The Role of Base Rates in Category Learning

John K. Kruschke

Department of Psychology and Cognitive Science Program
Indiana University, Bloomington IN 47405-4201 USA
e-mail: kruschke@indiana.edu

May 2, 1994

Indiana University Cognitive Science Program Research Report # 115

Copyright ©1994 by John K. Kruschke

The Role of Base Rates in Category Learning

John K. Kruschke
Department of Psychology and Cognitive Science Program
Indiana University, Bloomington IN 47405-4201 USA
e-mail: kruschke@indiana.edu

May 2, 1994

ABSTRACT: Previous researchers have discovered perplexing inconsistencies in how human subjects appear to utilize knowledge of category base rates when making category judgments. In particular, Medin and Edelson (1988) found an "inverse base rate effect" in which subjects tended to select a rare category when tested with a combination of conflicting cues, and Gluck and Bower (1988) reported apparent "base rate neglect" in which subjects tended to select a rare category when tested with a single symptom whose objective diagnosticity was equal for all categories. In this article I suggest that two principles underlie those effects: First, base rate information is learned and consistently deployed during all training and testing cases. Second, the dominant effect of base rates is to cause frequent categories to be learned before rare categories, so that the common categories are encoded in terms of their typical features, and the rare categories are encoded by whichever features distinguish them relative to the already-learned, common categories. Four new experiments are reported which confirm those principles. Experiment 1 replicates and extends the basic inverse base rate effect and establishes a standard for subsequent experiments and modeling. Experiment 2 shows that pre-training on a subset of categories has comparable effects as giving those categories high base rates during training. Experiment 3 shows that apparent "base rate neglect" is simply an attenuated case of the "inverse base rate effect," hence any model of one effect should account for the other. Experiment 4 uses probabilistic categories, and shows that apparent "base rate neglect" can be obtained even for cues that occur in less than 50% of the rare category's exemplars, and that cues that are not associated with any category decrease the base-rate bias. The principles are formalized in a new connectionist model that modifies attention to stimulus features based on the network's current knowledge. Quantitative fits to the empirical data are reported, and provide additional support for the principles.

Introduction

Research into the role of base rates in human decision making has evoked tremendous controversy (e.g., Koehler, 1993). Many studies have shown that people can be insensitive to base-rate information (e.g. Kahneman & Tversky, 1973; Bar-Hillel & Fischhoff, 1981), whereas other studies have shown that people can employ base-rate information, to some extent, when it is conveyed through experience (e.g. Christensen-Szalanski & Beach, 1982).

This article is concerned with the role of base rates in category learning, a type of decision making in which the base rates are directly experienced. I am specifically interested in explaining two perplexing patterns of results discovered by previous researchers, viz., the "inverse base rate effect" (Medin & Edelson, 1988) and "apparent base rate neglect" (Gluck & Bower, 1988). Medin and Edelson (1988 p.81) concluded that "The present experiments reveal that when base-rate information is conveyed through experience it does not influence decision making in some uniform manner. ... responses were independent of base rates, positively correlated with base rates, or even negatively correlated with base rates." Holyoak and Spellman (1993 p.285; Spellman 1993) also suggested that subjects use base-rate knowledge inconsistently, depending on whether the classification decision takes place during the training phase or during the testing phase. They suggest that in training, subjects use base rates accurately because the base rates are learned and used implicitly as part of the task, but, in testing, the base rates can be neglected because they must be accessed explicitly.

Contrary to those interpretations, this article argues that subjects in these category-learning experiments apply their knowledge of base rates in a uniform manner, regardless of the stimulus content or training vs. test phase. Subjects do underemphasize the base rates relative to normative, Bayesian prescriptions, but the underweighted knowledge is applied consistently to all cases.

This article proposes two main theses: First, the primary role of base rates in these experiments is to cause the high-frequency categories to be learned before the low-frequency categories. The fact that the high-frequency categories are learned before the low-frequency categories has been given little importance in previous reports, but here it is preeminent. It is a truism that what we learn depends on what we know (e.g. Lenat & Feigenbaum, 1987), and so what subjects learn about the rare categories should depend on what they have already learned about the more common categories. In particular, when learning about new (rare) categories, the learner should attend to the features that distinguish them from the already-learned (frequent) categories. Consequently, the common and rare categories are encoded asymmetrically, and this asymmetry causes the perplexing patterns observed in human classification decisions.

The other main thesis was already stated: A second role of base rates is to bias subjects to choose the more frequent categories, and the bias is applied consistently to all cases. However, when the bias is applied to asymmetric category representations, it can be obscured and appear to be inconsistent.

A corollary to those theses is that the base-rate neglect reported by Gluck and Bower (1988) is merely an attenuated case of the inverse base-rate effect reported by Medin and Edelson (1988), and the effects are caused by the same mechanisms. Therefore, any viable model of one effect should also account for the other. No previous model has yet been shown to thoroughly account for both effects (an extended discussion of previous models is provided later in the article).

This article provides both empirical and modeling evidence in support of the main theses. The first experiment replicates the inverse base-rate effect using fewer categories and cues than used by Medin and Edelson (1988), confirms that the frequent categories are learned before the rare categories, and confirms that subjects know the categories occur with different base rates. Experiment 2 exposes subjects to categories with equal base rates, after pre-training on a subset of categories. The results show a robust analogue of the inverse base-rate effect, with the later-learned categories being the preferred response in tests of conflicting cues. Experiment 3 is a hybrid design that includes minimally different sub-designs for the inverse base-rate effect and apparent base-rate neglect, and shows that base-rate neglect is an attenuated case of the inverse base-rate effect. Experiment 4 explores base-rate neglect using various cue probabilities, and demonstrates, among other things, that a cue can be encoded as distinctive of a category even if the cue occurs in fewer than 50% of the category instances, providing additional support for the main theses, and providing quantitative data for modeling.

The explanatory principles are formalized in a model, which is fit to the data from the four experiments. The model is an extension of the basic component-cue network used by Gluck and Bower (1988). There are two key additions: First, every input cue has an attention strength that is adjusted on the basis of error, such that distinctive features are more strongly attended when there is error, and existing knowledge is attended when response tendencies are correct but of insufficient magnitude. Second, classification probabilities are determined by mixing the prediction from the associative network with the learned base rates. The relative weight given to the base rates depends only on the number of cues in the input, not on the specific content of the input or on the prediction of the network.

At the very least, the empirical and modeling results indicate that the inverse base-rate effect and apparent base-rate neglect do not *necessarily* imply that people apply base-rate knowledge in some convoluted manner. That conclusion can only be reached, however, by virtue of the main thesis, i.e., rare categories are learned in terms of the features that distinguish them from the previously learned common categories. Thus, the major emphasis of the article is the learning mechanism that produces asymmetrically encoded categories.

Experiment 1: The inverse base rate effect

Medin and Edelson (1988) presented subjects with a fictitious disease diagnosis task. On each trial of a learning sequence, a list of symptoms was presented to the subject, who had to diagnose the hypothetical patient as having one of several possible fictitious diseases. The subject then was told the correct diagnosis, after which another list of symptoms was presented. The basic design involved a pair of diseases, designated C (for "common") and R (for "rare"), which occur in random order and have base rates with a 3:1 ratio. During training, every instance of disease C had two symptoms, labeled I and PC, and every instance of disease R had two symptoms, labeled I and PR. Symptom I occurred for both diseases and was an *imperfect* predictor of the two diseases; symptom PC was a *perfect predictor of the common disease*; and, symptom PR was a *perfect predictor of the rare disease*. In the original design, there were three pairs of diseases with this structure, hence six diseases all together, with nine possible symptoms.

After training, subjects were tested with combinations of symptoms not shown in training. When tested with the ambiguous symptom I, subjects tended to choose the common disease, consistent with the base rates. When tested with the ambiguous combination I+PC+PR, subjects again tended to choose the common disease (although less strongly). When presented with the conflicting symptoms PC+PR, however, subjects tended to choose the rare

disease, contrary (or "inverse") to the base rates.

The ideas described in the introduction explain the results as follows: During training, subjects first learn that symptoms I and PC are typical of disease C, because the case occurs so often. They also quickly learn that disease C occurs much more frequently than disease R. Subsequently, they try to learn the rare disease. They discover that the shared symptom I is misleading (they already think it indicates disease C) and instead pay attention to the distinctive symptom PR. When tested with PC+PR, subjects choose disease R because the symptom list contains the key distinctive symptom of disease R, but only one of the two typical symptoms for disease C. Subjects also incorporate their knowledge that disease R is rare and unlikely, but the distinctiveness of PR is so strong that there is a tendency to choose R nevertheless. When tested with symptoms I+PC+PR, subjects find that all symptoms of disease C and R are present, and the "tie vote" is broken by base-rate knowledge, which causes them to tend to choose the common disease.

Experiment 1 is a partial replication and extension of the primary experiment reported by Medin and Edelson (1988). Its purposes are (i) to establish that the inverse base rate effect is robustly obtainable when using the particular materials and procedure in my lab (for example, Shanks (1992) was unable to obtain an inverse base rate effect when using a 3:1 ratio of base rates), (ii) to establish a standard for the magnitude of the "inverse base rate effect," for comparison with subsequent experiments and for modeling, (iii) to extend the results of Medin and Edelson (1988) to a situation with a different number of categories, additional test cases, and explicit category frequency judgments, and (iv) to verify that subjects do learn the common disease before the rare disease, and have explicit knowledge that the base rates of the diseases are very different.

Method

Abstract training structure

The abstract structure of the training cases is shown in Table 1. The structure is the same as used by Medin and Edelson (1988), except that here only four categories and six symptoms were used, instead of six categories and nine symptoms. Notice that the structure of diseases C1 and R1 is the same as C2 and R2. Each pair of diseases has a shared symptom, labeled "I," that is an *imperfect* predictor of the diseases. Each disease also has a symptom that is unique, called its *perfect* predictor. The perfect predictor of the common disease is labeled "PC" and the perfect predictor of the rare disease is labeled "PR." Each row of Table 1 corresponds to a training instance, with presence of a symptom indicated by a "1," and absence by a "0."

Table 1: Abstract design of training stimuli in Experiment 1.

Symptoms						
I1	PC1	PR1	I2	PC2	PR2	Disease
1	1	0	0	0	0	C1
1	1	0	0	0	0	C1
1	1	0	0	0	0	C1
1	0	1	0	0	0	R1
0	0	0	1	1	0	C2
0	0	0	1	1	0	C2
0	0	0	1	1	0	C2
0	0	0	1	0	1	R2

Overall sequence of training and testing

Each block of training trials contained the eight cases shown in Table 1, randomly permuted for each block and each subject. Every subject was trained for 15 blocks (120 trials), with self-timed breaks every 5 blocks.

After training, the subject was shown novel combinations of symptoms, as specified in the results section. There were 18 test stimuli, shown twice each, in a random order, for a total of 36 test trials. Instructions appeared prior to the test cases that told the subject that they would be seeing combinations of symptoms they had not previously seen, and that they would not be told the correct diagnoses, but that it was very important that they make their best educated guess based on what they learned from the earlier part of the experiment.

After the test phase, the subject was asked to judge the frequencies of the four diseases during the training period. They indicated their estimates on a paper response sheet which was concealed inside an envelope during training and testing. The instructions on the paper told the subject that they should estimate the number of times each disease occurred during training, and that they should be careful to think about the actual number of occurrences, not the relative ease of learning or the number of times they got each one right or wrong, nor the apparent feedback or study time, etc. They were also told that the total of the four frequencies should be 120, and that if all four disease occurred equally often then each would have appeared 30 times. Subjects also had to copy a code word from the computer screen, which could be used later in analysis to determine the random permutation of disease labels for each subject.

The entire experiment lasted about 30 minutes.

Physical instantiation of the stimuli

Individual subjects sat before a PC-type computer in a sound-deadened, dimly-lit cubicle. Stimuli were presented on the computer monitor, and responses collected from the standard keyboard.

The six abstract symptoms in Table 1 were randomly assigned, for each subject, to the six symptom labels "ear aches," "skin rash," "back pain," "dizziness," "sore muscles," and "stuffy nose." The four abstract diseases were randomly assigned, for each subject, to the labels "F," "G," "H," and "J." Subjects pressed the corresponding keys on the computer keyboard to indicate their choice. The symptoms were presented as a vertical list on the computer screen, with the order of symptoms randomized on each trial. Thus, for example, the first case in Table 1 could appear in one trial with symptom I1 on the first line and symptom PC1 on the second line, and on another trial with PC1 on the first line and I1 on the second.

Sequence of events for an individual trial

On each training trial, a list of symptoms appeared, with a response prompt, "Diagnose as one of F, G, H or J." After the response, feedback appeared, as follows. If the response was wrong, a tone sounded along with the word "WRONG!" If the response was correct, there was no tone but the word "CORRECT!" If there was no response after 30 seconds, a warning appeared, "FASTER! You have only 30 seconds to make your diagnosis," accompanied by a tone. In all cases, the correct answer was then supplied: "This patient has disease [F,G,H,J]," and at the bottom of the computer screen there appeared the phrase, "After you have studied this case (up to 30 seconds), press the space bar to see the next one." If the subject took longer than 30 seconds of study time, a warning appeared, accompanied by a tone, and the next trial started automatically. There was approximately 750ms. blank screen between offset of the current trial's display and onset of the next trial. On test trials, subjects were not given corrective feedback, but told merely that "Your diagnosis has been recorded." As in training, they had to press the space bar to see the next case.

