Journal of Experimental Psychology: General
1988, Vol. 117, No. 1, 68-85

0096-3445/88/$00.75

Problem Structure and the Use of Base-Rate Information From Experience

Douglas L. Medin
University of Illinois at Urbana-Champaign

Stephen M. Edelson
Pitzer College

SUMMARY

This article is concerned with the use of base-rate information that is derived from experience in classifying examples of a category. The basic task involved simulated medical decision making in which participants learned to diagnose hypothetical diseases on the basis of symptom information. Alternative diseases differed in their relative frequency or base rates of occurrence. In five experiments initial learning was followed by a series of transfer tests designed to index the use of base-rate information. On these tests, patterns of symptoms were presented that suggested more than one disease and were therefore ambiguous. The alternative or candidate diseases on such tests could differ in their relative frequency of occurrence during learning. For example, a symptom might be presented that had appeared with both a relatively common and a relatively rare disease. If participants are using base-rate information appropriately (according to Bayes' theorem), then they should be more likely to predict that the common disease is present than that the rare disease is present on such ambiguous tests. Current classification models differ in their predictions concerning the use of base-rate information. For example, most prototype models imply an insensitivity to base-rate information, whereas many exemplar-based classification models predict appropriate use of base-rate information. The results reveal a consistent but complex pattern. Depending on the category structure and the nature of the ambiguous tests, participants use base-rate information appropriately, ignore base-rate information, or use base-rate information inappropriately (predict that the rare disease is more likely to be present). To our knowledge, no current categorization model predicts this pattern of results. To account for these results, a new model is described incorporating the ideas of property or symptom competition and context-sensitive retrieval.

Much of the recent work on decision making has contrasted optimal performance with the use of strategies and heuristics that may be efficient and satisfactory in some contexts but lead to systematic errors or judgmental biases in other contexts (for reviews, see Einhorn & Hogarth, 1981; Kahneman, Slovic, & Tversky, 1982). The catalog of human errors is quite extensive and, indeed, one problem with analyses of heuristic principles is that it is difficult to predict in advance which heuristics will be used (Nisbett & Ross, 1980). However, we wish to focus on the observation that both prescriptive and heuristic approaches to decision making tend to be interpreted largely in terms of explicit strategies or rules. In this article we argue that the processes associated with learning from experience can lead people to behave in a systematic or rule-like manner independent of explicit rules or strategies (see also McClelland & Rumelhart, 1985). In particular, our experiments can be viewed as tracing the consequences of learning for the accessibility of disease categories in response to symptom probes. Our data support the contention that analyses of experience-based decision making need to focus not only on potential strategies but also on the structure of the problem domain and the learning processes brought to bear on it. The present experiments are concerned with sources of base-rate information conveyed through experience. Participants learn to classify patterns of symptoms into disease categories where the probability or relative frequency of diagnostic categories is varied.

On some grounds one might expect people to use base-rate information routinely. Imagine a physician who has noted a symptom that is only associated with two diseases, one of which appears a hundred times more frequently than the other. It would be quite surprising if the physician thought that the odds associated with the relative likelihood of the two diseases were even. Yet that is precisely what the consensus

Portions of this article were presented at the 27th annual meeting of the Psychonomic Society, New Orleans, Louisiana, November 13, 1986.

This research was supported by National Library of Medicine Grant LM 04375 and National Science Foundation Grant 84-19576 to Douglas. L. Medin. Brian Ross, Gordon Logan, Elke Weber, Lawrence Barsalou, and William K. Estes provided valuable comments on earlier versions of this manuscript.

Correspondence concerning this article should be addressed to Douglas L. Medin, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

of research on the use of base-rate information suggests (Kahneman & Tversky, 1973; Lyon & Slovic, 1976; Tversky & Kahneman, 1974), even when the subjects in the experiments are physicians (Balla, 1982; Balla, Elstein, & Gates, 1983; Casscells, Schoenberg, & Graboys, 1978).

Although the normatively correct decision would need to take into account cost and benefits of alternative classifications, failure to use base-rate information clearly reduces the effectiveness of decisions. Bayes's theorem provides an explicit formula for combining prior odds or base-rate information with new diagnostic information. Consider the well-known cab problem associated with the work of Tversky and Kahneman (1980):

> A cab was involved in a hit-and-run accident: Two cab companies, the green and the blue, operate in the city. You are given the following data.
>
> 1. 85% of the cabs in the city are green and 15% are blue.
> 2. A witness identified the cab as a *blue* cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When the witness was presented with a sample of cabs (half of which were blue and half green) the witness made correct identifications in 80% of the cases and errors in 20% of the cases.
>
> Question: What is the probability that the cab involved in the accident was blue rather than green? (p. 63)

According to Bayes's Theorem the correct answer should be 12/29 or .41. In contrast, the modal answer in a typical experiment is .80, which suggests that the base-rate information simply was ignored. This result and others like it can be understood in terms of judgmental heuristics. For example, a combination of the lack of salience of base-rate information and the use of what is known as the representativeness heuristic (the tendency to make judgments by reference to what appears to be normal or typical) is consistent with base-rate information being ignored.

Role of Experience

Almost all of the research showing that people do not use base-rate information derives from paper-and-pencil tests like the cab problem. This is analogous to a physician's reading about disease frequency in a medical textbook. However, in many contexts base-rate information is conveyed through experience: in the case of a physician, through his or her practice. Although there has been only a handful of studies on base-rate information derived from experience, the results are quite consistent in showing that base-rate information is used (Carroll & Siegler, 1977; Christensen-Szalanski & Beach, 1982; Christensen-Szalanski & Bushyhead, 1981; Manis, Dovalina, Avis, & Cardoze, 1980).

There are two distinct, but not necessarily mutually exclusive, ways of viewing these results on base-rate information and experience: First, one can incorporate these findings into the generalization that base-rate information is used whenever it is made salient. Other operations that also appear to affect salience are (a) repeatedly presenting copies of a specific word problem that are identical except for the base rate (Fischhoff, Slovic, & Lichtenstein, 1979), (b) presenting only base-rate information in the word problem (Kahneman & Tversky, 1973), and (c) providing a causal linkage between the base rate and the problem outcome (Ajzen, 1977; Bar-Hillel & Fischhoff, 1981; Tversky & Kahneman, 1980).

According to the salience interpretation, experience makes base-rate information more distinctive. There is, however, one piece of data that is not entirely consistent with this view. Christensen-Szalanski and Beach (1982) gave base-rate experience in two different ways. One group of subjects was presented with base-rate information alone: they saw slides corresponding to patients and were told whether or not the disease was present. The other group of subjects was given the same base-rate information, but they were simultaneously given information about the results of a diagnostic test (that had an 80% accuracy). Only the latter group showed sensitivity to base-rate information on a later test. In other words, the base-rate information alone, which should have been salient, did not influence judgments. Base rate affected judgments only when subjects experienced the relation between the base rate and the diagnostic information. (Incidentally, this sensitivity to base rate did not transfer to a similar problem given in pencil-and-paper form.)

A second alternative approach to results on base-rate information and experience is to view them as the products of learning processes that allow people to perform in a manner reflecting sensitivity to base rates without explicitly or directly using base-rate information. To quote Christensen-Szalanski and Beach, "in some cases the use of perception and memory may be sufficient to behave in a 'Bayesian manner'" (Christensen-Szalanski & Beach, 1982, p. 277).

There are clear precedents for viewing base-rate effects in terms of processing mechanisms that implicitly incorporate base-rate information. For example, it is well known that high-frequency words are perceived more readily than low-frequency words. Many models which are addressed to this word frequency effect, such as Morton's logogen model (Morton, 1970), do not assume that these effects are mediated by an explicit Bayesian procedure but rather rely on indirect or implicit effects produced, for example, by the lowering of the threshold of detectors.

A central claim of this article is that the focus on explicit decision-making strategies and procedures has been associated with a neglect of the role of processing mechanisms that indirectly or implicitly incorporate base-rate information in judgmental tasks. Thus, there is a tendency or temptation to dismiss work on experience and decision making as "mere probability learning." We argue, however, that probability learning is far more complicated than one might guess (see also Estes, 1976) and that greater attention to understanding the learning mechanisms associated with the experience of base-rate information is needed.

Experience, Base Rate, and Categorization Theories

Research on the use of base-rate information from experience is also relevant to contemporary categorization models. For example, if categories are represented strictly in terms of prototypes, and classifications are based solely on distances from (similarity to) alternative prototypes, then one would not expect base rate (in the form of category size) to influence classification judgments. One could, of course, supplement a

prototype model with a decision process that explicitly takes category size into account (see Fried & Holyoak, 1984). In contrast, many exemplar models of classification that assume that an item is classified on the basis of its similarity to previous examples of a concept (e.g., Medin & Schaffer, 1978) allow differences in category size to influence classification without incorporating any process for using base-rate information explicitly. We shall see, however, that our results on the use of base-rate information are far more complicated than these categorization models anticipate. There is no simple answer to the question of whether or how base-rate information is used, and we believe that our results have important implications for current categorization models.

This article reports some initial experiments on the use of base-rate information that is conveyed through experience. Depending on the structure of the learning task and the nature of the test, people will appear to be sensitive to, ignore, or even inappropriately incorporate base-rate information. We argue that these results can be understood in terms of two basic processes: (a) *retrieval failure* associated with *changes in context* and (b) competition among predictors of a particular outcome. A major implication of these principles is that one cannot predict how base-rate information will influence performance without knowing the problem structure of the domain of interest.

