Journal of Experimental Psychology:
Learning, Memory, and Cognition
2001, Vol. 27, No. 6, 1385–1400

0278-7393/01/$5.00 DOI: 10.1037//0278-7393.27.6.1385

The Inverse Base-Rate Effect Is Not Explained by Eliminative Inference

John K. Kruschke
Indiana University Bloomington

The inverse base-rate effect is a phenomenon in which people learn about some common and some rare outcomes and in subsequent testing people predict the rare outcome for particular sets of conflicting cues, contrary to normative predictions (D. L. Medin & S. M. Edelson, 1988). P. Juslin, P. Wennerholm, and A. Winman (2001) suggested that the effect could be explained by eliminative inference, contrary to the attention-shifting explanation of J. K. Kruschke (1996a). The present article shows that the eliminative inference model exhibits ordinal discrepancies from previously published data and from data of 2 new experiments. A connectionist implementation of attentional theory fits the data well. The author concludes that people can use eliminative inference but that it cannot account for the inverse base-rate effect.

One of the more perplexing phenomena in human learning is the *inverse base-rate effect* reported by Medin and Edelson (1988). In Medin and Edelson's medical diagnosis task, people learned which symptoms indicated which fictitious diseases. On any given learning trial, people saw a short list of symptoms, made a diagnosis on the basis of what they had learned until that trial, and then were shown the correct answer. The left side of Figure 1 shows the core of the design. A relatively common (i.e., frequent) disease, labeled "C" in Figure 1, was always indicated by two symptoms, labeled "PC" and "I" in Figure 1. A rare disease, labeled "R," was always indicated by symptoms "PR" and "I." Notice that the structure of the diseases is symmetrical: The diseases share one symptom (I), and they each have a single distinctive symptom (PC and PR, respectively). The shared symptom is labeled "I" because it is an imperfect predictor of the diseases; the perfect predictor of the common disease is labeled "PC," and the perfect predictor of the rare disease is labeled "PR."

People can learn to correctly diagnose these patterns of symptoms fairly easily. Given the symmetry and simplicity of the disease structure, it is plausible that people's knowledge of the diseases would also be symmetric; that is, it is plausible that people's knowledge would accurately reflect the structure of the world. But when people are subsequently tested with novel symptom combinations, asymmetries occur. In particular, when tested with symptom I by itself, people tend to diagnose it as disease C. This actually makes sense, insofar as symptom I is ambiguous and the base rates favor the common disease. But when tested with the symptom pair PC.PR (the dot in the symptom label indicates co-occurrence), people tend to diagnose this case as the rare disease R. This tendency, along with other test results, is called the inverse base-rate effect.

The inverse base-rate effect has been found in other situations, so it is not highly restricted to particular base rates or to particular procedures. Several researchers have found the effect when using different base rates, varying base rates, or different category structures (Kruschke, 1996a; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992). The effect has been found when using a paired-associate memory paradigm with random words, instead of using simulated medical diagnosis (Dennis & Kruschke, 1998). The effect has also been found when using geometric (nonword) stimuli and responses (Fagot, Kruschke, Depy, & Vauclair, 1998).

Explanation by Attention Shifting

Kruschke (1996a) proposed that the inverse base-rate effect is caused by selective attention during learning. The common disease is learned before the rare disease because the common disease occurs so much more often. When learning the common disease, people tend to learn about both of its symptoms (or at least have equal probability of learning about both symptoms), because no other symptoms conflict with them. When subsequently learning the rare disease, people tend to shift their attention away from the shared symptom I, toward the distinctive symptom PR, because the shared symptom has already been learned to indicate the common disease, which is the wrong response in this case. To preserve and protect previous learning, and to accelerate new learning, people shift attention away from symptom I, toward symptom PR.

