Journal of Experimental Psychology:
Learning, Memory, and Cognition
1996, Vol. 22, No. 1, 3-26

0278-7393/96/$3.00

Base Rates in Category Learning

John K. Kruschke
Indiana University

Previous researchers have discovered perplexing inconsistencies in how people appear to utilize category base rates when making category judgments. In particular, D. L. Medin and S. M. Edelson (1988) found an *inverse base-rate effect,* in which participants tended to select a rare category when tested with a combination of conflicting cues, and M. A. Gluck and G. H. Bower (1988) reported apparent *base-rate neglect,* in which participants tended to select a rare category when tested with a single symptom for which objective diagnosticity was equal for all categories. This article suggests that common principles underlie both effects: First, base-rate information is learned and consistently applied to all training and testing cases. Second, the crucial effect of base rates is to cause frequent categories to be learned before rare categories so that the frequent categories are encoded by their typical features and the rare categories are encoded by their distinctive features. Four new experiments provide evidence consistent with those principles. The principles are formalized in a new connectionist model that can rapidly shift attention to distinctive features.

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

In this article I address 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, namely, the *inverse base-rate effect* (Medin & Edelson, 1988) and *apparent base-rate neglect* (Gluck & Bower, 1988). Medin and Edelson (1988) 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. (p. 81)

Holyoak and Spellman (1993, p. 285; Spellman, 1993) also suggested that people use base-rate knowledge inconsistently, depending on whether the classification decision takes place during the training phase or during the testing phase. They suggested that in training, people use base rates accurately because the base rates are learned and used implicitly as part of the task, but, in testing, the base rates may be neglected because they must be accessed explicitly.

Contrary to those interpretations, it is argued here that participants in these category-learning experiments apply their knowledge of base rates in a uniform manner, regardless of the stimulus content or training versus test phase. People do underemphasize the base rates relative to normative, Bayesian prescriptions, but the underweighted knowledge is consistently applied to all cases.

Two main theses are proposed in this article. First, the primary role of base rates is to cause the high-frequency categories to be learned before the low-frequency categories. The fact that the high-frequency categories are learned earlier than the low-frequency categories has been given little importance in previous reports, but here it is preeminent. What people learn depends on what they know (e.g., Lenat & Feigenbaum, 1987), and so what people 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 differently, and this difference 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 the decision maker to choose the more frequent categories, with the bias applied consistently to all cases. However, when the bias is applied to differently encoded categories, it can be obscured and appear to be inconsistent.

A corollary to these 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

This research was supported in part by National Institute of Mental Health First Award 1-R29-MH51572-01. Anastasia L. Kirkwood helped conduct several exploratory experiments on base-rate effects before those reported in this article. I thank Michael A. Erickson and Michael Kalish for useful discussion, and, for helpful comments on drafts of this article, I also thank James Chumbley, Robert Goldstone, Michael Kalish, Stephan Lewandowsky, Stephen C. McKinley, and Thomas J. Palmeri. For their assistance in conducting 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.

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychology, Indiana University, Bloomington, Indiana 47405. Electronic mail may be sent via Internet to kruschke@indiana.edu. John K. Kruschke's World Wide Web URL is http://www.indiana.edu/~kruschke/home.html.

also account for the other. No previous model has yet been shown to account thoroughly for both effects.

This article provides both empirical and modeling evidence consistent with the main theses. Experiment 1 replicated the inverse base-rate effect but used fewer categories and cues than those used by Medin and Edelson (1988), confirmed that the frequent categories are learned earlier than the rare categories, and confirmed that learners are aware of differential base rates. In Experiment 2 learners were pretrained on a subset of categories and then trained on a full set of categories with equal base rates. 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 was a hybrid design that included minimally different subdesigns for the inverse base-rate effect and apparent base-rate neglect, and the results are consistent with the idea that base-rate neglect is an attenuated case of the inverse base-rate effect. In Experiment 4 I explored base-rate neglect by using a previously untested combination of cue probabilities, and the results exhibit a pattern of data that no previous model of base-rate neglect can account for.

The explanatory principles are formalized in a model, which was 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 were two key additions: First, every input cue had an attention strength that was rapidly shifted by corrective feedback such that distinctive features were more strongly attended to when there was error. Second, classification probabilities were determined by mixing the prediction from the associative network with the learned base rates. The relative weight given to the base rates depended only on the number of cues in the input, not on the specific content of the input or on the prediction of the network.

The empirical evidence provided for the principles was indirect. None of the experiments showed directly the content of learner's category knowledge. Participants were not asked to list the learned features of the categories. Nevertheless, the empirical data are consistent with predictions of the principles for the standard category-learning paradigms used here. 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 was reached in these experiments by dint of the main thesis; that is, rare categories are learned in terms of the features that distinguish them from the previously learned common categories. The major emphasis of the article is the attention-shifting mechanism that produces differently encoded categories.

Experiment 1: The Inverse Base-Rate Effect

Medin and Edelson (1988) presented people with a fictitious disease diagnosis task. On each trial of a learning sequence, a list of symptoms was presented to the learner, who had to diagnose the hypothetical patient as having one of several possible fictitious diseases. The learner was then 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 occurred in random order and had 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, participants were tested with combinations of symptoms not shown in training. When tested with the ambiguous symptom I, people tended to choose the common disease, consistent with the base rates. When tested with the ambiguous combination I + PC + PR, people again tended to choose the common disease (although less strongly). When presented with the conflicting symptoms PC + PR, however, people 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, people first learn that symptoms I and PC are typical of disease C because that case occurs so often. They also quickly learn that disease C occurs much more frequently than disease R. Subsequently, they learn the rare disease. They discover that the shared symptom I is misleading, because they already associate it with disease C, and instead pay more attention to the distinctive symptom PR. When tested with PC + PR, people 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. People 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. (The meaning of the term *distinctive* is discussed at the end of the article.) When tested with symptoms I + PC + PR, people find that all symptoms of diseases 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 was a partial replication and extension of the primary experiment reported by Medin and Edelson (1988). Its purposes were (a) to establish that the inverse base-rate effect is robustly obtainable when using the particular materials and procedure in my lab (e.g., Shanks [1992, Experiment 1] did not obtain an inverse base-rate effect with a 3:1 ratio of base rates but did find it with a 7:1 base-rate ratio, and Medin and Edelson [1988, Experiment 4] found that the magnitude of the effect could be affected by instructions); (b) to establish a standard for the magnitude of the inverse base-rate effect, for comparison with subsequent experiments and for modeling; (c) to extend the results of Medin and Edelson to a situation with a different number of categories, additional test cases, and explicit category frequency judgments; and (d) to verify that participants do learn the common disease before the rare disease and have explicit knowledge that the base rates of the diseases are very different.

Method

Participants. Fifty-six students participated for partial credit in an introductory psychology course at Indiana University. None had participated in any other related experiment in my lab.

Design. The abstract structure of the training cases is shown in Table 1. The structure is the same as that 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 a *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.

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

The six abstract symptoms in Table 1 were randomly assigned, for each participant, 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 participant, to the labels *F, G, H,* and *J.* Participants 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. 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 in another trial with PC1 on the first line and I1 on the second.

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

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, the symptoms remained on the screen and 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 there was the word "CORRECT!" If there was no response after 30 s, 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 learner took longer than 30 s of study time, a warning appeared, accompanied by a tone, and the next trial started automatically. Between trials, the screen was blank for approximately 750 ms.

Table 1
Abstract Design of Training Stimuli in Experiment 1

Symptom						
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

Note. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare. 1 = presence of a symptom; 0 = absence of a symptom.

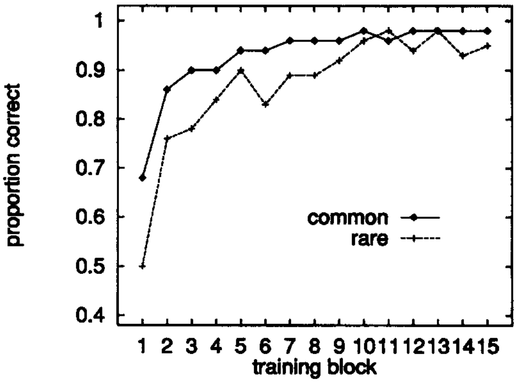

Figure 1. Training data from Experiment 1.

After training, the participant was shown novel combinations of symptoms, which are 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 before the test cases that told the participant that he or she would be seeing combinations of symptoms not previously seen and that he or she would not be told the correct diagnoses, but that it was very important to make the best guess on the basis of what he or she had learned from the earlier part of the experiment. When a participant made a response on a test trial, the computer displayed the phrase, "Your diagnosis has been recorded," where corrective feedback appeared in the training phase. As in training, the participant had to press the spacebar to see the next case.

After the test phase, the learner was asked to judge the frequencies of the four diseases during the training period. Estimates were indicated on a paper response sheet that was concealed inside an envelope during training and testing. The instructions on the paper told the participant to estimate the number of times each disease occurred during training and that he or she should be careful to think about the actual number of occurrences, not the relative ease of learning or the number of times he or she got each one right or wrong, nor the apparent feedback or study time, and so forth. The participant was also told that the total of the four frequencies should be 120 and that if all four diseases occurred equally often, then each would have appeared 30 times. Each participant 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 seen by that participant.

The entire experiment lasted about 30 min.

Results and Discussion

Training. The common diseases were learned much sooner than the rare diseases (see Figure 1). For each person, the difference between the proportion correct for the common and rare diseases was computed over the first third of training (40 trials, or 5 blocks). The mean difference was .858 − .750 = .108, $t(55) = 4.66$, $SE = 0.0231$, $p < .0001$.

It is worth noting here that participants attained 49% correct on the rare diseases in the first block of training, far above chance performance (25%). Each rare case occurred only once per block so that participants achieved that performance without ever seeing the rare cases before. Therefore, such accuracy in the first block cannot be based on learned associations between symptoms and diseases for those cases but must instead be based on some nonrandom guessing strategy. This will be especially relevant when the data are modeled.

By the end of training, the participants had thoroughly learned all of 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 small difference between common and rare remained significant, however, $t(55) = 2.46$, $SE = 0.0107$, $p = .017$. Every participant performed with at least 70% (7 out of 10 instances) accuracy 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 disease C2 is *Co*; when testing with PC2, disease C2 is *C* and disease C1 is *Co*. Each row in Table 2 is based on 224 observations (56 participants times two instantiations of each abstract symptom combination [one for each pair of common–rare diseases] times two presentations of each instantiation).

