The Impact of Category Knowledge on the Similarity of Instances

John P. Clapper and Gordon H. Bower
Stanford University

ADDRESS ALL CORRESPONDENCE TO:
John P. Clapper
Department of Psychology, Bldg. 420
Stanford University
Stanford, CA 94305

Running Head: CATEGORY KNOWLEDGE AND SIMILARITY OF INSTANCES
Abstract

It was hypothesized that similarity judgments between instances of a category would be influenced by subjects' prior norms for that category. Specifically, features that appear unusual or unexpected with respect to the category should be weighted heavily in subjects' judgments, while predictable features (category defaults) should tend to be ignored. In Experiment 1, we found that unusual properties (e.g., the unexpected absence of a default feature) affected similarity more than routinely encountered properties when they occurred as a difference between the stimuli, but not when they occurred as a common feature of the pair. In Experiment 2, subjects rated the similarity of instances from memory. Here, the surprising absence of an expected feature from both members of a pair increased their similarity, as originally predicted. The results are interpreted as evidence that the common or distinctive features of a stimulus pair are evaluated in terms of category-level expectations.
The Impact of Category Knowledge

on the Similarity of Instances

Similarity has long been considered a fundamental principle in the study of cognition. It has often been cited as the underlying basis for much of categorization, concept learning, generalization, and other forms of inductive learning. The similarity of a pair of stimuli can be described in set-theoretic terms as a function of their common and distinctive features weighted by their importance (Tversky, 1977). But the value of such a formulation requires some knowledge about how people decide what the features of an object are and which ones are most important or relevant for a given comparison. The simple view, that similarity is determined by an overall match between all features of the objects compared, fails to capture cases in which stimuli are considered highly similar due to sharing one or a few relatively subtle but critical properties, while much differentiating information about them is ignored as irrelevant. We will argue that the features considered relevant to a given comparison are strongly influenced by the subject’s background knowledge about the set of stimuli being compared.

How do people use their knowledge about a domain to decide which properties are most informative or relevant in a particular comparison? One important principle is suggested by people’s tendency to focus selectively on properties that seem surprising or unusual for objects of a given type (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986; Sokolov, 1976). We hypothesize that people employ an unusualness heuristic, by which the features of two objects being compared are weighted in terms of how unusual, unexpected, or surprising the features are with respect to an implicit reference category.
that contains both stimuli as instances (Garner, 1974; Kahneman & Miller, 1986). We assume that some such category is automatically evoked by the presentation of any pair of stimuli to be compared. This evoked category provides a frame of reference within which the instances will be evaluated and compared. The reference category presumably specifies relevant attributes (dimensions, roles, or "slots" of frames) according to which the stimuli can be characterized as matching or mismatching; the reference category also supplies norms for the values expected for each attribute. These norms determine whether the current instances are considered to have normal or unusual values on the relevant attributes. According to the unusualness heuristic, the lower the expected probability that a feature shared by the instances would have occurred by chance in a randomly-selected pair of instances from the reference category, the greater will be its importance in increasing their perceived similarity. A similar information analysis is applied to subjects' weighting of mismatching features of the members of the pair; the more unlikely a given difference is to have occurred by chance in the category, the more surprising it will appear and the greater will be its importance, thus greatly reducing the perceived similarity of the two instances.