Overview

The organization of the remainder of this article is as follows: We first describe a series of studies on the use of base-rate information conveyed through experience that can be thought of as either classification or decision-making tasks. The initial results are quite surprising and the first set of studies is designed to both clarify and rule out certain artifactual interpretations of these findings. Specifically, we find that for some tests people appear to use base rates in an inverse manner, that is, to predict that a rare disease rather than a common disease is more likely to be the correct diagnosis. The studies demonstrate, however, that rarity itself is far less important than the predictive value or validity of the individual symptoms associated with the disease categories. These observations lead us to consider variations on the idea that there is a competition among sources of information to predict some outcome (e.g., the correct classification). That is, is it useful to think of the symptoms associated with a disease as competing to predict that disease? A variation on Rescorla and Wagner's (1972) theory of condition is used to formalize the notion of competition. Although the competition principle is well supported, it is not sufficient to account for the entire pattern of results. Current categorization models also fail to predict at least one of our three major results. Consequently, we introduce a second principle, based on the notion of context change between learning and test conditions. The key idea is that the absence of a symptom that has been associated with two distinct diseases may impair access to one of the disease categories more than access to the other. Finally, we consider implications of competition and context change both for approaches to decision making and for particular classification theories.

General Method

The general paradigm used in all of our studies is a (very low-fidelity) simulated medical problem-solving procedure. Participants initially go through a learning phase in which they are presented with information about symptoms, asked to make a diagnosis, and then are told the correct diagnosis. Two symptoms, which may differ in their predictive value, are presented on each trial. The symptom information, choices of disease category, and feedback on diagnostic classifications are all presented on a television screen controlled by a small computer. A sample sequence of three learning trials is shown in Figure 1. The order in which a pair of symptoms appears is randomly determined on each trial as is the order of the set of diagnostic choices. Note that a given symptom (e.g., earaches in Figure 1) may appear with more than one disease.

Base-rate information is presented by means of variation in the relative frequency of different diseases. For example, "burlosis" may be three times as frequent as "namitis." In most studies there is no intended meaningful association between the two symptoms of a pair or between a symptom and a disease category. The exact assignment of symptoms to pairs and symptoms to category labels (diseases) is randomized for each individual subject, constrained only by the more abstract experimental design to be described shortly. Learning trials continue until a learning criterion is met or until a fixed number of learning trials has been given.

The learning phase is followed by a series of transfer tests, designed in part to evaluate the use of base-rate information. On these transfer tests participants are presented with tests

	Symptoms: earaches, dizziness
	1. burlosis
	2. namitis
	3. terrigitis
	4. coralgia
Choice?	5. althrax
3	6. buragamo
Correct diagnosis is coralgia	
	Symptoms: skin rash, sore muscles
	1. terrigitis
	2. burlosis
	3. althrax
	4. namitis
Choice?	5. buragamo
2	6. coralgia
That's correct!	
	Symptoms: back pain, earaches
	1. coralgia
	2. althrax
	3. buragamo
	4. terrigitis
Choice?	5. namitis
1	6. burlosis
Correct diagnosis is terrigitis	

Figure 1. A hypothetical sequence of three trials in the experiments showing the symptoms, diagnostic categories, and the form in which feedback was given.

involving either individual old symptoms or new combinations of old symptoms. In this phase of the experiment no feedback concerning correctness and incorrectness of classifications is given and participants are asked to make their diagnoses "based on what you learned before."

The abstract design of a typical base-rate experiment is shown in Table 1. Single letters substitute for symptoms and the numbers on the right stand for particular disease names. Note that the diseases vary in their relative frequency and that for each symptom pair, one symptom appears with more than one disease whereas the other is uniquely associated with a disease. Although symptoms of both types occur naturally in more realistic circumstances, the learning situation was quite contrived and we did not attempt to develop a high fidelity simulation of medical problem solving. (In addition, our subjects were college students, not medical personnel.) Our justification for the particular design is the more general issue of processing mechanisms associated with the use of base-rate information, and the main consequence of the simulated medical setting is that subjects find this to be a fairly engaging task. Later on, we shall offer some speculations concerning how expertise might alter performance in more realistic tasks.

The use of base-rate information was assessed on the transfer tests; three such types of tests are noted in Table 1. These derive from the distinction between perfect and imperfect predictors. A symptom was a perfect predictor if it is associated with only a single disease. For the *imperfect predictors* participants were presented with a single symptom that appeared with two different diseases. To the extent that base-rate information influences judgments, one might expect subjects to respond with the more frequent of the two diseases associated with a symptom. For example, for symptom *a* one would expect more disease 1 than disease 2 classifications. The other main test for base-rate information, referred to as *conflicting tests*, paired two symptoms that had been perfect predictors (for different diseases) but one of the associated diseases had appeared more often than the other. Again, to the extent that responses are derived from base-rate information, one would expect participants to predict that the more frequent disease is present. For example, on a *b,c* test, one would expect more disease 1 than disease 2 classifications. Similarly, one would expect more disease 1 than disease 4 responses on a *b,f* test. The third type of test, the *combined test*, was a simple combination of the first two types of tests. All the symptoms consistent with two different diseases were present and the subject had to choose which disease was more likely to be present. Finally, numerous other types of tests are typically given in order to disguise the primary purpose of the study.

The transfer tests may be somewhat more realistic than the learning trials in that physicians are often presented initially with information that may be both incomplete and ambiguous. Elstein, Shulman, and Sprafka (1978), using high-fidelity medical problem-solving tasks, found that physicians typically generate a set of candidate diagnoses from this initial information and work by elimination. If the correct diagnosis is not in this initial candidate set, then it is very unlikely that the correct diagnosis will be made. Because our subjects reported that they typically classified on the basis of the first disease that occurred to them, one can think of our transfer tests as an index of which disease classifications are likely to be at the top of a list of candidates. In some sense, the correct answer on conflicting and combined tests is that both diseases are present. It may be, however, that one disease category is more accessible than the other. In more realistic situations it is not uncommon to have conflicting symptoms that suggest multiple diseases even when multiple diseases are not present. Again, however, we wish to make no strong claims that we have captured this aspect of medical diagnosis.

The subjects in all our studies were male and female undergraduates who either participated in partial fulfillment of a course requirement or were paid for their services. All of the experiments lasted approximately 1 hr.

Table 1
Abstract Representation of a Typical Experiment

Learning: relative frequency	Symptoms	Disease
3	*a,b*	1
1	*a,c*	2
3	*d,e*	3
1	*d,f*	4
3	*g,h*	5
1	*g,i*	6
Transfer: Tests for base-rate information		

1. Imperfect predictors: *a, d, g*
2. Conflicting tests: *b, e,* or *h* vs. *c, f,* or *i* (e.g., *b,c*; *b,f*)
3. Combined test: *abc, def, ghi*

Note. The symptoms are represented in terms of single letters and the diseases in terms of numbers. The diseases differ in their relative frequency. See the text for concrete examples of symptoms and disease names.

Experiments 1, 2, 3, and 4

Our initial experiment used the design and test for base-rate information shown in Table 1. The puzzling results, shortly to be described, led to a series of follow-up studies (Experiments 2, 3, and 4) designed to clarify these initial findings.

Experiment 1

The first experiment assessed the use of base-rate information where base-rate information was conveyed through the relative frequency of different disease categories in a simulated medical problem-solving task.

Method

The abstract design of the learning materials was exactly that shown in Table 1. The nine symptoms were bad knees, body bruises, fallen arches, irritated eyes, jammed fingers, sore elbow, sprained ankle, strained shoulder, and swollen arms. The six disease names used were althrax, buragamo, burlosis, coralgia, midosis, and terrigitis. The symptoms and diseases were randomly assigned for each subject to the abstract structure in Table 1.

Learning consisted of up to 16 blocks of trials where a block is defined as the number of trials (12) needed to realize the relative frequency structure of Table 1. As shown in Figure 1, each trial began with the presentation of two symptoms and a set of disease categories (six, in this experiment) to choose from. After a classification decision was made, either the word *correct!* was displayed or, in the case of an error, the correct diagnosis was given. In addition to instructions concerning the general procedure, subjects were further told that by paying attention to the feedback, they could eventually be correct every time. The sequence of trials within a block was randomized. The learning phase was terminated when a participant was correct on every trial for 2 consecutive blocks or after 16 blocks. Responding was subject paced. There was no pause between blocks, but for each trial there was a self-paced feedback interval followed by a 500-ms intertrial interval.

Transfer tests immediately followed the learning phase. Subjects were told they would see new types of symptom conditions and that they should make what they thought was the correct diagnosis on the basis of what they had learned before, but that they would not be told whether or not their diagnosis was correct. Subjects were tested on the nine different symptoms presented individually, the nine possible pairs of conflicting symptoms differing in frequency, and the three combined tests listed in Table 1. In addition, 12 other paired tests were given consisting of a single perfect predictor (high or low frequency) combined with an imperfect predictor that had been associated with two other diseases (e.g., b, d). The order of the transfer tests and the order of the symptoms for a given test pair or triple were randomized.

After the 33 transfer trials, the subjects were given an extensive debriefing concerning the purpose of the study and asked for their informal observations. Altogether, 32 undergraduates participated in the study.

Results

All 32 participants met the learning criterion in 14 blocks or less (out of the maximum of 16). The principle transfer results are shown in Table 2.[1] The number of observations associated with a test ranged from 96 to 192 and the average standard error on the proportions in Table 1 was .033. Performance on the single perfect predictors was fairly good, with a nonsignificant trend for better performance on the perfect predictors having a lower relative frequency. Of greater interest, on the test involving imperfect predictors a large majority of the responses consisted of one of the two associated disease categories and there was a strong tendency for participants to prefer the more frequent disease category (relative proportion = .843), $t(31) = 9.93$, $p < .001$. This is what one would expect if subjects were using base-rate information appropriately. Responses on the paired conflicting test were surprising with respect to the use of base-rate information. Participants showed a strong overall trend to predict that the less frequent disease was present (relative proportion = .644), $t(31) = 4.03$, $p < .001$. Although one might expect that the combined test would yield an average of the imperfect predictor and conflicting test, responses on this test mapped onto relative frequencies almost as well as they did for the imperfect predictor tests (relative proportion = .716), $t(31) = 4.65$, $p < .001$. Finally, on the tests pitting perfect predictors against imperfect predictors there is additional evidence that low-frequency perfect predictors influenced choices more than high-frequency perfect predictors (a .594 proportion vs. a .385 proportion).