The results show a strong inverse base-rate effect for test cases PC + PR and PC + PRo, $\chi^2(1, N = 216) = 15.57$, $p < .001$, and $\chi^2(1, N = 205) = 10.78$, $p < .005$, respectively, computed from the response frequencies for the two corresponding diseases and expected values of 50:50; also, the magnitude of the effect did not differ significantly between those cases, $\chi^2(1, N = 421) = 0.173$, *ns*.[1]

For I + PC + PR, people showed a small base-rate consistency, preferring the common disease, $\chi^2(1, N = 220) = 7.273$, $p < .01$, using 50:50 for expected values, but for I + PC + PRo, there was a significantly larger preference for the common disease, $\chi^2(1, N = 428) = 16.48$, $p < .001$. Medin and Edelson (1988) collapsed those two cases in their data analysis, but they were evidently construed differently by the participants.

Frequency estimates. All 56 participants made frequency estimates that summed to 120. To have an approximately normal distribution of estimates for inferential statistics, I converted estimates to the natural logarithm of the ratio of common-to-rare frequencies, log[(C1 + C2)/(R1 + R2)]. If participants 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 was 0.7366, which is significantly greater than zero, $t(55) = 11.33$, $SE = 0.0650$, $p < .0001$. The antilog of the mean log ratio was 2.1; that is, the estimated frequency ratio was about 2.1:1 (as opposed to the actual 3:1). The obtained ratio should be interpreted 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 learners, overall, were strongly aware that the common diseases occurred more frequently than the rare diseases.

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

Note. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare; Co = the other common disease; Ro = the other rare disease.

Experiment 2: Pretraining a Subset of Categories Produces an Analogue of the Inverse Base-Rate Effect

If the explanatory principles are correct, then any manipulation that causes one category to be learned before another should also produce an effect analogous to the inverse base-rate effect. In particular, if people are trained on some categories before others, then the later categories should be encoded in terms of their distinctive features.

In Experiment 2 I used a design similar to Experiment 1, in that there were four diseases and six symptoms, with the same exemplars as Experiment 1, but there were two phases of training in Experiment 2. In the first, *early training* phase, only two diseases were trained, namely, the ones corresponding to the common diseases in Experiment 1. In the second *late training* phase, all four diseases appeared with equal base rates. I hypothesized that the perfect predictors of the later trained diseases would be more strongly encoded than the perfect predictors of the earlier trained disease.

Method

Participants. Fifty-three students participated for partial credit in an introductory psychology course at Indiana University. None had participated in any other related experiment in my lab.

Design. The abstract structure of the training cases is shown in Table 3. The two diseases that appear in the early training phase are labeled *E1* and *E2*, and the other two diseases, which appear only in the late training phase, are labeled *L1* and *L2*. Imperfectly predictive symptoms are labeled *I*, as before, and the perfect predictors of the early training or late training diseases are labeled *PE* or *PL*, respectively. As before, presence of a symptom is indicated by a 1, absence by a 0.

[1] The chi-square tests used throughout the article were conducted on the actual frequencies, not on the proportions reported in the summary tables. These chi-square tests are not strictly appropriate because the data include repeated measures from the same participants. That is, the data might violate the assumption of independence in the chi-square test. Fortunately, nearly all the effects of interest have such large chi-square values that even after they were conservatively reduced to reflect the violation of independence (Wickens, 1989, p. 28), they still exceeded conventional critical values.

Table 3
Abstract Design of Training Stimuli in Experiment 2

Symptom						
I1	PE1	PL1	I2	PE2	PL2	Disease
1	1	0	0	0	0	E1
1	1	0	0	0	0	E1
1	0	1	0	0	0	L1
1	0	1	0	0	0	L1
0	0	0	1	1	0	E2
0	0	0	1	1	0	E2
0	0	0	1	0	1	L2
0	0	0	1	0	1	L2

Note. For the symptoms, I = imperfect predictor of the two diseases; PE = perfect predictor of the earlier trained disease; PL = perfect predictor of the later trained disease. For the diseases, E1 and E2 represent the two diseases that appeared in the early training phase; L1 and L2 represent the two diseases that appeared in the late training phase. 1 = presence of a symptom; 0 = absence of a symptom.

Apparatus and stimuli. The apparatus and stimuli were the same as those used in Experiment 1.

Procedure. Participants were instructed that they would first be given preliminary training on two of the four diseases and that later they would learn all four. Early training consisted of nine blocks of eight trials, each block containing four instances of disease E1 and four instances of E2, in random order.

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

After training, participants were shown novel combinations of symptoms, directly analogous to the combinations tested in Experiment 1, and instructions and procedure identical to Experiment 1 were used in this experiment.

After the test phase, participants were asked to judge the frequencies of the four diseases during the late training period. The instructions indicated that the total of the four frequencies should be 72 and that if all four diseases occurred equally often, then each would have appeared 18 times. The procedure was otherwise identical to that in Experiment 1.

The entire experiment lasted about 30 min.

Results

Training. The results indicate that people remembered the early training diseases at the beginning of late training, as the difference in accuracy between the early and late training diseases in the first third of late training was significant (mean difference in proportion correct for the first 24 trials of late training was .974 − .867 = .107), $t(52) = 6.13$, $SE = 0.0174$, $p < .0001$.

Participants learned all diseases thoroughly by the end of training; mean proportion correct in the last third of late training was .973 for the early training diseases and .978 for the other late training diseases. All individuals performed with at least 83% (10 out of 12) accuracy on the non-early trained disease cases in the last third of late 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 early training disease E, the corresponding late training disease L, the other early training disease *Eo*, or the other late training disease *Lo*. Each row in Table 4 is based on 212 observations (53 participants times two instantiations of each abstract symptom combination [one for each pair of early and late training diseases] times two presentations of each instantiation).

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 people were tested with the conflicting pair PE + PL, they tended to choose the disease they learned later, χ^2 $(1, N = 204) = 24.02$, $p < .001$, using 50:50 for expected frequencies. The analogue to the inverse base-rate effect also appeared for the conflicting pair PE + PLo, χ^2 $(1, N = 199) = 8.45$, $p < .005$, and was not significantly less than the effect for PE + PL, χ^2 $(1, N = 403) = 2.05$, *ns*. The magnitude of the analogue to the inverse base-rate effect (case PE + PL) was not significantly less than the magnitude of the inverse base-rate effect itself (case PC + PR) in Experiment 1, χ^2 $(1, N = 420) = 0.644$, *ns*.

When tested with the ambiguous case I + PE + PL, people did not choose the early training disease significantly more often than the late training disease, χ^2 $(1, N = 208) = 0.481$, *ns*. That result can, nevertheless, be construed as a base-rate consistency effect because in Experiment 2 there were equal base rates in the late training phase, and it was plausible that participants' base-rate biases would be dominated by the more recent, late training phase. Participants did prefer the early training disease when tested with I + PE + PLo, χ^2 $(1, N = 196) = 51.02$, $p < .001$, and that preference was significantly stronger than for I + PE + PL, χ^2 $(1, N = 404) = 23.28$, $p < .001$. Again, it can be seen that the data do not allow the triple-symptom cases to be collapsed.

Table 4
Results From Test Trials of Experiment 2

	Choice proportion			
Symptoms	E	L	Eo	Lo
I	.802	.118	.033	.047
PE	.925	.014	.038	.024
PL	.028	.915	.019	.038
PE + PL	.316	.646	.019	.019
I + PE + PL	.514	.467	.014	.004
I + PEo	.379	.076	.502	.043
I + PLo	.208	.100	.014	.678
PE + PLo	.374	.047	.009	.569
I + PE + PLo	.698	.047	.028	.226

Note. For the symptoms, I = imperfect predictor of the two diseases; PE = perfect predictor of the early trained disease; PL = perfect predictor of the late trained disease. For the diseases, E = early trained disease; L = late trained disease; Eo = the other early trained disease; Lo = the other late trained disease.

Frequency estimates. Five of the 53 participants made frequency estimates that did not sum to 72: Two participants had sums of 62, 2 had sums of 82, and 1 had a sum of 80. Data from these participants were used anyway, as they were qualitatively indistinguishable from the others, and they did not affect any statistical conclusions.

The mean of log[(E1 + E2)/(L1 + L2)], across individuals, was 0.1913, which is significantly greater than zero, $t(52) = 3.09$, $SE = 0.0619$, $p < .0032$. The antilog of the mean log ratio was 1.2; that is, the estimated frequency ratio was about 1.2:1 (as opposed to the actual 1:1 in late training). That ratio was dramatically less than the 2.1:1 ratio obtained in Experiment 1, $t(107) = 6.08$, pooled $SE = 0.0632$, $p < .0001$, which had an actual base-rate ratio of 3:1.

Apparently, people thought that the early training diseases occurred (in the second phase of training) slightly more often than the late training diseases. It is plausible that memory for the early training instances was better than for the late training instances, especially early in the late training phase, and influenced the frequency estimates.

Discussion

Experiment 2 confirmed that learning one category before another produces an analogue of the inverse base-rate effect. This is consistent with the idea that one key role of base rates in the standard 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 participants' knowledge of base rates might play a role in their responses to the ambiguous test case, I + PC + PR (I + PE + PL), insofar as response tendencies corresponded to the magnitude of the estimated frequency ratios. Of course, participants in Experiment 2 were explicitly instructed to consider only the late phase of training when making their frequency estimates, and it is not known if that corresponds with the base-rate knowledge they applied in the test trials.

The two training phases of Experiment 2 can be construed as an extreme case of changing base rates during learning. In the early training phase, the base-rate ratio was ∞:1 (or 1:0); in the late training phase, the ratio was 1:1. The results are consistent with those obtained by Medin and Bettger (1991), who studied the influence of changing the base rates during learning. Using the basic symptom–disease pairings of Experiment 1, they changed the disease base rates at various points in training. Medin and Bettger (1991) found that

> in short, it appears that responses to common cue tests and inverse base-rate effects on conflicting tests are determined primarily by [the] relative frequency prevailing during learning. Relative frequency differences after learning is complete appear to produce more modest effects that take the form of direct base-rate effects for all three main types of tests. (p. 327)

Their results are nicely accommodated by the ideas proposed here: The base rates early in training influence the inverse base-rate effects by determining the order in which the categories are learned and hence their learned encoding; 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

In Experiment 3 I investigated the hypothesis 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, 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%.

Participants were trained on random sequences of cases generated from those probabilities and were then tested with single symptoms. Of primary interest was the result that, when presented with symptom s1 alone, people tended to choose the rare disease significantly more often than 50%. This 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).

This tendency has descriptively been labeled as *apparent base-rate neglect* because it could be accounted for by underestimating the extremeness of the base rates in Bayes's 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 decision makers have accurate knowledge that $p(\mathrm{s1}|\mathrm{R}) = .6$ and $p(\mathrm{s1}|\mathrm{C}) = .2$, and use Bayes's theorem, they will overestimate $p(\mathrm{R}|\mathrm{s1})$ if they underestimate the base-rate ratio. Recent explanations of apparent base-rate neglect do not invoke that account but appeal instead to error-driven learning in connectionist networks, which are described in the Modeling section.

