Memory & Cognition
2001, 29 (4), 587-597

An inverse base rate effect with continuously valued stimuli

MICHAEL L. KALISH
University of Western Australia, Nedlands, Western Australia, Australia

It is well known that people do not always make normative use of information about relative frequencies of categories when making categorical judgments. The "inverse base rate" effect (Medin & Edelson, 1988) is a typical example of this: Subjects violate normative reasoning principles by assigning certain ambiguous stimuli as belonging to the less frequent of two categories, rather than to the more common category. This effect has been explained as being due to the shifting of attention from shared stimulus features to distinctive features during learning. When stimuli are defined by values along continuous dimensions, rather than by the presence and absence of features, then attention could shift between dimensions or between values, or both. In three experiments, base rate differences were used to determine the way in which attention is shifted during learning about stimuli with continuously valued dimensions. Simulation modeling shows that the results are consistent with the movement of attention both between and within stimulus dimensions.

Categorization's central role in cognition has been matched by its prominence in theorizing in cognitive psychology. Artificial category learning experiments have produced a wide range of data, which have led to a large number of explanatory theories. Consideration of data across experimental paradigms has also led to more powerful theories—for example, when a single theory (ADIT; Kruschke, 1996) substituted for two theories in explaining both "apparent base rate neglect" (Gluck & Bower, 1988) and the "inverse base rate effect" (Medin & Edelson, 1988). ADIT has recently been extended to include a third class of experiments, in which stimulus features can take on continuous values (Kalish & Kruschke, 2000).

Many real-world categorization problems require attention to continuous dimensions, such as attending to size when sorting ponies from horses or when sorting bushes from trees. It seems desirable to explain these common problems together with those in which stimuli differ nominally—for example, when we categorize an illness as a flu rather than as a cold due to the presence of muscle aches. As I describe below, base rate effects (caused by learning about categories that are presented with different relative frequencies) have proven to be an important testing ground for theories of attention in the categorization of nominal stimuli. Given this, it is perhaps surprising that little is known about the action of attention during learning to categorize stimuli with continuous dimensions. The three experiments presented here applied the base rate manipulation to stimuli with continuous dimensions, on the assumption that this would lead to attention shifts within, rather than between, stimulus dimensions.

This research was supported by a Small ARC grant to the author. Thanks to Stephan Lewandowsky, Mark Gluck, Adriaan Tijsilling, and Steven Sloman for helpful comments on earlier drafts of the manuscript. Thanks also to Ben Nguyen for help with data collection. Correspondence should be addressed to M. L. Kalish, Department of Psychology, University of Western Australia, Nedlands, Western Australia 6009, Australia (e-mail: kalish@psy.uwa.edu.au).

Base Rates in Category Learning

In category learning experiments, subjects are typically presented with unlabeled instances (such as a set of symptoms without a diagnosis) and asked to judge which category they belong to (i.e., make a diagnosis). Feedback is then given (i.e., the correct diagnosis), which allows the subject to study the stimulus along with its correct label for a brief time. Although alternative approaches exist, from presenting labeled stimuli (Medin, Altom, Edelson, & Freko, 1982) to withholding feedback entirely (Ahn & Medin, 1992; Ashby, Queller, & Berretty, 1999; Medin, Wattenmaker, & Hampson, 1987), the supervised training method described above is particularly useful for manipulating and controlling both the order of presentation of items and categories and their relative frequency. Although order effects are highly diagnostic of the way people learn new categories (Goldstone, 1993; Kruschke, 1996; Shanks, 1991), for the purpose of this paper, I focus only on the effects of relative category frequency, or "base rates," and in particular on the inverse base rate effect and on apparent base rate neglect.

In apparent base rate neglect, subjects seem to fail to integrate the relative base rates into their decisions about category membership. For example, in the original finding of apparent base rate neglect, Gluck and Bower (1988) trained subjects to discriminate two categories, composed of stimuli with four dimensions. In a medical diagnosis framework, each dimension corresponded to a particular symptom, which was either present or absent for each stimulus (Gluck & Bower's, 1988, Experiments 1 and 2). On each trial, subjects were shown a list of symptoms

and asked to judge what fictitious disease that "patient" had. Each symptom was independent of all the others and depended only on the category from which the patient was drawn. Critically, one of the two categories was presented three times as often as the other. By setting the conditional probabilities of the symptoms given the diseases appropriately (i.e., so that one symptom occurred three times as often in the rare disease as in the common disease), Gluck and Bower constructed a situation in which the posterior probabilities of the two diseases given one particular symptom were equal. An ideal learner should thus be neutral about which of the two categories a stimulus with just that one feature belongs to. What Gluck and Bower found, however, was that when subjects were presented with this symptom in isolation after training, the rare category was chosen significantly more often than the common category. This amounts to base rate neglect, because it is consistent with considering only the conditional probabilities (which favor the rare disease 3:1), and ignoring the relative frequency of occurrence of the two diseases.

The second base rate related finding is termed the inverse base rate effect. The inverse base rate effect is said to occur when subjects classify an ambiguous stimulus that "ought" to belong to a more frequent category (by application of Bayes' theorem) as belonging to the less frequent category. Apparently, the subject believes the base rates to be the inverse of what they really are—thus the name of the effect. Medin and Edelson (1988) initially observed this effect (described below in detail), using binary predictor dimensions similar to those of Gluck and Bower (1988).

Despite the similarity between the experimental procedures, Gluck and Bower (1988) and Medin and Edelson (1988) each provided a separate account of their own data. Gluck and Bower showed that a connectionist implementation of prototype theory, the component cue model, could account for base rate neglect. Medin and Edelson, however, argued that a particular model based on instances instead of prototypes, known as the context model (Medin & Schaffer, 1978; Nosofsky, 1986), was needed to account for the inverse base rate effect. However, the context model could work only if provided with a mechanism for moving attention from one dimension to another. Kruschke (1992) implemented the generalized context model as a connectionist network and augmented it with a learning rule for adjusting the attention weights. This ALCOVE model captured a great many results in category learning, but initial claims that the model accounted for base rate neglect proved to be in error (Lewandowsky, 1995). For a time, then, it did not appear that any single model could explain the full range of base rate related effects in category learning.