Table 2
Transfer Results: Response Proportions in Experiment 1

Symptoms	HF	LF	Other
Single			
High-frequency perfect predictors	.812		.188
Low-frequency perfect predictors		.927	.073
Imperfect predictor test (e.g., *a*)	.781	.146	.073
Paired			
Conflicting test (e.g., *b,c*; *b,f*)	.323	.584	.094
Triples			
Combined test (e.g., *a,b,c*)	.708	.281	.010

		Imperfect		
Comparison	Perfect	HF	LF	Other
High-frequency perfect (e.g., *b-d*) vs. imperfect predictors	.385	.474	.068	0.73
Low-frequency perfect (e.g., *c-d*) vs. imperfect predictors	.594	.318	.052	.036

Note. Response proportions are broken down by type of test and according to whether the response was associated with a relevant high frequency category, relevant low frequency category, or some other category. HF = high frequency; LF = low frequency.

Combining tests into analyses of variance (ANOVAS) confirms the interactions noted in Table 2. An ANOVA with the variables of frequency and test type (imperfect, conflicting and combined) indicated that the effects of frequency, $F(1, 31) = 35.52$, $MS_e = .869$, $p < .01$; tests, $F(2, 62) = 3.77$, $MS_e = .072$, $p < .05$; and the Frequency × Tests interaction, $F(2, 62) = 32.50$, $MS_e = .974$, $p < .01$, all were significant. For the tests involving perfect versus imperfect predictors we combined the responses to the high-and low-frequency disease associated with the imperfect predictor. Neither of the main effects of frequency or test type was significant, but the Frequency × Tests interaction, $F(1, 31) = 11.11$, $MS_e = 3.75$, $p < .01$, was statistically reliable.

Discussion

The results were both mixed and unexpected from the point of view of base rates. On the tests involving imperfect predictors and on the combined tests, the disease category with the greater relative frequency was preferred, indicating appropriate use of base-rate information. On the conflicting test,

[1] Although the tables present response probabilities, all analyses were conducted with response frequency as the dependent variable. We use the .05 level for our tests of statistical significance. Because there are multiple tests associated with each experiment, there is a considerable risk of Type I errors. Our primary criterion for effects is that they be replicable across experiments, and the three major results reported in this article are repeatedly obtained both in the present experiments and in other studies conducted in our laboratory.

however, the opposite pattern of responding was observed: Subjects tended to predict that the *less* frequent disease was present. Furthermore on the other tests pairing perfect predictors with imperfect predictors, the low-frequency perfect predictors controlled classification significantly more than did the high-frequency perfect predictors. So we have two types of tests that show appropriate use of base-rate information and two types that show the opposite!

During the debriefing we asked subjects how they made their decisions on transfer tests. Although participants were aware that some diseases appeared more frequently than others, they did not use disease frequency (base rate) to justify their responses. A typical response was that when a test was given, a disease name would "pop" into awareness. Although subjects mentioned that they knew that more than one disease had been associated with a symptom or a new symptom pair, apparently one disease category tended to be more accessible than the others and the subjects said they responded with the first disease name that came to mind. Some subjects indicated that on conflicting tests, they gave the disease associated with the more severe symptom of the pair. Because the assignment of symptoms to diseases was randomized across subjects, this pattern of responding should only lead to equal response proportions across subjects, not a preference for the disease which had the lower relative frequency.

The results on the imperfect predictor tests were in the expected direction and are consistent with a considerable literature on probability learning (e.g., Binder & Estes, 1966). Actually, the results on the conflicting tests are not entirely unexpected, in that Binder and Estes reported this identical tendency for participants to give the response associated with the less frequent of two perfect predictors. They referred to this tendency as the "novelty effect" and speculated that stimuli that appeared less frequently capture attention because of their novelty. This would not, however, explain why the more frequent disease was chosen more often in the combined test. Before discussing that and related interpretations, however, we present a series of follow-up studies designed to clarify our initial results.

Experiment 2

In Experiment 1, every symptom pair contained a symptom that was an imperfect predictor of the disease category. Experiment 2 evaluated the role of these imperfect predictors or common symptoms in producing the inverse use of base rate information or "novelty effect" on the conflicting test. One rationale for the study is strictly empirical. In a study using a paradigm similar to that used by Binder and Estes (1966) but using stimulus compounds that did not have any common components, Medin and Robbins (1971) failed to observe the "novelty effect." Unfortunately, they did not include a comparison condition having common components. The present experiment used compounds of symptoms which either did or did not have symptoms which were imperfect predictors or common cues.

The abstract design of Experiment 2 is shown in Table 3. Triples rather than pairs of symptoms were used in order to vary the number of ambiguous symptoms in a compound. The top pair of diseases in Table 3 contains two imperfect predictors (*a* and *b*), the middle pair contains one (*e*), and the bottom pair has no imperfect predictors associated with it. The tests for base-rate information were the same as those used in the first study. If the presence of imperfect predictors is responsible for the inverse use of base-rate information, then the tendency to choose the less frequent symptom on conflicting tests should increase as the number of common cues paired with an imperfect predictor increases.

Table 3
Design of Experiment 2 and Transfer Tests for Use of Base-Rate Information

Relative frequency	Symptoms	Disease
3	*a,b,c*	1
1	*a,b,d*	2
3	*e,f,g*	3
1	*e,h,i*	4
3	*j,k,l*	5
1	*m,n,o*	6
Transfer: Tests for base-rate information		

1. Imperfect predictor tests: *a*, *b*, or *e*
2. Conflicting tests: *c*, *f*, *g*, *j*, *k*, or *l* vs. *d*, *h*, *i*, *m*, *n*, or *o* (e.g., *c,n*)
3. Combined test: *a,b,c,d*, *e,g,i*

Note. The symptoms are represented by letters and the disease categories by numbers. The diseases differ in their relative frequency. See the text for concrete examples of symptoms and disease names.

Method

The design of the learning materials conformed to that shown in Table 3. The 15 symptoms were back pain, constant fever, coughing, dizziness, dry mouth, dulled hearing, hair loss, nausea, puffy eyes, skin rash, slurred speech, stiff neck, sweaty palms, swollen tongue and tired easily. The disease names were buragamo, coralgia, gouphosis, midosis, namitis, and terrigitis. The symptoms and diseases were randomly assigned for each subject to the abstract structure of Table 3. The order in which the symptoms in a compound appeared was randomized for each trial.

Learning consisted of up to 16 blocks of trials in which a block is composed of the 12 individual trials to present the symptoms and diseases in accordance with the relative frequencies indicated in Table 3. Other details of the learning phase were exactly as in the first experiment.

Transfer tests immediately followed the learning phase and the general procedure was the same as in the first experiment. There were 60 different new types of transfer tests, consisting of 8 tests involving single symptoms, 30 with pairs of symptoms, 19 using triples of symptoms, and 3 involving quadruplets. The single-symptom tests involved all of the different logical types rather than each of the individual symptoms. For example, either symptom *a* or symptom *b* but not both would appear as a transfer test, because they are conceptually equivalent. The test pairs included the 15 possible combinations of the 6 different logical types (based on their relative frequency and compound type) of the symptoms that were perfect predictors. The full set of 60 tests is listed in Appendix A. Altogether, 40 undergraduates participated in the study.

Results

Of the 40 participants, 39 met the learning criterion within the 16-run limit and the average proportion of errors on the last block of training trials was .004.

The transfer test data of greatest interest are shown in Table 4. The number of observations for a given test ranged from 40 to 200 and the average standard error was .064. Although the overall performance was lower, the high-frequency category was preferred over the low-frequency category on the imperfect predictor test. This tendency was greater when there had been two common symptoms than when there had been only one. An ANOVA with number of common symptoms and frequency as variables showed that only the main effect of frequency was reliable, $F(1, 39) = 8.96$, $MS_e = .308$, $p < .01$. Tests on single, perfectly valid symptoms (i.e., perfect predictors) revealed a slight tendency for better performance on high-frequency symptoms than low-frequency symptoms (overall proportion correct, .708 vs. .667) and a somewhat greater tendency for performance to decrease as the number of common symptoms decreased. An ANOVA with frequency and number of common symptoms as variables indicated that neither of these effects was significant.

The conflicting tests showed that performance depended strongly on the presence of common symptoms. The results of the first experiment were replicated for the case where there were two common symptoms during training. When there were one or no common symptoms during training, however, there was a tendency for the high-frequency symptom to control responding. In general, as the number of common symptoms increased, control by the associated low-frequency perfect predictor increased and control by the high frequency perfect predictor decreased. This interaction between number of common symptoms and frequency was statistically reliable, $F(2, 78) = 3.28$, $MS_e = .398$, $p < .05$.