According to the principles propounded in this article, the effect is explained as follows: Participants learn the typical features of the common disease more quickly because the common disease happens so often. They learn that symptom s4 is very typical, s3 and s2 are somewhat typical, and s1 is only slightly typical. Participants then 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 does not occur very often and has already been encoded as somewhat typical of the common disease.) When people 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, 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 always and exclusively in the rare disease.

In Experiment 3, the abstract stimulus structure of Experiment 1 was 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 simultaneously observed in the same individuals. The relative strengths of the distinctive symptoms in the neglect situation and in the inverse situation can then directly be tested.

Method

Participants. Sixty-four students participated for partial credit in an introductory psychology course at Indiana University. None had participated in any other related experiment in my lab.

Design. The abstract design of the training stimuli differed from that of Experiment 1 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%; that is, $p(\mathrm{Rn}\,|\,\mathrm{PRn}) = .5 = p(\mathrm{Cn}\,|\,\mathrm{PRn})$.[2] Apparent base-rate neglect can then be tested when participants are presented with symptom PRn alone. The diseases and symptoms were labeled with an *n* suffix or an *i* suffix to indicate whether they correspond to the base-rate neglect substructure or to the inverse base-rate effect substructure, respectively. (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 the label was used to suggest the structural analogy to the inverse substructure.)

Apparatus and stimuli. The apparatus and stimuli were the same as those used in Experiment 1.

Procedure. Participants in Experiment 3 received a block of 21 test trials after every 5 training blocks instead of only at the end of all 15 training blocks, as in Experiment 1.

Results

Training. This design proved more difficult for people to learn than that in 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 the test results should reflect the decisions of people who ultimately did learn all of the diseases, further analysis excluded 8 participants (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 participants in Experiment 1 achieved at least that level of performance. Consequently, there are 56 participants included in the data reported below.

Table 5
Abstract Design of Training Stimuli in Experiment 3

Symptom						
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

Note. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare; n = base-rate neglect substructure; i = inverse base-rate effect substructure. 1 = presence of a symptom; 0 = absence of a symptom.

Participants learned the common diseases earlier 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 .829 − .675 = .154), $t(55) = 7.41$, $SE = 0.0208$, $p < .0001$. Participants also learned the diseases in the inverse substructure faster than those in the neglect substructure (mean difference .808 − .696 = .111), $t(55) = 5.02$, $SE = 0.0222$, $p < .0001$. There was no interaction between disease frequency and substructure (mean difference of differences = .0344), $t(55) = 0.88$, $SE = 0.0390$, *ns*.

The 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 .970 − .921 = .048), $t(55) = 3.53$, $SE = 0.0136$, $p = .0008$; inverse vs. neglect (mean difference .960 − .931 = .0292), $t(55) = 2.24$, $SE = 0.0130$, $p = .029$; and interaction (mean difference of differences = .0012), $t(55) = 0.05$, $SE = 0.0238$, *ns*.

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 participants times three presentations of each test case).

The results show a strong apparent base-rate neglect for symptom PRn, with people choosing disease Rn 77% of the time, as opposed to just 13% of the time for disease Cn, $\chi^2\,(1, N = 152) = 76.74$, $p < .001$. Thus, apparent base-rate neglect occurred even for these deterministic categories and did not depend on the probabilistic mapping in the work of Gluck and Bower (1988).

There was also a strong inverse base-rate effect for the conflicting symptoms PCi + PRi, with people preferring the

[2] The stimulus structure for Experiment 3 made all of the symptoms conditionally independent, as in the Gluck and Bower (1988) design and as in the design of Experiment 4 of this article. Formally, $p(Si + Sj\,|\,\mathrm{D}) = p(Si\,|\,\mathrm{D})p(Sj\,|\,\mathrm{D})$ for all symptoms *Si*, *Sj*, and disease D.

Table 6
Results From Test Trials of Experiment 3

	Choice proportion			
Symptom	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

Note. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare. n = base-rate neglect substructure; i = inverse base-rate effect substructure.

rare disease Ri to the common disease Ci, 64% to 32%, $\chi^2(1, N = 162) = 18.00, p < .001$.

There was no inverse base-rate effect, however, for the conflicting symptoms in the neglect substructure. Symptoms PCn + PRn elicited a small preference for the common disease (54% to 40%), $\chi^2(1, N = 158) = 3.65$, *ns*. Therefore, 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 confirmed that interpretation, for when people were presented with PRn + PRi, they preferred disease Ri to disease Rn (65% to 29%), $\chi^2(1, N = 158) = 22.78, p < .001$.

The base-rate consistency effect for the inverse substructure (test symptoms Ii + PCi + PRi), while in the expected direction, did not reach statistical significance $\chi^2(1, N = 149) = 1.134$. The effect may have been diluted by the relatively large proportion of Cn responses, which might in turn have been caused by a tendency to respond "Cn" whenever any three symptoms appeared.

Frequency estimation. Two of the 56 participants 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 substructures were initially considered separately. For each individual, 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 substructures did not differ (0.687 vs. 0.690). Therefore, the estimates for each individual were combined into a single transformed value, log[(Cn + Ci)/(Rn + Ri)]. Across individuals, its mean value was 0.6870, which is reliably larger than zero, $t(55) = 14.84$, $SE = 0.0463$, $p < .0001$, and corresponds to an estimated frequency ratio of 2.0:1 (compared with the actual ratio of 3:1). That frequency ratio was nearly equal to the 2.1:1 ratio obtained in Experiment 1, $t(110) = 0.62$, pooled $SD = 0.422, p = .54$.

There was, however, a difference in the total frequencies attributed to neglect and inverse substructures (57.4 vs. 62.8), $t(55) = 2.16$, $SE = 2.51, p = .035$, when using raw frequency estimates. This corresponds to the difference in learning speed for the two substructures; the inverse substructure was learned faster and was perceived as occurring more frequently.

Discussion

Experiment 3 can be interpreted as demonstrating that apparent base-rate neglect is an attenuated case of the inverse base-rate effect. The results are consistent with the idea that in both the inverse base-rate effect and in apparent base-rate neglect, people encode one of the symptoms as distinctive of the rare disease, but the magnitude of the distinctiveness differs.

The main implication is that if the same mechanism underlies both sets of phenomena, then any model that purports to account for apparent base-rate neglect should also account for the inverse base-rate effect simultaneously in a single experimental setting.

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. In Experiment 4 I expanded 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. In this experiment, symptom probabilities not explored by previously published experiments were used, demonstrating a pattern of results that could not be accounted for by any previous model of apparent base-rate neglect.

Method

Participants. Seventy-one students participated for partial credit in an introductory psychology course at Indiana University. None had participated in any other related experiment in my lab.

Design. Experiment 4 was again a medical diagnosis task, this time involving only 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 randomly chosen according to the base rates, and symptoms were then 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 simultaneously be present.

The design evokes 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%; that is, $p(\mathrm{R} \mid \mathrm{s2}) = p(\mathrm{R} \mid \mathrm{s3}) = .50$. It was of interest to discover whether both symptoms produce apparent base-rate neglect, and if so, to what extent.

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

The explanatory principles advocated in this article make the following predictions: People 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 (those symptoms have high conditional probabilities; see Table 7). 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 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 when it occurs, it will merely add uncertainty.

Apparatus and 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 there is no information, positive or negative, about that symptom. The same ambiguity affects computation of normative probabilities when using Bayes's theorem.

Substitutive features have been used by some experimenters (e.g., Gluck & Bower, 1988, Experiment 3; Nosofsky et al., 1992) to avoid the ambiguity of present–absent features. However, substitutive features introduce another potentially confounding factor that has not previously been addressed. In previous experiments, the substitutive features have been dimensionally labeled, designating alternative values on an explicit dimension, for example, stuffy nose versus runny nose. Suppose that a person has learned that stuffy nose is typical of disease C but has yet to associate runny nose with either disease. Some people 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.

Table 7
Conditional Probabilities 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

Note. s = symptom; C = common disease; R = rare disease.

To explore these issues, I used in Experiment 4 the probability structure of substitutive symptoms, as described in Table 7, but I used symptom labels that did not explicitly mark the mutually exclusive alternatives. The six symptom 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 participant. As a test of whether the dimensional polarities were undetected, as intended, the participants were asked at the end of the experiment to guess which symptoms were mutually exclusive.

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 the corresponding position in the vertical list.

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

Procedure. For each participant, a new random sequence of 200 training trials was generated. The sequence had to respect all of the conditional probabilities in Table 7, and the base rates of the diseases, to within a margin of error of ±3%.

The instructions did not indicate that symptoms occurred in mutually exclusive pairs but did forewarn the participant that two patients with the same symptoms might have different diseases. Participants worked uninterrupted through 200 training trials. Immediately after training, they read instructions telling them that they would see cases with some symptoms that were either "missing or unknown," marked with a ?, and that they should make their best guess on the basis of what they had learned before. Participants were also informed that they would not be told the correct diagnosis but that their best educated guess was very important, so they should not just respond randomly. Participants saw two repetitions of 27 test cases, generated by crossing three symptom values s, −s, and s-missing (denoted ?s) for the three symptom pairs. In particular, one of these combinations had all three symptoms missing, and participants were thereby probed for their choice base rates.

After testing, and while still in the computer cubicle, participants were instructed to fill out a paper questionnaire that was concealed inside an envelope beside the computer. In the questionnaire, the participant was asked to estimate the frequency of the diseases, and he or she was told that the total number of training cases was 200 and that equal frequencies would mean 100 cases of each disease. On the questionnaire, the participants were also asked to judge which symptoms were mutually exclusive. It listed the six symptom labels, and for each symptom, the participant was asked to write in another symptom that never occurred in the same patient. The participant 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 participant.

Each trial was the same as what has previously been described in Experiments 1–3, but with two minor changes. First, the maximum time to make a diagnosis was reduced from 30 s to 15 s. Second, when the learner 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 motivate the learner, as the probabilistic mapping could cause the participant to feel defeated and frustrated.

Results and Discussion

Training. The actual sample probabilities, averaged across 71 participants, were each within .006% of the design values in Table 7.

The primary question for the training phase is, Did participants learn about the common disease before they learned about the rare disease? In the previous experiments, the diseases were deterministically related to the symptoms, so it could be presumed that if participants learned anything about the diseases, their accuracy would improve. Indeed, their accuracy could reach 100% correct, and it did for many individuals. 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 were only probabilistically related to the symptoms, and accuracy was not necessarily a good indicator of whether the participant had learned anything. Here "learning" means a change from one pattern of responding to some other systematic pattern of responding, regardless of whether overall accuracy was thereby improved.