Subjects

56 subjects participated for partial credit in an introductory psychology course at Indiana University. None of the subjects had participated in any other related experiment in my lab.

Results

Training

The common diseases were learned much faster than the rare diseases. For each subject, the difference between the proportion correct for the common and rare diseases was computed over the first third of training (40 trials). The mean difference was .108, $t(55) = 4.66$, SE $= 0.0231$, $p < .0001$.

By the end of training, the subjects had thoroughly learned all the diseases. The mean proportion correct over the last third of training was .954 for the rare diseases and .980 for the common diseases. The difference between common and rare remained significant, however, $t(55) = 2.46$, SE $= 0.0107$, $p = .017$. Every subject got at least 70% (7 out of 10 instances) correct on the rare diseases over the last third of training.

Testing

Results from the test trials are shown in Table 2. The left column shows the abstract symptoms in the test stimulus, and the remaining columns show the proportion of diagnoses for the corresponding common disease C, the corresponding rare disease R, the *other* common disease Co, or the *other* rare disease Ro. For example, when testing with symptom PC1, disease C1 is "C" in Table 2 and C2 is "Co," and when testing with PC2, C2 is "C" and C1 is "Co." Each row in Table 2 is based on 224 observations (56 subjects, times two instantiations of each abstract symptom combination [one for each pair of common/rare diseases], times two presentations of each instantiation).

Table 2: Results from test trials of Experiment 1.

	Choice Proportion			
Symptoms	C	R	Co	Ro
I	.746	.174	.049	.031
PC	.933	.031	.031	.004
PR	.040	.911	.018	.031
PC+PR	**.353**	**.612**	.022	.013
I+PC+PR	**.580**	**.402**	.013	.004
I+PCo	.406	.080	.469	.045
I+PRo	.219	.085	.031	.665
PC+PRo	**.353**	.027	.058	**.563**
I+PC+PRo	**.719**	.036	.036	**.210**

The results show a strong inverse base rate effect for test cases PC+PR and PC+PRo ($\chi^2_{(1)} = 15.57$ $p < .001$, and $\chi^2_{(1)} = 10.78$ $p < .005$, respectively, using expected values of 50/50), and the magnitude of the effect did not dif-

fer significantly between those cases ($\chi^2_{(1)} = 0.173$ *n.s.*).[1]

For I+PC+PR, subjects showed a small base-rate consistency, preferring the common disease ($\chi^2 = 7.273$ $p < .01$, using 50/50 for expected values), but for I+PC+PRo there was a significantly larger preference for the common disease (interaction $\chi^2 = 16.48$ $p < .001$). Medin and Edelson (1988) collapsed those two cases in their data analysis, but they are evidently construed differently by the subjects. The model, described later in the article, also makes quite different predictions for the two cases of ambiguous triplets.

Frequency estimates

All 56 subjects made frequency estimates that summed to 120. In order to have an approximately normal distribution of estimates for inferential statistics, estimates were converted to the natural logarithm of the ratio of common to rare frequencies, log[(C1+C2)/(R1+R2)]. If subjects had judged the frequencies of the common and rare diseases to be equal, then the log of the ratio would be zero. The actual mean, across subjects, was 0.7366, which is significantly greater than zero, $t(55) = 11.33$, SE= 0.0650, $p < .0001$. The anti-log of the mean log ratio is 2.1, i.e., the estimated frequency ratio was about 2.1:1 (as opposed to the actual 3:1). One should interpret the obtained ratio very cautiously, however, as its exact value depends, no doubt, on the response scale used (0–120) and the anchoring at equal values (all 30) in the instructions. What can be concluded with certainty is that the subjects, overall, were strongly aware that the common diseases occurred more frequently than the rare diseases.

Discussion

Experiment 1 verified that the inverse base rate effect obtains in this variant of the Medin and Edelson (1988) paradigm. Extending the results of Medin and Edelson (1988), the present experiment also showed that the base rate consistency effect has different magnitudes for the ambiguous symptom triplets, I+PC+PR and I+PC+PRo. The experiment also verified that the common diseases are learned earlier than the rare diseases, and that subjects are keenly aware of base rate differences. It has thereby established a referent for subsequent experiments, and verified that the basic principles suggested in the introduction (that the more frequent categories are learned before the less frequent categories, and that relative base rates are learned) could indeed be involved. Finally, the results provide quantitative data for modeling.

[1]The χ^2 tests used throughout the article were conducted on the actual frequencies, not the proportions reported in the summary tables. These χ^2 tests are not strictly appropriate, because the data include repeated measures from the same subjects. That is, the data violate the assumption of independence in the χ^2 test. Fortunately, nearly all the effects of interest have such large χ^2 values that even after they are conservatively reduced to reflect the violation of independence (Rao & Scott, 1981), they still exceed conventional critical values. All χ^2 tests reported have 1 df, unless otherwise indicated, and so the subscript "(1)" is hereafter omitted.

Experiment 2: Pre-training a subset of categories produces an analogue of the inverse base rate effect

If it is correct that the basic mechanism responsible for the inverse base-rate effect is learning one category before the other, and consequently learning the later category in contrast to the earlier, then any manipulation which causes one category to be learned before another should also produce an effect analogous to the inverse base-rate effect. In particular, if subjects are trained on some categories before others, then the later categories should be encoded in terms of their distinctive features.

Experiment 2 used a design similar to Experiment 1, in that there were four diseases and six symptoms, with the same exemplars as Experiment 1, but Experiment 2 had two phases of training. In the first, "pre-training" phase, only two diseases were trained, namely, the ones corresponding to the common diseases in Experiment 1. In the second, "full-training" phase, all four diseases appeared with equal base rates. The hypothesis is that the perfect predictor of the later-trained disease will be more strongly encoded than the perfect predictor of the earlier-trained disease.

Method

Abstract training structure

The abstract structure of the training cases is shown in Table 3. The two disease that appear in the pre-training phase are labeled "P1" and "P2," and the other two diseases, which appear only in the full-training phase, are labeled "F1" and "F2." Imperfectly predictive symptoms are labeled "I," as before, and the perfect predictors of the pre-training or full-training diseases are labeled "PP" or "PF," respectively. As before, presence of a symptom is indicated by a "1," absence by a "0."

Overall sequence of training and testing

Subjects were instructed that they would first be given preliminary training on two of the four diseases, and that later they would learn all four. Pre-training consisted of nine

Table 3: Abstract design of training stimuli in Experiment 2.

Symptoms						
I1	PP1	PF1	I2	PP2	PF2	Disease
1	1	0	0	0	0	P1
1	1	0	0	0	0	P1
1	0	1	0	0	0	F1
1	0	1	0	0	0	F1
0	0	0	1	1	0	P2
0	0	0	1	1	0	P2
0	0	0	1	0	1	F2
0	0	0	1	0	1	F2

blocks of eight trials, each block containing four instances of disease P1 and four instances of P2, in random order.

Subjects were then instructed that they would be trained on all four diseases. Full training consisted of nine blocks of the the eight cases shown in Table 3, randomly permuted for each block and each subject.

After training, the subject was shown novel combinations of symptoms, directly analogous to the combinations tested in Experiment 1, and using the same instructions and procedure. Thus, there were 18 test stimuli, shown twice each, in a random order, for a total of 36 test trials. Instructions prior to the test cases were the same as in Experiment 1.

After the test phase, the subject was asked to judge the frequencies of the four diseases during the *full*-training period. Subjects were told that the total of the four frequencies should be 72, and that if all four disease occurred equally often then each would have appeared 18 times. The procedure was otherwise identical to that in Experiment 1.

The same symptoms and disease labels were used as in Experiment 1, with the same display procedure.

The entire experiment lasted about 30 minutes.

Subjects

53 subjects participated for partial credit in an introductory psychology course at Indiana University. None of the subjects had participated in any other related experiment in my lab.

Results

Training

It is possible, however improbable, that subjects would not remember the pre-trained diseases when starting the second phase of training. The results indicated that subjects did indeed remember the pre-training diseases, as the difference in accuracy between the pre- and full-training diseases in the first third of full training was significant (mean difference in proportion correct for the first 18 trials of full training was .107, $t(52) = 6.13$, SE $= 0.0174$, $p < .0001$).

Subjects learned all diseases thoroughly by the end of training; mean proportion correct in the last third of full training was .973 for the pre-training diseases and .978 for the other full-training diseases. All subjects got at least 83% (10 out of 12) of the non-pre-trained disease cases correct in the last third of full training.

Testing

Results from the test trials are shown in Table 4. The left column shows the abstract symptoms in the test stimulus, and the remaining columns show the proportion of diagnoses for the corresponding pre-trained disease P, the corresponding full-training disease F, the *other* pre-trained disease Po, or the *other* full-training disease Fo. Each row in Table 4 is based on 212 observations (53 subjects, times two instantiations of each abstract symptom combination [one for each pair of pre-/full-training diseases], times two presentations of each instantiation).

Table 4: Results from test trials of Experiment 2.

	Choice Proportion			
Symptoms	P	F	Po	Fo
I	.802	.118	.033	.047
PP	.925	.014	.038	.024
PF	.028	.915	.019	.038
PP+PF	**.316**	**.646**	.019	.019
I+PP+PF	**.514**	**.467**	.014	.004
I+PPo	.379	.076	.502	.043
I+PFo	.208	.100	.014	.678
PP+PFo	**.374**	.047	.009	**.569**
I+PP+PFo	**.698**	.047	.028	**.226**

Perhaps the most striking aspect of these results is how similar they are to those from Experiment 1. In particular, there was a strong analogue to the inverse base-rate effect: When subjects were tested with the conflicting pair PP+PF, they tended to choose the disease they learned later, in full training ($\chi^2 = 24.02$ $p < .001$, using 50/50 for expected frequencies). The analogue to the inverse base-rate effect also appeared for the conflicting pair PP+PFo ($\chi^2 = 8.45$ $p < .005$), and was not significantly less than the effect for PP+PF ($\chi^2 = 2.05$ *n.s.*). The magnitude of the analogue to the inverse base-rate effect (case PP+PF) was not significantly less than the magnitude of the inverse base-rate effect itself (case PC+PR) in Experiment 1 ($\chi^2 = 0.644$

n.s.).

When tested with the ambiguous case I+PP+PF, subjects did not choose the pre-training disease significantly more often than the full-training disease ($\chi^2 = 0.481$ *n.s.*). Indeed, that result can be construed as precisely a base-rate consistency effect, because in the present experiment there are equal base rates in the full-training phase, which the subjects perceived nearly as such (see the next section). However, subjects did prefer the pre-trained disease when tested with I+PP+PFo ($\chi^2 = 51.02$ $p < .001$), and that preference was significantly stronger than for I+PP+PF ($\chi^2 = 23.28$ $p < .001$). Again we see that the data do not allow the triple-symptom cases to be collapsed, and the model will also show large differences between them.

Frequency estimates

Five of the 53 subjects made frequency estimates that did not sum to 72: Two subjects had sums of 62, two subjects has sums of 82, and one subject had a sum of 80. Data from those subjects were used anyway, as they were qualitatively indistinguishable from the others, and none of the conclusions are affected.

The mean of log[(C1+C2)/(R1+R2)], across subjects, was 0.1913, which is significantly greater than zero, $t(52) = 3.09$, SE $= 0.0619$, $p < .0032$. The anti-log of the mean log ratio is 1.2, i.e., the estimated frequency ratio was about 1.2:1 (as opposed to the actual 1:1 in full training). That ratio is dramatically less than that obtained in Experiment 1, which was 2.1:1 ($t(107) = 6.08$, pooled SE $= 0.0632$, $p < .0001$).

One should interpret the obtained ratio cautiously, however, as its exact value depends, no doubt, on the response scale used (0–72) and the anchoring at equal values (all 18) in the instructions. What can be concluded is that subjects thought that the pre-training disease occurred slightly more often than the full-training diseases, despite the instructions that they should base their estimates on their experience of the full-training phase only and the objectively equal base rates in the full-training phase. Intuitively, it is plausible that memory for the pre-training instances was better than for the full-training instances, especially early in the full-training phase, and influenced the frequency estimates. I do not care to speculate further on the mechanism underlying these results; rather, I wish to emphasize that the estimated base rates were much more nearly equal than in Experiment 1, which corresponds with the greatly diminished "base rate consistency" effect in Experiment 2.

Discussion

Experiment 2 confirmed that learning one category before another will produce an analogue of the inverse base-rate effect, even when the perceived base rates are not much different from 1:1. This lends credence to the idea that one key role of base rates in the inverse base-rate effect is to cause the more frequent category to be learned before the rare category, and consequently for the later-learned category to be encoded in terms of its distinctive feature(s).

Experiment 2 also showed that subject's knowledge of base rates might also play a role in their responses to the ambiguous test case, I+PC+PR (I+PP+PF), insofar as response tendencies corresponded to the magnitude of the estimated frequency ratios. It is plausible that subjects have certain knowledge about the base rates of the diseases, which is applied in some uniform fashion to the diagnoses, on the one hand, and converted in some way to frequency estimates, on the other hand. This article presents a model for the application of base-rate knowledge in categorization, but presents no model for converting base-rate knowledge to frequency estimates. Only ordinal conclusions are drawn from the frequency-estimation data.