The combined tests show the same pattern noted in Experiment 1: a tendency to predict the more frequent disease. For the *a,b,c,d* test this trend was significant (relative proportion = .703), $t(39) = 2.64$, $p < .05$. We also included an *a,b* test involving two common symptoms. Performance again proved to be a direct function of relative frequency (relative proportion = .743) $t(39) = 3.18$, $p < .01$. Finally, an ANOVA was conducted with frequency and the three main types of tests (*a,b* vs. *c,d* vs. *a,b,c,d*) as variables. Neither the main effect of frequency, $F(1, 39) = 3.77$, $MS_e = .358$, $p < .06$, nor the main effect of tests ($F < 1$) was significant, but the interaction of these two variables was statistically reliable, $F(2, 78) = 9.42$, $MS_e = .399$, $p < .01$.

Table 4
Transfer Test Results: Response Proportions in Experiment 2

Test	HF	LF	Other
Single symptoms			
Imperfect predictor test			
2 common symptoms (e.g., *a* or *b*)	.575	.200	.225
1 common symptom (i.e., *e*)	.375	.225	.400
Perfect predictors			
High frequency			
2 common (i.e., *c*)	.800		.200
1 common (e.g., *f* or *g*)	.650		.350
0 common (e.g., *j*,*k*, or *l*)	.675		.325
Low frequency			
2 common (i.e., *d*)		.750	.250
1 common (e.g., *h* or *i*)		.675	.325
0 common (e.g., *m*, *n*, or *o*)		.575	.425
Conflicting			
2 common	.275	.625	.100
1 common	.400	.350	.250
0 common	.500	.375	.125
Combined			
Quadruplet (*a,b,c,d*)	.650	.275	.075
Triplet (*e,g,i*)	.475	.350	.175

Note: The data are broken down by type of test and responses classified by the relative frequency of the associated category or categories. HF = high frequency; LF = low frequency.

Discussion

The results replicate and extend the findings of the first experiment. Responses on the imperfect predictor tests were consistent with the appropriate use of base-rate information and responses on the conflicting test where two symptoms which were imperfect predictors had been associated with the perfectly valid symptom again revealed an inverse use of base-rate information. And again the combined test showed a direct use of relative frequency. That the inverse base-rate effect depended on the presence of two common cues suggests that symptom frequency or "novelty" alone is not responsible for this effect. In the absence of imperfect predictors there was no "novelty effect" or any overall effect of test responding. Thus, the present results are consistent with Medin and Robbins (1971) when no imperfect predictors are present and with Binder and Estes (1966) and Experiment 1 when more than one imperfect predictor is present. Therefore, our results suggest that the presence of common or imperfect cues in the training examples is necessary for the appearance of our inverse base-rate effect.

Although this pattern of results is puzzling with respect to frequency or novelty principles, it is at least partially consistent with a conceptualization growing out of a body of research on animal conditioning (e.g., Rescorla & Wagner, 1972; Rudy & Wagner, 1975). The main idea is that component cues associated with some outcome compete to predict that outcome. In the specific context of conditioning the claim is that a given unconditioned stimulus (US) has a fixed amount of strength that can be distributed among its predictors. The total predictive strength associated with cues is limited by this amount of US support. The better a predictor is, the more effectively it will compete for this fixed strength. Consider again the *a,b* symptom pair and the *a,c* pair in Table 1 in relation to this competition principle. Symptom *a* competes with symptom *b* to predict disease 1, and competes with symptom *c* to predict disease 2. Because symptom *a* is a better predictor of disease 1 than disease 2, it should compete more effectively with symptom *b* than symptom *c*. Because the total predictive strength is a constant, *b* will be weakened more by *a* than is *c*, and that therefore *c* will be stronger than *b*. Hence, on the *b,c* test one would expect *c* to be more likely to control responding. That is, the notion of competition leads naturally to behavior conforming to the inverse use of base-rate information on conflicting tests. Furthermore, the competition principle predicts that this effect will hinge on the presence of imperfect predictors in learning compounds. Because there are no common symptoms for diseases 5 and 6, each of the

symptoms is competing with other equally valid symptoms and should, therefore, have equal strength or predictive value. The results of Experiment 2 are consistent with this prediction.

The competition principle leads to a further, counterintuitive prediction. Consider again the *a,b* and *b,c* pairs in Table 1. According to the notion that subjects may have response biases, one might expect a positive correlation between disease 1 categorization responses on the *a* test and disease 1 responses on the *b,c* test. In contrast, the competition principles would predict a negative correlation. The data reveal a small but systematic pattern of negative correlations. For example, in Experiment 1 the correlations between responses on these two types of tests were −0.281, −0.012, and −0.054 for the three appropriate pairs of tests. Although we do not report these correlations elsewhere in this article, they are consistently negative.

There are some differences between the general competition principle and the formal model of conditioning offered by Rescorla and Wagner (1972). In Rescorla and Wagner's model, with extended training on *a,b* associated with disease *i* and *a,c* associated with disease *j*, *b* and *c* will have all of the predictive strength and the ambiguous symptom *a* will have none. Pre-asymptotically, however, the predicted pattern of symptom competition would be consistent with inverse base rate responding on a *b,c* test. Because we are not presently considering Rescorla and Wagner's model at this specific level of detail, we will content ourselves with noting the conceptual similarity in the conditioning domain and the present paradigm, and stick with the general (and somewhat vague) notion of competition among symptoms to predicted diseases. Gluck and Bower (1986) have recently reported some data on human categorization that fits very nicely with the Rescorla-Wagner competition model. The competition principle is, however, by itself incomplete and unable to account for all of our main results. Its most obvious failure is that it predicts no preference on the combined (e.g., *a,b,c*) test, whereas we consistently find a preference for the disease with greater relative frequency. Rescorla and Wagner's model cannot predict this result because it assumes that the total strength associated with each disease sums to a constant. So for now we will simply note that the general principle of competition can explain the inverse base-rate effect associated with conflicting tests. Of course, the idea that some components of a compound stimulus may receive more attention than others is not unique to animal conditioning studies and has received considerable attention in other human verbal learning contexts (e.g., Kroll & Grant, 1968; Saufley & Underwood, 1964). Experiments 3 and 4 provide further support for the competition principle and undermine a set of interpretations that would dismiss the significance of the inverse use of base-rate information on conflicting tests.

Experiment 3

So far the data show appropriate use of base-rate information on imperfect predictor tests and either no use or inappropriate use on conflicting tests. There are a number of interpretations of this pattern of data that rely on the notion that specialized strategies come into play on conflicting tests. For example, participants might decide that both diseases are present and that it is most important to detect the rare disease when it is present. Of course, one would have to explain why this same strategy would not apply on combined tests. Rescorla and Wagner's model ascribes no special significance to conflicting tests and, therefore, one ought to be able to observe competition effects that disrupt the appropriate use of base-rate information on imperfect predictor tests.

The design for the experiment is shown in Table 5. Note that not one of the symptoms associated with the first three diseases is a perfectly valid cue. Note also that for any given symptom (*a*, *b*, or *c*) the predictive value of the alternative symptom may differ. Consider symptom *c*. It is associated with disease 3 two thirds of the time, and on the basis of our earlier study, one might expect two thirds of the responses to an incomplete test on *c* to be disease 3 responses. The competition notion leads to somewhat different expectations. Although symptom *c* is associated with disease 2 one third of the time, it is competing against symptom *a*, which is associated with disease 2 but one fourth of the time. The competitor of *c* for predicting disease 3 is symptom *b*, which is associated with disease 3 two fifths of the time. Because *b* has a stronger association with disease 3 than *a* has with disease 2, one might expect that *c*'s association to disease 2 would benefit relative to its association to disease 3. This rather convoluted line of reasoning leads to the idea that responses on imperfect predictor tests of a, b, and *c* might map on to base rates to a lesser extent than in the earlier studies because in each case the unequal competition favors the association with the less frequent disease. In the absence of a specific quantitative model, it is not clear whether the unequal competition is sufficient to overcome the relative frequency differences. Therefore, we can expect a reduction in responding consistent with base rates but not necessarily a reversal. The conditions shown in the note in Table 5 allow a replication of our earlier results on conflicting tests.

Method

The design of the learning materials conformed to that shown in Table 5. The eight symptoms were back pain, constant fever, dizziness, dulled hearing, skin rash, slurred speech, nausea, and swollen

Table 5
Design of Experiments 3 and 4 and Associated Transfer Tests

Learning: Relative frequency	Symptoms	Disease
3	*a,b*	1
1	*a,c*	2
2	*b,c*	3
3	*d,e*	4
1	*d,f*	5
2	*g,h*	6
Transfer tests		
Imperfect predictor tests: *a*, *b*, *c*, or *d*		
Conflicting test: *e* vs. *f*		

Note: The symptoms are represented by single letters and the diseases by numbers. The diseases differ in their relative frequency. See the text for concrete examples of symptoms and disease names.

tongue. The disease names were buragamo, coralgia, gouphosis, midosis, namitis, and terrigitis. Details of randomization, stimulus presentation and learning criterion were as in the earlier experiments.

The 37 transfer test types were presented with the same procedure as in the first two studies. There were 7 single-symptom tests, 16 paired tests, and 14 tests involving triplets of symptoms. The single-symptom tests included each distinct type of symptom (either *g* or *h*, but not both, were presented) and the paired tests exhausted the logical types that did not include training pairs. The full set of 37 tests is listed in Appendix B. Altogether 45 undergraduates participated in the study.

Results

Of the 45 participants, 40 met the learning criterion and the average proportion of errors on the last block of training trials was .017. The few errors on the last block were fairly evenly scattered across the different symptom–disease pairings.