The explanatory principles suggest that people first learn the typical symptoms of the common disease and subsequently learn the distinctive symptoms of the rare disease. Therefore, people 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). One prediction of the hypothesis, therefore, is that people will tend to diagnose the case s1, −s2, −s3 as the common disease, early in training, and subsequently tend to diagnose the case −s1, s2, s3 as the rare disease.

Table 8 shows the proportion of common-disease diagnoses for the two key training cases (s1, −s2, −s3 vs. −s1, s2, s3) and the proportion of common-disease diagnoses averaged across all cases. The table also shows the number of instances of each case in each division of trials. The overall rate of *common* diagnoses was greater than 50% very early in training, which could be interpreted as indicating that participants quickly learned that the base rates of the diseases were unequal. People also learned that cases of symptoms s1, −s2, −s3 were usually the common disease (the programmed population probability was $p(\mathrm{C}|\mathrm{s1},-\mathrm{s2},-\mathrm{s3}) = .897$) and that cases of symptoms −s1, s2, s3 were usually the rare disease (the programmed population probability was $p(\mathrm{C}|-\mathrm{s1},\mathrm{s2},\mathrm{s3}) = .250$). Unfortunately, there were not enough occurrences of −s1, s2, s3, early enough in training, to reject the null hypothesis that the common and rare tendencies were learned at the same time (the programmed population base rate of the exemplar was only $p(-\mathrm{s1},\mathrm{s2},\mathrm{s3}) = .011$). Even in the first block of 20 trials, people were diagnosing −s1, s2, s3 as the rare disease significantly more often than would be expected by the baseline response rate. For the first block of 20 trials, the overall proportion of common-disease responses was .671. If that proportion is applied to case −s1, s2, s3, then the expected number of *rare* responses is $r = (1 - .671) \times 18 = 5.92$, as opposed to the actual $r = 11, p < .013$.

Table 8
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	1,412	.389	18
21–40	.892	740	.732	1,419	.214	14
41–60	.889	687	.718	1,419	.100	10
61–80	.913	710	.762	1,420	.454	11
81–100	.912	714	.749	1,419	.208	24
101–120	.912	725	.772	1,419	.500	10
121–140	.921	700	.798	1,419	.409	22
141–160	.923	726	.804	1,420	.250	12
161–180	.899	721	.773	1,420	.428	14
181–200	.903	699	.790	1,419	.500	14

Note. Symptoms s1, −s2, −s3 are predicted to be typical of the common disease; symptoms s2 and s3 are predicted to be distinctive of the rare disease; symptom −s1 is predicted to be neutral. C = common disease; N = the number of instances.

In Table 8 it is also made clear that people responded differently depending on the symptoms; that is, they did not just ignore the symptoms and respond randomly according to the disease base rates.

An unfortunate consequence of using just two diseases was that I could not be sure why people tended to choose the rare disease when presented with case −s1, s2, s3. Learners 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 just two categories have always been used.)

In summary, these data do not allow a confident conclusion that the rare disease was learned later than the common disease; however, the data do not necessitate the conclusion that the diseases were learned at the same time, either, because nonassociative inferences might have driven some of the early responses.

Testing. Table 9 shows the proportion of rare-disease choices for each test case, out of 142 trials per case (71 participants times two test trials per participant per case). The same proportions are graphically displayed in Figure 2. As shown in Table 9, when all three symptoms were missing, participants chose the rare disease 23% of the time, almost exactly matching the base rate of the category.

Participants also exhibited apparent base-rate neglect for both symptoms s3 and s2, choosing the *rare* category 58% of the time for s3, χ^2 $(1, N = 142) = 4.056, p < .05$, using 50:50 for expected values, and 68% for s2, χ^2 $(1, N = 142) = 17.606$, $p < .001$. The magnitudes of the apparent base-rate neglect for the two symptoms did not differ significantly from each other, χ^2 $(1, N = 284) = 2.554$, *ns*.

The data in Figure 2 also reveal influences of individual symptoms that appear to be consistent across all combinations of other symptoms. For example, symptom +s1 usually influences choices toward the common disease, in that rare-choice

Table 9
Results of Test Trials in Experiment 4

s2	s3	s1		
		−	?	+
−	−	.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

Note. Numbers are the proportion of responses for rare disease. s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present.

probabilities for +s1 are usually less than for ?s1 (compare the curves labeled *?s1* and *+s1* in Figure 2). More robust effects are shown by symptoms −s2 and −s3, both of which consistently influence choices toward the common disease, relative to choices in response to ?s2 and ?s3. Symptoms +s2 and +s3 consistently influence choices toward the rare disease, relative to ?s2 and ?s3.

Symptom −s1 has a more subtle consistency, unlike the other symptoms. The data in Table 9 are plotted in Figure 2, in which it can be seen that there is a cross-over interaction of the curves labeled *?s1* and *−s1*. Thus, whenever s1 is missing (i.e., the curve labeled *?s1*), and the other symptoms produce choices that tend toward the common disease (as in the cases toward the left side of Figure 2), then adding symptom −s1 makes the choice proportion shift toward the rare disease. However, when the other symptoms produce choices that tend toward the rare disease (as in the cases toward the right of Figure 2), then adding −s1 makes the choice proportion shift toward the common disease. The interaction was significant, χ^2 (9, N = 2,551) = 35.70, $p < .0001$, when using the 36 cells involving ?s1 and −s1, but what makes it interesting was the cross-over. My interpretation of that result is that symptom −s1 was not actively encoded as part of either disease and merely added uncertainty when it appeared, shifting the choice toward random guessing.

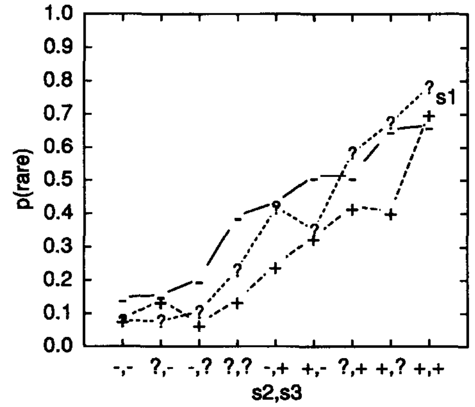

Figure 2. Test-trial data from Experiment 4. p(rare) = proportion of rare-disease choices; s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present.

Frequency estimation. As in the previous experiments, the value of log(C/R) was computed for each participant, 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.

Judgments of mutually exclusive symptoms. Several participants spontaneously reported to the experimenter that they were just guessing about mutually exclusive symptoms, and several participants left some response fields blank. No participant wrote in the same symptom as the probe symptom, so apparently they did understand the task.

For each symptom, the observed response frequency for the mutually exclusive symptom was tested against the expected value, which was determined by the total response frequency collapsed across all responses. For example, when probed with symptom s1, 14 participants indicated that −s1 was the mutually exclusive symptom, and 53 participants indicated other symptoms (none of which was s1 itself). Of 307 non-s1 responses overall, 51 were −s1; hence, the expected frequency of responding −s1 was 11.13, and the expected frequency of other responses was 55.87, yielding χ^2 (1, N = 67) = 0.8875, *ns*. For the other probe symptoms, chi-square values for the correct response were as follows: −s1, χ^2 (1, N = 65) = 9.90, $p < .005$; s2, χ^2 (1, N = 64) = 0.238, *ns*; −s2, χ^2 (1, N = 65) = 1.26, *ns*; s3, χ^2 (1, N = 65) = 2.21, *ns*; and −s3, χ^2 (1, N = 65) = 5.15, $p < .025$. The symptom with the largest chi-square value, −s1, was the symptom with the lowest overall base rate, and its complementary symptom, s1, had a very small chi-square value. Moreover, symptom −s3 had a significant chi-square value, but its complement, symptom s3, did not.

The significant chi-square value for −s1 may have been generated by a response strategy as follows: "I have rarely seen −s1, so its 'opposite' must occur a lot. So I will choose the most common symptom, s1." Analogous logic would not necessarily generate a significant chi-square value for s1 because participants 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 will choose one of those."

These results can be interpreted as indicating that participants most likely did not know which symptoms were mutually exclusive and, therefore, did not respond in training or testing by using nonassociative inference on the basis of dimensional polarity, as the design intended.

In summary, Experiment 4 demonstrated that apparent base-rate neglect can be obtained even for symptoms that occur in less than 50% of the cases of the rare disease and for nondimensionally labeled symptoms. Experiment 4 also demonstrated that the presence of symptoms that have not been associated with any disease, such as −s1, 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 consistent with the main theses of this article, that differential base rates in the inverse base-rate effect and apparent base-rate neglect (a) 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 (b) consistently bias responding to match the base rates, to the extent that other cues are absent.

In this section of the article I describe a new connectionist model that implements those principles. The logic of this modeling effort is the same as that used in the empirical studies: If the model fits the data reasonably well, then there is confirmatory evidence that the principles it embodies, and their formalization in the model, are accurate representations of human learning. The emphasis is on the explanatory power of the principles, not on the particular formalisms in the model. The principles gain cogency if they can successfully be formalized to fit data reasonably well. If the principles are correct, then it should be possible to implement them in formalisms other than those used here (e.g., production systems) and still match the data fairly well. 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*.[3]

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). The activation of the *i*th input node is determined by the presence or absence of the corresponding feature:

$$a_i^{\text{in}} = \begin{cases} 1 \text{ if feature } i \text{ is present} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

Every input node has a direct connection to every output node, with an initial connection weight of 0.0. A connection weight represents the learned degree of association between the feature and the category; that is, the weights represent the long-term category knowledge of the network. The activation of the *k*th output node represents how strongly the network classifies the current input into category *k*:

$$a_k^{\text{out}} = \sum_i w_{ki} a_i^{\text{in}}, \quad (2)$$

where w_{ki} is the connection weight from input node *i* to output node *k*.

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 have already learned 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 those aspects of the stimulus that are consistent with my current knowledge of X because those aspects produced the correct response. On the other hand, if I erroneously classify the stimulus (e.g., 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 and alter my knowledge about Category Y to include that distinctive feature (viz., Feature C). That is, whatever I learn about Category Y should deemphasize the features that caused me to choose the wrong category. Finally, if I classify the stimulus merely by guessing, I should modify my knowledge by using the entire stimulus.

Those intuitions can be 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

$$a_k^{\text{out}} = \sum_i w_{ki} \alpha_i a_i^{\text{in}}, \quad (3)$$

where α_i is the attention strength on input node *i*.