The two training phases of Experiment 2 can be construed as an extreme case of base rate changes during learning. In the pre-training phase, the base rate ratio is ∞:1 (or 1:0), and in the full-training phase the ratio is 1:1. The results are consistent with those obtained by Medin and Bettger (1991), who studied the influence of changing base rates during learning. They found that the inverse base-rate effect was affected primarily by base rates early in training, but that the base-rate consistency effects were also influenced by base rates later in training. Medin and Bettger's (1991) results are nicely accommodated by the ideas proposed here: The base rates early in training indirectly affect the encoding of the categories by influencing the order in which the categories are learned, whereas the base rates later in training affect the magnitude of base-rate bias applied in the test cases.

Experiment 3: Apparent base rate neglect is an attenuated case of the inverse base rate effect

Experiment 3 will emphasize that the apparent base-rate neglect observed by Gluck and Bower (1988) is an attenuated case of the inverse base rate effect. In the basic design studied by Gluck and Bower (1988), there were two diseases, designated C and R, that occurred with 3:1 base rates. There were four symptoms that could be present or absent in each patient. If the patient had the common disease C, then the conditional probabilities of having symptoms s1–s4 were .2, .3, .4, and .6, respectively. If the patient had the rare disease R, then the conditional probabilities of having symptoms s1–s4 were .6, .4, .3, and .2, respectively. The probabilities were selected so that

the true probability of the rare disease, given symptom s1, was 50%.

Subjects were trained on random sequences of cases generated from those probabilities, and then tested with single symptoms. Of primary interest was the result that, when presented with symptom s1 alone, subjects tended to choose the rare disease significantly more often than 50%. That result has been replicated several times (Cobos, López, Rando, Fernández, & Almaraz, 1993; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Myers, Lohmeier, & Well, 1994; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990b).

The tendency has been descriptively labeled as "apparent base rate neglect" because it can be accounted for by underestimating the extremeness of the base rates in Bayes' theorem: $p(\mathrm{R}|\mathrm{s1}) = p(\mathrm{s1}|\mathrm{R})p(\mathrm{R})/[p(\mathrm{s1}|\mathrm{R})p(\mathrm{R}) + p(\mathrm{s1}|\mathrm{C})p(\mathrm{C})]$. That is, if subjects have accurate knowledge that $p(\mathrm{s1}|\mathrm{R}) = .6$ and $p(\mathrm{s1}|\mathrm{C}) = .2$, and use Bayes' theorem, they will overestimate $p(\mathrm{R}|\mathrm{s1})$ if they underestimate the base rate ratio.

The principles propounded in this article explain the effect as follows: Subjects first learn the typical features of the common disease, because the common disease happens so often. They learn that symptom s4 is very typical, and s3 and s2 are somewhat typical, and s1 is only slightly typical. Then they learn about the distinctive features of the rare disease. The most distinctive feature is symptom s1: It occurs frequently in cases of the rare disease, and has not been encoded as a very typical symptom of the common disease. (Symptom s2 might be encoded as somewhat distinctive, but not very much because it doesn't occur very often and has already been encoded as somewhat typical of the common disease.) When subjects are tested with the single symptom s1, they tend to respond that it is a case of the rare disease because they have encoded it as the distinctive symptom of that disease, and despite the fact that they are also influenced by base-rate bias to choose the common disease.

Apparent base-rate neglect is thereby explained in the same way as the inverse base-rate effect. In both situations, one of the symptoms is encoded as distinctive of the rare disease. The only difference is in magnitude: In the inverse base-rate effect, the distinctive symptom is extremely distinctive, as it occurs for every instance of the rare disease and for no instance of the common disease; in apparent base-rate neglect, the distinctive symptom is only somewhat distinctive, as it does not occur exclusively in the rare disease.

In Experiment 3, the abstract stimulus structure of Experiment 1 is modified slightly to produce a circumstance in which $p(\mathrm{R}|\mathrm{s}) = .50$ for one of the rare disease symptoms, thereby allowing apparent base rate neglect and the inverse base rate effect to be observed simultaneously in the same subjects. We can then directly test the relative strengths of distinctive symptom in the neglect situation and the distinctive symptom in the inverse situation.

Method

Abstract training structure

Experiment 3 was identical to Experiment 1 in all respects but two. First, the abstract design of the training stimuli differed in a single training instance, as shown in Table 5. The third exemplar (third row of Table 5) contained all three symptoms, In+PCn+PRn, instead of just two symptoms In+PCn. The purpose of including the third symptom was to make the conditional probability of the rare disease Rn, given the symptom PRn, equal to 50%; i.e., $p(\mathrm{Rn}|\mathrm{PRn}) = 0.5 = p(\mathrm{Cn}|\mathrm{PRn})$.[2] We can then test for apparent base-rate neglect when subjects are presented with symptom PRn alone. The diseases and symptoms are labeled with an "n" suffix or an "i" suffix to indicate whether they correspond to the base-rate neglect sub-structure of the inverse base-rate effect sub-structure. (In the present situation it is a misnomer to label the third symptom "PRn," which inaccurately implies that it is a perfect predictor of the rare disease. Nevertheless I use the label to suggest the structural analogy to the inverse sub-structure.)

Table 5: Abstract design of training stimuli in Experiment 3.

Symptoms						
In	PCn	PRn	Ii	PCi	PRi	Disease
1	1	0	0	0	0	Cn
1	1	0	0	0	0	Cn
1	1	**1**	0	0	0	Cn
1	0	1	0	0	0	Rn
0	0	0	1	1	0	Ci
0	0	0	1	1	0	Ci
0	0	0	1	1	0	Ci
0	0	0	1	0	1	Ri

The second difference between Experiment 3 and Experiment 1 is that subjects in Experiment 3 received a block of 21 test trials after every five training blocks, instead of only at the end of all 15 training blocks as in Experiment 1. I was interested to see if the base-rate neglect effect developed at a different pace than the inverse base-rate effect.

[2]Notice that the stimulus structure for Experiment 3 also makes all the conditional probabilities, of symptoms given diseases, mutually independent, as in the Gluck and Bower (1988) design.

Subjects

64 subjects participated for partial credit in an introductory psychology course at Indiana University. None of the subjects had participated in any other related experiment in my lab.

Results

Training

This design proved more difficult for subjects to learn than Experiment 1, probably because of the added ambiguity of symptom PRn and because of the interruptions in training every five blocks (for testing trials). Because I wanted the test results to reflect the decisions of subjects who ultimately did learn all the diseases, I excluded from further analysis eight subjects (out of 64) who failed to achieve at least 70% correct on the rare diseases in the last third of training. The 70% criterion was selected because all subjects in Experiment 1 achieved at least that level of performance. Consequently, there are 56 subjects included in the data reported below.

Subjects learned the common diseases faster than the rare diseases. In the first third of training (40 trials), the difference between the proportion correct for the common and rare diseases was significant (mean difference $= .1541$, $t(55) = 7.41$, SE $= .0208$, $p < .0001$). Subjects also learned the diseases in the inverse sub-structure faster than those in the neglect sub-structure (mean difference $= .1114$, $t(55) = 5.02$, SE $= .0222$, $p < .0001$). There was no interaction between disease frequency and substructure (mean difference of differences $= .0344$, $t(55) = 0.88$, SE $= .0390$, $n.s.$).

Those same effects persisted into the final third of training, although all proportions correct were very high, ranging from .907 for the rare disease in the neglect substructure, to .985 for the common disease in the inverse substructure (common vs. rare: mean difference $= .0482$, $t(55) = 3.53$, SE $= .0136$, $p = .0008$; neglect vs. inverse: mean difference $= .0292$, $t(55) = 2.24$, SE $= .0130$, $p = .029$; interaction: mean difference of differences $= .0012$, $t(55) = 0.05$, SE $= .0238$, $n.s.$).

Testing

There did not appear to be any notable difference in the onset times of apparent base-rate neglect and the inverse base-rate effects, as both were already evident in the first test block (after 40 trials of training). Therefore, data from the three test blocks were combined and are shown in Table 6. The left column shows the abstract symptoms in the test stimulus, and the remaining columns show the proportion of diagnoses for each disease. Each row in Table 6 is based on 168 observations (56 subjects times three presentations of each test case).

Table 6: Results from test trials of Experiment 3.

	Choice Proportion			
Symptoms	Cn	Rn	Ci	Ri
In	.637	.268	.030	.065
Ii	.036	.060	.780	.125
In+Ii	.359	.174	.365	.102
PCn	.833	.113	.018	.036
PCi	.036	.030	.893	.042
PCn+PCi	.488	.054	.411	.048
PRn	**.131**	**.774**	.042	.054
PRi	.006	.030	.030	.935
PRn+PRi	.024	**.292**	.036	**.649**
PCn+PRn	**.542**	**.400**	.018	.042
PCi+PRi	.012	.024	**.321**	**.643**
In+PCn+PRn	.875	.107	.012	.006
Ii+PCi+PRi	.089	.024	.482	.405
In+PCi	.327	.155	.500	.018
Ii+PCn	.536	.048	.345	.071
In+PRi	.232	.089	.030	.649
Ii+PRn	.084	.506	.265	.145
PCn+PRi	.293	.030	.006	.671
PCi+PRn	.077	.458	.423	.042
In+PCn+PRi	.696	.036	.000	.268
Ii+PCi+PRn	.137	.173	.643	.048

The results show a strong apparent base-rate neglect for symptom PRn, with subjects choosing disease Rn 77.4% of the time, as opposed to just 13.1% choices for disease Cn ($\chi^2 = 76.74$ $p < .001$). Thus, apparent base rate neglect occurs even for these deterministic categories, and does not depend on the probabilistic mapping in the work of Gluck and Bower (1988).

There is also a strong inverse base-rate effect for the conflicting symptoms PCi+PRi, with subjects preferring the rare disease Ri to the common disease Ci, 64.3% to 32.1% ($\chi^2 = 18.00$ $p < .001$).

There was no inverse base-rate effect, however, for the conflicting symptoms in the neglect sub-structure. Symptoms PCn+PRn elicited a small preference for the common disease, 54.2% to 40.0% ($\chi^2 = 3.65$ $n.s.$). Thus it seems that the symptom PRn was strongly enough associated with the rare disease Rn to produce apparent base-rate neglect, but not strongly enough to produce an inverse base-rate effect. A direct comparison of the strengths of symptoms PRn and PRi confirms that interpretation, for when subjects were presented with PRn+PRi, they preferred disease Ri to disease Rn, 64.9% to 29.2% ($\chi^2 = 22.78$ $p < .001$).

The data in Table 6 also show an unexpected tendency for subjects to choose disease Cn whenever there are three symptoms present. For example, for symptoms Ii+PCi+PRi, there seems to have been a shift away from choosing disease Ci (base rate consistency) to disease Cn. Similar shifts can be seen in the last two rows of Table 6, which both involve stimuli with three symptoms. That tendency might explain the huge base-rate consistency effect observed for In+PCn+PRn ($\chi^2 = 100.85$ $p < .001$), which was one of the training exemplars, and the weak base-rate consistency effect observed for Ii+PCi+PRi ($\chi^2 = 1.134$ *n.s.*).

In retrospect, that shift can be handily explained as a response strategy that subjects easily could induce. In the training stimuli (see Table 5), only one of the eight exemplars had three symptoms; moreover, that particular stimulus was relatively difficult to learn because of its similarity to the corresponding rare disease. A subject in search of a distinguishing feature for that case would easily notice that there are three symptoms instead of the usual two, and thereby induce the response tendency, "If there are three symptoms, then diagnose the case as disease Cn."

Frequency Estimation

Two of the 56 subjects did not give frequency estimates that summed to 120, but their data were included in the analysis because they were qualitatively indistinguishable from the others.

Frequency estimates for the base-rate neglect and inverse base-rate effect sub-structures were initially considered separately. For each subject, the frequency estimates were transformed into log(Cn/Rn) and log(Ci/Ri) to yield approximately normal distributions. The means for the "neglect" and "inverse" sub-structures did not differ (0.687 vs. 0.690). Therefore the estimates were combined into a single transformed value for each subject, log((Cn+Ci)/(Rn+Ri)). Its mean value was 0.6870, which is reliably larger than zero ($t(55) = 14.84$, SE $= 0.0463$, $p < .0001$), and corresponds to a frequency ratio of 2.0:1. That frequency ratio is nearly equal to the 2.1:1 ratio obtained in Experiment 1 ($t(110) = 0.62$, pooled stdev $= 0.422$, $p = 0.54$).

There was, however, a difference in the total frequencies attributed to neglect and inverse substructures (57.4 vs. 62.8, $t(55) = 2.16$, SE $= 2.51$, $p = .035$). This reflects the difference in learning between the two sub-structures; the inverse sub-structure was learned faster, and was perceived as occurring more frequently.

As stated before, the exact value of the estimated frequency ratio should be interpreted cautiously, as it is affected by the response scale and anchoring. What can be concluded is that subjects were well aware of the different base rates of the categories, for both the neglect and inverse sub-structures, and the perceived base rate ratios were the same for both sub-structures.

Discussion

Experiment 3 can be interpreted as demonstrating that apparent base-rate neglect is an attenuated case of the inverse base-rate effect. In both the inverse base-rate effect and in apparent base-rate neglect, the subject encodes one of the symptoms as distinctive of the rare disease, but the magnitude of the distinctiveness differs.

The main implication is that the same mechanism underlies both sets of phenomena, and that any model which purports to account for apparent base-rate neglect should also account for the inverse base-rate effect. Indeed, this experiment demands that any such model must account for both effects simultaneously, within the same experiment and same subjects.