In response to this situation, Kruschke (1996) provided a unified account of both apparent base rate neglect and the inverse base rate effect, showing that a prototype model with rapidly shifting dimensional attention (called *ADIT*) could quantitatively explain both phenomena. Like Gluck and Bower's (1988) component cue model, ADIT has a set of input nodes, each of which is active when a particular feature (symptom) is present and inactive when it is absent. Unlike Gluck and Bower's model, these nodes are connected by two weights to each of the category nodes, which represent the disposition to choose that category. One set of weights is slowly adapted in response to error and represents what is learned about the categories, whereas the other set is rapidly adapted and represents a shifting attentional bias for or against individual dimensions. The attention strengths are also reset at the beginning of each trial, so that associative weights alone carry long-term effects of learning. Additionally, attention strengths are normalized after each shift, so that, unlike associative weights, when attention to one dimension grows, attention to other dimensions must diminish.

A major limitation of this model is that it is restricted to discrete, separable dimensions. In all three of the experiments in Kruschke (1996), stimuli were composed of multiple dimensions; however, each dimension could take only one value, and that value was either present or absent on each trial. An alternative design for the experiment is to allow each dimension to take one of two values and then present only one of the values on each trial. Gluck and Bower's (1988) Experiment 3 (see also Nosofsky, Kruschke, & McKinley, 1992) showed that these "substitutive" features produce base rate neglect, just as do the present/absent features. A model such as ADIT can accommodate substitutive features by representing each stimulus pair (such as "dry skin" and "oily skin") with two input nodes, each of which signals the presence or absence of one symptom (Gluck & Bower, 1988; Kruschke, 1996). Essentially, the model does not identify the substitutive features as belonging to a common dimension, nor do the subjects. Base rate effects are produced when attention moves between stimulus features, just as in the present/absent case.

A similar but more extreme change in the design of the experiment is to move from the present/absent stimulus dimensions to a continuous (rather than dichotomous) scale. One important difference between substitutive features and continuous scales is that a given stimulus value on a continuous scale contains other (lesser) values, in the way that a 10-cm line includes 1-cm lines. This difference requires a substantial change in the formulation of ADIT, which Kalish and Kruschke (2000) have recently provided. The extension is substantial, because, in addition to a different representation of the input, new commitments must be made about the way attention can shift over the stimulus representation. In essence, the extended model proposed that people are able to shift attention within a single dimension, instead of only between dimensions. Suppose, for example, that a stimulus appears that is both small and bright, and the subject gives the wrong categorical response. In the original ADIT, the only adaptive change in attention a subject could make would be to shift attention to just the size (or just the luminance) of the stimulus. In Kalish and Kruschke's generalized model (called *Corner*), attention can shift instead to the value of size (the smallness), a difference that is identifi-

Table 1
Schematic Organization of the Basic Inverse Base Rate Design

Dimension			
I_k	PC_k	PR_k	Category
1	1	0	C_k
1	1	0	C_k
1	1	0	C_k
1	1	1	R_k

Note—The subscript k denotes one set of dimensions and categories. The total number of dimensions and categories depends on the number of replications of this design within the experiment, denoted K. In Experiments 1–3, $K = 2$; in Medin and Edelson (1998), $K = 3$.

able the next time a large stimulus appears. ADIT predicts that much has been learned about the size dimension, whereas Corner predicts that nothing has been learned about largeness.

Base rates change the way people represent contrasting categories, when the stimuli have binary dimensions. When the stimuli have continuous dimensions, it is unknown how (or even if) the representation changes. The Corner model predicts that changes will occur within a dimension, because of the way attention shifts. In this paper, I ask whether the Corner model or the original ADIT can best predict the effects of base rate differences on the knowledge gained during learning with continuous dimensioned stimuli. This analysis of attention shifts derives from extensions of the inverse base rate design, which I now describe in detail.

Modifying the Inverse Base Rate Design

Medin and Edelson (1988) obtained the inverse base rate effect under the following paradigm. Subjects were trained to discriminate six categories, based on nine-dimensional inputs (see Table 1). The input dimensions were framed as the symptoms of fictitious diseases, whereas the category labels were the fictitious disease names. The six categories were made up of three common (C) and three rare (R) categories, with the common categories presented three times as often as the rare categories. Each common category was paired with a single rare category; for each pair, only three of the symptoms might be shown. The other symptoms could be shown only when one of the other common/rare category pairs was presented. For any one common/rare pair, there was one symptom that was presented only when the common disease was the correct "diagnosis," one symptom that was present only when the correct diagnosis was the rare disease, and one symptom that was always present, regardless of the correct diagnosis. These are called the *perfect predictor of the common disease* (PC), the *perfect predictor of the rare disease* (PR), and the *imperfect predictor* (I), respectively. Because the three category pairs were all formally the same, discussion of the inverse base rate effect centers on the roles of PC, PR, and I.

Subjects were trained until they made the correct diagnosis for every member of the training set—that is, when they correctly identified that the presentation of I and PC together meant the patient had the common disease and that the presentation of I and PR together meant the patient had the rare disease. After this training, subjects were presented with PC and PR together and were asked to make a diagnosis. Subjects selected the rare disease, even though the common disease is more likely in this case. This is the inverse base rate effect, so labeled because, in the absence of information—the {PR,PC} combination is completely uninformative—subjects went against the base rate.

In each of the following three experiments, the difference between the original present/absent features and the stimuli was systematically increased. In Experiment 1, features distributed binomially along a clear continuum were used. Because the inverse base rate effect is sensitive to changes in materials (Shanks, 1991), this manipulation provided a necessary control condition. In Experiment 2, stimulus values were drawn from uniform distributions, but the logical structure of Experiment 1 was preserved. In Experiment 3, stimulus values were drawn from overlapping Gaussian distributions, and a probabilistic classification problem was produced.

EXPERIMENT 1
Binomial Distributions

In Experiment 1, several small modifications to the classic inverse base rate effect paradigm were presented. First, following Kruschke (1996), I reduced the design to two sets of common/rare disease pairs and, thus, to only six symptom dimensions. Second, for the sake of generality, the disease names were replaced with other arbitrary category labels—in this case, animal names. Third, the symptoms were presented as six continuous dimensions.

Method

Subjects. Fifteen first-year students enrolled in psychology at the University of Western Australia participated for course credit.