When a stimulus is presented to the network, attention initially is evenly distributed over all of the features that are present. 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

$$\alpha_i = N^{-1/\eta}, \quad (4)$$

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 $r = \eta$. That is, the normalization causes $(\Sigma_i \alpha_i^\eta)^{1/\eta} = 1.0$. 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.[4]

After the appropriate feature nodes are activated and the attention strengths are normalized, the output nodes are

[3] ADIT stands for attention to distinctive input, and, like other models that I have proposed, such as ALCOVE and AMBRY, it represents just a small part of the mental architecture. (An adit is an entrance to a mine.)

[4] The explanatory principles are not committed to this particular formalization for attention normalization. When η is arbitrarily set to 1.0 so that each feature gets an attention strength of 1/*N*, or when the attention strengths are not normalized at all ($\eta \rightarrow \infty$), the model still showed qualitatively correct results for the major effects.

activated according to Equation 3. 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 participants. Output node k receives a "teacher" value, t_k, which is determined by

$$t_k = \begin{cases} \tilde{1} = \max(1, a_k^{\text{out}}) & \text{if } k \text{ is correct} \\ \tilde{0} = \min(0, a_k^{\text{out}}) & \text{otherwise.} \end{cases} \tag{5}$$

Those teacher values are "humble" insofar as the activation of a node can exceed its target value (either less than 0.0 or greater than 1.0) without penalty (see Kruschke, 1992). Both the attention strengths and the association weights are adjusted in the direction that maximally decreases the error:

$$\mathrm{E} = .5 \sum_k (t_k - a_k^{\text{out}})^2. \tag{6}$$

The attention and weight adjustments are not simultaneously conducted, however. After receiving feedback, the model first decides which features in the stimulus are relevant to the present case, rapidly shifts attention to those features, recomputes its category predictions, and only then adjusts its knowledge on the basis of those attended features.

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

$$\begin{aligned} \Delta\alpha_i &= -\lambda_\alpha \partial \mathrm{E}/\partial\alpha_i \\ &= \lambda_\alpha \left[\sum_k (t_k - a_k^{\text{out}}) w_{ki} \right] a_i^{\text{in}} \qquad (7) \\ &= \lambda_\alpha \left[\underbrace{(\tilde{1} - a_c^{\text{out}}) w_{ci}}_{A} - \underbrace{\sum_{k \neq c} (a_k^{\text{out}} - \tilde{0}) w_{ki}}_{B} \right] a_i^{\text{in}}, \qquad (8) \end{aligned}$$

where λ_α is a freely estimated constant of proportionality, called the *attention shift rate,* and c is the index of the correct category. If attention strengths are driven to negative values by Equation 7, 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 renormalized by using the Minkowski-r metric, with $r = \eta$. Attention shifts in the model are intended to reflect large and rapid attention shifts in humans, and as a consequence, the shift rate, λ_α, can be quite large compared with typical "learning rates" in other gradient descent algorithms that are intended to converge gradually to a minimum.

Equation 8 captures the intuitions, expressed earlier, about what should be done with the knowledge after getting corrective feedback. Recall that the intuitions had two parts: one saying that features consistent with existing knowledge should be attended to when the response is correct, and the other part saying that distinctive features should be attended to when the response is in error. The A term in Equation 8 amplifies attention to features consistent with existing knowledge of the correct category. The B term reduces attention to features that match knowledge of erroneous categories, thereby leaving distinctive features attended to. This interpretation of Equation 7, in terms of the decomposition in Equation 8, is not post hoc; in fact, I started with some other 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.

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

$$\begin{aligned} \Delta w_{ki} &= -\lambda_w \partial \mathrm{E}/\partial_{w_{ci}} \\ &= \lambda_w (t_k - a_k^{\text{out}}) \alpha_i a_i^{\text{in}}, \end{aligned} \tag{9}$$

where λ_w is a freely estimated constant of proportionality, called the *weight learning rate,* and the indices k and i range over all category and feature nodes, respectively. To reiterate, the reason for adjusting weights after shifting attention is that the model, like a human, should learn about only those features to which it is attending.

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 by using the same variant of the Luce (1963) choice rule 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:

$$p_x = \exp(\phi a_x^{\text{out}}) \Big/ \sum_k \exp(\phi a_k^{\text{out}}), \tag{10}$$

where ϕ is a freely estimated scaling constant.

The network probabilities are then mixed with the base rates. The model assumes 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 should not have much influence, but if there are few or no features present, then the base rates should dominate. The formalism used to express this is as follows. Let b_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 b_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

$$p'_x = p_x b_x^{\beta/(\beta+N)} \Big/ \sum_k p_k b_k^{\beta/(\beta+N)}, \tag{11}$$

and β is a freely estimated constant called the *base-rate bias,* and N is the number of features in the current stimulus. Therefore, if there are many features in the stimulus, then N is large, and the base rates have a relatively small influence on the choice. If there are few features in the stimulus, then N is small, 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.[5]

The most important aspect of this method for combining base rates with choice probabilities is that base-rate knowledge is uniformly applied 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 model fits human performance, there is evidence that base-rate knowledge might 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 b_k are initialized to equal values, that is, one over the number of categories. They are then adjusted as follows:

$$\Delta b_k = \frac{1}{1 + T}([t_k] - b_k), \tag{12}$$

where T is the trial number ($T \geq 1$), and $[t_k]$ is the teacher value from Equation 5 clipped at 1 or 0. The values of b_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. The formalism is an instantiation of "direct coding" of frequencies, in which every occurrence of a category automatically adjusts an explicit representation of base rate (relative frequency). There is evidence that humans use both direct coding and "indirect coding," in which there is no explicit representation of frequency, but frequency information is retrospectively constructed by an accounting of distinct memory traces (e.g., Jonides & Jones, 1992; Jonides & Naveh-Benjamin, 1987). All that matters for the present model is that base-rate knowledge is acquired so it can be used in Equation 11. 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 normalization power (η in Equation 4), the attention shift rate (λ_α in Equation 7), the weight learning rate (λ_w in Equation 9), the choice probability scaling constant (ϕ in Equation 10), and the base-rate bias (β in Equation 11).

To recapitulate the sequence of events in the model: When a stimulus occurs, the input nodes are activated corresponding to features that are present, and attention is evenly distributed over all of the activated feature nodes, normalized according to the number of features present. 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 rapidly and extensively shifted away from features that cause error and toward features that reduce error. The attention is renormalized. The network output is recomputed on the basis of the new distribution of attention, and then the long-term association weights are adjusted to reduce the error further, on the basis of the features attended to. The process then repeats for the next trial. Notice that the attention strengths are equally distributed across all presented features at the beginning of every trial. The accumulated connection weights reflect which features tend to be attended to for each category.

In summary, the main mechanism that allows the model to address the inverse base-rate effect and apparent base-rate neglect is the attention-shifting mechanism. Attention is shifted so that features that match existing knowledge are attended to, to the extent that the response is correct, and distinctive features are attended to, to the extent that the response is erroneous. 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 consistently applied 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 was 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 participants (i.e., 200 different random orderings of training trials). Fits were measured as the root-mean-square 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 $\phi = 4.16$, $\beta = 0.268$, $\eta = 2.63$, $\lambda_w = 0.324$, and $\lambda_\alpha = 2.35$, yielding RMSD $= 0.0308$. Quantitative predictions and differences between predicted and empirical values are shown in Table 10.

The model showed a robust inverse base-rate effect for PC + PR and a base-rate consistency effect for I + PC + PR. Moreover, the model showed 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.

It is instructive to consider the association weights that developed by the time of testing. Table 11 shows the mean weights in the last training block before testing (averaged across eight training trials and 200 simulated participants). The inverse base-rate effect was produced by the fact that the

[5] The formula in Equation 11 is ad hoc, and the explanatory principles are not committed to its specific form. For example, a variant in which N is raised to the power $1/\eta$, instead of the fixed power of 1.0 in Equation 11, fits essentially as well.

Table 10
Best Fit of ADIT to Test Trials of Experiment 1

	Choice proportion							
	C		R		Co		Ro	
Symptom	P	Δ	P	Δ	P	Δ	P	Δ
I	.745	−.001	.137	−.037	.066	.017	.052	.021
PC	.903	−.030	.015	−.016	.046	.015	.036	.032
PR	.010	−.030	.947	.036	.024	.006	.019	−.012
PC + PR	.326	−.027	.553	−.058	.064	.042	.057	.043
I + PC + PR	.588	.008	.355	−.047	.030	.016	.027	.023
I + PCo	.347	−.059	.098	.018	.531	.062	.023	−.021
I + PRo	.244	.025	.069	−.016	.019	−.012	.668	.003
PC + PRo	.346	−.006	.015	−.012	.018	−.040	.620	.058
I + PC + PRo	.723	.004	.019	−.017	.012	−.024	.247	.037

Note. ADIT = attention to distinctive input model. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare; Co = the other common disease; Ro = the other rare disease. P = prediction; Δ = difference between predicted and empirical values.

weight from the distinctive symptom of the rare disease, PR, was much greater than the weight from the distinctive symptom of the common disease, PC. The difference in weights was large enough to override the base-rate influence.

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

Table 11 also reveals that there were no nonzero weights across pairs of diseases; for example, the weight from PC to Co is zero. That is a consequence of the fact that the model only adjusts weights for symptoms that are present (and attended to). 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).

The model was not fit to the training data but showed qualitative performance comparable to human learners, insofar as the common categories were learned earlier than the rare categories. At the end of training, the model attained 96% correct on the common categories and 87% correct on the rare diseases, as opposed to 98% and 95%, respectively, attained by human learners. In early training, however, the model performed much worse than human learners; for example, in the first block of training, human learners attained 68% and 49% correct for the common and rare diseases, respectively, but the model showed only 54% and 16% correct, respectively.

As noted previously, early training performance on the rare diseases might be elevated in humans because of mechanisms not implemented in the model. In the first block, participants performed well-above chance on rare diseases for which they had never before seen a case and, therefore, must have been relying on some nonrandom guessing strategy. It is also likely that humans performed relatively well in early training on the common diseases because of short-term memory across consecutively repeated trials, which was not implemented in the model. Evidence for this comes from other experiments run in my lab, in which some symptom–disease cases always occurred in clusters of three consecutive repetitions. Results from early blocks of training showed that accuracy on the first occurrence in a triplet was relatively poor; whereas performance on the second and third occurrences was excellent.

Fit to Experiment 2. Recall that in Experiment 2 I examined the effects of early training with two of the four diseases, followed by training with all four diseases occurring 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 only slight preference for the early trained diseases.