Experiment 4: Further exploration of apparent "base rate neglect"

The emphasis of the previous three experiments was on the learning of distinctive features as the underlying mechanism of both the inverse base-rate effect and apparent base-rate neglect. Experiment 4 explores more subtle combinations of symptom probabilities, with several purposes in mind. First, it demonstrates that a feature can be learned as distinctive of a category even if the feature rarely occurs in the category. Second, it explores the role of features that are not strongly associated with any category. Third, it provides data for constraining models of the interaction between base rate knowledge and feature-category knowledge. Fourth, it expands the generality of the empirical phenomena, and the explanatory principles, by investigating probabilistic categories, as opposed to the deterministic mappings used in Experiments 1–3. Fifth, it uses non-dimensional substitutive features (defined below), as opposed to the present/absent features in the previous experiments.

Method

Abstract training structure

Experiment 4 was again a medical diagnosis task, this time involving just two diseases and three pairs of symptoms. The common disease C occurred in 75% of the training trials, and the rare disease R occurred in the remaining 25% of the training trials. On any given training trial, a disease was chosen randomly according to the base rates, and then symptoms were selected according to the conditional probabilities shown in Table 7. Symptoms occurred

in mutually exclusive pairs, so that on any given trial either symptom "s" would appear or symptom "-s" would appear, but not both. Thus every training trial had three symptoms. This type of feature structure has been called "substitutive" (Tversky, 1977) because symptoms s and -s can substitute for each other, but cannot be present simultaneously.

Table 7: Conditional probability of symptoms given diseases in Experiment 4.

	Disease	
Symptom	C	R
s1	.8667	.8667
-s1	.1333	.1333
s2	.2000	.6000
-s2	.8000	.4000
s3	.1333	.4000
-s3	.8667	.6000

The design evokes several interesting questions that could not be addressed in the probability structure used by previous researchers. Notice first that for *both* symptoms s2 and s3, the normative probability of the rare disease is 50%; i.e., $p(\mathrm{R}|\mathrm{s2}) = p(\mathrm{R}|\mathrm{s3}) = .50$. It is of interest to discover if both symptoms produce apparent base-rate neglect, and to what extent. If one supposes that subjects confuse the probability of the disease given the symptom, $p(\mathrm{D}|\mathrm{S})$, with the probability of the symptom given the disease, $p(\mathrm{S}|\mathrm{D})$, as suggested by some research (for a brief review see Hamm (1994)), then subjects will exhibit apparent base-rate neglect for symptom s2, because $p(\mathrm{s2}|\mathrm{R}) > .50$, but subjects will *not* exhibit apparent base-rate neglect for symptom s3, because $p(\mathrm{s3}|\mathrm{R}) < .50$.[3]

A second question regards what subjects will learn about symptoms s1 and -s1. Both of those symptoms occur with equal conditional probabilities in the two diseases. Will subjects associate each symptom equally with the two diseases? What will subjects learn about symptom -s1, which occurs fairly rarely overall?

A third question regards comparisons of symptoms with equal conditional probabilities. For example, symptoms s1 and -s3 both have conditional probability of 0.8667 for the common disease. Will the two symptoms be equally strongly associated with the common disease? Similarly, symptoms s2 and -s3 both have conditional probability of 0.60 for the rare disease. Will they be equally strongly associated with the rare disease?

[3]This discussion is based on the assumption that subjective probabilities of symptoms given diseases match the objective probabilities, which is probably not true (e.g. Cobos et al., 1993).

The explanatory principles espoused in this article make the following predictions: Subjects should first learn the typical symptoms of the common disease. That is, they will learn that symptoms s1, -s2, and -s3 are typical of the common disease. Subsequently, they will look for distinctive features of the rare disease, and learn that symptoms s2 and s3 are distinctive of the rare disease, with the distinctiveness of s2 encoded a bit more strongly than that of s3, because s2 occurs more often. Symptom -s1, however, is so rare that it will not be actively encoded for either disease, and so when it occurs, it will merely add uncertainty to the subject's decision. It will be seen that the uncertainty decreases the influence of the base rates, rather than increasing it.

Physical instantiation of the stimuli

There has been some discussion in the literature on the relative merits of present/absent and substitutive features in these types of experiments (e.g. Gluck & Bower, 1988; Nosofsky et al., 1992; Shanks, 1990a, 1990b). One potential problem with present/absent features is that in test trials it might not be clear whether the lack of a symptom means that there is positive information that the patient does not have that symptom, or whether the lack of the symptom means merely that we have no information, positive or negative, about that symptom. The same ambiguity affects computation of normative probabilities when using Bayes' Theorem.

Substitutive features have been used by some experimenters (e.g., Gluck & Bower 1988, Exp. 3; Nosofsky et al. 1992) to avoid the ambiguity of present/absent features. But substitutive features introduce another potentially confounding factor that has not been previously addressed. In previous experiments, the substitutive features have been *dimensionally labeled*, designating alternative values on an explicit dimension; e.g., stuffy nose vs. runny nose. Now, suppose that a subject has learned that stuffy nose is typical of disease C, but has yet to associate runny nose with either disease. Some subjects might infer, by virtue of the dimensional polarity, that if stuffy nose is a symptom of one disease, then runny nose must be a symptom of the other disease. That type of inference goes beyond the basic associative learning principles presumably being studied in these experiments, and, in the interests of simplicity in design and theory, is to be avoided.

On the other hand, it is presumptuous to believe that subjects treat the substitutive features as substitutive at all. Rather, it is quite possible that subjects interpret "stuffy nose" and "runny nose" as two, distinct, present/absent symptoms. Indeed, I believe that to be the case for most subjects. Thus, a subject could, in principle, be presented with a test case that includes both stuffy nose and runny nose, make a diagnosis as in any other case, and only sub-

sequently realize that two of the symptoms were mutually exclusive.

To explore these issues, Experiment 4 used the probability structure of substitutive symptoms, as described in Table 7, but used symptom labels that did not explicitly mark the the mutually exclusive alternatives. The six symptoms labels were "nose bleeds," "spastic knee," "blurred vision, "hair loss," "swollen tonsils," and "stomach cramps." The assignment of labels to abstract symptoms was randomly permuted for each subject. At the end of the experiment, the subjects were asked to guess which symptoms were mutually exclusive, as described below. Whether or not this approach disambiguates the issues broached above, it is, nevertheless, a novel design with results that should be addressed by candidate theories.

On each trial, the symptoms were presented in a vertical list, as in the previous experiments. The display order was also randomized on each trial, so that on one trial symptom s1 might appear above symptom s2, but on another trial symptom s2 would appear above symptom s1. Every training trial had three symptoms displayed (because of the three pairs of mutually exclusive symptoms). Test trials, which often involved missing symptoms, displayed the missing symptoms as a "?" at that position in the vertical list. The location of missing symptoms was also random, so that a "?" could appear at any position in the list of symptoms.

Category labels were "F" or "J," randomly assigned to the common or rare diseases for each subject.

Overall sequence of training and testing

For each subject, a new random sequence of 200 training trials was generated. The sequence had to respect all the conditional probabilities in Table 7, and the base rates of the diseases, to within a margin of error of $\pm 3\%$. For example, the intended base rate of the common disease was 75%, and a candidate random sequence was rejected if it had more than 78% or fewer than 72% common disease trials. Thus each subject, individually, was trained on a sequence that respected all the ordinal probability relationships in the intended design.

The instructions did not indicate that symptoms occurred in mutually exclusive pairs, but did forewarn the subject that two patients with the same symptoms might have different diseases. Subjects received 200 training trials, uninterrupted. Immediately after training, they received instructions telling them that they would see cases with some symptoms "missing or unknown," marked with a "?", and that they should make their best guess based on what they learned before. They were also told that they would not be told the correct diagnosis, but that their best educated guess was very important, so they shouldn't just respond randomly. Subjects were shown two repetitions of 27 test cases, generated by crossing three symptom values s, -s, and s-missing for the three symptom pairs. In particular, one of those combinations had all three symptoms missing, and subjects were thereby probed about their choice base-rates.

After testing, and while still in the computer cubicle, subjects were instructed to fill out a paper questionnaire, concealed until that time inside an envelope beside the computer. The questionnaire asked the subject to estimate the frequency of the diseases, and they were told that the total number of training cases was 200, and that equal frequencies would mean 100 cases of each disease. The questionnaire also asked them to judge which symptoms were mutually exclusive. It listed the six symptom labels, and for each symptom, asked the subject to write in another symptom that never occurred in the same patient. The subject also had to fill in a code word, displayed on the computer screen, that the experimenter could use to determine the random permutation of disease labels and symptom labels for that subject.

Sequence of events for an individual trial

Each trial was the same as described in previous experiments, with two minor changes. First, the maximum time to make a diagnosis was reduced from 30 seconds to 15 seconds. Second, when the subject made an incorrect response, the feedback message was a bit more elaborate: "WRONG! Try to be as accurate as possible for BOTH diseases." This was intended to better motivate the subject, as the probabilistic mapping can be rather defeating and frustrating.

Subjects

71 subjects participated for partial credit in an introductory psychology course at Indiana University. None of the subjects had participated in any other related experiment in my lab.

Results and Discussion

Training

The actual sample probabilities, averaged across 71 subjects, are shown in Table 8. Sample base rates were $p(\mathrm{C}) = .7486$ and $p(\mathrm{R}) = .2514$. Along with the constraints on individual sequence generation described earlier, we can be confident that subjects, both individually and collectively, were exposed to a representative sample of the intended probabilities.

The primary question for the training phase is, Did subjects learn about the common disease before they learned about the rare disease? In the previous experiments the diseases were deterministically related to the symptoms,

Table 8: Actual sample probabilities of training stimuli in Experiment 4.

	Disease	
Symptom	C	R
s1	.8636	.8683
-s1	.1364	.1317
s2	.1990	.6000
-s2	.8010	.4000
s3	.1322	.4056
-s3	.8678	.6944

so we could *presume* that if subjects learned anything about the diseases, their accuracy would improve. Indeed, their accuracy could reach 100% correct, and did for many subjects. Because accuracy on both the common and rare diseases eventually rose to the same, nearly perfect, level, it was reasonable to measure the relative learning by the relative accuracies.

In the present experiment, however, the diseases are only probabilistically related to the symptoms, and accuracy is not necessarily a good indicator of whether the subject has learned anything. By "learning" I mean a change from one pattern of responding to some other systematic pattern of responding, regardless of whether overall accuracy is thereby improved.

The theory suggests that subjects first learn the typical symptoms of the common disease, and subsequently learn the distinctive symptoms of the rare disease. Thus, subjects should first learn that symptoms s1, -s2, and -s3 are typical of the common disease, and subsequently learn that symptoms s2 and s3 are distinctive of the rare disease (with symptom -s1 associated with neither disease). The hypothesis, therefore, is that subjects will tend to diagnose s1,-s2,-s3 as the common disease, early in training, and subsequently tend to diagnose -s1,s2,s3 as the rare disease.

Table 9 shows the proportion of common-disease diagnoses for the two key training cases, and across all cases. The table also shows the number of instances of each type to which subjects responded. Evidently, subjects quickly learned that the base rates of the diseases were unequal, insofar as their overall rate of "common" diagnoses was greater than 50%. Subjects also learned that cases of symptoms s1,-s2,-s3 were usually the common disease (the programmed, "population" probability was $p(\text{c}|\text{s1,-s2,-s3}) = .897$), and that cases of symptoms -s1,s2,s3 were usually the rare disease (the programmed, "population" probability was $p(\text{c}|\text{-s1,s2,s3}) = .250$). Unfortunately, there were not enough occurrences of -s1,s2,s3, early enough in training, to reject the null hypothesis that the common and

Table 9: Proportion of cases diagnosed as the common disease in Experiment 4.

	s1,-s2,-s3		overall		-s1,s2,s3	
trials	p(C)	N	p(C)	N	p(C)	N
1–20	.844	706	.671	1412	.389	18
21–40	.892	740	.732	1419	.214	14
41–60	.889	687	.718	1419	.100	10
61–80	.913	710	.762	1420	.454	11
81–100	.912	714	.749	1419	.208	24
101–120	.912	725	.772	1419	.500	10
121–140	.921	700	.798	1419	.409	22
141–160	.923	726	.804	1420	.250	12
161–180	.899	721	.773	1420	.428	14
181–200	.903	699	.790	1419	.500	14

rare tendencies were learned at the same time (the programmed, "population" base rate of the exemplar is only $p(\text{-s1,s2,s3}) = .011$). Even in the first block of 20 trials, subjects are diagnosing -s1 s2 s3 as the rare disease significantly more often than would be expected by the baseline response rate (expected $r = (1-.671)\times 18 = 5.92$, actual $r = 11$, $p < .013$).

Table 9 also makes it clear that subjects responded differently depending on the symptoms; i.e., they did not just ignore the symptoms and respond randomly according to the disease base rates.

An unfortunate consequence of using just two diseases is that we cannot be sure *why* subjects tended to choose the rare disease when presented with symptoms -s1,s2,s3. Subjects might have actively encoded those symptoms as indicative of the rare disease, or instead might have simply inferred that the symptoms were not indicative of the already-learned common disease. All previous research on apparent base-rate neglect suffers the same ambiguity because it has always used just two categories. Future experiments would do well to include three or more categories.

Testing

Table 10 shows the proportion of rare choices for each test case, out of 142 trials per case (71 subjects times 2 test trials per subject per case). As shown in Table 10, when all three symptoms were missing, subjects chose the rare disease only 23.2% of the time, almost exactly matching the base rates of the categories.