An implication of this framework is that features shared by most typical instances (i.e., highly expected properties such as wings for birds) will be largely ignored when people judge instances of the same category; that is, shared defaults will do little to increase the similarity of specific pairs. Such expected defaults should be taken for granted by the subject and treated as a "background" within which most comparisons take place.
The rationale underlying the unusualness principle is that objects in the world do not usually arise from random combinations of independent properties; rather, the "generation process" follows principles of a natural, lawful order. In an orderly world, objects have correlated properties (e.g., similar animal species have close genetic ancestors). Therefore, when we notice a highly improbable feature that two objects share, it is a good bet that this is diagnostic of their being related in other significant respects, such as sharing a common generating cause. Biological taxonomists and other scientists involved in constructing classification schemes often seem to follow this heuristic (Mayr, 1982). For instance, paleontologists construct taxonomies and family groupings from the sketchy and often ambiguous fossil record left by plant and animal species from the remote past. If two such species share some unusual structure not found in others of the same general type, then paleontologists take this as evidence for the species' close evolutionary relatedness. Thus, certain predatory dinosaurs are thought to be close relatives of modern birds due to subtle skeletal properties uniquely shared by birds and these dinosaurs, despite the lack of resemblance in their overall appearance (e.g., Bakker, 1986). Analogously, in criminal investigations, crimes may be attributed to the same individual because of some distinctive feature ("MO") or "trademark" shared by cases that might differ in other respects, e.g., a murderer who kills by an unusual method or leaves a distinctive mark on the victims. Importantly, the credibility of such attributions seems related to the novelty or unusualness of the shared feature, i.e., the improbability that the feature would have occurred independently in two separate instances merely by chance.
Besides dominating comparisons between explicitly-presented instances, noticing an unusual property in an object or situation may often remind us of a previous instance in which the same unusual property was also present. Schank (1982) has proposed that such spontaneous *reminders* often result from violations of default expectations shared by two episodes. For example, encountering waiters on roller skates while dining at a restaurant would probably remind us of another restaurant that shared this exceptional property. By contrast, people are less likely to be reminded of a specific episode by routine or default features of an event, perhaps because such features, being associated with many episodes in memory, provide poor retrieval cues for any particular one. Such reminders may be a key factor in learning new categories and routine scripts, and in peoples' ability to recall and use previously encountered examples in solving novel problems (see, e.g., Ross, 1984; Schank, 1982).

The goal of our experiments is to demonstrate that a feature's unusualness or improbability within a reference category increases its impact on judgments of pairwise similarity. We will use the framework of Tversky's (1977) contrast model, wherein the similarity of two objects increases with their common (shared) features and decreases with their distinctive features (differences). Features may vary widely in their prominence or *salience*, and thus in their impact on similarity judgments. Tversky noted that a feature's salience could be influenced by both *intensive* (i.e., physical/perceptual salience) and *diagnostic* (partitioning utility) factors. Our notion of unusualness is related to Tversky's use of diagnosticity, since both relate to subjects' background knowledge of categories. However, diagnosticity is concerned with how the immediate context of a stimulus pair affects the partitioning of instances into the same or different
categories, which in turn affects their similarity. For example, the pair <Austria - Hungary> will quite appear similar when presented in the set <Austria - Hungary - Sweden - Norway>, because people will naturally group the countries by geographical location (Scandinavian versus central European), and Austria and Hungary are assigned to the same category. By contrast, the same pair appears much less similar when presented in the set <Austria - Hungary - Sweden - Poland>, because people divide this set into communist versus free-world countries, and Austria and Hungary now appear to be in different categories (see Tversky, 1977). Such context effects on subjects' classifications will be distinguished from the unusualness heuristic, since the latter presumably affects the weighting of features posterior to the objects' categorization. In other words, the unusualness heuristic describes the criteria by which a particular reference category determines feature salience; it does not describe how the immediate context determines which reference category is evoked in the first place.

We predict that objects that share an unusual value of a given attribute will be rated more similar than an otherwise-equivalent pair of objects that share a more familiar or expected value. In some cases, the unusualness heuristic may cause subjects to assign similarity ratings in a manner contrary to the results ordinarily obtained in similarity experiments. For example, previous research has shown that adding a common feature to a pair of objects increases their similarity, compared to an equivalent pair in which this feature is absent from both instances (e.g., Gati & Tversky, 1984). In such experiments, the target feature was present and absent equally often in the stimulus sets. We predict that the opposite result -- an increase in similarity due to removing the same feature from two objects -- could be obtained in situations in which subjects had learned to expect the
feature to be present in all typical instances of the reference category. When a feature's presence is the default for instances within a given category, then removing this feature from a specific instance should cause subjects to notice this absence and consider it a highly salient property of that instance. (For example, not being able to fly is a salient property of ostriches and emus in the class of birds). When this absence is common to both members of a pair, then this shared anomaly should strongly increase their perceived similarity, making them more similar than when the less-salient default is present. This prediction arises because in our approach "common features" are defined relative to subjects' expectations about the stimuli rather than relative to physical manipulations of them.