ADIT was fit to the test-trial data of Experiment 2 in the

Table 11
Association Weights for Best Fit to Experiment 1

	From symptom					
To disease	I	PC	PR	Io	PCo	PRo
C	.583	.717	−.206	.000	.000	.000
R	.225	−.224	.939	.000	.000	.000
Co	.000	.000	.000	.584	.717	−.206
Ro	.000	.000	.000	.225	−.224	.940

Note. For the diseases, C = common; R = rare; Co = the other common disease; Ro = the other rare disease. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease.

Table 12
Best Fit of ADIT to Experiment 2

	Choice proportion							
	E		L		Eo		Lo	
Symptom	P	Δ	P	Δ	P	Δ	P	Δ
I	.783	−.019	.129	.011	.044	.011	.044	−.003
PE	.929	.005	.014	−.000	.029	−.009	.028	.005
PL	.004	−.025	.969	.054	.014	−.005	.014	−.024
PE + PL	.246	−.070	.636	−.010	.059	.040	.059	.040
I + PE + PL	.514	.000	.416	−.051	.035	.021	.035	.030
I + PEo	.346	−.033	.102	.026	.521	.019	.030	−.013
I + PLo	.258	.049	.076	−.023	.015	.000	.652	−.026
PE + PLo	.360	−.014	.021	−.027	.014	.004	.605	.037
I + PE + PLo	.727	.029	.028	−.019	.011	−.017	.233	.007

Note. ADIT = attention to distinctive input model. For the symptoms, I = imperfect predictor of the two diseases; PE = perfect predictor of the early trained disease; PL = perfect predictor of the late trained disease. For the diseases, E = early trained disease; L = late trained disease; Eo = the other early trained disease; Lo = the other late trained disease. P = prediction; Δ = difference between predicted and empirical values.

same way as it was for Experiment 1, with only one change: When learning the base rates of the categories, the model reinitialized its trial number (T in Equation 12) to one when starting the second phase of learning. This reflects the fact that participants were explicitly told that two more diseases would be introduced in the second phase. Best fitting parameter values were $\phi = 4.40$, $\beta = 0.523$, $\eta = 1.78$, $\lambda_w = 0.173$, and $\lambda_\alpha = 1.22$, yielding RMSD = 0.0272. Best fitting predictions and the differences between predicted and empirical values are shown in Table 12.

The model showed a robust analogue of the inverse base-rate effect for PE + PL. The model also showed the base-rate consistency effect for I + PE + PL, which indicates relatively small preference for the early trained disease. Moreover, the model showed a strong preference for disease E in the case of I + PE + PLo.

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. Thus, after several trials in the first phase, the model's learned base rates were .5, 0, .5, and 0 for the four diseases, respectively, and after several trials in the second phase, the learned base rates were .25, .25, .25, and .25, respectively.

In fact, the model's base-rate knowledge was "too good" when compared with human participants, who estimated that the early training diseases occurred slightly but significantly more often than the late training diseases. The model could easily be modified so that its base-rate learning mechanism has more "inertia" from the first phase of training (e.g., by resetting the value of T in Equation 12 to a moderately high value at the beginning of the late training phase), but this would entail more free parameters, and the improvement in fit would probably be only slight for this particular experimental design.

Table 13 shows the mean weights in the last training block before testing (averaged across eight training trials and 200 simulated participants). The weights were similar to those obtained in simulating Experiment 1.

Fit to Experiment 3. Recall that in Experiment 3 I modified the design of Experiment 1 so it contained a combination of the designs for apparent base-rate neglect and the inverse base-rate effect. The basic results were that the distinctiveness of symptom PRn for the rare disease Rn in the neglect subdesign was strong enough to produce apparent base-rate neglect but was 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 300 simulated participants were used because of the additional combinatorial possibilities in the hybrid design. Best fitting parameter values were $\phi = 4.16$, $\beta = 0.420$, $\eta = 17.8$,[6] $\lambda_w = 0.344$, and $\lambda_\alpha = 4.78$, yielding RMSD = 0.0472. Best fitting predictions are shown in Table 14.

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

Table 15 shows the mean weights in the last training block before testing (averaged across eight training trials and 300 simulated participants). Of importance, symptom PRn has a

[6] The relatively large value of η does not necessarily indicate a deficiency for the model but reflects a characteristic of the formalism. The attention normalization scheme uses the Minkowski-r metric, and small increases in attention strength on each dimension can require large increases in the value of r. For example, with three present features, to raise the initial attention strength on each from .75 to .95 requires raising η from 3.82 to 21.42. Apparently, to learn the case In + PCn + PRn, the model must attend to all three symptoms, which requires a relatively large value for η.

Table 13
Association Weights for Best Fit to Experiment 2

	From symptom					
To disease	I	PE	PL	Io	PEo	PLo
E	.654	.788	−.303	.000	.000	.000
L	.245	−.160	.955	.000	.000	.000
Eo	.000	.000	.000	.654	.788	−.304
Lo	.000	.000	.000	.246	−.160	.956

Note. For the diseases, E = early trained disease; L = late trained disease; Eo = the other early trained disease; Lo = the other late trained disease. For the symptoms, I = imperfect predictor of the two diseases; PE = perfect predictor of the early trained disease; PL = perfect predictor of the late trained disease; Io = imperfect predictor of the other two diseases; PEo = perfect indicator of the other early trained disease; PLo = perfect indicator of the other late trained disease.

relatively weak association with the rare disease Rn, compared with the association weight from symptom PRi to disease Ri.

Although the model did capture the primary effects of interest, it also noticeably deviated from the data at a few points. The largest deviation was for the training pattern In + PCn + PRn. The model had difficulty achieving a high accuracy on that pattern because two of its three symptoms constituted the pattern for a different disease. Humans might surmount that problem by using mechanisms not available to the model. For example, humans might encode exemplars, that is, particular conjunctions of symptoms and diseases, instead of encoding only a prototype of attended symptoms, as ADIT does. Humans might also, with difficulty (Hearst, 1991), positively encode the absence of symptom PCn in the corresponding rare case, In + PRn. ADIT has no way of positively encoding the absence of a symptom.

The next largest deviations, for case Ii + PCi + PRn and case Ii, might be the result of anomalies in the data rather than the result of model deficiencies, at least to some extent. For the case Ii + PCi + PRn, the data showed a relatively large tendency to choose disease Cn, but that disease never appeared in training with any of the symptoms in that test case. For the case Ii, the data showed a relatively large tendency to choose Ci, apparently indicating that Ii was more strongly associated with Ci than In was associated with Cn. However, if that were true, then the test case In + Ii should show more responses for Ci than for Cn, which did not obtain.

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 participants appeared to have little knowledge of which symptoms were mutually exclusive, so this representation was reasonable.

The model was fit to the test-trial data of Experiment 4 by using the same 71 training sequences as were shown to human participants. Best fitting parameter values were $\phi = 1.62$, $\beta = 1.40$, $\eta = 1.75$, $\lambda_w = 0.152$, and $\lambda_\alpha = 3.06$, yielding

Table 14
Best Fit of ADIT to Experiment 3

	Choice proportion							
	Cn		Rn		Ci		Ri	
Symptom	P	Δ	P	Δ	P	Δ	P	Δ
In	.629	−.008	.210	−.058	.094	.064	.068	.002
Ii	.095	.059	.068	.009	.673	−.107	.164	.039
In + Ii	.375	.016	.131	−.042	.387	.022	.107	.005
PCn	.859	.025	.038	−.075	.060	.042	.043	.008
PCi	.074	.038	.053	.024	.829	−.064	.043	.002
PCn + PCi	.549	.061	.025	−.028	.402	−.009	.024	−.024
PRn	.167	.036	.705	−.069	.074	.033	.054	.000
PRi	.033	.027	.024	−.006	.023	−.006	.920	−.015
PRn + PRi	.036	.012	.334	.042	.016	−.020	.615	−.034
PCn + PRn	.569	.027	.382	−.017	.027	.009	.022	−.020
PCi + PRi	.031	.020	.026	.002	.286	−.035	.656	.013
In + PCn + PRn	.742	−.133	.241	.134	.009	−.003	.008	.002
Ii + PRi + PRi	.012	−.078	.010	−.014	.570	.087	.408	.004
In + PCi	.320	−.008	.111	−.044	.537	.037	.033	.015
Ii + PCn	.615	.080	.029	−.018	.279	−.066	.076	.005
In + PRi	.172	−.060	.057	−.032	.019	−.011	.752	.103
Ii + PRn	.084	−.001	.545	.039	.293	.028	.079	−.065
PCn + PRi	.360	.066	.014	−.016	.016	.010	.610	−.060
PCi + PRn	.071	−.006	.491	.032	.413	−.009	.025	−.017
In + PCn + PRi	.704	.007	.013	−.023	.007	.007	.276	.008
Ii + PCi + PRn	.019	−.118	.211	.039	.758	.115	.012	−.035

Note. ADIT = attention to distinctive input model. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = predictor of the rare disease. For the diseases, C = common; R = rare. n = base-rate neglect substructure; i = inverse base-rate effect substructure; P = prediction; Δ = difference between predicted and empirical values.

Table 15
Association Weights for Best Fit to Experiment 3

	From symptom					
To disease	In	PCn	PRn	Ii	PCi	PRi
Cn	.427	.704	.021	.000	.000	.000
Rn	.141	−.070	.655	.000	.000	.000
Ci	.000	.000	.000	.458	.605	−.080
Ri	.000	.000	.000	.133	−.089	.933

Note. For the diseases, C = common; R = rare. For the symptoms, I = imperfect predictor of the two diseases; PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. n = base-rate neglect substructure; i = inverse base-rate substructure.

RMSD = 0.0435. Best fitting predictions are shown in Table 16 and are plotted in Figure 3.

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

The model also nicely showed the cross-over interaction for symptoms −s1 and ?s1, which was exhibited in the human data (see Figure 3). In cases in which other symptoms led responses to the common disease, adding −s1 increased the tendency toward the rare disease; however, when other symptoms led responses to the rare disease, adding −s1 decreased the tendency toward rare.

The model produced that cross-over interaction for symptoms −s1 and ?s1 because the magnitude of the base-rate bias was modulated by the number of features in the input. When −s1 was added to the input, the base-rate bias was decreased, and −s1 itself added no strong preference to the outcome.

Association weights for the best fit to Experiment 4 are shown in Table 17. The weights reflect the predictions made earlier by the general principles: The common disease is represented by symptoms s1, −s2, and −s3; the rare disease is represented by symptoms s2 and s3, with symptom −s1 being only weakly, and nearly equally, associated with both diseases.