On the other hand, subjects also exhibited apparent base-rate neglect for both symptoms s3 and s2, choosing the rare category 58.4% of the time for s3 ($\chi^2 = 4.056$ $p < .05$, using 50/50 for expected values), and 67.6%

for s2 ($\chi^2 = 17.606$ $p < .001$). The magnitudes of the apparent base-rate neglect for the two symptoms do not differ significantly from each other ($\chi^2 = 2.554$ *n.s.*).

Table 10: Results of test trials in Experiment 4. Note: Table shows proportion of responses for rare disease.

	s1		
s2, s3	-	?	+
-, -	.148	.078	.084
?, -	.157	.078	.141
-, ?	.204	.106	.071
?, ?	.394	**.232**	.141
-, +	.433	.418	.246
+, -	.514	.352	.331
?, +	.514	**.584**	.422
+, ?	.655	**.676**	.408
+, +	.667	.782	.704

The data also reveal influences of individual symptoms that appear consistent across all combinations of other symptoms. For example, symptom +s1 seems consistently to influence choices toward the common disease, in that rare-choice probabilities for +s1 are consistently less than for ?s1 (compare the third and fourth columns of Table 10). More robust effects are shown by symptoms -s2 and -s3, both of which also consistently influence choices toward the common disease. On the other hand, symptoms +s2 and +s3 consistently influence choices toward the rare disease.

Symptom -s1 has a more subtle consistency, unlike the other symptoms. Whenever s1 is missing (i.e., column ?s1), and the other symptoms make subjects' choices tend toward the common disease (as in the first several rows of Table 10), then adding symptom -s1 makes the choice proportion shift toward the *rare* disease. But, when the other symptoms make subjects' choices tend toward the rare disease (as in the bottom rows of Table 10), then adding -s1 makes the choice proportion shift toward the *common* disease. The interaction is significant ($\chi^2_{(8)} = 20.848$ $p < .0125$), but what makes it interesting is the *crossover*. My interpretation of that result is that symptom -s1 is not actively encoded as part of either disease, and merely adds uncertainty when it appears, shifting the choice toward random guessing.

Frequency Estimation

As in the previous experiments, the value of log(C/R) was computed for each subject, and the mean was 1.012, which is significantly different from zero ($t(70) = 17.70$, SE = 0.0571, $p < .0001$), and corresponds to a frequency ratio of 2.7:1. Again, that ratio should be interpreted cautiously, as its exact value is affected by the scale (0-200) and anchoring of the estimates. In any case, it is clear that subjects were well aware of a difference in base rates.

Judgments of Mutually Exclusive Symptoms

Subjects found it rather difficult to guess the mutually exclusive symptoms, and several subjects spontaneously reported to the experimenter that they were just guessing, and several subjects left some response fields blank. No subject wrote in the same symptom as the probe symptom, so apparently they *did* understand the task.

For purposes of inferential statistics, the expected response frequency for each symptom was estimated as the average across all probe symptoms. Of the 391 responses (6 probe symptoms times 71 subjects, minus 35 blank responses), 84 were s1, 51 -s1, 57 s2, 70 -s2, 54 s3, 75 -s3. Those response frequencies are highly positively correlated with the base rates of the symptoms in the training trials, which were, respectively, .288, .045, .100, .233, .067 and .266 (98% of the variance in response probabilities is accounted for by the linear correlation with the base rates of the symptoms). To the extent that responses match the base rates, we have evidence that subjects were indeed just guessing. (If subjects judged the mutually exclusive symptoms perfectly, then all symptoms would be generated equally frequently in the responses.) For each symptom, the observed response frequency for the mutually exclusive symptom was tested against the expected value given by the overall frequencies. For example, when probed with symptom s1, 14 subjects indicated that -s1 was the mutually exclusive symptom, and 53 subjects indicated other symptoms (none of which were s1 itself). The expected frequency of responding -s1 is 11.13, and the expected frequency of other responses is 55.87, yielding $\chi^2_{(1)} = 0.8875$, n.s. For the other probe symptoms, χ^2 values for the correct response were as follows: -s1 $\chi^2 = 9.90$ $p < .005$; s2 $\chi^2 = 0.238$ *n.s.*; -s2 $\chi^2 = 1.26$ *n.s.*; s3 $\chi^2 = 2.21$ *n.s.*; -s3 $\chi^2 = 5.15$ $p < .025$.

The symptom with the largest χ^2, -s1, was the symptom with the lowest overall base rate, and its complementary symptom, s1, had a very small χ^2 value. Moreover, symptom -s3 had a significant χ^2 value, but its complement, symptom s3, did not. It appears, therefore, that some subjects might have had a weak sense of which symptoms were mutually exclusive, but most subjects did not. It is quite possible that subjects had no explicit knowledge that symptoms s1 and -s1 were mutually exclusive, but instead used a response strategy for -s1 as follows: "I've rarely seen -s1, so its 'opposite' must occur a lot. So I'll choose the most common symptom, s1." The converse logic wouldn't necessarily generate a significant χ^2 for s1,

because subjects had not actively encoded -s1: "Symptom s1 occurs a lot, so its 'opposite' is probably rare. The rare symptoms *I know about* are s2 and s3, so I'll choose one of those." I interpret these results as indicating that subjects most likely did not respond in training or test by using non-associative inference based on dimensional polarity.

In summary, Experiment 4 demonstrated that apparent base-rate neglect can be obtained even for symptoms that occur in less that 50% of the cases of the rare disease. Experiment 4 also demonstrated that the presence of symptoms that have not been associated with any disease seems merely to increase the tendency toward random guessing. Experiment 4 also provided a rich data set for quantitative modeling.

Modeling

The experiments provided converging, confirmatory evidence supporting the main theses of this article, that the roles of base rates in the inverse base-rate effect and apparent base-rate neglect are (i) to cause the common category to be learned first, and consequently cause the rare category to be learned in terms of the features that distinguish it from the already-learned category, and (ii) to consistently bias responding to match the base rates, to the extent that other cues are absent.

In the remainder of the article I describe a new connectionist model that implements those principles. The model is fit to the data from the four experiments. The logic of this exercise is the same as that used in the empirical studies: If the model fits the data reasonably well, we have confirmatory evidence that the principles it embodies, and their formalization in the model, are accurate representations of human learning.

If the principles are correct, then it should be possible to implement them in other formalisms (e.g., production systems) and still match the data fairly well. Thus, the explanatory power comes from the underlying principles embodied in the model, not necessarily from their particular formal expression. Nevertheless, this particular formal model is the only model, of which I am aware, that can quantitatively fit both the inverse base-rate effect and apparent base-rate neglect. The model is called ADIT[4]

A new connectionist model

The model is an extension of the component-cue model of Gluck and Bower (1988). In the component-cue model, there is one input node per feature/symptom, and one output node per category/disease. If a feature is present, the input node has activation 1.0, otherwise it has activation 0.0. The activation of the k-th output node represents how strongly the network classifies the current input into category k. All input nodes have connections to all output nodes, with initial connection weights of 0.0. A connection weight represents the learned degree of association between the feature and the category; i.e., the weights represent the long-term category knowledge of the network.

There are two main extensions of the component-cue model in ADIT. The first extension implements the principle of shifting attention to distinctive features. The second extension implements the influence of base-rate knowledge.

The attention mechanism is motivated by the following intuitions: Suppose I've learned, somehow, that features A and B are typical of category X. If a new stimulus appears, containing features A and C, and I correctly classify the stimulus as category X, then I should amplify my current knowledge of X. That is, I should *not* necessarily modify my knowledge on the basis of the current stimulus (which includes feature C but not feature B) but I should instead just amplify my current knowledge (involving just A and B), because my current knowledge produced the correct response. On the other hand, if I erroneously classify the stimulus (it is actually an instance of category Y), then I should try to determine what is distinctive about the present stimulus relative to the category I incorrectly predicted (namely, feature C), and alter my knowledge about category Y to include that distinctive feature. If I classify the stimulus merely by guessing, I should modify my knowledge by using the entire stimulus.

Those intuitions can formalized as follows. Each input node in ADIT has an *attention* strength, which determines how much the corresponding feature influences the classification decision. Specifically, the k-th output node has activation determined by

$$a_k^{out} = \sum_i w_{ki} \alpha_i a_i^{in} \tag{1}$$

where w_{ki} is the connection weight from input node i to output node k, α_i is the attention strength on input node i, and a_i^{in} is the activation of the i-th input node.

When a stimulus is presented to the network, the nodes representing features that are present have their activation set to 1.0, and the attention strengths are normalized so that as more features are present, the attention given to each one is lessened. Normalization of attention can be formalized many different ways; ADIT assumes that if the stimulus has N features, then the attention given to each feature is

$$\alpha_i = N^{-1/\eta} \tag{2}$$

[4] ADIT stands for Attention to Distinctive InpuT, and, like other models I've proposed such as ALCOVE and AMBRY, represents just a small part of the total mental architecture.

where η is a freely estimated parameter ($\eta > 0$). (This formula for attention strength normalization causes the "length" of the attention vector to be 1.0 when measured using a Minkowski-r metric, where $r = \eta$.) If η is large, each attention strength is nearly 1.0, even when many features are present. If η is small, the attention strength for individual features drops off rapidly as the number of features increases. I am not especially committed to this particular, ad hoc formalization, but I do believe some form of attention normalization is necessary to capture the empirical results.

After the appropriate feature nodes are activated and the attention strengths are normalized, the output nodes are activated according to Equation 1. Choice probabilities for each category are then determined by the relative activation of each node and the base-rate bias, according to a formula described later.

After the output nodes are activated, the network receives corrective feedback, like human subjects. Output node k receives a "teacher" value t_k determined by:

$$t_k = \begin{cases} \max(1, a_k^{out}) & \text{if } k \text{ is correct} \\ \min(0, a_k^{out}) & \text{if } k \text{ is erroneous} \end{cases} \tag{3}$$

Those teacher values are "humble" insofar as the activation of a node can exceed its teacher value (either less than 0.0 or greater than 1.0) without penalty (see Kruschke, 1992). Discrepancies between the teacher value and the actual output activations are used to modify attention and long-term association weights. Indeed, both the attention strengths and the weights are adjusted in the direction that maximally decreases the *error*

$$E = .5 \sum_k (t_k - a_k^{out})^2. \tag{4}$$

The attention and weight adjustments are not conducted simultaneously, however. First attention is adjusted, then category activations are re-computed using those new attention strengths, and then the weights are adjusted. That sequence is based on my intuition that subjects, after receiving feedback, first decide which features in the stimulus are relevant to the present case, and only then adjust their knowledge on the basis of those attended features.

Attention strengths are adjusted in the direction (opposite) of the gradient of the error:

$$\begin{aligned} \Delta\alpha_i &\propto -\partial E/\partial\alpha_i \\ &= \lambda_\alpha \left[\sum_k \left(t_k - a_k^{out}\right) w_{ki} \right] a_i^{in} \end{aligned} \tag{5}$$

where λ_α is a freely estimated constant of proportionality, called the attention learning rate. If attention strengths are driven to negative values by Equation 5, then they are reset to zero, because negative attention strengths might not have a clear psychological interpretation. After the attention strengths are adjusted, they are *not* renormalized, because the network must be allowed to attenuate its attention, overall, if its knowledge conflicts entirely with the current stimulus and feedback. This is based on my intuition that when subjects are confronted with trials that contradict previous learning, they suspend judgment, or reduce their commitment to the information in the current trial.

Equation 5 captures rather precisely the intuitions expressed earlier about what one should do with one's knowledge after getting corrective feedback. Recall that the intuitions had two parts, one saying that existing knowledge should be attended when the response is correct, and the other part saying that distinctive features should be attended when the response is in error. These two parts can be seen by de-composing Equation 5 as follows. If feature i is present, then $a_i^{in} = 1.0$, and so

$$\Delta\alpha_i \propto \left(\tilde{1} - a_c^{out}\right) w_{ci} - \sum_{e \neq c} a_e^{out} w_{ei} \tag{6}$$

where c is the index for the *correct* category, e is an index ranging over the *erroneous* categories, and $\tilde{1}$ is shorthand for the humble teacher for the correct category (as defined in Equation 3). The first term in Equation 6 amplifies attention to the existing knowledge of the correct category, to the extent that the correct category was not fully activated. The second term reduces attention to features that match knowledge of erroneous categories, to the extent that those erroneous categories are activated, thereby leaving distinctive features attended to. The attention-shifting mechanism actively suppresses attention to features that are shared with incorrect categories (second term in Equation 6), but amplifies attention to distinctive features only indirectly, by adding attention to features already in the existing knowledge about the correct category (first term in Equation 6).[5]

Once the attention strengths have been modified to reflect the relevance of the features for the present trial, the category activations are recomputed to give (new) predictions based on the current category knowledge and the attended features. At that point, the connection weights are adjusted, again in the direction (opposite) of the gradient on error:

$$\begin{aligned} \Delta w_{ci} &\propto -\partial E/\partial w_{ci} \\ &= \lambda_w \left(t_c - a_c^{out}\right) \alpha_i a_i^{in} \end{aligned} \tag{7}$$

[5]The interpretation of Equation 5 offered here is not post hoc; in fact I started with some other ad hoc formalizations of the two-part intuition about how attention should be adjusted, addressing each part separately, only to realize later that they could be unified in the gradient descent formalization explained here.

where λ_w is a freely estimated constant of proportionality, called the weight learning rate. Only the weights to the correct category are adjusted. This reflects my intuition that human subjects only modify their knowledge about the category indicated by the corrective feedback, especially in situations when there are multiple categories. For example, when told that the current stimulus is a case of category X, it is difficult to imagine that subjects then also actively conclude that it's not a U and it's not a V and it's not a W and it's not a Y, etc. Thus, ADIT assumes that subjects use negative feedback to shift attention away from erroneously predicted categories, but not to adjust long term knowledge of those erroneous categories. (It is possible that subjects do learn about more than one category per trial in some situations; this is an avenue for future research.)