The transfer test data of greatest interest are shown in Table 6. The number of observations for each test was 45 and the average standard error for the proportion in Table 6 was .058. The results are somewhat mixed. Although symptom *a* had "unequal competitors" and symptom *d* had equal competitors, responses to *a* and *d* were very similar. On the other hand, responses to symptom *b* were the inverse of what one would expect on the basis of relative frequency, $t(44) = 2.63$, $p < .05$, and responses to symptom *c* showed only the slightest tendency to favor the disease with greater relative frequency, $t(44) = .31$ *ns*. Responses to the conflicting test were in the same direction as in the earlier studies but the effect was much smaller. For the *a,b,c* triplet test the response proportions showed no tendency to map onto relative frequencies. An ANOVA with the variables of frequency and test type (imperfect, conflicting, combined) for symptoms *d*, *e*, and *f* showed a significant effect of test type, $F(2, 88) = 3.71$, $MS_e = .028$, $p < .05$, but neither the main effect of frequency, $F(1,44) = 3.10$, $MS_e = .53$, $p = .09$, nor the interaction of frequency and test type, $F(2, 88) = 2.33$, $MS_e = .42$, $p = .11$, proved to be reliable. The entire set of test responses is given in Appendix B.

Table 6
Transfer Test Results: Response Proportions for Experiment 3

Test	HF	MF	LF	Other
Single symptoms				
Imperfect predictors				
a (3:1)	.644		.289	.067
b (3:2)	.289		.644	.067
c (2:1)	.467		.422	.111
d (3:1)	.600		.267	.133
Perfect predictors				
1 common				
High frequency	.867			.133
Low frequency			.889	.111
0 common	.822[a]			.178
Pairs of symptoms				
Conflicting test (e.g., *e* vs. *f*)	.444		.511	.067
Triplets of symptoms				
Combined test (*d,e,f*)	.600		.400	.000
Combined imperfect predictors (*a,b,c*)	.333	.378	.289	.000

Note: Relative frequencies of the diseases associated with the imperfect predictor tests are shown in parentheses. HF = high frequency; MF = medium frequency; LF = low frequency.
[a] Correct.

Discussion

The results give weak support to the view that failure of classification responses to map onto base rates or relative frequencies is not confined to conflicting tests. Of the three imperfect predictor tests associated with unequal competition, one showed responses consistent with base rates, one showed no effect of base rate, and the third showed an inverse base-rate effect. The other results were consistent with earlier findings but the trends were also somewhat weaker. Because the results were not so clear-cut as they might be, we decided to replicate this experiment with an instructional manipulation designed to either increase or decrease the hypothesized competition process.

Experiment 4

Method

Experiment 4 was identical to Experiment 3 except that subjects were randomly assigned to one of two instructional conditions. In the *focus* condition participants were told that symptoms varied in their predictive value and that they should find the most reliable or best predictor for each disease. In the *complete* condition subjects were told that symptoms varied in their predictive value but that they should learn about all the symptoms because later on they might be given tests in which only partial information was given. The focus instructions were designed to increase competition among symptoms and the complete learning instructions were designed to minimize competition. Thirty subjects were tested in the complete condition and 28 subjects were tested in the focus condition.

Results

Every subject in the complete condition met the learning criterion as did all but two subjects in the focus condition. The complete group had no errors on the last block, whereas the focus group had an average proportion of .022 errors.

The main results of interest are given in Table 7. The number of observations for each test was 30 for the focus group and 28 for the complete group (except for the perfect predictor test with 0 common symptoms, which had twice as many observations). The average standard error on these observations was .073 for the focus group and .076 for the complete group. The focus and complete groups show different patterns of responding on the imperfect predictor tests. Specifically, classifications of the complete group follow the relative presentation frequencies to a greater extent than the focus group on symptoms *b* and *c*. An ANOVA on the set of imperfect predictor tests indicated that the main effect of

Table 7
Transfer Test Results: Response Proportions for Experiment 4

Test	HF	MF	LF	Other
Single symptoms				
Imperfect predictors				
a (3:1)				
Focus	.536		.321	.143
Complete	.567		.400	.033
b (3:2)				
Focus	.464		.500	.035
Complete	.600		.300	.100
c (2:1)				
Focus	.429		.536	.036
Complete	.600		.367	.033
d (3:1)				
Focus	.500		.250	.250
Complete	.667		.200	.133
Perfect predictors				
1 common				
High frequency				
Focus	.714			.286
Complete	.867			.133
Low frequency				
Focus			.893	.107
Complete			.800	.200
0 common				
High frequency				
Focus	.857[a]			.143
Complete	.767[a]			.233
Pairs of symptoms				
Conflicting test (*e* vs. *f*)				
Focus	.179		.750	.071
Complete	.333		.567	.100
Triplets of symptoms				
Combined test (*d,e,f*)				
Focus	.429		.500	.071
Complete	.633		.300	.067
Combined imperfect predictors (*a,b,c*)				
Focus	.286	.500	.214	.000
Complete	.433	.300	.200	.067

Note. Relative frequencies of the diseases associated with imperfect predictor tests are in parentheses. HF = high frequency; MF = medium frequency; LF = low frequency.
[a] Correct.

frequency, $F(1, 56) = 10.14$, $MS_e = 1.72$, $p < .01$, and the interaction of frequency and instructions were significant, $F(1, 56) = 3.25$, $MS_e = 3.25$, $p < .05$, by a one-tailed test. On the *a,b,c* triplet test shown in the bottom of Table 7 the complete group's response proportions were .43, .30, and .20 for relative frequencies of 3 to 2 to 1, whereas the focus group's corresponding proportions were .29, .50, and .21. This interaction did not, however, prove to be reliable ($p > .20$).

On the conflicting tests both groups showed the inverse base-rate effect observed in the earlier experiments. This tendency is clear in both groups but it is considerably larger in the focus group than in the complete group, as would be expected if the focus instructions increased the competition among symptoms. On the combined (*d,e,f*) test, the complete group responded directly in proportion to relative frequency, whereas the focus group showed a slight inverse frequency effect. An ANOVA with the variables of instruction, frequency, and test types revealed that the interaction of frequency and instruction, $F(1, 56) = 4.85$, $MS_e = .46$, $p < .05$, and the test type by frequency interaction, $F(2, 55) = 14.26$, $MS_e = .36$, $p < .01$, were significant. No main effect or other interaction was reliable.

Discussion

The results provide further support to the competition principle. Responses on the conflicting test are consistent with symptom competition occurring in both groups but were larger for the instructional manipulation (focus instructions), which was aimed at increasing competition.

In addition to this further evidence for competition effects, the results from the imperfect predictor tests undermine the interpretations of the results that assume that different processes come into play for imperfect predictor and conflicting tests. Although the statistical support for competition effects on imperfect predictor tests is not as strong as we would like, the results are at least roughly consistent across Experiments 3 and 4. In addition, we have never obtained inverse base-rate effects on imperfect predictors when unequal competition was not set up by design. According to the competition principle, the effects of symptoms competing to predict diseases ought to be manifest on both conflicting and imperfect predictor tests, and the results of both Experiment 3 and Experiment 4 are generally consistent with this claim. The competition principle still cannot explain the direct use of relative frequency information on the combined test (although the focus group of Experiment 4 did not show this effect). Note also that explanations based on the notion that infrequent perfect predictors become especially salient also fail to predict the results on complete tests, because they imply an inverse rather than a direct base-rate effect on such tests as well as on conflicting tests. The latter result obtains, but not the former.

Rescorla and Wagner's model assumes that predictive strengths sum to a constant. One could develop competition models that give up this assumption and consequently do not necessarily predict equal responding on combined tests. Consider our standard paradigm where *a,b* is associated with category 1 and appears three times as often as *a,c*, which is associated with category 2. Assume further that at the end of training the predictive strengths are as follows: *a* for category 1 = 9; *a* for category 2 = 3; *b* for category 1 = 2; *c* for category 2 = 4. If we further assume that responses are proportional to predictive strengths then the following proportion of high-frequency category responses is expected on the base-rate tests: $a = 9/(9 + 3) = .75$; $b,c = 2/(2 + 4) = .33$; $a,b,c = (9 + 2)/(9 + 3 + 2 + 4) = .61$. This is roughly the pattern we observed.

This type of model, which allows strengths associated with a disease not to sum to a constant is, however, inadequate for two major reasons. One is that it implies that the imperfect predictor has greater predictive value than the perfect predictors, an implication not supported by other tests which com-

bine these types of predictors with other symptoms for paired tests. In general, perfect predictors are more likely to control responding than imperfect predictors. The other main problem with the predictive strength model is that when the perfect predictors are tested alone, performance on the less frequent perfect predictor is not clearly better than performance on the more frequent perfect predictor. In all of our experiments the differences are small, and in Experiment 2 and the complete condition of Experiment 4 the results favor the more frequent perfect predictor. We have experimented with a variety of other competition models but have been unsuccessful in developing any model relying solely on competition that can account for the full set of results on singles, pairs, and triples of symptoms.

Summary of Empirical Results

The first four experiments provide a consistent set of results that will need to be addressed by any theory attempting to give an account of the use of base-rate information conveyed through experience. It is clear that the pattern of performance depends both on the structure of the training experience and the nature of the test. First of all, responding directly in proportion to relative frequencies appears to occur only when common cues or imperfect predictors form part of the learning condition. Second, the results depend on how base-rate information is tested. Usually, on imperfect predictor tests responses map directly onto base rates, whereas on conflicting tests they map inversely on to base rates. (This generalization concerning imperfect predictor vs. conflicting tests is not really appropriately stated in terms of type of test, because it is possible to produce departures from the appropriate use of base-rate information on imperfect predictor tests by varying the structure of the training stimuli; see Experiments 3 and 4. Third, on combined tests, responses once again map directly onto base rates. Finally, the inverse frequency effect on conflicting tests does not derive from low-frequency perfect predictors being better learned than high-frequency perfect predictors, because performance on these two types of symptoms presented alone is comparable.