By a similar line of reasoning, we can predict that if one stimulus of a pair violates default expectations whereas the other does not, then that difference should be unusually salient. Because exceptional values are more salient than routinely-encountered ones, they should have a larger impact as distinctive features than would other values of that attribute that are within the expected range of variation. If two instances mismatch in a familiar way, such a difference would be regarded as ordinary for members of that category. Thus, the perceived difference between the two cases should be less than if one instance displayed a default value while the other instance violated this default. For example, a white horse would probably seem less similar to a green than to a brown horse, because green is an unusual color for horses whereas brown and white are normal. In the white-green comparison, the unusualness or "surprise value" of the color difference should increase its salience over the white-brown comparison, causing a greater reduction in rated similarity. These predictions about the
impact of unusual absence of a default feature were tested in Experiment 1.

Experiment 1

In this experiment, subjects rated the similarity of pairs of realistic drawings of fictitious insects, all of which shared some consistently occurring features (defaults) and differed along several variable attributes. After observing several pairs of bugs, subjects were expected to learn the norms of the stimuli, specifying the default versus variable features. In the midst of this uniform training series, occasional stimulus pairs were presented in which one or both insects violated the category expectations; the exceptional bugs would either (1) lack an expected default, or (2) possess an extra feature not usually present in the other instances.

Because of their unusualness, pairs sharing a missing default should appear more similar than otherwise equivalent pairs having this default present. Also, adding an unexpected component to both instances should increase their similarity more than should adding a variable feature, because an unfamiliar component should appear more unusual (hence, informative) than a familiar component. In addition, surprising differences between instances (i.e., when one has the expected value and the other has an exceptional value) should reduce their similarity more than a corresponding difference between two routine values of a variable attribute.
Method

Subjects

Twenty-two undergraduate students of Stanford University participated in this experiment to fulfill a service requirement for their Introductory Psychology course.

Procedure

Groups of one to three subjects were tested for a single session lasting approximately 50 minutes. The stimuli were presented in booklets; one pair of insects appeared side-by-side on each page. The similarity ratings were made on a 20-point scale, from 1 (very low similarity) to 20 (very high similarity). The instructions suggested that subjects think of the insects as a collection of newly-discovered species from a remote island, and themselves as biologists assessing similarities to help build classification trees, trace evolutionary patterns, and so on. Subjects spent 12 seconds rating each pair; their pace was controlled by an automatic beeper which indicated when they should turn the page to the next pair. The experiment comprised a total of 220 trials; subjects were given three one-minute rest periods, following trials 60, 120, and 180.

Materials and Design

The stimuli were realistic line drawings of fictitious insect-like creatures (see Figure 1).
The insects all had a common "base" structure (body, legs, eyes, and so on), and varied along nine binary attributes. Six of these were *additive* attributes, defined by the presence or absence of a particular feature (e.g., wings); the remaining three were *substitutive* attributes, defined by the presence of one of two alternate values (e.g., white or black eyes).

*Training Stimuli.* For four of the six additive attributes, a particular value occurred in all the training instances, while the other value never occurred; these will be referred to as *consistent* attributes. Of these, two attributes were consistently absent while the other two were consistently present. For example, all the bugs in one set had wings and antennae but lacked the pincer mouthparts and tails that were characteristic of some insects in the sets seen by other subjects. The remaining two additive attributes and the three substitutive attributes had two values that occurred equally often in the stimulus patterns; these will be referred to as *variable* attributes. Thus, the instances in each stimulus set varied in five two-valued attributes, defining 32 distinct instances per set. Figure 1 depicts instances from two different sets, illustrating their within-groups consistencies and differences.

*Exceptional Test Stimuli.* These stimuli differed from the training stimuli by having an exceptional value on one of their consistent attributes. When the presence of a given component is the norm and this component is deleted from a particular stimulus, it
will be called an *absent* feature. When absence is the norm and the component is present, we will refer to it as an *added* feature. A particular deviant instance contained only a single exceptional value. Within the stimulus set seen by each group of subjects, two features were consistently present in instances and two others were consistently absent. For each of these consistently present or absent features, fifteen exceptional instances were constructed in which that default expectation was violated, i.e., in which the typically-present feature would be absent or the typically-absent feature present.

*Counterbalancing.* Three stimulus sets were constructed, each of which was presented to a different group of subjects. The stimulus sets differed in which particular features were consistently present, absent, or variable. Within each set, features were randomly assigned to these three norm conditions, except that it was required that each additive feature be consistently present in one set, consistently absent in another, and variable in the third.