A caricature of what people learn, according to this theory, is indicated in the right side of Figure 1. The diagram indicates that both symptoms of the common disease are learned to have moderate associative strength with it, but predominantly only the distinctive symptom PR has a strong association with the rare disease. In subsequent tests, therefore, symptom pair PC.PR makes a moderate strength prediction of C but a stronger prediction of R. The main idea illustrated by the right side of Figure 1 is the asymmetry of the learned structure: Cue I is more strongly associated with outcome C than with outcome R, and cue PR is more

This research was supported in part by National Science Foundation Grant BCS-9910720. For comments on a previous version of this article, I thank Peter Juslin, Thomas Wallsten, and two anonymous reviewers. For assistance administering the experiments, I thank Siegal Eli, Kristin Eller, Melody Lynch, and Caroline Reading. For additional developments, see my World Wide Web page at http://www.indiana.edu/~kruschke/.

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychology, Indiana University, 1101 East 10th Street, Bloomington, Indiana 47405-7007. Electronic mail may be sent to kruschke@indiana.edu.

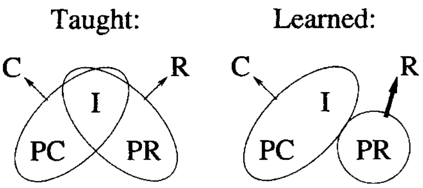

Figure 1. Left side: The core design of the inverse base-rate effect. C = common disease; R = rare disease; PC = a perfect predictor of the common disease; PR = a perfect predictor of the rare disease; I = an imperfect predictor. Right side: Caricature of what is learned according to attentional theory.

strongly associated with outcome R than cue PC is associated with outcome C.

In general, attention shifting can be highly adaptive, despite the fact that it can generate side effects such as the inverse base-rate effect. Learners want to make correct choices in as few training trials as possible. This need for speed can be accommodated with attention shifting. A number of other apparently irrational learning phenomena have been explained by this approach (Kruschke & Blair, 2000; Kruschke & Johansen, 1999).

Explanation by Eliminative Inference

Juslin, Wennerholm, and Winman (1999, 2001) proposed an alternative explanation of the inverse base-rate effect. In this approach, there is no asymmetry in people's knowledge, in that if people know about the rare disease they know about both of its symptoms, and if people know about the common disease they know about both of its symptoms. Instead, the inverse base-rate effect is a consequence of a response strategy for unrecognized stimuli. According to this strategy, called *eliminative inference,* if the current stimulus is not a known stimulus, then the response should not be one of the known responses. In other words, if a stimulus seems to be novel, then eliminate the familiar responses and select a novel response. According to this alternative account, people simply have a lower probability of knowing about the rare disease than of knowing about the common disease. During test trials with a novel combination of symptoms, people recognize that it is not a combination they know about and therefore infer that it cannot be a disease that they know about. Because people are more likely not to know about the rare disease, they are more likely to choose the rare disease in this case.

More specifically, according to this theory, people can learn the rules I.PC → C and I.PR → R. Notice that the rule about the rare disease includes the shared symptom I, as does the rule about the common disease. In this sense the knowledge is symmetric, unlike the asymmetry predicted by attentional theory. When presented with the symptom pair PC.PR, if both rules are known, then *constrained induction* (described in more detail later) uses both partially matched rules, which results in diseases C and R being chosen equally often. If, however, the common rule is known but the rare rule is not known, then eliminative inference causes the known (common) response to be rejected and the unknown (rare) response to be selected. A probabilistic mixture of these various knowledge states generates the choice proportions observed in human learners.

The eliminative inference theory suggests that the inverse base-rate effect occurs because of a response strategy. The attention-shifting theory suggests that the inverse base-rate effect occurs because of a learning mechanism. Juslin et al. (2001) presented experimental evidence that people do spontaneously use eliminative inference in these sorts of experiments. For example, when supplied in testing with a new disease, people will often select that disease for novel symptom combinations. Kruschke and Bradley (1995) showed the same type of behavior, but they called it strategic guessing. The two mechanisms (i.e., response and learning) are not mutually exclusive, and the issue is whether eliminative inference is critical for obtaining and explaining the inverse base-rate effect.