Table 16
Predictions of ADIT for Experiment 4

		s1					
		−		?		+	
s2	s3	P	Δ	P	Δ	P	Δ
−	−	.145	−.003	.098	.021	.088	.004
?	−	.202	.045	.119	.042	.114	−.027
−	?	.215	.011	.134	.029	.123	.052
?	?	.352	−.042	.253	.020	.168	.027
−	+	.395	−.037	.368	−.050	.266	.020
+	−	.425	−.089	.405	.053	.289	−.042
?	+	.576	.062	.623	.038	.390	−.032
+	?	.631	−.024	.700	.024	.445	.036
+	+	.735	.068	.782	.000	.607	−.097

Note. ADIT = attention to distinctive input model. s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present; P = prediction; Δ = difference between predicted and empirical values.

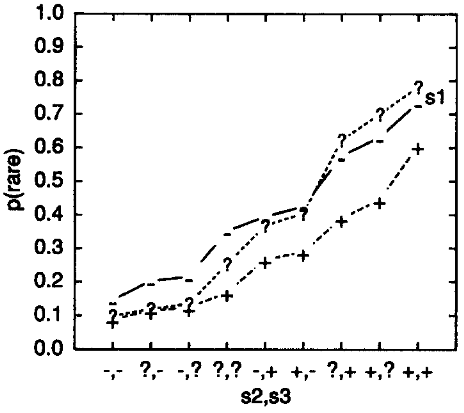

Figure 3. Predictions of ADIT for Experiment 4. p(rare) = proportion of rare-disease choices; s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present. ADIT = attention to distinctive input model.

Summary of fits by ADIT. The ADIT model provided a reasonably good fit to the test-trial data from all four experiments. It robustly showed the inverse base-rate effect, base-rate neglect, base-rate consistency, and the effects of features with no strong association to any category. The model showed some deviations from the training data for early training blocks and had difficulty learning the three-symptom pattern in Experiment 3, but probably because human participants used additional mechanisms not implemented in the model. Overall, the performance of the model is additional evidence that the underlying explanatory principles might be correct.

Previous Models

A variety of category-learning models previously have been proposed to address the inverse base-rate effect, apparent base-rate neglect, or both, but none of them has been shown to capture adequately all of 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.

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 the same mechanism with which it produces classical "blocking" effects (Rescorla & Wagner, 1972). In a simple blocking situation, there are two

Table 17
Association Weights for Best Fit to Experiment 4

	From symptom					
To disease	s1	−s1	s2	−s2	s3	−s3
C	.772	.208	−.003	.771	.027	.839
R	.063	.203	.918	−.036	.728	−.040

Note. C = common disease; R = rare disease. s and −s are mutually exclusive symptoms.

symptoms, s1 and s2, and a disease, X. If a participant is initially trained with only s1 → X and is subsequently trained on s1 + s2 → X, then the second symptom, s2, will not develop a very large association with the disease. That is, the prior learning of the s1 → X association blocks learning of the s2 → X association. Blocking occurs in the component cue model because association weights are error driven, and there is little or no error after the first phase of training.

The base-rate neglect situation is comparable: Consider a minimalist base-rate neglect structure, with two exemplars, s1 + s2 → C and s2 → R, that occur randomly mixed with equal base rates. Normatively, $p(\mathrm{R} \mid \mathrm{s2}) = .50$, but the model creates a stronger association weight from s2 to R than from s2 to C. The reason is that the change in weight from s2 on s1 + s2 → C trials is blocked by the learning of the s1-to-C association: As s1 becomes a good predictor of C, there is little error, and so the weight from s2 to C does not grow to the same extent as 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, 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 preasymptotically, which was not pointed out by Markman.

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* (ACM), in which cues (symptoms) develop different saliences depending on their base rates, or expectancies. Using a formalism inspired by Wagner (1978), Shanks (1992) defined the change in expectancy of feature *i* as

$$\Delta e_i = \lambda_e(a_i^{\text{in}} - e_i), \tag{13}$$

where e_i is the expectancy of feature i, and λ_e is a freely estimated constant of proportionality, called the *expectancy learning rate.* Initially, $e_i = 0$, for all i. Asymptotically, for small expectancy learning rates, $\mathrm{E}(e_i) = p(a_i^{\text{in}} = 1)$; that is, the expectancy of a feature is the base rate of the feature.

ACM uses the expectancies to modulate the learning of connection weights from individual features such that unexpected features have faster learning:

$$\Delta w_{ki} = \lambda_w(\mathrm{t}_k - a_k^{\text{out}})(1 - \mathrm{e}_i)a_i^{\text{in}}, \tag{14}$$

where λ_w is a freely estimated constant of proportionality, called the *weight learning rate.* In the simulations reported below, the expectancies were adjusted for a given trial (using Equation 13) before the weights were adjusted (using Equation 14).

ACM can account for the inverse base-rate effect, at least qualitatively, because the expectancy of symptom PR is lower than that of PC. 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 it provide quantitative fits, as its declared goal was to demonstrate robust qualitative effects.

The experiments described in this article provide new challenges for ACM. In particular, the results from Experiment 4 prove to be impossible for ACM to capture.

To fit ACM to the data, it was extended to include the base-rate learning and bias, and choice probability mapping, used in ADIT. Note that when the base-rate bias approaches zero ($\beta \rightarrow 0.0$, Equation 11), the base rates have no effect on the choice probabilities, except for the null stimulus ($N = 0$), in which case the choice probabilities match the learned base rates. Therefore, the extension of ACM can be reduced to ACM, except for the null case, for which it gains the advantage of base-rate matching. Moreover, the original component cue model is a special case of ACM, for which $\lambda_e = 0.0$. Therefore, if the extended ACM cannot fit the data, then neither can the original ACM nor can the original component cue model.

The extended ACM was fit to the test-trial data of Experiment 4 by using a hill-climbing parameter search, starting with $\beta = 0.001$ and $\lambda_e = 0.0$, so that it was effectively the original component cue model at the beginning of the search and could recruit higher values of β and λ_e if needed. The best fitting parameter values were $\phi = 3.44$, $\beta = 1.12$, $\lambda_w = 0.309$, and $\lambda_e = 0.0869$, yielding RMSD = 0.0810, far worse than ADIT.

Test-trial predictions of the extended ACM are shown in Table 18 and are plotted in Figure 4. Whereas the extended ACM showed some apparent base-rate neglect for symptoms s2 and s3, it badly misfit the effects of symptoms s1 and −s1. Symptom s1 did not shift choices far enough toward the common disease, and symptom −s1 shifted choices too far toward the common disease, relative to human performance. Table 19 shows that the association weights from symptom −s1 favored the common disease slightly more than the weights from symptom s1, quite unlike the corresponding weights in ADIT (Table 17).

The association weights for symptoms s1 and −s1, shown in Table 19, were the result of the learned expectancies of the symptoms s1, −s1, s2, −s2, s3, and −s3 and were .857, .143, .300, .700, .199, and .801, respectively. The learned expectancies accurately reflect the true base rates of the symptoms but caused learning of association weights from symptom s1 to be too slow and learning of association weights from symptom −s1 to be too fast. By contrast, the association weights learned by ADIT (Table 17) are quite different for symptoms s1 and −s1.

It is worth reiterating that the failure of the ACM entails the failure of the original component cue model. That is especially interesting because Experiment 4 had the same type of design

Table 18
Predictions of the Extended Attentional Connectionist Model for Experiment 4

		s1					
		−		?		+	
s2	s3	P	Δ	P	Δ	P	Δ
−	−	.071	−.077	.063	−.014	.053	−.032
?	−	.177	.020	.177	.099	.158	.017
−	?	.135	−.069	.130	.024	.117	.046
?	?	.290	−.104	.253	.020	.283	.142
−	+	.278	−.154	.310	−.109	.274	.027
+	−	.382	−.133	.446	.094	.389	.058
?	+	.487	−.027	.546	−.038	.510	.087
+	?	.537	−.117	.615	−.061	.568	.160
+	+	.665	−.002	.747	−.035	.707	.002

Note. s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present; P = prediction; Δ = difference between predicted and empirical values.

that the component cue model was shown to fit by Gluck and Bower (1988). The major difference between Experiment 4 and previous designs is the particular choice of conditional probabilities of symptoms given diseases: No previously published experiment had any symptoms with conditional probabilities equal for different diseases, as symptoms s1 and −s1 did in Experiment 4.

In closing this discussion of ACM, it should be emphasized that the attention strengths in ACM are very different than those in ADIT. In ACM, attention directly reflects the relatively long-term base rates of individual cues and is only used in learning. In ADIT, attention reflects the extent to which a cue is incorporated into learning and classification 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. As a final demonstration of the importance of the attention-shifting mechanism in ADIT, Table 20 shows the best fit of ADIT, with no attention shifting to the data from Experiment 1. Best fitting parameter values, with λ_α fixed at 0.0, were $\phi = 4.60$, $\beta = 0.01$ (lowest allowed value in the parameter search), $\eta = 0.75$, and $\lambda_w = 0.517$, yielding RMSD = 0.0946. Table 20 indicates clearly that no inverse base-rate effect was produced. Hence, an extension of the component cue model with attention normalization and base-rate bias, but without attention shifting, was not sufficient to show an inverse base-rate effect.

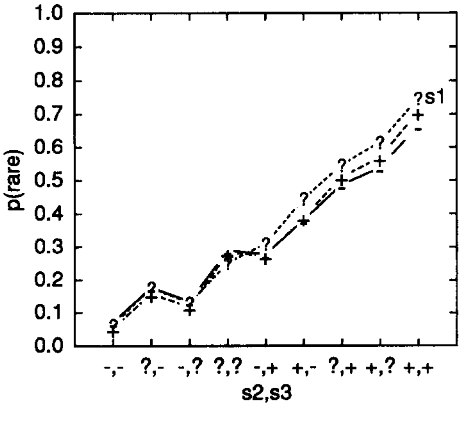

Figure 4. Predictions of the extended attentional connectionist model for Experiment 4. p(rare) = proportion of rare-disease choices; s = symptom; ? = symptom missing; + = symptom present; − = mutually exclusive symptom present.

Table 19
Association Weights of the Extended Attentional Connectionist Model for Experiment 4

	From symptom					
To disease	s1	−s1	s2	−s2	s3	−s3
C	.244	.256	−.013	.367	.043	.316
R	.121	.109	.322	−.058	.284	.010

Note. C = common disease; R = rare disease; s and −s are mutually exclusive symptoms.

Context model and extensions. The *context model* (Medin & Schaffer, 1978) uses an exemplar-based representation of category content, with classification decisions that are based on summed similarity to all exemplars of each category. In the generalized context model (Nosofsky, 1986), similarity is computed by using a city-block metric and exponentially decaying generalization gradient (for highly discriminable stimuli on separable dimensions). 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. In this scheme, 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 would have much lower attention. 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 uses a single set of attention strengths for all exemplars, however, and was therefore unable to account for the inverse base-rate effect (Kruschke, 1992; Nosofsky & Kruschke, 1992).