The training trials were constructed by randomly sampling (without replacement) pairs from the 496 possible pairs formed from each 32-instance stimulus set, proceeding until 140 training trials were specified. The 40 test trials were constructed in the same way, except that each pair was modified by substituting an exceptional value for a default in one or both instances. Four types of test trials were defined according to the type of exceptional value (absent versus added) and whether it was a common or distinctive feature of the instances. One trial from each of these four conditions was presented in each block of twenty trials. Within a given block, each of the four test trials presented an exceptional value of a different consistent feature.
Four yoked control trials were also presented in each block, one for each test trial. Each yoked control trial consisted of the same stimuli as a specific, corresponding test trial, except that the unusual aspect presented in the test trial was eliminated in the yoked control trial. By equating the test and yoked control trials in all respects except for the critical violations, we could directly assess how different exceptional features influenced subjects’ similarity ratings.

The order of trials was randomized with the constraints that (1) test trials were separated from one another by at least two intervening non-test trials, and (2) a test trial was separated from its yoked control trial by at least one intervening trial. The first twenty trials comprised a "practice block," and contained no test trials.

*Results and Discussion*

The effects of unusualness on the impact of distinctive features (differences) were consistent with our original hypotheses. To assess this effect, we subtracted the similarity of pairs which mismatched on a particular attribute from equivalent pairs which matched on that attribute. The resulting difference score provided an estimate of the impact of that particular distinctive feature (see Table 1). The reliability of each difference was evaluated with a matched-samples t-test.

----------------------------------

Insert Table 1 about here

----------------------------------
As expected, adding distinctive features reduced similarity in all conditions. A baseline distinctive-feature effect was estimated by subtracting (1) the similarity of normal pairs in which one member had a particular variable feature that was lacking in the other, from (2) the similarity of other normal pairs in which that variable feature was lacking in both members. The average reduction in similarity in this condition was approximately 1.42 points on a 20-point scale, a highly significant effect, $t(21) = 10.33, p < .001$. The effect of unusual distinctive features was expected to be larger than this baseline effect. Consistent with this prediction, the reduction in similarity caused by deleting an expected default from one member of a pair averaged approximately 2.60 points, $t(21) = 7.20, p < .001$. This (2.60) effect was almost twice the size of the baseline effect (1.42) of adding a distinctive variable feature, a statistically significant difference, $t(21) = 3.20, p < .01$. The effects of adding a novel distinctive feature to one pair member were also large, reducing similarity by an average of 3.02 points ($t(21) = 8.85, p < .001$). This effect was over twice that of adding a variable distinctive feature, $t(21) = 4.30, p < .001$; the effect did not differ significantly in magnitude from the absent-default effect, $t(21) = 1.03, p > 0.50$.

Turning to the results for common features, these effects were measured similarly to those for distinctive features, viz., by subtracting the similarity of pairs in which the common component was absent from the similarity of equivalent pairs in which that component was present. The resulting difference estimates the effect of adding that component as a common feature (see Table 1). The reliability of each such difference was tested with a matched-samples t-test.
In typical similarity experiments, adding a common feature to a pair of stimuli usually increases their similarity. Our data showed only a small effect of adding a common (variable) component to both pair members (an average of 0.22 points, $t(21) = 2.84, p < .01$). Thus, in this experiment common features had a much smaller effect in increasing subjects' similarity ratings than did distinctive features in decreasing them. In fact, the decrease in similarity due to adding a distinctive component to one pair member was over six times as large as the increase in similarity due to adding a common component to both stimuli (1.42 versus 0.22 points, $t(21) = 10.80, p < .001$). While adding common features had a significant effect, this last comparison shows that subjects' similarity judgements were dominated by differences between the stimuli rather than by their shared features.

Contrary to our predictions, removing the same default component from two instances did not increase their similarity relative to a yoked control in which the default was present in both pair members. In fact, deleting a common default reduced similarity slightly (by 0.07 points), although this effect did not approach statistical significance, $t(21) = 0.48, p > .50$. A related failed prediction was that the surprise value which should have accrued by adding a novel common feature did not increase similarity any more than did adding a common variable feature. Adding such novel components did increase similarity (by 0.49 points, $t(21) = 2.96, p < .01$), but the size of this effect was not reliably greater than that due to adding a variable common feature (which was 0.22 points), $t(21) = 1.49, p > .20$. However, the common feature effect in this condition was marginally larger than in the absent default condition (a difference of about 0.43 points, $t(21) = 1.82, p < .05$, one-tailed). This indicates that subjects' expectations may have
somewhat affected their weighting of common features. Overall, our manipulations of common features affected similarity in the expected direction, but they did not do so at the expected magnitude.

