conditions.

For similarity judgments, there were 18 cases in which the prominent item had more distinctive features listed, 9 cases in which the nonprominent item had more dis-

tinctive features, and one tie. For difference judgments, the corresponding cases were 17 and 11. Neither of these differences is statistically reliable. There are, however, consistent differences in distinctive features within a pair. For 22 out of 29 pairs, the member with more distinctive features for similarity also had more distinctive features for difference judgments ($p < .01$, by a binomial test).

The consistent difference in distinctive features listed for items of the pairs suggests yet another possible predictor of asymmetry. Are asymmetries of judgment systematically related to distinctive feature differences? The answer appears to be no for both similarity and difference judgments. Using the 22 pairs where the distinctive features agreed in both similarity and difference judgments, a t test found no significant differences in ratings [$t(21) = 0.535, p > .10$, for similarity ratings, and $t(21) = 0.735$, $p > .10$, for difference ratings]. The same pattern of results held for the ratings that did not ask for feature listings [$t(21) = 0.558, p > .10$, for similarity ratings, and $t(21) = -0.283, p > .10$, for difference ratings].

Common feature analysis. Our feature-listing data have shown that distinctive features are more closely associated with one item of the comparison (more distinctive features listed for the subject of the comparison). The next natural question is whether the same holds true for common features. Medin et al. (1993) found that common features were more closely associated with the referent term, rather than with the subject term. Overall, there was no reliable trend for biased common features to be preferentially associated with either the subject or the referent terms of comparisons. A biased common feature refers to features that are common to both items of the comparison but independently judged to be biased toward one of the concepts. For example, in comparing dogs and cows, people often list the feature *found on farms* as a common feature, but independent judges rated this feature as more true of cows than of dogs.

Directly judged prominence also did not predict the frequency of biased common features. There were, however, systematic patterns favoring one member of a pair over the other in the assignment or attribution of common features (as Ortony, 1979, and Glucksberg & Keysar, 1990, might predict). Regardless of the direction of the comparison, one of the concepts consistently had more common features listed as biased toward it than did the other concept. This pattern held for both similarity and difference judgments (similarity, sign test $Z = 5.33, p < .01$; difference, sign test $Z = 2.37, p < .05$). Furthermore, for 23 of the 28 pairs showing item differences in the number of biased common features within a pair, the difference was in the same direction for similarity and difference. These data suggest that common features may be more central or closely linked to one concept than to the other. Furthermore, this concept (the one with more biased common features listed) was also preferred as the referent of the comparison. This trend was evident for both similarity and difference comparison order and was statistically reliable for difference judgments (sign test $Z = 2.46$, $p < .01$). These data are consistent with the referent model.

Discussion

Although we employed the same measures of prominence as Tversky (1977), we again failed to observe rating asymmetries. An important difference between our data and Tversky's is that our judged prominence and comparison frame preference measures were only weakly correlated. Furthermore, a regression analysis found that these prominence measures did not reliably predict asymmetries in either similarity or difference ratings.

The feature-listing data do provide support both for the contrast model and for the idea that people give more weight to the referent of comparisons. More distinctive features were listed for the subjects of comparisons, and, for similarity, people preferred that concepts with more distinctive features be placed in the subject position (as would be expected in accordance with the contrast model). The idea that similarity comparisons begin with a focus on the referent also received some support. This perspective suggests that properties of the referent are evaluated with respect to the subject and may represent candidate inferences that people may be biased to adopt (Medin et al., 1993). Independent judges reliably rated common features as being biased toward one member of pairs. Furthermore, people reliably preferred to place the member with a greater number of biased common features in the referent position. This fits with the referent model. An issue for both the contrast model and the referent perspective is that these prominence effects were not translated into rating asymmetries.

Our null results on rating asymmetries do not appear to be easily dismissable as being due to a weak experimental manipulation. The problem with suggesting that subjects treated the judgment task as nondirectional is that both the feature-listing data and other measures of prominence show clear effects of directionality. That is, the preconditions necessary for observing judgment asymmetries were in place. As a final effort to produce rating asymmetries, in Experiment 3 we shifted from a between-subjects to a within-subjects design in which subjects made similarity (or difference) judgments across two directions of comparison.

EXPERIMENT 3

At this point, it appears that asymmetries of comparison can be reliably produced and linked to prominence judgments but that rating asymmetries are themselves much less robust. Our final study aimed to produce rating asymmetries by more strongly emphasizing the directionality of the comparison. In Experiment 3, direction was used as a within-subjects factor, so that all the subjects rated one direction and its reverse. We used stimuli from the previous experiment that showed strong agreement in judged prominence within a word pair as well as between judged

prominence and comparison frame preference. The general idea was that a within-subjects comparison of directions would serve to further highlight directionality.