Many of the foregoing results are consistent with the idea that symptoms compete to predict a disease, but the competition principle is not sufficient to account for the entire pattern of results. In the next section of this article we present some ideas that embody the competition principle but further add a second principle, retrieval failure induced by changes in context, to address our results. The new key idea is that performance on combined tests derives from the fact that the high-frequency perfect predictor suffers more from the change in context between learning and test than does the low-frequency perfect predictor. This situation may arise from plausible assumptions concerning the details of the learning process.

Competition, Context Change, and the Use of Base-Rate Information

Earlier we mentioned that classification models could readily be applied to our data and that certain of these models implied that base-rate information would be used. Consider the Medin and Schaffer (1978) context model, which assumes that classification is based on the retrieval of category examples that are similar to a probe. The greater the category frequency (base rate), the more likely it would be that an example of that category would be accessed in response to a probe and become the basis for categorization. Access is assumed to be based on similarity. Similarity along a symptom dimension is assumed to vary between 0 and 1, with 1 representing identity or maximum similarity. Overall similarity of a transfer item is assumed to be a multiplicative function of the individual symptom similarities. In the present experiments a symptom is either present or absent and differences between the status of symptoms in the probe and stored exemplars drive the similarity computation. To illustrate this, imagine that symptoms a and b have been associated with disease 1, symptoms a and c with disease 2, and that disease 1 has appeared three times as often as disease 2. There should then be three times as many a,b exemplar representations as a,c representations. Because the context model uses a ratio rule to predict classification performance, the probability that a disease 1 categorization would be made on the incomplete test a is

$$p(\text{disease } 1/a) = \frac{3Sb}{3Sb + Sc}, \tag{1}$$

where Sb and Sc represent the similarities that come into play when the probe and stored exemplar differ in symptom status on b and c, respectively.

If Sb and Sc have the same similarity value, then we expect three fourths of the classification responses to be consistent with the more frequent disease. On a conflicting b,c test the prediction equation would be

$$p(\text{disease } 1/b,c) = \frac{3Sa \cdot Sc}{3Sa \cdot Sc + Sa \cdot Sb}. \tag{2}$$

Of course, Sa can be cancelled out, and if Sb and Sc have the same value then again we expect appropriate use of base-rate information, contrary to our results. Finally on the combined test, the equation is

$$p(\text{disease } 1/a,b,c) = \frac{3Sc}{3Sc + Sb}. \tag{3}$$

Equations 2 and 3 are equivalent regardless of the values of Sb and Sc and, therefore, the simple extension of the context model predicts the same results on conflicting and complete tests. This was, of course, not the pattern of results we observed and this version of the context model is inadequate. Actually, the original context model did not directly address processes taking place during learning (but see Nosofsky, 1986, 1987). The modified context model represents one attempt to do so and we show that building a notion of competition indirectly into the learning assumptions provides some plausible account of our otherwise puzzling results.

Modified Context Model

The main idea of the context model is that when an item is presented to be classified, that item acts as a retrieval cue

to access information associated with stored exemplars. The model does not, however, necessarily assume that a distinct representation is set up for each exemplar, but rather allows for the possibility that selective attention may lead people to encode more information concerning certain attributes or symptoms than others.

As one attempt to describe how selective attention might work we outline a pair of processing models which assume that the retrieval processes assumed by Medin and Schaffer (1978) to take place during transfer test also take place during learning. This will entail the assumption that similarity parameters are associated with specific exemplar representations rather than (symptom) dimensions in their entirety, which is a second difference from Medin and Schaffer's original model.

Representation-Specific Similarity Parameters

As just noted, although the original context model assumed that classification is based on specific item information, it was not assumed that these representations were necessarily distinctive. For example, in some traditional concept task where only a single dimension is relevant to the classification, the representations might include very little else other than information along this dimension. Because, depending on the learning circumstances, representations can be fairly detailed or fairly abstract, we shall replace the term exemplars with the more neutral term, *representational units*, in the following discussion.

In the original context model, a given symptom (e.g., *a*) would have the same salience in all representations. There are both empirical (e.g., Elio & Anderson, 1981; Ortony, Vondruska, Foss, & Jones, 1985) and more intuitive reasons for abandoning this assumption. For example, on the frequent *a,b* learning trial, a subject may attend to *a* more than *b* but attend to *c* more than *a* on the infrequent *a,c* learning trial. It seems implausible that *a* would be equally salient in these two representations.

Imagine that people do attend relatively more to *c* in the *a,c* training pairs than they do to *b* in the *a,b* training pairs. Then at the end of the training the following representational units may exist with the possible similarity values. Small similarity values correspond to high salience and increase sensitivity to differences between probes and stored units for a particular symptom (see Table 8).

Consider again the performance on the three main types of probes. On the incomplete *a* test from Equation 1 we predict 82% choices of the more frequent category. Equation 2 must be adjusted because *Sa* has a different value in the *a,c* representation than in the *a,b* representation. When this is done, the predicted percentage of disease 1 choices on the *b,c* conflicting test is .60/(.60 + .80), or 43%. Finally, on the combined test, the disease 1 responses are predicted to occur 67% of the time. These predictions are in qualitative agreement with our data.[2]

Table 8
A Hypothetical Pattern of Representational Units and Associated Similarity Values for a Study Varying Where the Symptom Pair a,b Appears Three Times as Often as the Symptom Pair a,c

Representational unit	*Sa*	*Sb*	*Sc*	Disease
a,b	.20	.15		1
a,b	.20	.15		1
a,b	.20	.15		1
a,c	.80		.10	2

It may not be obvious from the similarity parameters, but this version of the context model implies that two factors are involved in producing the pattern of results we observe. First of all, the parameter values are consistent with the idea that symptoms compete for attention on the basis of their ability to correctly predict the associated disease. Symptom *a* is a much better predictor of disease 1 than of disease 2, and its similarity parameter has been represented as being much smaller in the former case than the latter. This sensitivity to difference is the second key factor. The reason that responses consistent with base rates do not occur on conflicting tests is that the *a,b* representational units suffer much more than does the *a,c* representational unit from the absence of symptom *a* in the probe. This predicted pattern is reversed on the combined *a,b,c* test when symptom *a* is present. Therefore, this model implies that both competition and retrieval failure produced by changes in context (differences between the probe and stored representation) are responsible for the observed pattern of results. Another way to think about this differential context effect is as follows: *a* and *b* are more unitized and form a better integrated pattern than *a* and *c*.

We still have not presented anything like a process model for learning. The following is a brief sketch of two related possibilities. The common idea is that, over the course of learning, representational units corresponding to *a,b* and *a,c* will be set up with each of the components having some similarity parameter associated with it. The key learning assumptions are as follows.

1. Each time a representation unit is retrieved and leads to a correct classification, properties (symptoms) in common between the probe and the retrieved unit are rehearsed or strengthened. This rehearsal is assumed to lead to better encoding of the symptoms (and reduce the value for the similarity parameters associated with these symptoms) in the representational units.

2. When a retrieved unit leads to an error, properties of the probe that are *different* from the retrieved unit are rehearsed or strengthened. The difference in salience of the two symptoms of a pair will depend on the number of occasions where both symptoms are rehearsed equally relative to the number of occasions where one symptoms is differentially rehearsed.

[2] Of course, one need not assume that there are exactly three times as many *a,b* representations as *a,c* representations. We have done some preliminary examination of variants of the context model where similarity is influenced both by matching (they increase similarity) and mismatching values. In these versions each type is assumed to have only a single representational unit or one assumes that units are set up only on trials in which errors occur. Because these variants yield no qualitative differences in the present context, they will not be described in detail.

Furthermore, it seems plausible to assume that more rehearsal takes place after an error than after a correct response, although this idea is not essential to our present development of the model. Because *a,b* appears much more often than *a,c*, there will be more times when an *a,c* probe successfully retrieves an *a,b* unit rather than an *a,c* unit, relative to the number of times an *a,b* probe successfully retrieves an *a,c* rather than an *a,b* unit. Consequently, differential rehearsal of *a,c* probes (favoring *c*) will occur relatively more often than differential rehearsal of *a,b* probes (favoring *b*). As a consequence, *a* and *b* will be linked more closely in an *a,b* unit than *a* and *c* are linked in an *a,c* unit. This is exactly what was assumed in the foregoing example.

At present, the revised model exists as a computer simulation or, more accurately, as a set of simulations. It is possible to simulate these learning assumptions and to produce representational units with at least roughly the desired properties. We feel far from closure, however, and will not present a more detailed simulation. There are many possibilities that remain to be explored that are consistent with the general processes of competition and context change. For example, one might assume that rehearsal after errors is biased toward more novel symptoms and that rehearsal after correct responses is biased toward more frequent symptoms. In general, surprising events (such as errors) might tend to be associated with novel events in an organism's environment. At least this seems like a plausible heuristic. Nor is it clear that overt errors are needed to produce differential rehearsal. A subject might retrieve a representation corresponding to *a,b* on an *a,c* trial, note the difference, and then differentially rehearse symptom *c* when the feedback is provided. Because of these and other issues, it is perhaps more appropriate to approach the revised model more abstractly.

The revised context model incorporates the two processes of context change and competition. Context change is a transparent aspect of the model, and the idea of competition is embodied through the mechanism of differential rehearsal that favors one symptom of a pair over the other one.

Experiment 5 tests the differential rehearsal idea associated with the revised context model concerning the presence of feedback.

Experiment 5

The design of the study is shown in Table 9. The major change is that subjects are not always told whether their classification was correct or incorrect. For example, feedback was given only one half of the time on *a,b* trials and one fourth of the time on *a,c* trials. The first two disease categories equate presentation frequency and vary feedback frequency, whereas the remaining six categories incorporate a basic relative frequency design but systematically vary the probability that feedback will be given after a high-frequency trial.