In summary, surprise occurring on distinctive features seemed to affect perceived similarity much more strongly than when it occurred on common features. In interpreting the results, it is important to recall that the baseline effect of adding a common feature was far smaller in this experiment than was the effect of adding a distinctive feature. Thus, common features played a much smaller role in our subjects’ judgement strategies than did distinctive features. Although subjective reports were not systematically collected in this study, the subjects we interviewed at the end of their sessions almost always reported ignoring common attributes and merely counting differences, arriving at numerical ratings based only on these differences. Unlike more wholistic comparisons, such a difference-counting strategy would be insensitive to common features regardless of their information value.

Given these extenuating arguments, then, the general pattern of results are not inconsistent with our original framework: the predicted results were obtained for distinctive features and we can rationalize the weak common feature effects in a reasonable way. Nonetheless, it is disappointing that the predicted results were not obtained for common features. Therefore, a second experiment was designed to pursue these common feature effects.

Experiment 2
We suspected that many subjects in the previous experiment may have used a "differences-only" strategy for computing the similarity of instances. Those we interviewed said they did so. We hypothesized that more wholistic or intuitive judgments would be induced by asking subjects to rate the similarity of instances based on their memory of them (i.e., using labels for memorized instances). Since people's learning of, and memory for, instances may well be dominated by their unusual features within the set, the memory-based similarity judgments may show greater sensitivity to the unusualness of the instances' features. Therefore, when two instances share an unusual feature, their memories may seem far more similar than was revealed by the differences-only ratings of presented stimulus pairs in Experiment 1.

Experiment 2 consisted of two phases. First, subjects learned to associate a specific label (a CVC nonsense syllable) with each of ten instances from a single category of insects. Within this stimulus set, eight of the instances possessed a target default value that was absent from the remaining two. Next, subjects rated the similarity of specific pairs of instances learned earlier, referred to only by their CVC labels. One of these pairs referred to the two deviant instances that were lacking the target default. We expected to find that this pair would be rated more similar, on average, than a yoked control pair in which the default was present in both instances.
Method

Subjects

The subjects were seventeen students of Stanford University, who participated to fulfill a service requirement for their Introductory Psychology course.

Materials and Design

As in Experiment 1, the stimuli in this experiment were realistic line drawings of a single category of fictitious insects. These insects consisted of a common "base" structure (i.e., body, legs head, etc.), and varied along eight binary attributes (e.g., different types of wings, antennae, and so on). Here, all eight attributes of the stimuli were substitutive, i.e., defined by the presence of one of two alternative values. Four different stimulus sets were constructed; each was presented to a different group of subjects. Within each set, the instances varied along three of the eight possible substitutive attributes, while the other five maintained a single, consistent value across all the instances. Eight such "normal" training instances were constructed for each group of subjects by using all possible combinations of values of the three variable attributes.

In addition to these normal training instances, two "deviant" instances were also constructed for each group. The deviants were each identical to one of the normal instances, except that they lacked a specific default feature that was present in the normal instances; within a given stimulus set, the same default was missing from both deviants. To illustrate, in one set the eight normal insects were all characterized by wings of a
consistent appearance; the deviants in this set, by contrast, had no wings. The two deviants within each group were constructed to be as dissimilar as possible, that is, they had contrasting values on all three of their variable attributes. Combining normal and deviant instances, a total of ten different stimuli were presented to each group of subjects.

Four of the eight attributes were used as "target" (additive) features, a different one for each group. The four stimulus sets also differed in which specific attributes were variable across instances, and which had consistent default values. Within each set, it was randomly decided which attributes would be variable or consistent. However, this random assignment was subject to the constraint that each of the eight attributes should be variable in at least one group and consistent in at least one other. This balancing was undertaken to ensure that the results would not occur merely as an artifact of idiosyncratic combinations or configurations of feature assignments.