Method

Subjects. The subjects from this study were recruited from the Chicagoland area. The subjects volunteered their time. The task took less than 5 min to complete.

Materials, Design, and Procedure. The stimuli used in this experiment were 12 word pairs from the previous experiment (Experiment 2) that showed strong agreement in prominence within a word pair as well as between two different measures of prominence (direct ratings of prominence and natural order comparison in a similarity frame). The subjects were approached and asked whether they would like to participate in a psychology study. They were asked to assess the similarity (or difference) of 12 word pairs (two directions for each pair). The similarity was rated on a scale from 1 to 9, where 9 represents *maximal similarity* (or *maximal difference*).

The word pairs were presented in a sentence form. For each comparison, one order was presented, and immediately following that comparison, its reversed order was presented. All the orders were counterbalanced and randomized. The subjects read the instructions and completed the task at their own pace.

Results and Discussion

The similarity ratings revealed a significant difference in ratings across directions. The contrast model predicts that similarity ratings will be greater when the prominent item is in the referent position, rather than the reversed direction. The data support this prediction [4.22 and 3.97; $t(11) = 2.30, p < .05$].[5] The predicted difference appeared in 9 of the 12 pairs.

The contrast model also predicts that difference asymmetries should follow with opposite signs. That is, the comparison with the prominent item in the subject position should be more different from the comparison with the prominent item in the referent position. The data also agree with this prediction; the difference rating when the prominent item was in the referent position ($M = 5.66$) was less than when the prominent item was in the subject position ($M = 5.84$). Although this difference is small, a t test shows it to be reliable [$t(11) = -2.57, p < .01$]. This difference held for 7 of the 12 pairs, and there were three ties.

In brief, the within-subjects manipulation was successful in producing rating asymmetries for both similarity and difference comparisons. This pattern held for a considerable majority of the pairs but was far from unanimous. On the other hand, Tversky's (1977) original finding of asymmetries in similarity ratings only held for 15 of the 21 pairs, which is essentially the same proportion as we observed (9 of 12). We now turn to the implications of our findings as a whole.

GENERAL DISCUSSION

Although most of our findings were couched in terms of nonreplication of rating asymmetries, we believe that a closer look at the context of these findings is considerably more informative. First, consider some positive findings. People do show consistent patterns of preferred comparison order that are correlated with feature-listing measures. For example, distinctive features are more accessible for the subjects than are the referents of comparisons. In addition, certain *common* features tend to be more reliably associated with one member of a pair than with the other, and people prefer to place items with a greater number of biased common features in the referent position of comparisons. These feature-listing and preference effects are not, however, associated with rating asymmetries. That is, the measures of prominence appear to be more robust than the presumed consequences of prominence differences.

Other aspects of our data provide some support for more recent ideas of the processes of similarity. Both Ortony (1979) and Glucksberg and Keysar (1990) have suggested that features shared by two concepts may be more closely associated with one concept than with the other. Medin et al. (1993) raised the additional possibility that similarity comparisons are biased toward the referent (properties of the referent may be evaluated with respect to the subject) and that, therefore, common features may tend to be more closely associated with the referent than with the subject of comparisons.

The above view requires two factors to produce asymmetries of ratings. One is differential salience of common features, and the other is that the particular common features entering into a comparison not vary substantially as a function of the direction of the comparison. North Korea will be more similar to Red China than Red China is to North Korea if the most important common feature remains *communism* for each comparison. If the comparison of Red China with North Korea shifts to a common feature salient for North Korea (e.g., hostility toward the United States), ratings may not show any asymmetry.

Medin et al. (1993) did report a shift favoring common features salient for the referent term, but they did not look for main effects of individual items. The present results do show this main effect and are consistent with the notion of biased common features. We also find that people prefer to place items with a greater number of biased common features into the referent position of comparisons. The relative weakness of rating asymmetries would have to be explained after the fact by the claim that the features entering into the comparison shifted as a function of comparison direction. A problem for this position is that asymmetries should have appeared when the dimension or feature (e.g., size) was explicitly mentioned in the comparison statement, but they did not. Again, the weak link is between notions of prominence and ratings.

Now let us return to the contrast model. In some respects, its successes are complementary to those of the referent model. First of all, people listed common features more often when making similarity judgments than when making difference judgments. More important, our feature-listing data were consistent with differential attention to the subject term, in that more distinctive features were listed for subjects than for referents of comparisons. This finding should have set the stage for rating asym-

metries, but we failed to observe them when comparison order was varied between subjects. Finally, the contrast model correctly predicts the direction of asymmetries that we did observe in Experiment 3, where comparison order was varied in a within-subjects design.