The second main extension of the component-cue model involves the influence of base rates on choice probabilities. There are two components involved. One component converts the category-node activations to "unbiased" choice probabilities, and the second component modifies those probabilities according to the learned base rates of the categories.

Category-node activations are mapped to "unbiased" choice probabilities using the same mechanism as used in ALCOVE (Kruschke, 1992) and its predecessors. Category x receives a high choice probability, p_x, to the extent that its *relative* activation is high:

$$p_x = \exp(\phi a_x^{out}) \Big/ \sum_k \exp(\phi a_k^{out}) \tag{8}$$

where ϕ is a freely estimated scaling constant.

The network probabilities are then mixed with the base rates. The motivating intuition is that the influence of the base rates should depend on how much potential information is in the stimulus, so that if there are many features present, the base rates shouldn't have much influence, but if there are few or no features present, then the base rates should dominate. The formalism I used to express that is as follows. Let r_k be the proportional base rate of category k, as currently learned by the network. (The learning method for the base rate is described later. For now it can be assumed that r_k is an accurate reflection of the true base rate.) Let N be the number of features in the current stimulus. Then the "biased" choice probability is given as:

$$pr_x = p_x\, r_x^B \Big/ \sum_k p_k\, r_k^B \tag{9}$$

where

$$B = \beta/(\beta + N^{1/\eta}) \tag{10}$$

and β is a freely estimated constant called the base-rate bias. (Recall that η is the attention normalization power in Equation 2.) Therefore, when there are many features in the stimulus, N is large, so B is small, and the base rates have a relatively small influence on the choice. When there are few features in the stimulus, N is small, so B is large, and the base rates have a relatively large influence on the choice. In particular, when N is zero, the base rates completely determine the choice.

The formulas in Equations 9 and 10 are ad hoc, and I am not committed to their specific forms. I do believe, however, that the influence of the base rates must be modified in some way depending on the information available from other sources. What is most important about the present method for combining base rates with choice probabilities is that base-rate knowledge is applied uniformly across all training and test trials, and without regard to the content of the stimulus other than its number of features. It might ultimately turn out that this is not an accurate reflection of human decision making, but insofar as the present models fits human performance, we have evidence that base rate knowledge could be applied in this relatively simple manner.

The final detail that needs to be described is how the base rates are learned by the model. The values of r_k are initialized to equal values, i.e., one over the number of categories. They are then adjusted as follows:

$$\Delta r_k = \begin{cases} (1 - r_k)/(1 + T) & \text{if } k \text{ is correct} \\ (0 - r_k)/(1 + T) & \text{if } k \text{ is erroneous} \end{cases} \tag{11}$$

where T is the trial number. The values of r_k converge to the correct base rates very rapidly, within two training blocks (16 trials) in the simulations reported here. This formalism is undoubtedly an inaccurate reflection of psychological mechanism, and is probably too sensitive to early base-rates and too insensitive to changes in base rates in later trials, relative to humans. This formalism would probably have to be modified to fit accurately results from Medin and Bettger (1991), for example. All that matters for the present model is that base rate knowledge is acquired so it can be used in Equations 9 and 10. The formalism is attractive for the present purposes because it adds no new free parameters to the model, and makes no assertions about systematic distortions from the true base rates.

All told, there are five freely estimated parameters in the model: the attention learning rate (λ_α in Equation 5), the weight learning rate (λ_w in Equation 7), the attention normalization power (η in Equation 2), the base-rate bias (β in Equation 10), and the choice probability scaling constant (ϕ in Equation 9).

To recapitulate the sequence of events in the model: When a stimulus is presented, the nodes representing present features are activated, and attention is evenly distributed over all the presented feature nodes, normalized according to the number of features presented (Equation 2).

Table 11: Best fit of ADIT to test trials of Experiment 1.

	Choice Proportion			
Symptoms	C	R	Co	Ro
I	.746 (.001)	.167 (-.008)	.054 (.005)	.034 (.002)
PC	.911 (-.022)	.025 (-.006)	.040 (.009)	.025 (.020)
PR	.020 (-.021)	.948 (.038)	.020 (.002)	.012 (-.019)
PC+PR	**.356** (.003)	**.590** (-.022)	.033 (.010)	.022 (.009)
I+PC+PR	**.550** (-.031)	**.427** (.026)	.014 (.000)	.009 (.005)
I+PCo	.347 (-.059)	.107 (.027)	.514 (.045)	.032 (-.013)
I+PRo	.256 (.037)	.079 (-.006)	.035 (.004)	.630 (-.035)
PC+PRo	**.356** (.003)	.022 (-.005)	.033 (-.025)	**.590** (.027)
I+PC+PRo	**.739** (.020)	.035 (-.001)	.018 (-.018)	**.207** (-.002)

Activation is then propagated to the output nodes, and a category prediction is made according to the probabilities computed from the network, mixed with probabilities from the base rates. Corrective feedback is then supplied, and attention is consequently shifted away from features that cause error and toward features that reduce error, for that particular case (Equation 5). The network output is recomputed based on the new distribution of attention values, and then the weights for the correct category are adjusted to further reduce the error. The process then repeats for the next trial.

In summary, the main mechanism that allows the model to address the inverse base-rate effect and apparent base-rate neglect is the attention learning mechanism. Attention is adjusted so that features that match existing knowledge are attended to when the response is correct, and distinctive features are attended to when the response is erroneous. The attention adjustment is formalized as simple gradient descent on error. The other mechanism most important in the present context is the manner in which base rate knowledge is combined with other category knowledge. Base rates are applied consistently in both training and testing phases, and their influence depends only on the total number of features in the stimulus, not on the particular features or categories involved.

Fit to Experiment 1

Recall that Experiment 1 was a partial replication and extension of the inverse base-rate effect discovered by Medin and Edelson (1988). The goal of the model is to capture not only the inverse base-rate effect but also all the other test cases.

Model predictions were made by taking the mean test-trial choice probabilities of 200 simulated subjects (i.e., 200 different random orderings of training trials). Fits were measured as the root-mean-squared deviation (RMSD) between the 36 predicted choice probabilities and the 36 empirical probabilities. An initial grid search of parameter space was conducted, and then a hill-climbing search began at the best point found in the grid search. Best fitting parameter values were $\phi = 3.56$, $\beta = 0.761$, $\eta = 2.48$, $\lambda_w = 0.310$, and $\lambda_\alpha = 2.01$, yielding RMSD= 0.0216. Thus, the fit is excellent. Quantitative predictions are shown in Table 11, where numbers in parentheses show the difference between predicted and empirical values..

The model shows a robust inverse base-rate effect for PC+PR, and a base-rate consistency effect for I+PC+PR. Moreover, the model shows the increased base-rate consistency for I+PC+PRo. The model made those test-trial predictions with full knowledge of the base rates, as its internal values for the base rates stabilized at the actual base rates within two training blocks.

Table 12: Association weights for best fit to Experiment 1

to	from symptom					
disease	I	PC	PR	Io	PCo	PRo
C	.738	.884	.000	.000	.000	.000
R	.448	.000	1.213	.000	.000	.000
Co	.000	.000	.000	.738	.884	.000
Ro	.000	.000	.000	.448	.000	1.212

It is instructive to consider the association weights that developed by the time of testing. Table 12 shows the mean weights in the last training block prior to testing (averaged across 8 training trials and 200 simulated subjects). The inverse base-rate effect is produced by the fact that the weight from the distinctive symptom of the rare disease, PR, is much greater than the weight from the distinctive

Table 13: Best fit of ADIT to Experiment 2.

Symptoms	Choice Proportion			
	P	F	Po	Fo
I	.799 (-.003)	.172 (.054)	.014 (-.019)	.014 (-.033)
PP	.958 (.034)	.014 (.000)	.014 (-.024)	.014 (-.010)
PF	.004 (-.024)	.988 (.073)	.004 (-.015)	.004 (-.034)
PP+PF	**.305** (-.011)	**.647** (.001)	.024 (.005)	.024 (.005)
I+PP+PF	**.518** (.004)	**.456** (-.011)	.013 (-.001)	.013 (.008)
I+PPo	.381 (.002)	.151 (.075)	.434 (-.068)	.034 (-.009)
I+PFo	.256 (.048)	.102 (.002)	.023 (.009)	.619 (-.059)
PP+PFo	**.305** (-.069)	.024 (-.023)	.024 (.014)	**.647** (.078)
I+PP+PFo	**.718** (.020)	.055 (.008)	.018 (-.010)	**.209** (-.017)

symptom of the common disease, PC. The difference in weights is large enough to override the base-rate influence.

The model generates the asymmetric association weights because of its attention shifting mechanism. The common category occurs more frequently, so, early in training, the weights from I and from PC to C grow faster than the weights from I and from PR to R. When the rare training pattern I+PR is presented, the initial prediction (with equally distributed attention) is that disease C is likely, because C is activated by the weight from I to C. The feedback then indicates the correct diagnosis, R, and so attention to symptom I is decreased, by subtracting out the weights from I and PC to C, and attention to symptom PR is increased, by adding any existing weight from PR to R. The weights are then adjusted using the new distribution of attention.

Table 12 also reveals that none of the weights are negative. For example, there is no negative weight from symptom PR to disease C. And, importantly, there are no nonzero weights between pairs of disease; e.g., the weight from PC to Co is zero. That is a consequence of the fact that the model only adjusts weights for the correct category; i.e., weights are never decreased, only increased, up to the point when error is zero. To my intuition that accurately reflects subject's knowledge in these tasks; they learn about (some of) the symptoms present in the diseases, not about which symptoms are absent from the diseases. Indeed, it is well known that noticing the absence of a feature is far more difficult than noticing the presence of a feature (e.g. Hearst, 1991).

Fit to Experiment 2

Recall that Experiment 2 examined the effects of pretraining on two of the four diseases, following by full training on all four diseases, with equal base rates. The main results were that an analogue of the inverse base-rate effect occurred, and that base-rate consistency also obtained for the combined symptoms, where "consistency" in this case means approximately equal preference for the diseases.

ADIT was fit to the test-trial data of Experiment 2 in the same way as for Experiment 1, with only one change: When learning the base rates of the categories, the model re-initialized its trial number (T in Equation 11) to zero when starting the second phase of learning. This reflects the fact that subjects were explicitly told that two more diseases would be introduced in the second phase. Best fitting parameter values were $\phi = 4.72$, $\beta = 1.02$, $\eta = 1.36$, $\lambda_w = 0.273$, and $\lambda_\alpha = 1.09$, yielding RMSD= .0344. Best fitting predictions are shown in Table 13.

The model shows a robust analogue of the inverse base-rate effect for PP+PF. The model also shows the base-rate consistency effect for I+PP+PF, where that means approximately equal responding because the base rates are equal in this case. Moreover, the model shows the preference for disease P in the case of I+PP+PFo.

Table 14: Association weights for best fit to Experiment 2

to disease	from symptom					
	I	PP	PF	Io	PPo	PFo
P	.839	.880	.000	.000	.000	.000
F	.518	.000	1.128	.000	.000	.000
Po	.000	.000	.000	.838	.879	.000
Fo	.000	.000	.000	.518	.000	1.129

The model made those test-trial predictions with full knowledge of the base rates, as its internal values for the base rates stabilized at the actual base rates within two training blocks in each phase of learning. In fact, the model's base rate knowledge is "too good" compared with

Table 15: Best fit of ADIT to Experiment 3.

Symptoms	Choice Proportion			
	Cn	Rn	Ci	Ri
In	.642 (.005)	.231 (-.037)	.072 (.043)	.054 (-.011)
Ii	.070 (.034)	.052 (-.007)	.624 (-.155)	.254 (.129)
In+Ii	.385 (.026)	.119 (-.054)	.354 (-.011)	.142 (.040)
PCn	.878 (.045)	.036 (-.077)	.049 (.031)	.036 (.001)
PCi	.056 (.020)	.042 (.012)	.860 (-.032)	.042 (.000)
PCn+PCi	.503 (.015)	.019 (-.034)	.458 (.048)	.019 (-.028)
PRn	**.213** (.082)	**.687** (-.087)	.058 (.016)	.043 (-.010)
PRi	.031 (.025)	.023 (-.007)	.031 (.001)	.915 (-.020)
PRn+PRi	.053 (.029)	**.330** (.039)	.017 (-.019)	**.600** (-.049)
PCn+PRn	**.576** (.034)	**.392** (-.007)	.018 (.000)	.014 (-.028)
PCi+PRi	.012 (.000)	.009 (-.015)	**.331** (.010)	**.647** (.004)
In+PCn+PRn	.742 (-.133)	.250 (.143)	.004 (-.007)	.003 (-.003)
Ii+PCi+PRi	.004 (-.086)	.003 (-.021)	.511 (.029)	.482 (.078)
In+PCi	.317 (-.011)	.098 (-.057)	.561 (.061)	.025 (.007)
Ii+PCn	.599 (.064)	.023 (-.024)	.269 (-.076)	.108 (.036)
In+PRi	.203 (-.029)	.059 (-.030)	.021 (-.009)	.716 (.068)
Ii+PRn	.108 (.024)	.505 (-.001)	.277 (.012)	.110 (-.034)
PCn+PRi	.351 (.057)	.013 (-.017)	.017 (.011)	.620 (-.051)
PCi+PRn	.090 (.012)	.433 (-.025)	.457 (.035)	.020 (-.022)
In+PCn+PRi	.704 (.008)	.023 (-.013)	.007 (.007)	.267 (-.001)
Ii+PCi+PRn	.017 (-.120)	.161 (-.012)	.805 (.163)	.017 (-.031)

human subjects, who estimated the pre-trained diseases to occur slightly but significantly more often than the full-training diseases. The model could easily be modified so that its base-rate learning mechanism has a bit more "inertia" from the first phase of training (e.g, by re-setting the value of T in Equation 11 to a moderately high value at the beginning of the full-training phase), but that would entail more free parameters, and the improvement in fit would probably be only slight.