A specific CVC nonsense syllable was assigned as a verbal label for each instance. The letter combinations were generated randomly from a pool of ten consonants and five vowels, except English words were not allowed. A different set of ten labels was generated for each stimulus set. Within each group, the assignment of labels to instances was randomized.

Procedure

Subjects were tested individually for a single session of two phases lasting a total of approximately 50 minutes. In Phase I, subjects learned to associate a specific verbal label with each training instance. In Phase II, they rated the similarity of specific pairs of
instances from memory, given only their verbal labels.

**Phase I: Instance Learning.** This phase consisted of two subphases, *familiarization* followed by *name-recall*. In the initial familiarization, the experimenter sat facing the subject with a stack of 5 by 8" index cards, each of which had a picture of a specific insect on one side and the insect's name printed on the other side. The experimenter shuffled the deck, and then presented the cards one at a time to the subject. For each card, the experimenter began by showing the subject the insect picture for about ten seconds, after which the card was turned over to show its name for another three seconds; this was repeated twice for each card, after which it was removed and the next one was presented. The experimenter cycled through all ten instances twice in this fashion, reshuffling the deck after each cycle.

Following this initial familiarization, subjects learned the picture-name associations by a cued name-recall procedure. The insect pictures were presented one at a time in random order and the subject tried to recall its correct name from memory. Following this response (or after 10 seconds had elapsed), the card was turned over to show the correct name for that instance. If the subject had responded with an incorrect label for a given instance, the experimenter would show them the correct name for three seconds, turn the card over and show them the picture for twenty seconds, and then the name again for another three seconds. Following this study period, the card was removed and the next one presented. This procedure continued until the subject had correctly labeled all the instances on two consecutive cycles, or until a total of 15 cycles had been completed.
Phase II: Similarity Ratings. In this phase, subjects rated the similarity of specific pairs of instances from memory, given only their names learned in Phase I. Each name pair was printed on a separate index card, shown one at a time. Subjects rated the similarity of each pair on a 20-point scale, from 1 (very low similarity) to 20 (very high similarity).

Subjects from each group rated 11 such pairs. One of these, the target pair, consisted of the two deviant instances. A second, control pair contained normal training instances that were identical to the deviants except that the target default was present in these instances. In addition to the target and control pairs, nine normal pairs were also presented; these differed in their degree of mismatch on the three variable attributes. Three of the pairs had one mismatching variable attribute, three others had two mismatching variables, and the rest mismatched on all three. If subjects were behaving systematically and remembering the instances learned in Phase I, then pairs that mismatched on more attributes should be rated less similar.

The pairs were presented in a different random order to each subject, within the constraints that the target and control pairs were separated by at least two intervening normal trials; neither the target nor the control pair occurred in the first two trials of the series, nor on the final trial.
Results

Distinctive Feature Effects

The normal pairs varied in whether the instances differed along one, two, or all three of their variable attributes. This factor strongly affected subjects’ similarity ratings. Pairs with one mismatching attribute were assigned an average similarity rating of 13.63, compared to 9.72 for pairs with two mismatches, and 6.59 for pairs with three differences. This decreasing trend in rated similarity was significant \( t(16) = 10.42, p < .01 \), and was shown by all 17 subjects. The pairwise comparisons between the individual conditions were also highly reliable: \( t(16) = 6.02, p < .01 \) for one versus two differences, and \( t(16) = 6.25, p < .01 \) for two versus three.

Common Feature Effects

To compute the effect of surprise on the weighting of a common feature, we subtracted the similarity of the control pair in which the target default was present from that of the target pair in which it was absent. The resulting difference score estimates the impact of the absent default as a common feature of the stimuli, relative to the baseline condition in which the default was present.

Removing a default from both pair members increased their similarity by an average of 3.35 points, relative to the control pairs. This effect was statistically significant, \( t(16) = 2.67, p < .02 \). This difference was nearly the same magnitude as the impact of a mismatching attribute (described above), although the effect on similarity
was in the opposite direction. Thus, each mismatch between pair members decreased their similarity by an average of 3.52 points, an effect roughly equal to the 3.35-point increase due to shared absence. Although the instances in the target pairs differed on three variable attributes, their mean similarity of 10.0 was significantly greater than that of normal pairs differing by three attributes (mean similarity of 6.59; t(16) = 3.29, p < .01) and roughly equal to those that differed by only two (9.72). Thus, the results strongly support the hypothesis that an absent default can function as a highly salient common feature and increase rated similarity. This is the effect for which we were searching.