Apparatus. The subjects were seated in individual, brightly lit stations, each sound-masked by a ventilation fan. At each station, a PC-type computer displayed the stimuli on a VGA monitor and collected responses using a standard keyboard and mouse.

Procedure. The subjects were asked to read instructions that detailed the procedure. On each trial, six vertically oriented bars of different colors were displayed side by side. The subjects were told that this "graph" represented levels of different "blood proteins" (this label appeared below the bars as well). On the basis of the colored bars, the subjects were to choose one of four native Australian animal names (*quokka*, *koala*, *wallaby*, *numbat*). The names were

Table 2
Schematic Organization of Experiment 1

Dimension						
I	PC	PR	I	PC	PR	Category
1	1	0	0	0	0	A
1	1	0	0	0	0	A
1	1	0	0	0	0	A
1	0	1	0	0	0	B
0	0	0	1	1	0	C
0	0	0	1	1	0	C
0	0	0	1	1	0	C
0	0	0	1	0	1	D

Note—Tall bars are labeled "1"; short bares are labeled "0."

Table 3
Results of Experiment 1 and Predictions of the Full Corner Model and the Same Model With No Attention Shifting

	Response				Corner				No Shifting			
Stimulus	C	R	Co	Ro	C	R	Co	Ro	C	R	Co	Ro
I	.58	.33	.05	.03	.55	.21	.12	.12	.38	.35	.14	.14
PR	.05	.87	.05	.03	.06	.82	.06	.06	.05	.73	.11	.11
PC	.68	.15	.05	.12	.69	.10	.11	.11	.74	.04	.11	.11
PC + PR	.33	.60	.05	.02	.31	.55	.07	.07	.37	.35	.14	.14
I + PR + PC	.53	.38	.05	.03	.52	.38	.05	.05	.45	.40	.07	.07
I + PRo	.18	.18	.02	.62	.24	.11	.06	.59	.23	.22	.04	.51
I + PCo	.42	.07	.43	.08	.33	.14	.45	.09	.22	.21	.53	.04
I + PC + PRo	.75	.03	.10	.12	.56	.06	.04	.33	.67	.05	.02	.26
PC + PRo	.28	.03	.12	.57	.32	.06	.06	.55	.48	.03	.03	.45

displayed in individual boxes on the computer monitor and were selected by clicking the mouse button over them.

The schematic organization of the training stimuli is shown in Table 2. On each trial, two of the vertical bars were tall, extending 100 mm above their bases. The remaining four bars were short, extending only 10 mm above their bases. The association of pairs of bars with the response categories was deterministic; the task of the subject was to learn that association.

On each trial, after the subject chose a category label, the correct response was provided by the computer. The correct response, together with the subject's response and the stimulus display, remained available for study for up to 1.5 sec. Once the subject acknowledged the feedback, again using the mouse, the next trial appeared following a brief (500-msec) delay.

Trials were grouped together into blocks. In each block, each of the two common categories was presented three times, and each of the two rare categories was presented once. Trial order was randomized within each block. Each subject completed 15 contiguous blocks. The relative horizontal order of each colored bar within the graph was randomized between trials, so that bar color was not confounded with bar location. Similarly, the assignment of animal names and bar colors to categories was randomized between subjects to overcome any confounding stimulus effects.

Following the 120 training trials, the subjects entered into the transfer phase. Eighteen test stimuli, comprising two isomorphic sets of nine items (see Table 3), were shown twice each, in a random order. Recall that there were six dimensions, only three of which were involved in discriminating any one pair of categories. Isomorphic sets are drawn by comparing across these pairs: For example, "I" in Table 3 represents two bars, one for the first category pair and a different bar for the second pair. On each of the 36 transfer trials, no corrective feedback was given.

Results and Discussion

The total number of responses of each category given to each of the nine transfer stimulus types is shown in Table 3. The two responses of each subject to the isomorphic transfer items were taken to be independent, as were the two replications of the item types, and so there were 60 independent observations for each stimulus type. The perfect common and perfect rare predictors (PC and PR) were both strongly associated with their respective categories. When the PC dimension was large, there were more "C" responses than "R" responses [$\chi^2(1, N = 50) = 19.22, p < .05$] ($N = 50$, rather than 60, because only the responses to the correct category pair, C and R, are considered in this analysis), whereas the PR dimension produced the opposite pattern [$\chi^2(1, N = 55) = 41.89, p = .05$]. Large values of the imperfect predictor dimension, I, were associated with the common category marginally more than the rare category [$\chi^2(1, N = 55) = 3.56, p = .06$]. Thus, the subjects appear to have learned the basic nature of the classification task.

When all three predictors were presented together, the subjects seemed to choose the common category more often than the rare category, although this was not significant [$\chi^2(1, N = 55) = 1.16, p = .28$]. This suggests that the PR dimension is similar in strength to the combination of {I,PC}. However, when the I and PC predictors were associated with a category pair different from the PR predictor—so that it was the "PR" of the "other" category, denoted "PRo"—the subjects chose the common category predicted by the two [I + PC + PRo, $\chi^2(1, N = 52) = 26.32, p < .05$]. Taken together, these results show that the combination of {I,PC} is stronger than PR alone.

Critically, when the two perfect predictors alone were placed in competition, the subjects chose the rare category significantly more than the common category. This was true both when the two predictors were associated with the same category pair [PR + PC, $\chi^2(1, N = 56) = 4.02, p < .05$] and when they were associated with different pairs of categories [PC + PRo, $\chi^2(1, N = 51) = 5.02, p < .05$]. The strength of the perfect rare predictor was also evident when the PR of one category pair was placed in conflict with the imperfect predictor of the other category pair (I + PRo). The total number of responses to the rare category predicted by PRo was marginally greater than the number of combined responses to both members of the pair of categories predicted by the imperfect predictor [$\chi^2(1, N = 59) = 3.32, p = .07$]. In contrast, when I of one category pair was placed in opposition with PC or the other pair, there was nearly equal division of responses between the two categories predicted by the I cue and the common category predicted by the PC [I + PCo, $\chi^2(1, N = 55) = 0.07, p > .5$].