Another finding was that the different measures of prominence were not strongly convergent. Although Tversky (1977) took two measures of prominence and showed virtually unanimous agreement, identical measures of prominence in our study showed very little overlap, and their correlation was nonsignificant.

Our data suggest that rating asymmetries are generally quite weak and only evidenced under circumscribed conditions. They are considerably less robust than we would have thought at the beginning of this line of work. The only evidence we have seen that would suggest otherwise is a recent experiment by Catrambone, Beike, and Neidenthal (1996). Their study used countries varying in their familiarity as stimuli. In the main condition of interest, similarity statements were directional (where the subject and the referent were made explicit), much as in the Tversky (1977) study and in our studies. They found small, but reliable, asymmetries, with similarity being higher when an unfamiliar country was compared with a familiar country (consistent with Tversky's model).

Catrambone et al. (1996) used a procedure in which direction of comparison was varied across subjects. That is, half of the subjects saw familiar countries compared with unfamiliar countries, and half were given comparisons in the opposite direction. In our work, we had always varied comparison direction (across pairs) within subjects, and a conjecture is that their procedure made the asymmetries more salient. To test this idea, we ran a replication of the Catrambone et al. experiment, using their stimuli and with both their procedure and ours (for different subjects) and with groups sizes comparable with those of Catrambone et al. Unfortunately, we found no reliable asymmetries in either condition. For both procedures, the means were in the right direction, but the difference favoring unfamiliar to familiar was less than a 10th of a rating point. We have no speculations as to why our exact replication failed to yield reliable asymmetries (they are not even reliable if we collapse across conditions to produce twice the number of subjects run by Catrambone et al.). Our nonreplication does reinforce the idea that rating asymmetries are very sensitive and can only be produced under circumscribed conditions.

We should emphasize that there is no particular reason to focus on rating asymmetries as the sole or even the privileged measure of asymmetries. The feature-listing data, ratings of biased common features, and preferred comparison order represent measures of directionality in similarity and difference comparisons that models of similarity must address. If similarity or difference ratings constitute a less sensitive index of comparison asymmetries, perhaps attention should shift to other measures of directionality. As we have seen, although these measures did not strongly converge, they did provide evidence consistent with both the contrast model and the referent model. Overall, our findings pose further challenges with respect to a unitary notion of salience or prominence. Salience appears be a multifaceted construct that we are only beginning to understand.

REFERENCES

BATTIG, W. F., & MONTAGUE, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80** (3, Pt. 2), 1-46.

CATRAMBONE, R., BEIKE, D., & NEIDENTHAL, P. (1996). Is the self-concept a habitual referent on judgments of similarity? *Psychological Science*, **7**, 158-163.

CLEMENT, C. A., & GENTNER, D. (1991). Systematicity as a selectional constraint in analogical mapping. *Cognitive Science*, **15**, 89-132.

GLUCKSBERG, S., & KEYSAR, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, **97**, 3-18.

HENLEY, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning & Verbal Behavior*, **8**, 176-184.

KRUMHANSL, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, **85**, 445-463.

MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

NOSOFSKY, R. M. (1991). Stimulus bias, asymmetric similarity and classification. *Cognitive Psychology*, **23**, 94-140.

NOSOFSKY, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, **43**, 25-53.

ORTONY, A. (1979). Beyond literal similarity. *Psychological Review*, **86**, 161-180.

SCHÖNEMANN, P. H. (1990). Psychophysical maps for rectangles. In H.-G. Geissler (Ed.), *Psychophysical explorations of mental structures* (pp. 149-164). Toronto: Hogrefe & Huber.

SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

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

TVERSKY, A., & GATI, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 79-98). Hillsdale, NJ: Erlbaum.

NOTES

1. Technically speaking, the above equation only represented a special case of the contrast model. Tversky's (1977) more general model allows three distinct mappings of the three feature sets into real numbers.

2. The means are presented so that the comparison with the prominent item in the referent position is the first mean and the comparison with the prominent item in the subject position is the second reported mean.

3. .72 or greater was used because, in the Tversky, 1977, paper, this is the prominence proportion he obtained for most of his comparisons.

4. There was also no effect of feature listings on ratings [$F(1,248) = 0.31$, $MS_e = 5.10$, $p > .1$].

5. There was a small number of people who gave the same ratings for both directions of comparison. There were 3 of these in each condition (similarity and difference). The analyses done without these subjects were still significant.

(Manuscript received September 8, 1997;
revision accepted for publication April 13, 1998.)