Discussion

These results are interesting in several respects. First, they substantiate the hypothesis that people attend to the features of a stimulus according to their unusualness or "surprise value" with respect to a reference category. Second, the results support the more general point that the features of a stimulus are not derived solely from perceptual data (i.e., the physical structure of the stimulus); rather, features can arise on the basis of subjects’ background expectations as they are applied to interpret the stimulus. Third, the results validate the similarity-from-memory procedure, and suggest its use for further investigations of concept learning and how past learning influences similarity judgments. We discuss each of these issues in turn.

First, the results support the claim that a feature’s salience is strongly influenced by its unexpectedness. Our result must be contrasted to the usual result in similarity
experiments, namely, that the salience of an absent feature is typically less than that of its presence. That is, when a feature is present and absent equally often in the stimulus set, removing a common feature from a pair of instances typically reduces their similarity (e.g., Gati and Tversky, 1984). This standard effect was reversed in Experiment 2, in which the surprising absence of a default had a larger effect as a common feature than did its presence. Since subjects' expectations were sufficiently powerful to reverse the usual outcome of such comparisons, we may conclude that such expectations are important in controlling subjects' attention and weighting of features in similarity judgments.

Besides showing that subjects notice the unexpected absence of a default and weight this absence heavily in comparisons, the present results suggest that such absence would have a similar impact when instances are presented individually in tasks not involving explicit comparison. Although a value's unusualness increased its weight when it occurred as a difference between stimuli in Experiment 1, a critic could argue that subjects only noticed absent defaults there due to the contrasting presence of that component in the other pair member. That experiment did not produce strong evidence that absence (or other unusual values) would be noticed in pairs in which both instances were lacking the target default, since shared absence did not increase similarity in that experiment. If the critic's explanation were correct, then subjects should also overlook the absent default in individually presented instances, since the required contrast with a normal instance is also unavailable in that condition. But the Experiment 2 results demonstrate just the opposite, that explicit contrast is not required for the absent default to be noticed and considered highly informative. This result increases our confidence that the high salience of surprising values would generalize to situations other than the
specific task investigated here.

The impact of a reference category on the evaluation of these stimuli also supports a more general thesis, namely, that the features of an object may arise from a largely top-down process of fitting previously-learned category models to the perceptual input. Thus, features cannot be conceived simply in terms the stimulus evoking some fixed vocabulary of a priori perceptual primitives; rather, the description of a stimulus should be considered the product of the subject's theory-based interpretation. The fact that our subjects focused on absent features in their similarity judgments is particularly important for this claim. Although an infinite number of features are technically missing from any object, people are likely to notice and encode the absence of only those features that they would otherwise expect to be present. For example, it is felicitous to say that a penguin is a flightless bird but not that a dog is a flightless mammal, even though both statements are technically correct (see Grice, 1975). An absent feature exists solely with respect to the background expectations used to interpret the stimulus; in this sense, an absent default is a purely conceptual feature, not a component of the physical object itself. The fact that subjects interpreted instances in terms of such absent features demonstrates the role of the reference category in specifying the set of "features" by which objects are described.

The results also support the validity of the experimental procedure introduced here. Most previous similarity experiments (as did our Experiment 1) have asked subjects to compare pairs of stimuli (e.g., pictures or verbal descriptions) available for their inspection at the time the rating is made, or to judge the similarity of familiar, real-
world concepts referred to by name (e.g., "robin-sparrow", "dog-cat"). This experiment was similar to the latter task in referring to objects by name, but unusual in teaching subjects about the objects and their arbitrary labels prior to the rating phase. Relative to many prior similarity experiments, this one used a fairly small sample of subjects (seventeen), and, in particular, collected a rather small number of data points from each (a total of eleven). Despite such limitations, the data were extremely orderly and the predictions were soundly confirmed. Every subject gave lower similarity ratings to pairs as the number of mismatches increased; this factor was included largely to check that subjects had been able to remember the instances and render sensible judgments, which they plainly did. These results suggest that the similarity-from-memory task may prove useful for further investigations of how conceptual knowledge influences similarity.