Table 14 shows the mean weights in the last training block prior to testing (averaged across 8 training trials and 200 simulated subjects). The weights are very similar to those obtained in simulating Experiment 1.

Fit to Experiment 3

Recall that Experiment 3 modified the design of Experiment 1 so it contained a hybrid of the paradigms for apparent base-rate neglect and the inverse base-rate effect. The base results were that the distinctiveness of symptom PRn for the rare disease Rn in the neglect sub-design was strong enough to produce apparent base-rate neglect, but not strong enough to produce an inverse base-rate effect.

The model was fit to the test-trial data of Experiment 3 in the same way as for Experiment 1, this time using 300 simulated subjects because of the additional combinatorial possibilities in the hybrid design. Best fitting parameter values were $\phi = 4.50$, $\beta = 0.36$, $\eta = 39.0$, $\lambda_w = 0.480$, and $\lambda_\alpha = 3.38$, yielding RMSD= .0511. Best fitting predictions are shown in Table 15.

The model shows a strong apparent base-rate neglect for symptom PRn, and shows that symptom PRn is not as strongly associated with disease Rn as symptom PRi is associated with disease Ri (see test case PRn+PRi). Moreover, the model shows a strong inverse base rate effect for the inverse sub-design (see test case PCi+PRi) but not for the neglect sub-design (test case PCn+PRn).

Table 16 shows the mean weights in the last training block prior to testing (averaged across 8 training trials and 300 simulated subjects). Importantly, symptom PRn has a non-zero association weight (.114) with the common disease Cn, and consequently has a relatively weak association with the rare disease Rn, compared to the association weights from symptom PRi to diseases Ci and Ri.

The model is obviously a bit strained to fit some aspects of the test data (and some of the parameters are distorted to extreme values). That is mostly because of the poor fits to test cases with three symptoms. Earlier it was argued that subjects might have learned that a distinctive feature of one

Table 16: Association weights for best fit to Experiment 3.

to	from symptom					
disease	In	PCn	PRn	Ii	PCi	PRi
Cn	.439	.681	**.114**	.000	.000	.000
Rn	.210	.000	**.608**	.000	.000	.000
Ci	.000	.000	.000	.452	.672	**.000**
Ri	.000	.000	.000	.287	.000	**.880**

of the cases of disease Cn was that it had three symptoms; hence, subjects would tend to choose Cn whenever confronted with a case with three symptoms. The model has no mechanism to implement that strategy. The other test case in which the model misfits the data is Ii, for which subjects chose Ci much more than the model predicts. In this case, however, it seems that the datum might be anomalous, not necessarily that the model is deficient, because if subjects truly associated symptom Ii with disease Ci very strongly, then they should choose Ci more often than Cn when tested with In+Ii, but they did not.

Fit to Experiment 4

In Experiment 4, there were three pairs of mutually exclusive symptoms. Those were represented in the model as six present/absent symptoms, with no explicit representation of their exclusivity. This is the same scheme used by Gluck and Bower (1988) and by Nosofsky et al. (1992). Recall that subjects appeared to have little knowledge of which symptoms were mutually exclusive, so this representation is reasonable.

The model was fit to the test-trial data of Experiment 4 by using the same 71 training sequences as were shown human subjects. Best fitting parameter values were $\phi = 1.20$, $\beta = 0.958$, $\eta = 1.54$, $\lambda_w = 0.315$, and $\lambda_\alpha = 1.09$, yielding RMSD= .0468. Best fitting predictions are shown in Table 17.

All of the major effects seen the data are captured by the model. When presented with an empty list of symptoms, the model matches the base rates of the categories. When presented with symptoms s2 or s3 alone, it shows apparent base rate neglect, with a stronger tendency toward the rare diagnosis for symptom s2 than for s3.

The model also nicely shows the cross-over interaction for symptom -s1 exhibited in the human data. When other symptoms lead responses to the common disease, adding -s1 increases the tendency toward rare; but, when other symptoms lead responses to the rare disease, adding -s1 decreases the tendency toward rare.

The model produces that cross-over interaction for symptom -s1 because the magnitude of the base-rate bias is modulated by the number of features in the input. When -s1 is added to the input, the base-rate bias is decreased, and -s1 itself adds no strong preference to the outcome.

Table 17: Predictions of ADIT for Experiment 4.

	s1		
s2, s3	-	?	+
-, -	.148 (.000)	.101 (.023)	.087 (.002)
?, -	.191 (.034)	.111 (.034)	.104 (-.037)
-, ?	.226 (.021)	.148 (.042)	.123 (.052)
?, ?	.363 (-.032)	**.251** (.019)	.173 (.032)
-, +	.393 (-.040)	.379 (-.040)	.270 (.023)
+, -	.406 (-.108)	.395 (.043)	.281 (-.050)
?, +	.567 (.053)	**.640** (.056)	.392 (-.031)
+, ?	.629 (-.025)	**.726** (.050)	.455 (.047)
+, +	.713 (.046)	.777 (-.004)	.589 (-.116)

Association weights for the best fit to Experiment 4, are shown in Table 18. The weights reflect the predictions made earlier by the general principles: The common disease is represented by symptoms s1, -s2 and -s3, and the rare disease is represented by the symptoms s2 and s3, with symptom -s1 being only weakly, and nearly equally, associated with both diseases.

In summary, the ADIT model provides a good fit to the test-trial data from all four experiments. It robustly shows the inverse base-rate effect, base-rate neglect, base-rate consistency, and the effects of features with no strong association to any category. It had only one notable deficiency in modeling three-symptoms cases in Experiment 3, which I suggested was caused by subjects inventing a response strategy not implemented in the model.

Previous models

A variety of category-learning models previously have been proposed to address the inverse base-rate effect and/or apparent base-rate neglect, but none of them has been shown adequately to capture all the test-trial data associated with both effects. The models are briefly reviewed in this section, with indications of how their explanatory principles differ from those embodied in ADIT.

Context Model and Extensions

The *context model* (Medin & Schaffer, 1978) uses an exemplar-based representation of category content, with classification decisions based on summed similarity to all exemplars of each category. In the generalized context model (Nosofsky, 1986), similarity is computed using a

Table 18: Association weights for best fit to Experiment 4.

to	from symptom					
disease	s1	-s1	s2	-s2	s3	-s3
C	1.389	.362	.256	1.247	.231	1.543
R	.377	.321	1.556	.167	1.171	.199

city-block metric and exponentially decaying generalization gradient. Each input feature is differentially weighted by an attention strength, so that more strongly attended features affect the similarity computation more strongly.

Medin and Edelson (1988) suggested that the inverse base-rate effect could be accommodated by a modified context model if individual exemplars had individual attention strengths. Thus, symptoms in training exemplar I+PC would each have moderate attention strengths, but in training exemplar I+PR, symptom PR would have much higher attention, and symptom I much lower. When tested with conflicting symptoms PC+PR, the high attention to PR would cause a better match to the rare-disease exemplar, thereby producing the inverse base-rate effect. The basic intuitions underlying that modified context model are very similar to those underlying ADIT. Unfortunately, no process model was introduced by which the different attention strengths could be established in the context model, nor were quantitative predictions made for the full range of test cases.

The generalized context model was supplied with a learning mechanism for attention strengths and association weights in ALCOVE (Kruschke, 1992, 1993a, 1993b). The attentional learning in ALCOVE allows it to fit a number of human learning phenomena. ALCOVE used a single set of attention strengths for all exemplars, however, and is therefore unable to account for the inverse base-rate effect (Kruschke, 1992; Nosofsky & Kruschke, 1992).

On the other other hand, quantitative fits showed that ALCOVE was able to produce apparent base-rate neglect (Kruschke, 1992; Nosofsky et al., 1992), whereas the (generalized) context model could not. Unfortunately, that modeling result was subsequently shown to be a consequence of the particular training sequence employed in the experiment (originally used by Estes et al. (1989)), and ALCOVE cannot show apparent base-rate neglect, on average, unless additional modification are made (Kruschke, 1994; Lewandowsky, 1994).

One should *not* conclude from those failures that no exemplar-based model can exhibit apparent base-rate neglect and the inverse base-rate effect. On the contrary, I believe that an exemplar-based model such as ALCOVE could show the appropriate effects if (1) it uses the attentional learning mechanisms described in this article and (2) uses a more appropriate *feature-based* input representation and similarity function instead of the dimension-based representation and similarity function presently in the model.

Rational Model

Closely related to the context model is the *rational model* of Anderson (1990). A special case of the rational model stores individual exemplars and computes normative probabilities of each category on the basis of how similar the input is to the stored exemplars, and is basically identical with the context model (Nosofsky, 1991). More generally, the rational model can collapse some exemplars together into clusters represented by cluster prototypes, and compute normative classification probabilities based on the similarity of the input to those multiple prototypes.

Anderson (1990, pp.117–120) showed that the rational model can produce some aspects of the inverse base-rate effect, despite the model's normative probability computations. In particular, he showed that the model can show a preference for the rare disease when tested with symptoms PC+PR, but show a preference for the common disease when tested with symptoms I+PC+PR. The model produces those results by creating a similarity gradient on the common-disease training exemplar, I+PC, that is very tall but drops off rapidly with distance, and a similarity gradient on the rare-disease training exemplar, I+PR, that is relatively short but drops off very gradually with distance. Therefore, test cases such as PC+PR that are far from both exemplars are dominated by the less-rapidly decaying rare-disease similarity, but test cases such as I+PC+PR that are near both exemplars are dominated by the higher common-disease similarity.

Unfortunately, Anderson (1990) did not address the remaining test data reported by (Medin & Edelson, 1988); e.g., symptom combinations such as I+PCo. In that case, the model predicts persistent inverse base-rate effects, because the test case is far from both training exemplars, and the more slowly decaying rare-disease similarity will dominate. The data, on the contrary, show that subjects do not exhibit an inverse base-rate effect, instead choosing disease C much more than disease R.

The rational model cannot exhibit apparent base-rate neglect, either. Anderson (1990, pp.120–125)

tried to dismiss the effect by arguing that subjects in Gluck and Bower's (1988) experiment were confusing p(disease|symptom) with p(symptom|disease), and generally that the behavior of subjects in the test phase is not to be weighed very heavily; instead, it is the training data that matter. The rational model predicts probability matching, and, insofar as the training data exhibit probability matching, the rational model fits well. However, after numerous replications of apparent base-rate neglect manifested in different test procedures, failure to model the effect can no longer be dismissed.

In summary, the rational model cannot properly address either the inverse base-rate effect or apparent base-rate neglect. The rational model probably needs some form of attentional learning, like that implemented in ADIT, to accommodate these effects (cf. Kruschke, 1993b).

Exemplar Fragment Storage (CLEM)

Myers et al. (1994) recently proposed an exemplar-based model that accounts for some aspects of the inverse base-rate effect and apparent base-rate neglect. Their model is an extension of Chumbley's (1986) CLEM (concept learning by exemplar memorization) model. The model probabilistically remembers fragments of the current exemplar when it has made an error, or guessed, or made a correct response from weak evidence. The model probabilistically forgets memory traces whenever they match an input. (The interested reader can consult the original article for details.)

Myers et al. (1994) showed that a revised version of CLEM can account for some aspects of results from the inverse base-rate paradigm; in particular, for test cases I, PC+PR, and I+PC+PR. The key aspect of CLEM that causes the inverse base-rate effect is greater forgetting of PC$\mapsto$C traces than of PR$\mapsto$R traces. Simultaneously, the base-rate consistency effect is produced by a greater number of I$\mapsto$C traces than I$\mapsto$R traces, produced by encoding when a correct response is based on weak evidence. Myers et al. (1994) did not discuss any other test cases. It would be informative to know whether or not CLEM can fit the data for all categories in all test cases. The experiments reported in this article provide additional challenges for CLEM.

If CLEM can fit the data presented here, then there is a need to develop new experimental designs to distinguish it from ADIT, because the models are based on rather different explanatory principles. In ADIT, both the inverse base-rate effect and apparent base-rate neglect are caused by shifting attention to distinctive features when the predicted category is wrong. Indeed, one of the main points of this article is that apparent base-rate neglect is just an attenuated case of the inverse base-rate effect. In CLEM, apparent base-rate neglect is caused by storing exemplar fragments primarily when there is an erroneous prediction, so that the distinctive symptom of the rare disease is stored in more traces for the rare disease than traces for the common disease. The inverse base-rate effect, on the other hand, is caused by a subtle trade-off of forgetting more frequently accessed traces and encoding exemplar fragments when correct responses are based on weak evidence. CLEM thereby treats the effects as stemming from different mechanisms.