The inverse base rate effect was thus observed, with the single cues associated with the rare categories having greater response strength than the single cues associated with the common categories. These results were obtained using stimuli that varied along a continuum. However, the

Table 4
Results of Experiment 2 and Predictions of the Full Corner Model and the Same Model With No Attention Shifting

	Response				Corner				No Shifting			
Symptom	C	R	Co	Ro	C	R	Co	Ro	C	R	Co	Ro
I	.48	.30	.08	.13	.46	.29	.12	.12	.38	.35	.14	.14
PR	.07	.72	.10	.12	.06	.79	.07	.07	.05	.73	.11	.11
PC	.58	.17	.12	.13	.73	.08	.09	.09	.74	.04	.11	.11
PC + PR	.25	.60	.10	.05	.37	.49	.07	.07	.37	.35	.14	.14
I + PR + PC	.47	.43	.10	.00	.49	.41	.05	.05	.45	.40	.07	.07
I + PRo	.33	.17	.05	.45	.22	.16	.06	.56	.23	.22	.04	.51
I + PCo	.27	.17	.47	.10	.27	.16	.50	.08	.23	.21	.53	.04
I + PC + PRo	.63	.07	.07	.22	.57	.08	.04	.30	.67	.05	.02	.26
PC + PRo	.32	.05	.05	.58	.40	.06	.06	.49	.48	.03	.03	.45

stimuli in Experiment 1 took only two of the infinite possible values of bar height. The presence of the inverse base rate effect with these stimuli is thus to be expected given the similarity between present/absent and substitutive features. Subjects apparently treat large and small bars in just the same way that they treat present/absent verbal features.

EXPERIMENT 2
Continuous Analog of the Inverse Base Rate Design

The challenge in extending this design to truly continuous dimensions is to reconstruct a sense in which one dimension can perfectly predict one category. The solution chosen here is to set a criterial value for the dimension. Then, when a pattern has a value on that dimension that is greater than the criterial value, that pattern will always be a member of only one category. When the dimension has a value lower than the critical value, the pattern can belong to any of the other categories. Imperfect prediction is provided by a dimension that takes on extreme values for a pair of categories.

Method

Subjects. Fifteen first-year students enrolled in psychology at the University of Western Australia participated for course credit. None had participated in Experiment 1.

Apparatus. The apparatus used in Experiment 2 was the same as that used in Experiment 1.

Procedure. The subjects were given the same instructions used in Experiment 1. The stimulus display was identical to that in Experiment 1, with one exception. Each vertically oriented bar could take on a range of values, determined both by the category that was being presented and by a measure of uniformly distributed random noise. Each perfect predictor took on a value in the range of 60–100 mm when the category it signaled was presented, and it took on a value in the range of 10–50 mm otherwise. The two imperfect predictors took on the larger values when either of their two predicted patterns were present, and they took on the smaller values otherwise. Each subject completed 120 trials of training, 45 each of the two common categories and 15 each of the two rare categories. Stimuli were randomly selected and were presented in a different random order (within 15 contiguous blocks) for each subject. The spatial arrangement of colored bars was randomized between trials, whereas assignment of bar colors and responses to categories was randomized between subjects.

Following the training phase, the subjects were shown the same 38 transfer stimuli as in Experiment 1. Values of the dimensions of the transfer stimuli were either 10 mm (short) or 100 mm (long).

Results and Discussion

The overall proportion of responses given to each of the nine transfer stimulus types are shown in Table 4. As in Experiment 1, the two perfect predictors (PC and PR) were both strongly associated with their respective categories [PC, $\chi^2(1, N = 45) = 12.80, p < .01$; PR, $\chi^2(1, N = 47) = 30.72, p < .01$]. The imperfect predictor was slightly more strongly associated with the common category than with the rare category, but not significantly so [$\chi^2(1, N = 47) = 2.13, p = .14$]. When all three predictors were presented together, the subjects chose the two categories about equally often [$\chi^2(1, N = 54) = 0.019, p > .8$]. When the I and PC predictors were related to one category pair, and the PR was related to the other category pair, the subjects chose the common category predicted by I and PC significantly more than the rare category predicted by PRo [$\chi^2(1, N = 51) = 11.29, p < .05$].

As in Experiment 1, when the two perfect predictors were placed in competition, the subjects chose the rare category significantly more than the common category. This was true both when the two predictors were associated with the same category pair [PR + PC, $\chi^2(1, N = 51) = 7.84, p < .01$] and when they were associated with different pairs of categories [PC + PRo, $\chi^2(1, N = 54) = 4.17, p < .05$].

Despite this, the PR predictor was not quite as strong as it was in Experiment 1. When the perfect rare predictor of one category pair was placed in conflict with the imperfect predictor of the other category pair (I + PRo), the number of responses to the rare category predicted by PRo was nearly the same as the number of responses to the pair of categories predicted by the imperfect predictor [$\chi^2(1, N = 57) = 0.07, p > .7$]. Similarly, when an imperfect and PC predictor of different category pairs were placed in opposition, there was nearly equal division of responses between the two categories predicted by the I cue and the common category predicted by the PC [I +

Table 5
Results of Experiment 3 and Predictions of the Full Corner Model, and the Same Model With No Attention Shifting

	Response				Corner				No Shifting			
Symptom	C	R	Co	Ro	C	R	Co	Ro	C	R	Co	Ro
I	.52	.35	.12	.02	.66	.23	.05	.05	.56	.30	.07	.07
PR	.03	.85	.08	.03	.06	.80	.07	.06	.05	.72	.12	.11
PC	.80	.05	.08	.07	.84	.05	.05	.05	.84	.04	.06	.06
PC + PR	.48	.37	.10	.05	.48	.41	.05	.05	.47	.33	.10	.09
I + PR + PC	.70	.20	.05	.05	.70	.26	.02	.02	.68	.28	.02	.02
I + PRo	.38	.15	.08	.38	.39	.14	.05	.42	.41	.22	.02	.34
I + PCo	.37	.28	.30	.05	.36	.14	.46	.05	.32	.17	.48	.03
I + PC + PRo	.93	.05	.00	.02	.74	.08	.02	.15	.89	.03	.01	.08
PC + PRo	.50	.07	.03	.40	.49	.05	.05	.42	.63	.03	.02	.31

PCo, $\chi^2(1, N = 54) = 0.02, p > .8$]. Finally, when the I and PC predictors were associated with a category pair different from the PR predictor, the subjects chose the common category predicted by the two [I + PC + PRo, $\chi^2(1, N = 51) = 11.29, p = .01$].