What is especially interesting in the work of Juslin et al. (2001) is the claim that eliminative inference plays a major role in the inverse base-rate effect and that attention shifting, which is a form of cue competition, might have little or no role. Juslin et al. (2001) stated that "the results do, however, demonstrate that cue competition may not be the sole, or even the main, explanation of the inverse base-rate effect" (p. 857) and that "the implementation of eliminative inference presented in this article (ELMO) serves to illustrate that [this] kind of mechanism *alone* has the potential to explain the whole pattern of data observed in the Medin and Edelson design" (p. 865).

Preview of This Article

This article is organized as follows: First, a detailed description of the eliminative inference model (ELMO) is provided. Then a detailed description of a connectionist implementation of attention-shifting theory, called EXIT (Kruschke, in press), is supplied. Then both models are fit to data from two previously published experiments. One of these experiments is the simplest published design that has shown a robust inverse base-rate effect (Kruschke, 1996a, Experiment 1). The other experiment is a phased-training analogue of the basic design, which again has shown robust effects (Kruschke, 1996a, Experiment 2). ELMO nicely produced some of the important aspects of the inverse base-rate effect for both of these experiments. Nevertheless, systematic discrepancies from the data are pointed out, which highlight an asymmetry in human choices suggestive of the attentional theory. EXIT fits these data much better.

Next, a new experiment is presented in which people learned about (a) one pair of common and rare diseases that shared a symptom, as in the standard design of the inverse base-rate effect, and (b) another pair of common and rare diseases that had no shared symptom. Medin and Edelson (1988, Experiment 2) reported a similar experiment of more complicated design. Because ELMO bases its predictions of an inverse base-rate effect on the relative base rates of diseases, it predicts an inverse base-rate effect for both pairs of diseases, whether or not there was a shared symptom during training. The attentional theory bases its predictions on shifts of attention away from shared symptoms. ELMO again shows systematic discrepancies from the new data. Some of the test items have never been used in published experiments, and these items produce an inverse base-rate effect in human choice that ELMO does not accommodate. EXIT again fits the data better than ELMO.

A second new experiment is then presented, in which the base rates of the diseases are reversed during the later part of training. This is supposed to encourage strong learning of the diseases that were initially rare. Medin and Bettger (1991) reported experiments in which the base rates changed during training, but none had the same design as the experiment reported here. Critical new test items are also used. The results of this experiment show a very robust inverse base-rate effect, but ELMO is unable to generate any such effect. EXIT, on the other hand, fits the data very well.

The article concludes by acknowledging a role for eliminative inference in these sorts of experiments but rejecting it as the primary cause of the inverse base-rate effect. Juslin et al. (2001) provided good evidence that people do sometimes respond according to eliminative inference. Kruschke and Bradley (1995) previously showed the same sort of robust strategic guessing, which they implemented in a connectionist model. Nevertheless, on the basis of the kinds of data and model fits presented in this article, it is concluded that eliminative inference does not explain the inverse base-rate effect. Instead, people appear to have asymmetric knowledge that can be explained by attention-shifting theory.

Eliminative Inference and the ELMO Model

In this section, a detailed description of the ELMO model is presented, along with examples of its application in an actual experiment. Table 1 shows the design of what is perhaps the simplest published example of the inverse base-rate effect, from Experiment 1 of Kruschke (1996a). There were two common diseases and two rare diseases, with each of the common diseases occurring three times as often as the rare diseases. Thus, a block of training trials consisted of three occurrences of each of the common disease cases, $I_1.PC_1 \rightarrow C_1$ and $I_2.PC_2 \rightarrow C_2$, and one occurrence of each of the rare disease cases, $I_1.PR_1 \rightarrow R_1$ and $I_2.PR_2 \rightarrow R_2$. There were 15 blocks of training, and then the testing phase presented the nine types of items shown in Table 1. Each of the test types was instantiated two ways. For example, type PC.PR included cases $PC_1.PR_1$ and $PC_2.PR_2$. Types with a subscript *o* indicate combinations of symptoms from other disease pairs. For example, type $PC.PR_o$ included cases of $PC_1.PR_2$ and $PC_2.PR_1$.