What effect should the feedback frequency have, according to the differential rehearsal idea? It seems that differential rehearsal of a probe should be less likely on trials where no feedback is given. As feedback decreases, there should be fewer opportunities for differential rehearsal to occur when on a high-frequency trial. Therefore, the inverse base-rate effect should *increase* as feedback on high-frequency trials *decreases*. This follows from the idea that the inverse frequency effect is produced by relatively more differential rehearsal on low-frequency diseases than on high-frequency diseases. This prediction is not necessarily expected. For example, in the previous studies presentation, frequency and feedback frequency were perfectly confounded and it may well be that inverse frequency effects on conflicting tests derive from feedback frequency. On that interpretation, one would expect the inverse base-rate effect to decrease as feedback decreases and actually revert to a positive base-rate effect when feedback for the high-frequency disease occurs less often than feedback for the low-frequency disease.

Table 9
Design of Experiment 5 and Associated Transfer Tests

Relative frequency	Feedback frequency	Symptoms	Disease
4	2	*a,b*	1
4	1	*a,c*	2
4	1	*d,e*	3
2	2	*d,f*	4
4	2	*g,h*	5
2	2	*g,i*	6
4	4	*j,k*	7
2	2	*j,l*	8
	Transfer tests		

Imperfect predictor tests: *a*, *d*, *g*, or *j*
Perfect predictor tests: *b*, *c*, *e*, *f*, *h*, *i*, *k*, or *l*
Conflicting test: *b*, *e*, *h*, or *k* vs. *c*, *f*, *i*, or *l*

Note. Letters refer to symptoms and the associated diseases (indicated by numbers) differ both in presentation frequency and in the frequency with which feedback is given after a classification response.

Method

The abstract design of the learning materials was exactly that shown in Table 9. The 12 symptoms were back pain, constant fever, dizziness, hair loss, nausea, puffy eyes, skin rash, slurred speech, sweaty palms, swollen tongue, tired easily, and vomiting. The disease names were althrax, buragamo, burlosis, coralgia, gouphosis, midosis, namitis, and terrigitis. Details of randomization and stimulus presentation were identical to the earlier experiments with the following exception: The feedback frequency varied in the manner indicated in Table 9. For example, feedback was given at random on two of the four presentations of the *a,b* pair in a particular block. Subjects were told that just as physicians did not always immediately find out about their diagnosis they also would only be told if their diagnosis was correct part of the time. The learning criterion was one errorless block (24 trials) or a maximum of eight runs (192 trials). The transfer tests consisted of 12 single symptom presentations, 28 paired symptom tests and 4 triple symptom tests (exhausting all combinations of the perfect predictors, *b*, *c*, *e*, *f*, *h*, *i*, *k*, and *l*), for a total of 44 tests. The full set of tests is listed in Appendix C. Altogether, 40 undergraduates participated in the study.

Results

Only 7 of the 40 subjects failed to meet the learning criterion. The average proportion of errors on the last run was .035. The errors were fairly equally scattered across the learn-

ing pairs, although there was a tendency for more errors to occur on pairs with smaller feedback frequencies. The transfer tests of greatest interest are the conflicting tests summarized in Table 10. They show that the use of the more frequently presented symptom as the basis of classification decreased, and the use of the less frequently presented symptom increased, as the feedback frequency of the more frequently presented symptom decreased. This pattern of responding is precisely that expected from the differential rehearsal interpretation associated with the revised context model. An ANOVA on the key conflicting tests revealed a significant effect of presentation frequency, $F(1, 39) = 4.78$, $MS_e = 6.15$, $p < .05$, and a significant interaction of presentation frequency and feedback frequency, $F(2, 78) = 4.97$, $MS_e = 3.47$, $p < .01$. The tests involving triplets were less informative; the proportions of high- and low-frequency responses were as follows: *abc*, .475:.525; *def*, .475:.475; *ghi*, .600:.350; and *jkl*, .525:.499. The imperfect predictor tests showed responses roughly in proportion to presentation frequencies. The full set of results is given in Appendix C.

Discussion

The results show that the inverse base-rate effect associated with conflicting tests does not derive from the fact that low presentation frequencies mean that there will be low feedback frequencies. The present study broke the correlation between presentation and feedback frequencies and found that as feedback frequency decreased for high-frequency diseases, control by low-frequency category (the inverse base-rate effect) actually increased. This result is consistent with the differential rehearsal interpretation associated with an alternative version of the context model.

It is perhaps a mistake to impute these predictions to any particular model at this stage. The main point is that there is a set of models incorporating the context change and competition principles that can give a coherent account of a variety of findings concerning the use of base-rate information derived from experience.

General Discussion

We suggested at the beginning of this article that in addition to explicit strategies and rules there may be other processes associated with learning to categorize that might influence the use of base-rate information from experience. The present experiments reveal that when base-rate information is conveyed through experience it does not influence decision making in some uniform manner. Instead, our studies show a coherent but complicated pattern. Whether responses map directly on to base rates must be qualified in terms of particular learning strategies, category structures, and types of tests. Depending on the particular pattern of these factors, responses were independent of base rates, positively correlated with base rates, or even negatively correlated with base rates.

We have argued that it is possible to make sense of this diverse set of rules in terms of two underlying principles, competition among symptoms to predict a disease and retrieval failure associated with changes in context. The first principle largely derives from a body of work in the animal learning and conditioning literature (e.g., Medin, 1975; Rescorla & Wagner, 1972) and the second is ubiquitous in studies of both human and animal memory. Each principle is associated with a large set of potential processing models. For example, many of the recent distributed-memory models use a delta rule for changing connection strengths, and the delta rule is formally similar to the learning rule associated with Rescorla and Wagner's model (Sutton & Barto, 1981). In our simulation of the revised context model, competition derives from selective rehearsal of symptoms, which is assumed to be activated by an inappropriate reminding. This idea is similar to Schank's (1982) notion of failure-driven reminding, except that we assume that reminding is automatic and that what varies as a function of surprise or failure is the pattern of attention to symptoms. A second key assumption is that the salience of a symptom is not independent of the disease with which it appears. Ortony et al. (1985) have argued persuasively that models of similarity must assume that the salience of a property can vary across objects and that the salience of a property for a given object can vary across contexts.

Our results raise difficulties for current theories of categorization and decision making in that, to our knowledge, no previous theory is able to predict this complex pattern of interactions. On the other hand, the entire set of results is rendered coherent in terms of two basic processes: competition and context change. The competition process is assumed to result in sets of representations which are differentially sensitive to changes in context between learning and test.

Having some understanding of our base-rate phenomena, we can suggest training techniques to improve decision making. We found that the effects of competition were diminished (but not eliminated) when we asked participants to attend to both symptoms associated with a disease. In other work we have used symptoms that are meaningfully related (in terms of causal linkages) to the associated disease. So far it appears that these results can be described by the assumption that causal linkages lead participants to attend to both symptoms associated with a disease. That is, the results are very similar to those obtained when we directly ask participants to learn about both symptoms.

Although a version of the context model was able to account qualitatively for the present results, it may offer an oversimplified view of symptom relations. In the model all symptoms are associated with the disease with varied consequences depending only on the similarity parameters. One can imagine, however, that some symptoms might be more central and

Table 10
Results of Conflicting Tests in Experiment 5

Symptom	Proportion of responses: Feedback frequency on high-frequency item		
	1	2	4
High-relative-frequency (e.g., *e*, *h*, & *h* vs. others)	.336	.432	.482
Low-relative-frequency (e.g., *f*, *i*, & *l* vs. others)	.561	.546	.443

directly linked to a disease than others, and that some symptoms might be used to explain others. For example, if *overweight* and *skin rash* are associated with the fictitious disease, *burlosis*, then one might conceive of burlosis as a glandular disease producing obesity and skin rash might be thought of as secondary. For instance, one might imagine an obese person wearing a tight belt which produced a rash. For some other disease *skin rash* might well be a primary symptom. If *skin rash* is presented as a probe, one might be more likely to be reminded of the disease in which it is a primary symptom than the one in which it is a secondary system. These relations as well as other possible patterns of competition and coordination of symptoms that would vary with the underlying causal structure can be only (very) roughly approximated by variations in similarity values. Therefore, although we think the general notion of competition and context change will have considerable generality, we do not have a corresponding process model which is powerful enough to do what we think it should do.

The results and associated theories are less important in themselves perhaps than they are to the underlying basic point that processes that are not necessarily strategic or explicit can exert a major effect on decision making. Our optimistic presentation of the revised context model to the contrary, these effects are not well understood and are badly in need of further investigation. We are not arguing in favor of any sort of unconscious learning. Rather the claim is that processes used for one purpose (e.g., learning the relations among symptoms and diseases) can propagate to influence performance in a different task (e.g., transfer response and use of base-rate information).

In a sense the present approach is not at all incompatible with the recent body of work on judgmental heuristics. Indeed, according to our interpretation of the present results we have been studying the underpinning associated with the availability heuristic. What is perhaps most important is that the process model that embodies the availability heuristic does not map onto base rates in a simple way. If one tried to account for the present results in terms of an abstract principle of availability with category availability increasing with relative frequency, the present results would be a jumble. Again, this consideration suggests that greater attention should be directed toward nonexplicit factors that influence decision making.

We are left with more questions than when we began. Although our subjects did not report using explicit decision-making strategies, people certainly do in other contexts. A central question is how such explicit decision-making processes interact with basic learning processes, such as those studied here, to determine performance.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*, 303–314.

Balla, J. I. (1982). The use of critical cues and prior probability in decision-making. *Methods of Information in Medicine, 21*, 9–14.