As in Experiment 1, these results display a clear inverse base rate effect. The effect was slightly less than with the discrete stimuli of Experiment 1, with the I + PC combination slightly stronger.

Experiments 1 and 2 involved continuous stimulus dimensions, together with categories that differed in their base rates. In both experiments, category membership was uniquely determined by a single stimulus dimension taking on an exceptionally large value. Thus, the similarity between the results of the two experiments is perhaps not surprising. The issue, taken up below, is how to account for this similarity formally.

With discrete stimuli, the introduction of uncertainty in the mapping from stimulus patterns to categories results in a greatly attenuated form of the inverse base rate effect, known as base rate neglect (Gluck & Bower, 1988; Kruschke, 1996). Probabilistic categorization comes from allowing the same stimulus to be associated with multiple categories. With continuous stimulus dimensions, this can be accomplished by overlapping the range of values that stimuli from different categories can have. In Experiment 3, this manipulation was used in an attempt to reduce the inverse base rate effect.

EXPERIMENT 3
Overlapping Continuous Dimensions

Method

Subjects. Fifteen first-year students enrolled in psychology at the University of Western Australia participated for course credit. None had participated in either of the first two experiments.

Apparatus. The apparatus used in Experiment 3 was the same as that used in Experiments 1 and 2.

Procedure. The subjects were again given the same instructions as were used in the first two experiments. Stimulus presentation was identical, with the exception that stimuli were drawn from overlapping normal distributions along all six stimulus dimensions. Dimensional validity for individual categories was manipulated by changing the mean of the distribution. Stimulus values for the predictive dimensions ({I,PC} or {I,PR}, although no dimension remained perfectly predictive) had a mean of 90 mm and a standard deviation of 6 mm, whereas values for the other relevant dimension and the three unrelated dimensions were drawn from a distribution with mean of 45 mm and a standard deviation of 25 mm. Thus, there was an approximately 10% chance that a stimulus value drawn from the predictive distribution would have a value less than that of a stimulus drawn from one of the unrelated (or irrelevant) distributions.

The actual set of 200 training stimuli presented to each subject was selected randomly, according to the distribution parameters. Stimuli were presented in a random order, within 25 contiguous blocks in which each category occurred either three times (for common categories) or once (for rare categories). The spatial arrangement of colored vertical bars varied randomly between trials, and the assignment of bars and labels to categories varied randomly between subjects. On approximately 40 of the 200 trials, at least one of the two large stimulus values would be less than at least one of the four small stimulus values.

Following the training trials, the subjects moved into a transfer phase. Transfer stimuli were identical to those used in Experiments 1 and 2, and, again, no feedback was given.

Results and Discussion

The overall proportion of responses given to each of the nine transfer stimulus types are shown in Table 5. As in Experiment 1, the two perfect predictors (PC and PR) were strongly associated with their respective categories [PC, $\chi^2(1, N = 51) = 37.96, p < .05$; PR, $\chi^2(1, N = 53) = 43.47, p < .05$]. Unlike in Experiments 1 and 2, the imperfect predictor was not significantly more strongly associated with the common category than with the rare category [$\chi^2(1, N = 52) = 1.56, p = .21$]. When all three predictors were presented together, the subjects strongly preferred the common category [$\chi^2(1, N = 54) = 15.57, p < .05$]. This was also true when the I and PC dimensions were associated with one pair, and the PR dimension was associated with the other [I + PC + PRo, $\chi^2(1, N = 57) = 51.16, p < .05$].

Unlike in Experiments 1 and 2, the strength of the PR cue was not sufficient to lead to a preference for the rare category when the two perfect predictors were placed in competition. Instead, the subjects chose the two categories about equally often. This was true both when the two predictors were associated with the same category pair [PR + PC, $\chi^2(1, N = 51) = 0.71, p = .40$] and when they were associated with different pairs of categories [PC + PRo, $\chi^2(1, N = 54) = 0.46, p = .50$]. The inverse base rate

effect seen in the first two experiments was clearly greatly attenuated, so that the subjects were effectively treating the PR and PC features equally. The subjects neglected the base rates, failing to prefer the common category when PR and PC were placed in competition. A between-experiments comparison supported this: There was a significant difference in the choice frequencies of rare and common categories across all three experiments when PC + PR or PC + PRo were presented [$\chi^2(1, N = 313) = 14.92, p < .01$]. Over 99% of this difference was between Experiment 3 and the total of Experiments 1 and 2 [$\chi^2(1, N = 313) = 14.91, p < .01$], showing that Experiment 3 essentially eliminated the inverse base rate effect.

GENERAL DISCUSSION

Theoretical Modeling

The clear results of these three experiments is that the strength of the PR cue diminished as the range of stimulus values presented went from binary to continuous (but deterministic) to continuous and overlapping. The question that theoretical modeling can answer is, Why? The explanation when stimuli are binary-valued is that people shift their attention between stimulus features in order to reduce classification error. I explore that possibility here, when the stimuli are continuously valued.

Kruschke (1996) used the concept of attention shifts to explain the inverse base rate effect. The theory states that people react to categorization errors by shifting their attention away from stimulus features that cause error to stimulus features that do not cause as much error. This adaptive shift of attention has two effects. First, the perceived value of a stimulus feature is amplified by the devotion of attention to the feature: The feature is effectively more salient than it was before. Second, associative learning that connects features to categories is also modulated by attention. Shifts of attention thus determine which features are learned about on each trial and how rapidly that learning proceeds. This theory was formalized in the ADIT model for present/absent features (as discussed in the introduction). For the ADIT model, a "feature" is defined as a stimulus dimension. In the present context of continuously valued stimulus dimensions, this definition is no longer obviously correct. With continuous stimulus dimensions, a feature could be either a dimension (such as color) or a particular value on a dimension (such as redness).

Kalish and Kruschke (2000) provided a formulation of attention shifting in which the definition of a feature is taken to refer to a stimulus value, rather than a stimulus dimension. The Corner model uses a vector of input nodes to represent the value of each stimulus dimension, so that the dimension is broken up into a set of discrete features. By using a vector to represent each dimension, Corner can readily model the shift in attention between stimulus features as the shift in the distribution of attention over the "value" elements of all of the dimension vectors, regardless of which dimension the elements belong to.