On the other hand, quantitative fits showed that ALCOVE was able to produce apparent base-rate neglect (Nosofsky et al., 1992; Kruschke, 1992); whereas the (generalized) context model could not. That result subsequently was shown to be a consequence of the particular training sequence used in the experiment (originally used by Estes et al., 1989), and, on

Table 20
Predictions of ADIT Without Attention Shifts for Experiment 1

	Choice proportion							
	C		R		Co		Ro	
Symptom	P	Δ	P	Δ	P	Δ	P	Δ
I	.663	−.083	.317	.142	.010	−.039	.010	−.021
PC	.998	.065	.000	−.031	.001	−.031	.001	−.004
PR	.000	−.040	.997	.086	.002	−.016	.002	−.030
PC + PR	.470	.117	.350	−.262	.090	.068	.090	.077
I + PC + PR	.499	−.082	.355	−.047	.073	.060	.073	.069
I + PCo	.194	−.212	.144	.064	.651	.182	.011	−.034
I + PRo	.232	.013	.173	.088	.013	−.018	.582	−.083
PC + PRo	.562	.209	.010	−.017	.010	−.048	.419	−.143
I + PC + PRo	.696	−.023	.055	.019	.025	−.011	.225	.015

Note. ADIT = attention to distinctive input model. For the symptoms, I = imperfect predictor of the two diseases. PC = perfect predictor of the common disease; PR = perfect predictor of the rare disease. For the diseases, C = common; R = rare; Co = the other common disease; Ro = the other rare disease. P = proportion; Δ = difference between predicted and empirical values.

average, ALCOVE cannot show apparent base-rate neglect (Lewandowsky, 1995).

The failure of ALCOVE to show apparent base-rate neglect should not be interpreted as implying that exemplar-based models in general are unable to show that effect because there are at least three important differences between ALCOVE and ADIT aside from exemplar versus prototype representation of categories. First, the attention strengths in ALCOVE are adjusted gradually and cumulatively across trials; whereas in ADIT, the attention strengths are redistributed and given potentially large shifts on every training trial. Second, the input representation in ALCOVE is of continuous dimensions; whereas the input representation in ADIT is nondimensionalized present–absent features. Third, ALCOVE does not explicitly learn base rates and mix them with network predictions in the way ADIT does.

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 formally equivalent to 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 on the basis of 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, it was shown that the model can exhibit a preference for the rare disease when tested with symptoms PC + PR, but it shows 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 and Edelson (1988), for example, symptom combinations such as I + PCo. In this 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 people 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) argued that participants in Gluck and Bower's (1988) experiment were confusing $p(\text{disease}\,|\,\text{symptom})$ with $p(\text{symptom}\,|\,\text{disease})$ and generally that the behavior of participants 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 attention shifting, 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) concept learning by exemplar memorization (CLEM) 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 degrades memory traces whenever they match an input. (The interested reader can consult Chumbley, 1986, 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 degradation of PC → C traces than of PR → R traces. Simultaneously, the base-rate consistency effect is produced by a greater number of I → C traces than I → R traces, generated when a correct response is based on weak evidence. Myers et al. did not discuss any other test cases. CLEM produces apparent base rate because of a different mechanism, the storing of 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. It would be informative to know whether CLEM can fit the data presented in this article.

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 different explanatory principles. In ADIT, both the inverse base-rate effect and apparent base-rate neglect are caused by shifting attention to distinctive features. 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. CLEM, on the other hand, treats the effects as stemming from different mechanisms.

Conclusion

Summary

In this article I have suggested that category base rates play two major roles in the experimental paradigms of the inverse base-rate effect and apparent base-rate neglect. First, the categories are learned at different times and, consequently, are encoded differently: The more frequent categories are learned earlier, in terms of their typical features, and the rare categories are subsequently learned predominantly in terms of their distinctive features. Second, people consistently favor the more frequent categories: Differential base rates are learned, and base-rate knowledge is used 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. The apparent inconsistencies in base-rate utilization observed in the inverse base-rate effect are merely apparent; people 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 participants do learn the common categories before the rare categories and that participants do know which categories are more frequent. Experiment 1 replicated the basic inverse base-rate effect and used fewer symptoms and diseases than the original study (Medin & Edelson, 1988), and the results showed that test case I + PC + PR did not produce the same response proportions as case I + PC + PRo. Experiment 2 verified that pretraining on a subset of categories produces effects comparable to the inverse base-rate effect. In Experiment 3 I used a hybrid design in which participants manifested 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.

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 rapid gradient descent on error with respect to the attention strengths. The attention shift also accentuates stimulus features that are consistent with current knowledge when current knowledge is correct. The second major extension of the component cue model was the use of base-rate knowledge in mapping network activations to choice probabilities. 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), there is evidence that both the underlying principles embodied in the model, and their specific formalization, are correct. On the other hand, the principles do not demand the particular formalizations described herein, and the extension of the component cue model used here was selected only by virtue of relative simplicity.

The Meanings of Typical and Distinctive

The terms *typical* and *distinctive* have been used loosely throughout this article because the general explanatory principles do not depend on specific definitions of them. Nevertheless, it may be useful to discuss their intended meanings at this point to clarify the claims being made and to indicate some directions for future research.

Distinctiveness can be defined as a physical measurement or as a psychological construct. Physically, a feature is distinctive of a target category, relative to a specified contrast category, to the extent that the feature occurs in the target category but not in the contrast category. A feature is psychologically distinctive of a target category, relative to some psychologically relevant contrast category(ies), to the extent that the feature is noticed to occur in the target category but is not known to occur in the contrast category(ies). Typicality is a degenerate case of distinctiveness, for which the relevant contrast categories are taken as null sets, that is, neutral background knowledge. All features are distinctive relative to neutral background knowledge, that is, typical, to the extent that they are noticed to occur. These definitions are imprecise but provide the basis on which the qualitative predictions made earlier in the article were derived.

The hypotheses about what people learn, in the experiments reported in this article, can be restated from this perspective: Participants encode the common diseases mostly by their distinctive symptoms relative to neutral background knowledge and encode the rare diseases mostly by their distinctive symptoms relative to the (already learned) common categories. This is also an accurate description of learning in ADIT.

The distinctiveness of a feature is reflected in ADIT on two different time scales. On the short time scale, the attention strengths indicate distinctiveness for a single trial, in that they indicate which features are distinctive of the current stimulus' category relative to other categories that are known at that moment. On the long time scale, the association weights for a category indicate which features tend to be distinctive of the category across many trials.

Future Research

The experiments reported in this article provided only indirect evidence that the rare categories are predominantly encoded by their distinctive features; future research could probe learned category knowledge more directly, perhaps by asking participants to list the features of the learned categories.

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 category 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. 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 terminology, the common categories are "positively defined concepts," and the rare categories are "negatively defined concepts.") In other experiments in my lab I found that simply giving participants 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 contrast) is sufficient to produce a weak analogue of the inverse base-rate effect.

There is still much to be learned about the psychology of typicality and distinctiveness. Goldstone (1993) suggested there is a continuum of category encoding from typicality to distinctiveness. One open question for future research is, Do human learners have separate representations for the typical features of a category and for its distinctive features relative to some contrast categories? From the perspective of a general-purpose adaptive learning system, it might be useful to retain knowledge of all the typical features of categories and compute their distinctive features relative to particular contrast categories "on the fly" as different situations imply different contrast categories. On the other hand, the process of learning all typical features of every category might be too costly, and the system would have to concentrate on learning the distinctive features of the categories in the particular learning situation. There is evidence that human novices emphasize distinctive features of categories; whereas experts have additional knowledge about nondistinctive features (Murphy & Wright, 1984).

There is also much more to be learned about how people acquire base-rate knowledge in category-learning tasks and how they utilize it. Contrary to the simplistic base-rate learning mechanism in ADIT, Experiments 2 and 3 suggested that frequency estimation is not affected by presentation frequency alone but also by how well the symptom-category associations are learned. This suggests that base-rate knowledge might depend on indirect, retrospective accounting of memory for distinct instances (e.g., Jonides & Jones, 1992; Jonides & Naveh-Benjamin, 1987), or the results might be accounted for by direct coding of frequencies, driven in part by the associative strength of the categories. The ADIT model also assumes that the magnitude of base-rate bias is only modulated by the number of features in the stimulus, but not by their content. Future research might force a revision of that assumption.

Whereas ADIT captured the data from the present experiments fairly well, it is applicable only to linearly separable categories. Future modeling will need to formalize the principles in more complex internal representations, such as exemplars, to address an even broader spectrum of category-learning data.

References

Anderson, J. R. (1990}. *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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

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.

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 15th 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 A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes* (Vol. 1, pp. 169–199). Hillsdale, NJ: Erlbaum.

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* (Report No. 88). Bloomington: Indiana University, Cognitive Science Program.

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

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. In L. W. Porter & M. R. Rosenzweig (Eds.), *Annual review of psychology* (Vol. 44, pp. 265–315). San Diego, CA: Academic Press.

Jonides, J., & Jones, C. M. (1992). Direct coding of frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 368–378.

Jonides, J., & Naveh-Benjamin, M. (1987). Estimating frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 230–240.

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

Koehler, J. J. (1993). The base rate fallacy myth [27 paragraphs]. *Psycoloquy* [On-line serial], *4*(49). Available World Wide Web: gopher://gopher.Princeton.EDU:70/1ftp%3Aprinceton.edu@/pub/harnad/Psycoloquy/1993.volume.4/psyc.93.4.49.base-rate.1.koehler

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychology 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 G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines: The psychology of learning and motivation* (Vol. 29, pp. 57–90). San Diego, CA: Academic Press.

Lenat, D. B., & Feigenbaum, E. A. (1987). On the thresholds of knowledge. In J. McDermott (Ed.), *Proceedings of the 10th joint conference on artificial intelligence* (pp. 1173–1182). San Mateo, CA: Morgan Kaufmann.

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

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

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.

Murphy, G. L., & Wright, J. C. (1984). Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 144–155.

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 D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 207–250). San Diego, CA: Academic Press.

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.

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

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 [7 paragraphs]. *Psycoloquy* [On-line serial], *4*(61). Available World Wide Web: gopher://gopher.Princeton.EDU:70/0ftp%3Aprinceton.edu@/pub/harnad/Psycoloquy/1993.volume.4/psyc.93.4.61.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 S. H. Hulse, H. Fowler, & W. H. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 177–210). Hillsdale, NJ: Erlbaum.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum.

Received May 26, 1994
Revision received January 12, 1995
Accepted February 1, 1995 ■