Balla, J. I., Elstein, A., & Gates, P. (1983). Effects of prevalence and test diagnosticity upon clinical judgments of probability. *Methods of Information in Medicine, 22*, 25–28.

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

Binder, A., & Estes, W. K. (1966). Transfer of response in visual recognition situations as a function of frequency variables. *Psychological Monographs, 80* (23, Whole No. 631).

Carroll, J. S., & Siegler, R. S. (1977). Strategies for the use of base-rate information. *Organizational Behavior and Human Performance, 19*, 392–402.

Casscells, B. S., Schoenberg, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 229*, 999–1000.

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

Christensen-Szalanski, J. J. J. & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 928–935.

Einhorn, H. J., & Hogarth, R. M. (1981) Behavioral decision therapy: Processes of judgment and choice. *Annual Review of Psychology, 32*, 53–88.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 397–417.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37–64.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23*, 339–359.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distribution: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–258.

Gluck, M. A., & Bower, G. H. (1986, November). *Category learning, judgment, and the Rescorla-Wagner Model (aka the Delta-Rule).* Paper presented at the 27th Annual Meetings of the Psychonomic Society, New Orleans, LA.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

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

Kroll, N. E. A., & Grant, D. A. (1968). Cue selection in paired-associate concept learning paradigms. *Journal of Verbal Learning and Verbal Behavior, 7*, 64–71.

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40*, 287–298.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38*, 287–298.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159–188.

Medin, D. L. (1975). Theories of discrimination learning and learning set. In W. K. Estes, (Ed.), *Handbook of learning and cognitive processes* (pp. 131–169). Hillsdale, NJ: Erlbaum.

Medin, D. L., & Robbins, D. (1971). Effect of frequency on transfer

performance after successive discrimination training. *Journal of Experimental Psychology, 87*, 434–436.

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

Morton, J. (1970). A functional model for memory. In D. A. Norman (Ed.), *Models of human memory* (pp. 203–260). New York: Academic Press.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 87–108.

Ortony, A., Vondruska, R. J., Foss, M. A., & Jones, L. E. (1985). Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language, 24*, 569–594.

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

Rudy, J. W., & Wagner, A. R. (1975). Stimulus selection in associative learning. In W. K. Estes (Ed.) *Handbook of learning and cognitive processes* (Vol. 2, pp. 369–303). Hillsdale, NJ: Erlbaum.

Saufley, W. H., Jr., & Underwood, B. J. (1964). Cue-selection inference in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 3*, 474–479.

Schank, R. C. (1982). *Dynamic memory: A theory of learning in people and computers.* Cambridge, MA: Harvard University Press.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135–170.

Tversky, A., & Kahneman, P. (1974). Judgment under uncertainty: Heuristics and biases. *Sciences, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

Appendix A

Detailed Results on Transfer Tests in Experiment 2

	Disease response					
Symptom(s)	1	2	3	4	5	6
c	32	2	2	0	2	2
a	23	8	3	2	3	1
d	1	30	3	3	2	1
f(g)	2	1	26	6	2	3
e	3	3	15	9	6	4
h(i)	2	2	5	27	1	3
j(k,l)	0	2	3	6	27	2
m(n,o)	2	3	2	7	3	23
c,d	11	25	1	1	2	0
c,f	18	03	14	02	1	2
c,h	13	4	4	14	0	5
c,j	12	4	1	5	18	0
c,m	11	0	2	5	0	22
a,c	31	3	2	0	3	1
c,e	19	1	8	5	3	4
d,g	1	24	11	0	2	2
d,i	0	22	1	14	1	2
d,k	2	24	0	1	11	2
d,n	0	21	2	1	3	13
b,d	2	23	7	6	1	1
d,e	2	23	7	6	1	1
f,h	1	2	16	14	2	5
f,l	2	2	19	3	12	2
g,o	2	2	14	5	1	16
a,g	10	7	18	2	1	2
e,g	1	3	27	6	1	2
h,k	3	0	5	17	11	4
i,n	0	1	2	22	1	14
b,i	11	6	1	15	4	2
e,i	1	0	6	28	3	2
j,o	0	0	3	2	29	15
b,k	9	2	1	4	22	2
e,l	2	3	11	6	15	3
a,m	11	4	1	5	0	19
e,n	0	1	11	6	4	18
b,e	15	4	8	4	5	4
a,c,d	18	19	1	0	2	0

(*continued on next page*)

Appendix A (*continued*)

Symptom(s)	Disease response 1	2	3	4	5	6	7	8
f,h,i	1	3	8	25	0	3		
f,g,i	0	0	28	9	1	2		
e,g,h	2	1	19	14	2	2		
f,g,h,i	1	0	19	18	0	2		
e,g,h,i	2	1	10	24	1	2		
l,n,o	1	2	0	3	10	24		
j,k,o	0	3	1	2	25	9		
c,d,e	12	22	3	2	0	1		
f,e,l	0	1	21	6	12	0		
b,c,i	23	4	1	10	2	0		
b,d,l	4	25	1	2	5	3		
e,h,m	1	0	4	20	0	15		
a,e,j	7	2	8	6	15	2		
a,e,m	8	5	5	5	2	15		
c,i,j	8	2	3	9	16	2		
d,i,k	0	19	0	7	13	1		
a,b	26	9	2	0	2	1		
a,b,e	22	6	6	2	2	2		
a,b,c,d	26	11	0	0	1	2		
c,d,k	10	14	2	1	12	1		
c,h,n	9	2	1	16	1	11		
a,b,o	20	8	1	1	2	8		

Note. Refer to Table 4 for basic design.

Appendix B

Detailed Results on Transfer Tests in Experiment 3

Symptom(s)	Disease response 1	2	3	4	5	6
a	29	13	10	0	0	2
b	13	3	29	0	0	0
c	1	19	21	1	2	1
d	0	2	2	27	12	2
e	1	2	1	40	0	1
f	0	2	0	2	40	1
g(h)	2	5	4	1	4	74
a,b,c	15	13	17	0	0	0
e,f	0	1	0	20	23	1
d,e,f	0	0	0	27	18	0
a,d	15	7	6	8	8	1
a,e	16	6	0	23	0	0
a,f	10	5	2	1	27	0
a,g(h)	28	11	3	3	1	44
b,d	11	1	10	15	7	1
b,e	6	2	12	24	1	0
b,f	5	3	6	1	30	0
b,g(h)	12	6	19	3	2	48
c,d	2	12	10	13	8	0
c,e	0	6	16	22	0	1
c,f	1	3	8	1	31	1
c,g(h)	4	15	17	0	2	52
d,g(h)	3	0	1	27	16	43
e,g(h)	1	5	0	43	1	40
f,g(h)	0	0	4	2	51	33
a,b,h	25	1	7	1	1	10
a,c,g	8	21	4	0	0	12
b,c,g(h)	2	11	54	0	1	22
d,e,g	1	0	0	32	1	11

Appendix B (*continued*)

Symptom(s)	Disease response 1	2	3	4	5	6
d,f,h	2	1	0	4	31	7
a,d,g	7	7	0	9	4	18
b,d,h	4	1	8	10	3	19
a,c,g	9	5	0	15	1	15
b,e,g	4	2	10	12	1	16
c,e,g	0	3	9	19	0	14
b,f,h	2	1	4	1	23	14
c,f,h	0	3	6	0	20	16

Note. Refer to Table 6 for basic design.

Appendix C

Detailed Results on Transfer Tests in Experiment 5

Symptom(s)	Disease response 1	2	3	4	5	6	7	8
a	16	20	2	0	1	0	1	0
b	33	2	1	0	0	1	3	0
c	1	33	0	1	0	3	0	2
d	0	1	22	13	1	1	2	0
e	1	0	33	1	0	1	0	3
f	0	2	1	36	1	0	0	0
g	0	1	3	1	20	14	0	1
h	1	0	0	2	34	0	1	2
i	0	1	0	1	1	35	1	0
j	0	2	3	0	0	0	22	13
k	1	0	1	1	0	3	34	0
l	0	2	2	2	0	1	1	32
b,c	18	20	0	1	0	1	0	0
e,f	1	0	9	28	0	1	1	0
l,i	0	0	1	1	17	20	1	0
k,l	0	1	0	1	0	0	21	17
a,b,c	19	21	0	0	0	0	0	0
d,e,f	0	0	19	19	0	0	0	2
g,h,i	0	0	0	0	24	14	1	1
j,k,l	1	0	0	1	0	1	21	16
b,e	23	0	14	1	1	0	1	0
b,f	15	0	0	25	0	0	0	0
b,h	22	1	2	1	13	0	0	1
b,i	15	0	1	1	0	20	1	2
b,k	20	1	0	0	0	0	19	0
b,l	18	0	1	1	0	0	0	20
c,l	1	19	17	0	0	1	0	2
c,f	0	12	0	26	1	0	0	1
c,h	1	16	1	0	22	0	0	0
c,i	1	17	0	0	1	21	0	0
c,k	0	15	3	1	22	0	1	1
c,l	0	15	3	0	0	1	0	21
e,h	0	0	16	0	22	0	1	1
e,i	0	0	9	3	0	27	0	1
e,k	1	0	15	0	0	2	22	0
e,l	0	1	14	3	0	0	1	21
f,h	0	1	0	21	17	1	0	0
f,i	0	0	1	17	0	20	1	1
f,k	1	1	0	18	0	0	19	1
f,l	1	1	0	22	0	1	3	12
h,k	0	1	1	0	16	1	21	0
h,l	2	1	1	0	14	3	0	19
i,k	0	1	1	0	0	22	16	0
i,l	0	0	0	2	0	23	1	14

Note. Refer to Table 9 for basic design.

Received February 2, 1987
Revision received July 31, 1987
Accepted August 6, 1987 ■