Importantly, Kalish and Kruschke (2000) showed that Corner is simply an extension of ADIT; if the number of nodes per dimension drops to 1, the two models are functionally identical. In order to highlight the difference of attention shifts within versus between stimulus dimensions, I show here that Corner can also be reduced to ADIT through another method—one that introduces a new parameter and allows hierarchical models to be constructed that test hypotheses about what constitutes a feature over which attention can shift. In order to introduce this parameter, I first set out Corner formally.

Corner

A stimulus is presented to the network as a vector x of values $\{x_1, \ldots, x_N\}$ along each of the N stimulus dimensions. These values are represented internally by a set of nodes, so that the value of node i along dimension j is given by

$$a_{ij}^{\text{in}} = \begin{cases} 1 & \text{if } (x_j - \mu_{ij}) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where μ_{ij} is the location of node i along dimension j. This is a form of "thermometer" coding (Anderson, Silverstein, Ritz, & Jones, 1997), a representation that embodies the principle that large values must contain small ones.

Attention is of limited capacity and so gets applied to the representation of the stimulus in two steps. First, the attention devoted to any given value, α_{ij}, is set equal to the activation of that value node, a_{ij}^{in}. Then, attention is normalized:

$$\alpha_{ij} := \frac{\alpha_{ij}}{\left(\sum_j \sum_i \alpha_{ij}^{\eta}\right)^{1/\eta}}, \tag{2}$$

where η is a free parameter ($\eta > 0$) that represents the attention capacity available to the network.

After attention is applied, activity is multiplied by attention and weights (modifiable associations) to produce activation at each of the K category nodes:

$$a_k^{\text{out}} = \sum_i \sum_j \alpha_{ij} w_{ijk} a_{ij}^{\text{in}}, \tag{3}$$

where w_{ijk} is the weight from feature ij to category node k.

During learning, the model compares these output values (which represent dispositions to categorize the stimulus into each category) with the correct category label, as supplied by the teacher. Teacher values are given by

$$t_k = \begin{cases} \max(1, a_k^{\text{out}}) & \text{if } x \text{ is in category } k \\ \max(0, a_k^{\text{out}}) & \text{otherwise.} \end{cases} \tag{4}$$

The comparison results in error at each category nodes:

$$E_k = t_k - a_k^{\text{out}}. \tag{5}$$

The error is then passed back down to the input nodes, where attention strengths are shifted to reduce error:

$$\Delta\alpha_{ij} = -\gamma_{\text{value}} \frac{\delta E}{\delta \alpha_{ij}}$$
$$= -\gamma_{\text{value}} a_{ij}^{\text{in}} \sum_k E_k w_{ijk}, \quad (6)$$

where γ_{value} is the rate at which attention shifts between dimension value nodes. After the shift is completed, negative attention strengths are set to zero because negative attention is not possible in the model.

At this point, I can introduce the new parameter that allows Corner and ADIT to nest within a larger model framework. First, the error sent back to the input units can be summed across all nodes which code a dimension, j,

$$E(\alpha_{\cdot j}) = \sum_i \sum_k E_k w_{ijk}. \quad (7)$$

This dimensional error can be used to adjust all the attention strengths applied to values on the dimension equally:

$$\Delta\alpha_{ij} = -\gamma_{\text{dimension}} \frac{\delta E}{\delta \alpha_{\cdot j}}$$
$$= -\gamma_{\text{dimension}} a_{ij}^{\text{in}} E(\alpha_{\cdot j}), \quad (8)$$

where $\gamma_{\text{dimension}}$ is the dimensional attention shift rate.

These two sources of adjustment (by attention to values and by attention to dimensions) can be combined into a single equation

$$\Delta\alpha_{ij} = -\left(\gamma_{value} a_{ij}^{\text{in}} \sum_k E_k w_{ijk} + \gamma_{\text{dimension}} a_{ij}^{\text{in}} \sum_i \sum_k E_k w_{ijk} \right). \quad (9)$$

The parameter $\gamma_{\text{dimension}}$ is new to this version of Corner. If γ_{value} is zero, then Corner reduces to a form of ADIT, regardless of the number of nodes coding each dimension. ADIT* (the new form) shifts attention between dimensions only, not within a dimension.

After attention has been shifted, it is then renormalized (by Equation 2). Activation again passes up to the category nodes, and error again passes back down. Attention strengths are now left unchanged, and the model updates the association weights:

$$\Delta w_{ijk} = \lambda a_{ij}^{in} \alpha_{ij} E_k, \quad (10)$$

where λ is the association learning rate parameter. Because attention is shifted before weights are updated, weight changes occur only between category nodes and features that are the focus of attention.

Two final steps are required to map network outputs onto responses. The first is to use a choice rule to turn activations into probabilities. The probability that the model will choose any category K is given by

$$p_\kappa = \frac{\exp(\phi a_\kappa^{\text{out}})}{\sum_k \exp(\phi a_k^{\text{out}})}, \quad (11)$$

where ϕ is a constant reflecting the certainty with which categories are selected as responses (Kalish & Kruschke, 2000; Luce, 1963).

Up to this point, the only role for category base rates has been in determining the order of presentation of stimuli to the model (via sampling). However, it is possible that subjects in the experiment actually infer the base rates and explicitly use this information to alter their response selection. This possibility is considered by including a constant, β, that reflects the value that subjects place on (accurate) base rate information relative to stimulus-specific information:

$$p'_\kappa = \frac{p_\kappa b_\kappa^{\beta/(\beta+N)}}{\sum_k p_k b_k^{\beta/(\beta+N)} s}, \quad (12)$$

where N is the number of stimulus dimensions presented (which is a constant value in these experiments), and b_k is the relative base rate ($\sum b_k = 1$) of category k.

The complete Corner model thus has six parameters: η, the level of total available attention; λ, the learning rate of the association weights; ϕ, the scaling constant in the choice rule; β, the relative strength of the subjective base rate information; and γ_{value} and $\gamma_{\text{dimension}}$, the two attention shift rate parameters. For convenience, I refer to the family of models as *Corner*, the six-parameter version as *full Corner*, the five-parameter version from Kalish and Kruschke (2000) as *Corner**, and the five-parameter version of ADIT as *ADIT**.