Component-Cue Model and Extensions

Gluck and Bower (1988) proposed the component-cue model to account for apparent base-rate neglect. The component-cue model produces apparent base-rate neglect by exactly the same mechanism with which it produces classical "blocking" effects (Rescorla & Wagner, 1972). In a simple blocking situation, there are two symptoms, s1 and s2, and a disease, X. If a subject is initially trained with only s1$\mapsto$X, and is subsequently trained on s1+s2$\mapsto$X, then the second symptom, s2, will not develop a very large association with the disease. That is, the prior learning of the s1$\mapsto$X association *blocks* learning of the s2$\mapsto$X association. The reason is that after the first phase of training, the symptom s1 predicts disease X, so that there is no error when presented with s1+s2, and consequently, because learning is error-driven, no association is established between s2 and X. In a minimalist base-rate neglect structure, consider two exemplars, s1+s2$\mapsto$C and s2$\mapsto$R, which occur randomly mixed with equal base rates. Normatively, p(R|s2) = .50, but the model creates a stronger association weight from s2 to R than from s2 to C. The reason is that the weight change for s2 on s1+s2$\mapsto$C trials is *blocked* by the learning of the s1-to-C association: As s1 becomes a good predictor of C, there is no error, and so the weight from s2 to C does not grow to the same extent and the weight from s2 to R.

Despite its success with apparent base-rate neglect, the component-cue model cannot account for the inverse base-rate effect. Markman (1989) showed that the component-cue model can exhibit some aspects of the inverse base-rate effect if the absence of a symptom is represented on the input nodes by a value of -1 instead of 0. Unfortunately, that solution is incomplete: First, as pointed out by Markman (1989), it does not make proper predictions for test cases that cross disease pairs, such as I+PCo. Second, even on the test cases for which it does work, it works only pre-asymptotically, which was not pointed out by Markman (1989).

Gluck (1992) proposed another variant of the component-cue model, that incorporates the distributed cue representation of stimulus-sampling theory (Atkinson & Estes, 1963). He showed that such a model can exhibit some aspects of the inverse base-rate effect. However,

the model shows only slight preference for the rare disease on PC+PR tests, and only slight preference for the common disease on I+PC+PR tests. No quantitative fits were presented. Moreover, no predictions for other test cases, such as I+PCo, were presented. It remains unclear, then, to what extent the stimulus-sampling version of the component-cue model is viable.

Shanks (1992) proposed a variant, called the "attentional connectionist model" or ACM, in which cues (symptoms) develop different saliences depending on their base rates, or expectancies. Using a formalism inspired by Wagner (1978), the learning rate for an individual cue was multiplied by 1.0 minus the probability of the cue occurring. Because symptom PR occurs relatively infrequently, its learning rate is multiplied by a relatively large value. Consequently, the weight from PR to R grows larger than the weight from PC to C, and the model prefers R when tested with PC+PR. Shanks (1992) did not describe predictions for other test cases such as I+PCo, nor did he provide quantitative fits (his ostensive goal was to demonstrate robust qualitative effects). The experiments described in the present article provide additional challenges for the ACM. In particular, the results from Experiment 2, involving pre-training, might prove difficult, because of the decrease in the base-rate "consistency" effect for I+PC+PR (I+PP+PF). Moreover, no quantitative fits of the ACM to base-rate neglect data have yet been provided.

Should the ACM be able to capture the results of the present experiments, then new experiments will need to be conducted to distinguish it from ADIT, because they are based on different explanatory principles. In the ACM, attention directly reflects the relatively long-term base rates of individual cues, and is used only in learning. In ADIT, attention reflects the extent to which a cue is incorporated into classification and learning on a given trial, and is *not* directly related to base rates.

As described earlier, ADIT is itself an extension of the component-cue model. It has two main mechanisms beyond the original component-cue model. First, is adjusts attention to input cues on each trial, so that distinctive features are amplified when errors are made, and current knowledge is amplified when correct predictions are made. Second, it uses a response rule in which category base rates are mixed with network predictions, such that when fewer input features are present, the base rates have a stronger influence.

Summary and Conclusion

This article suggests that category base rates play two major roles in the experimental paradigms of the inverse base-rate effect and apparent base-rate neglect. The first role is causing the categories to be learned at different times: The more frequent categories are learned first, in terms of their typical features, and the rare categories are subsequently learned in terms of their distinctive features. The second role is causing subjects to consistently favor the more frequent categories: Differential base rates are learned, and that knowledge is deployed consistently in all training and test cases, with the relative influence of the base rate knowledge governed by only the number of cues in the stimulus. Thus, the apparent inconsistencies in base-rate utilization observed in the inverse base-rate effect are merely apparent; subjects are consistently applying base-rate knowledge to asymmetric category representations. Moreover, apparent base-rate neglect in the Gluck and Bower (1988) paradigm and the inverse base-rate effect in the Medin and Edelson (1988) paradigm are seen to be manifestations of the same mechanisms, with apparent base-rate neglect being an attenuated case of the inverse base-rate effect.

Evidence for these claims was provided in two forms. Experiments with human learners provided empirical confirmation of several implications of those principles, and a new model formalized those principles and fit the data well. The experiments verified that subjects do learn the common categories before the rare categories and that subjects do have clear (ordinal) knowledge about which categories are more frequent. Experiment 1 replicated the basic inverse base-rate effect using fewer symptoms and diseases than the original study (Medin & Edelson, 1988), and found that test case I+PC+PR did not produce the same results as case I+PC+PRo. Experiment 2 verified that pre-training on a subset of categories produces effects comparable to the inverse base-rate effect, even when the full-training base rates are all equal. Experiment 3 used a hybrid design in which subjects displayed both the inverse base-rate effect and apparent base-rate neglect, and its results were interpreted to mean that apparent base-rate neglect and the inverse base-rate effect are both caused by the learned distinctiveness of one of the symptoms for the rare disease, with the distinctiveness simply less strong in the case of apparent base-rate neglect. Experiment 4 showed that apparent base-rate neglect obtains even for cues that occur in fewer than half the rare cases, and also explored the influence of cues that occur rarely in all categories; such cues only decrease the influence of base-rate knowledge, but contribute little else toward the classification decision.

The causal chain from differential base rates to asymmetric category coding has at least two links. The first link is that differential base rates cause the common category to be learned before the rare category. The second link is that learning one before the other causes the first category to be learned in terms of its typical features, and the second category to be learned in terms of the features that distinguish it from the first category. In principle, therefore, an analogue of the inverse base-rate effect could be generated

by any influence that causes one category to be learned by its typical features, and the other category to be learned by its distinctive features. Neither differential base rates nor different learning times are necessary. Thus, for example, it might be sufficient to label one category as "A" and the other as "not A," so that "A" is learned in terms of its typical features and "not A" is learned in terms of the features that distinguish it from "A" (Goldstone, 1993). (In Goldstone's (1993) terminology, the common categories are "positively defined concepts," and the rare categories are "negatively defined concepts.") In other experiments in my lab, I have found that simply giving subjects corrective feedback that states "Think about the typical symptoms of this disease" or "Think about the distinctive features of this disease" (without indicating which disease is an appropriate referent) is sufficient to produce a weak analogue of the inverse base rate effect.

A new model, called ADIT, formalized the explanatory principles in a connectionist framework. The basic component-cue model of Gluck and Bower (1988) was extended in two main ways. First, each input node was given an attentional gate, which reflected the extent to which an input cue would be used in the classification decision. The idea of attending to distinctive features, when current knowledge leads to error, was formalized as gradient descent on error with respect to the attention strengths. The attention shift not only accentuates distinctive features when current knowledge is wrong, it also accentuates current knowledge (not the input) when current knowledge is correct. The gradient descent formalism also has the attractive property of being consistent with the associative weight learning mechanism, which is also gradient descent on error. The second major extension of the component-cue model is in the use of base-rate knowledge in the mapping from network activations to choice probabilities. The model also has several less central, but important other extensions. In particular, only the correct category has its association weights adjusted. Insofar as the model fits the data (and this model is the only one yet proposed to fit such extensive data from the neglect and inverse paradigms), we have evidence that both the underlying principles embodied in the model, and their specific formalization, are correct. On the other hand, I am not wedded to the particular formalizations described herein, and the explanatory principles could probably also be implemented in exemplar-based models, and in production systems. The extension of the component-cue model used here was selected only by virtue of relative simplicity. Future modeling will have to pursue those more complicated models in order to address a broader spectrum of category-learning data.

The research presented here establishes two roles of base rates in category learning. Future empirical work could investigate details of those separate influences, to better constrain models and obviate the various ad hoc modeling assumptions made here. In particular, there is much more to be learned about how people acquire base-rate knowledge, as the experiments reported here indicate that it is not affected by presentation frequency alone, but also by how well the symptom-category associations are known.

Acknowledgments

This research was supported in part by NIMH FIRST Award 1-R29-MH51572-01. Anastasia L. Kirkwood conducted several exploratory experiments on base-rate effects prior to those reported here. Thanks to Michael A. Erickson and Michael Kalish for useful discussion. For helpful comments on a previous draft of this article, I thank Robert Goldstone, Michael Kalish, Stephen C. McKinley and Thomas J. Palmeri. For their assistance in running the experiments, I am grateful to Kevin Adamick, Rebecca Baldwin, Michael A. Erickson, Gilbert Hyatt, Don Lyon, Timothy Magill, Anthony Masterson, Jane Ream, and Michael Rush.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ.

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In Luce, R. D., Bush, R. R., & Galanter, E. (Eds.), *Handbook of Mathematical Psychology*. Wiley, New York.

Bar-Hillel, M., & Fischhoff, B. (1981). When do base rates affect predictions?. *Journal of Personality and Social Psychology*, *41*, 671–680.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance*, *29*, 270–278.

Chumbley, J. I. (1986). CLEM: Concept learning by exemplar memorization.. Unpublished manuscript, University of Massachusetts at Amherst.

Cobos, P. L., López, F. J., Rando, M. A., Fernández, P., & Almaraz, J. (1993). Connectionism and probability judgment: Suggestions on biases. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 342–346 Hillsdale, NJ. Erlbaum.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory

storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 556–576.

Gluck, M. A. (1992). Stimulus sampling and distributed representations in adaptive network theories of learning. In Healy, A., Kosslyn, S., & Shiffrin, R. (Eds.), *Essays in Honor of William K. Estes, Volume 1*, pp. 169–199. Erlbaum, Hillsdale, NJ.

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. (1993). Positively and negatively defined concepts. Research Report 88, Cognitive Science Program, Indiana University, Bloomington.

Hamm, R. M. (1994). Underweighting of base-rate information reflects important difficulties people have with probabilistic inference. *PSYCOLOQUY*, *5*(3). base-rate.7.hamm.

Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79*, 432–443.

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. In *Annual Review of Psychology*, Vol. 44, pp. 265–315. Academic Press, San Diego.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Koehler, J. J. (1993). The base rate fallacy myth. *PSYCOLOQUY*, *4*(49). base-rate.1.koehler.

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

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1993b). Three principles for models of category learning. In Nakamura, G. V., Taraban, R., & Medin, D. L. (Eds.), *Categorization by Humans and Machines: The Psychology of Learning and Motivation*, Vol. 29, pp. 57–90. Academic Press, San Diego.

Kruschke, J. K. (1994). Reply to Lewandowsky (1994). *Psychological Review*, *00*, 000–000. in press.

Lenat, D. B., & Feigenbaum, E. A. (1987). On the thresholds of knowledge. In McDermott, J. (Ed.), *Proceedings of the Tenth Joint Conference on Artificial Intelligence*, pp. 1173–1182 San Mateo, CA. Morgan Kaufmann.

Lewandowsky, S. (1994). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review*, *00*, 000–000. in press.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General*, *118*, 417–421.

Medin, D. L., & Bettger, J. G. (1991). Sensitivity to changes in base-rate information. *American Journal of Psychology*, *104*, 311–332.

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., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Myers, J. L., Lohmeier, J. H., & Well, A. D. (1994). Modeling probabilistic categorization data: Exemplar memory and connectionist nets. *Psychological Science*, *5*, 83–89.

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, *2*, 416–421.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In Medin, D. L. (Ed.), *The Psychology of Learning and Motivation*, Vol. 28, pp. 207–250. Academic Press, San Diego.

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

Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, *76*, 221–230.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement.. In Black, A. H., & Prokasy, W. F. (Eds.), *Classical Conditioning: II. Current Research and Theory*, pp. 64–99. Appleton-Century-Crofts, New York.

Shanks, D. R. (1990a). Connectionism and human learning: Critique of Gluck and Bower (1988). *Journal of Experimental Psychology: General*, *119*, 101–104.

Shanks, D. R. (1990b). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology*, *42A*, 209–237.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3–18.

Spellman, B. A. (1993). Implicit learning of base rates. *PSYCOLOQUY*, *4*(61). base-rate.4.spellman.

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

Wagner, A. R. (1978). Expectancies and the priming of STM. In Hulse, S. H., Fowler, H., & Honig, W. H. (Eds.), *Cognitive processes in animal behavior*. Erlbaum, Hillsdale, NJ.