General Discussion

Earlier, we hypothesized that people employ an unusualness heuristic to select informative or relevant properties in comparing stimuli. Thus, when two objects share an unusual property, their similarity should be greatly increased. Conversely, when a surprising difference is observed (as when one stimulus lacks an expected default present in the other), then this discrepancy should greatly reduce the similarity of the pair. Both predictions were supported by the preceding experiments. Experiment 1 found a significant effect of unusualness on the impact of a distinctive feature. However, in that experiment, common features had a much smaller effect on subjects' judgments than did distinctive features, possibly due to a specific judgment strategy (of only counting differences) induced by the experiment. In Experiment 2, when subjects rated the
similarity of instances from memory, the absence of an expected default in both instances increased their similarity, as had been predicted.

At first glance, the demonstration that removing common features can sometimes increase similarity seems to contradict a basic assumption of Tversky’s (1977) contrast model, which expects similarity to always be a positive function of the common features of a pair. However, our result can be accommodated within the contrast model simply by assuming that the absence of an expected default would be noticed and recorded by subjects as an informative, salient property of an instance, while its presence would normally be taken for granted and hence ignored in subjects’ judgments. Thus, what counts as a "feature" in subjects’ internal representations depends on their prior expectations with respect to the reference category. As a result, the physical operations of adding (or subtracting) parts of a stimulus pattern may sometimes have opposite psychological effects. The contrast model does not explicitly assume that stimulus pairs are judged within the framework of an evoked reference category. However, such assumptions are not inconsistent with the model; they are merely one of many possible representational frameworks that could be developed within it.

In this paper, we have described unusualness as a function of a value’s infrequency within a particular reference category. However, this is not meant to imply that infrequency is the only possible determinant of unusualness. A variety of other factors are likely to play a role, such as a value’s recency, vividness, duration of prior exposures, meaningfulness, relation to a context, and so on. Moreover, other properties of an attribute’s frequency profile besides the infrequency of a specific value may affect
its perceived unusualness -- for instance, whether the value violates a specific default or is merely one of many equally improbable alternatives. These issues await future research.

The tasks developed here may prove useful for investigating many issues regarding stimulus representation, categorization, and similarity. Besides influencing ratings of similarity, category-level expectations should also influence other performances that involve explicit or implicit comparisons, such as the selection of previously-encountered examples to guide the solution of a current problem and spontaneous reminders across episodes. We need principles to describe how prior experience in a domain affects the representation of stimuli, in part by controlling which features are attended to and which are backgrounded. Such principles could improve the predictive power of cognitive theories in several domains of applications.
References


Footnotes

1. To settle upon our terminology, we are discussing a procedure in which pairs of stimuli are presented to subjects to be judged for similarity. Each pair can be thought of as sharing some common features and differing on some distinctive features. A feature is a value of an attribute or dimension of stimulus variation. The comparison stimuli are presumed to evoke a reference category which itself is characterized by norms that specify the likelihood of different values of each relevant attribute. The unusualness of a given value of an attribute in a presented stimulus is given by its improbability in the norms of the reference category.

2. For common feature effects, Table 1 shows the average similarity of pairs in which the feature was present minus that of pairs in which the feature was absent. This difference score would have been negative in the absent default condition had our predictions been confirmed.
Author Notes

This research was supported by the Air Force Office of Scientific Research Grant 87-0282.

We would like to express our appreciation to Amos Tversky, Barbara Tversky, Misha Pavel, and Patrick Suppes for their valuable comments and advice on the empirical research described in this article.

Correspondence concerning this article should be sent to John Clapper, Department of Psychology, Bldg. 420, Stanford University, Stanford, CA 94305.
Table 1

*Difference Scores for Experiment 1*

<table>
<thead>
<tr>
<th></th>
<th>Common</th>
<th>Distinctive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent Defaults</td>
<td>0.07</td>
<td>2.60</td>
</tr>
<tr>
<td>Novel Attributes</td>
<td>0.49</td>
<td>3.02</td>
</tr>
<tr>
<td>Variable Attributes</td>
<td>0.22</td>
<td>1.42</td>
</tr>
</tbody>
</table>
Figure Caption

*Figure 1.* Sample instances from two stimulus sets shown to different groups of subjects in Experiment 1. In one set (upper pair), wings and antennae are consistently present, tails and jaw pincers are consistently absent, and forelimbs and leg bristles are variable. These assignments are changed in the other set (lower pair).