Nested Model Tests

Corner produces base rate effects when association weights are formed only after attention is shifted either within or between stimulus dimensions, or both. In order to test the possibility that this could account for the differences in base rate effects between the three experiments, I fit Corner to all three data sets simultaneously. Each subject's training series (the set of input and teacher values) was used to train the network, which was then tested on the set of transfer stimuli. The goodness of fit on the transfer stimuli determined the selection of Corner's parameters, via a hill-climbing algorithm. I used 10 input nodes for each dimension; similar results obtain with more nodes.

The full Corner fit exceptionally well, with a non-significant lack of fit [$G^2(75) = 95.1, p > .05, N = 1{,}620$]. This suggests that attention shifts are important for explaining these results, and so comparison with the three reduced forms of the model could be informative. First, if γ_{value} is set to zero, then the full model can be compared with ADIT*. If $\gamma_{\text{dimension}}$ is set to zero, the full model can be compared with Corner*. Because ADIT* and Corner* have the same number of free parameters, they can also be directly compared. Finally, the role of attention shifting per se can be assessed by comparing versions of Corner to a null model, in which both attention shift parameters are set to zero. Because the models are all nested, likelihood ratio tests can be used to determine the significance of any

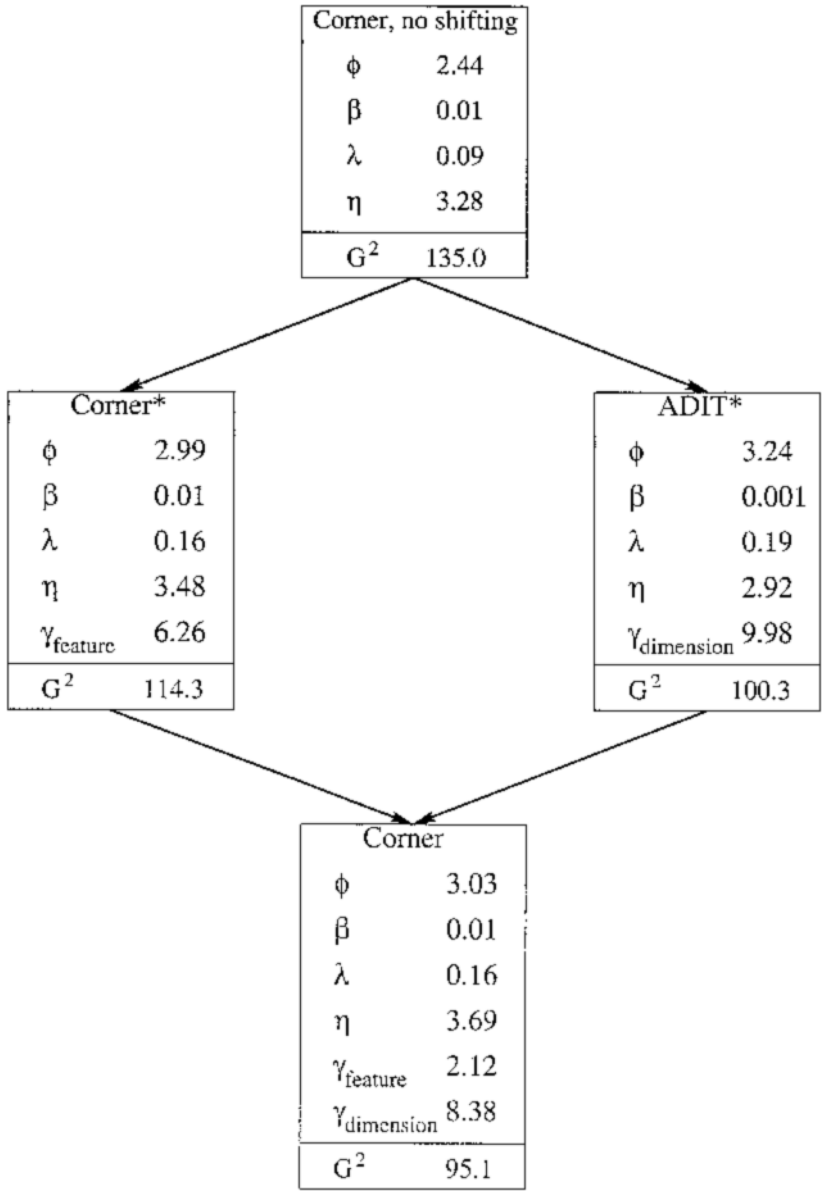

Figure 1. The set of nested models used to fit the results of all three experiments. Parameters are defined in the text.

increases in goodness of fit provided by the addition of new free parameters. The best-fitting parameter values of the four Corner models are shown in Figure 1.

The four-parameter Corner without any attention shifts fits quite well [$R^2(U) = .986$, relative to the homogeneous model] but not so well that it cannot be rejected as a plausible source of these data [$G^2(77) = 135.0, p < .01, N = 1{,}620$]. Treating this as the null model, the five-parameter Corner* accounts for 15.3% of the remaining variance, which is a significant improvement [$\chi^2(1) = 20.7, p < .01$]. Similarly, ADIT* accounts for 25.7% of the remaining variance, which is also a significant improvement [$\chi^2(1) = 34.7, p < .01$]. ADIT* also fits better than Corner*, since they both have five parameters. Against ADIT*, the full Corner is nonetheless statistically superior [$\chi^2(1) = 5.2, p < .05$]. This means that attention shifts within stimulus dimensions provide significant explanatory power, even after factoring in attention shifts between stimulus dimensions.

The predictions of the full Corner and the null Corner are shown along with the observed response probabilities in Tables 3–5. Without attention shifts, the null Corner cannot produce an inverse base rate effect in any of the three experiments. The full Corner correctly produces the effect in Experiments 1 and 2, but, like the subjects, not in Experiment 3. Since Corner makes its predictions without changing parameters between experiments, the differ-

ences in the strengths of the predictors between experiments must be entirely due to the contingencies in effect during training.

Corner explains the effect of the training regimes as follows. In Experiment 1, the subjects were trained with a direct continuous analog of the standard inverse base rate design (Kruschke, 1996; Medin & Edelson, 1988). On each trial, the set of input nodes that represent the dimensions with large values have more active nodes than do the set that represent the dimensions with smaller values. These extra nodes are available to be associated with the response categories. Very early in training, associations develop between the value nodes of the I and PC dimensions and the common category. When a rare category item is presented, attention is shifted away from the I dimension value nodes, because they are causing a misprediction in favor of the common category. The activated value nodes of the PR dimension thus receive the bulk of the "blame" for the categorization error, and the associative weights to those nodes increase fastest. As in the original ADIT, the model holds that people shift their attention exclusively to the PR dimension and use that information to predict the appearance of the rare category. With all the attention flowing to a single dimension, association weights become stronger between PR and rare than between PC (or I) and common, where attention is divided between two dimensions.

In Experiment 2, the level of variability between trials increases because, on each trial, the two primary predictors, such as {I,PC} or {I,PR}, are of variable magnitude. The irrelevant predictors are also of varying magnitude, with the difference between the two types of predictors always visible, but not always large. The model is affected by this variability, because the predictive validity of each dimension decreases and because the number of perfectly valid values also decreases. For example, on any given trial, the value of the I dimension may be greater than that of the PC dimension. This makes the "large" value nodes of I perfect predictors of the common category. Association weights to the (more valid) "large" nodes of PC thus do not have as much of an opportunity to grow early in training, on average, as they do in Experiment 1. Similarly, the presence of activity in the PR dimension nodes (due to a random PR value on some given trial) decreases the validity of the "moderate" PR value nodes, so that they are no longer perfect predictors of the rare category. With this situation holding, attention tends to shift less between dimensions than in Experiment 1, due to less early learning, and tends to shift more within dimensions, due to a distribution of validity among the dimension values.

In Experiment 3, the level of variability from trial to trial is even higher. On any given trial, any one dimension may have the largest value, drawing the most initial attention. For example, when a rare category stimulus is presented, it is possible that the PRo predictor (which signals the "other" rare category) may have a large value. Because association weights do not develop instantly, the spurious predictors are often magnets for attention, since they do not mispredict the category label. Attention is, on average, distributed over value nodes on many different dimensions, which means that early learning does not produce strong biases in attention. Without early associations between {I,PC} and the common category, I is not ignored when {I,PR} appears, and so there is no asymmetry between PC and PR. The preference for the common category is still far less than the base rates alone would predict, however, and thus represents a form of base rate neglect.

Conclusions

These experiments showed that (1) the inverse base rate effect occurs even when stimulus dimensions vary continuously, (2) the effect is significantly weakened when categorization is made probabilistic instead of deterministic, and (3) these results are consistent with the base rates affecting the nature of the category representations that subjects form, by virtue of the shifting of attention among stimulus features during learning. The existence of attention shifts has been confirmed in a number of recent studies (e.g., Kruschke, 1996; Kruschke & Johansen, 1999; Lewandowsky, Kalish, & Griffiths, 2000), but, in all prior cases, attention was seen to shift only between stimulus dimensions. Corner allows attention to shift within a dimension, and simulation modeling confirmed that these shifts occurred in these experiments.

Shifts of attention during learning are apparently useful, because they increase the speed with which new categories can be acquired (Kruschke & Johansen, 1999). However, as a consequence of attention shifts, people do not learn to be optimal classifiers—instead, people learn to ignore some information that might be relevant and to attend to other information that might be misleading. When attention shifts within a stimulus dimension, people are led to accentuate the distinctive attributes of the stimulus. A stimulus that is a bit larger than average is seen as very large, one that is redder than average is seen as very red, and so on. Goldstone (1988) showed that biases such as these can occur spontaneously: Subjects in his experiment accentuated the redness of a stimulus in one context and accentuated the orangeness of the same stimulus in a different context. Systematic biases of this sort are also consistent with the performance of experts. For example, the height of an average tree is believed to be greater than it really is (Medin, Lynch, Coley, & Atran, 1997), perhaps because height makes trees distinctive from bushes. Firefighters devalue the role of wind in driving back burns (Lewandowsky & Kirsner, 2000), perhaps because they are generally lit in light winds, making them distinctive from wind-driven wild fires. Models of attention have focused on dimensional attention, whereas these phenomena appear to require attention shifts within a dimension.

Distinguishing intradimensional shifts from interdimensional shifts of attention is not a straightforward matter, however. If interdimensional shifts are made on an item-specific basis, the two very different psychological processes may produce very similar outcomes. For exam-

ple, firefighters might shift their attention away from wind to other variables (such as the slope of the terrain) for some fires, but not others. Exemplar-specific dimensional attention has proved to be a useful concept in explaining biases in category learning experiments (Kruschke, in press); distinguishing this from intradimensional shifts may prove difficult. Indeed, even in the present set of experiments, ADIT* and Corner* make very similar predictions. Dunn and Kalish (2000) reported an analysis of these models that revealed that intradimensional and interdimensional shifts can mimic each other over a wide range of parameter values.

Complex models, even when formally specified, can be difficult to tell apart in practice. Despite the difficulty in distinguishing dimensional attention from attention to features, the nested model fitting approach used here was able to show that both types of attention shifts appear to be working simultaneously in these experiments.

REFERENCES

Ahn, W. K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, **16**, 81-121.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, **84**, 413-451.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, **61**, 1178-1199.

Dunn, J., & Kalish, M. L. (2000, August). *Comparing models of categorization using signed difference analysis.* Paper presented at the 31st Annual Meeting of the Society for Mathematical Psychology, Kingston, Ontario, Canada.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227-247.

Goldstone, R. L. (1988). Perceptual learning. *Annual Review of Psychology*, **49**, 585-612.

Goldstone, R. L. (1993). *Positively and negatively defined concepts* (Research Rep. 88). Cognitive Science Program, Indiana University, Bloomington.

Kalish, M. L., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research*, **64**, 105-116.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 3-26.

Kruschke, J. K. (in press). Toward a unified model of category learning. *Journal of Mathematical Psychology.*

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1083-1119.

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review*, **102**, 185-191.

Lewandowsky, S., Kalish, M. L., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1666-1684.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & Cognition*, **28**, 295-305.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, P. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 37-50.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, **117**, 68-85.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, **32**, 49-96.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, **19**, 242-279.

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

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 211-233.

Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science*, **3**, 143-162.

(Manuscript received April 27, 2000;
revision accepted for publication November 2, 2000.)