In ELMO, every training case has the potential of being learned as a rule for that disease. Therefore, in the present situation, ELMO might know some of the following four rules: $I_1.PC_1 \rightarrow C_1$, $I_1.PR_1 \rightarrow R_1$, $I_2.PC_2 \rightarrow C_2$, and $I_2.PR_2 \rightarrow R_2$. Rules in ELMO's memory are denoted the same way as training instances in the experiment design, because the only rules that ELMO can learn are, essentially, copies of the training instances. ELMO's knowledge state is abbreviated as a vector of 1s and 0s that indicate knowing or not knowing the corresponding rule (in the sequential order C_1, R_1, C_2, R_2). For example, ELMO might know only the two rules for the two common diseases, and this knowledge state is denoted by the vector 1010.

Table 1
Design of Experiment 1 From Kruschke (1996a)

Phase	Symptoms → Disease
3:1 base-rate training	(3×) $I_1.PC_1 \rightarrow C_1$ (3×) $I_2.PC_2 \rightarrow C_2$ (1×) $I_1.PR_1 \rightarrow R_1$ (1×) $I_2.PR_2 \rightarrow R_2$
Testing	I PC PR PC.PR I.PC.PR $I.PC_o$ $I.PR_o$ $PC.PR_o$ $I.PC.PR_o$

Note. C = common disease; R = rare disease; PC = perfectly predictive symptom of a common disease; PR = perfectly predictive symptom of a rare disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. A subscript "o" indicates a symptom from the other pair of diseases (e.g., $I.PC_o$ indicates cases of $I_1.PC_2$ and $I_2.PC_1$ collapsed).

The common disease rules are known with a probability denoted p_C, and the rare disease rules are known with a probability denoted p_R. Presumably $p_R < p_C$ so that eliminative inference can produce rare-disease responses, but these values are freely estimated parameters. Knowledge of each rule is assumed to be independent of knowledge of the other rules; therefore the probability of a knowledge state can be computed as the product of the probabilities of knowing individual rules. For example, the probability of being in knowledge state 1010 is $p_C(1 - p_R)p_C(1 - p_R)$. Table 2 lists the 16 possible knowledge states in ELMO applied to Experiment 1 of Kruschke (1996a).

When a set of symptoms, that is, the *probe,* is presented to ELMO for diagnosis, ELMO compares the probe with the rules it knows and then chooses a response strategy accordingly. As a first step, the similarities of the probe to each of the conditions of the known rules are computed. If the stimulus is highly similar to one or more known rules, then *direct induction* (described in more detail later) uses these rules and selects a disease from among the implied diseases, with choice probability proportional to the similarity of the probe to the rule's condition.

Similarity is computed as in the well-known context model of Medin and Schaffer (1978). If the stimulus and the rule condition match exactly, then their similarity is 1.0. For any mismatched feature, however, the similarity is multiplied by a factor $s < 1$, which thereby reduces the similarity. Different features can have different degrees of impact on the similarity, depending on the salience or the learned importance of the feature. A larger value of s indicates a smaller importance, because a larger value of s reduces the similarity less. Therefore the factor s is here called the *insignificance* of the feature.

It is important to note that the insignificance of symptom PC is the same as the insignificance of symptom PR. This equality maintains the symmetry of the known rules and is one of the crucial distinctions between ELMO and the attentional theory. If, instead, the insignificance of PC was greater than the insignificance of PR (i.e., if $s_{PC} > s_{PR}$), then ELMO would implement a central assumption of attentional theory. If ELMO also allowed the insignificance of symptom I to depend on which rule it was part of, then ELMO would implement another central assumption of attentional theory.

The general formalization of the similarity rule is as follows. Let a probe or a rule condition be represented by a vector of values indicating the absence or presence of the symptoms. For notational purposes, we set the symptoms in the order I_1, PC_1, PR_1, I_2, PC_2, PR_2. The vector [1, 1, 0, 0, 0, 0], for example, represents the presence of I_1 and of PC_1 and the absence of all other symptoms. Then the overall similarity of a probe $\mathbf{x} = [\ldots, x_i, \ldots]$ to a rule condition $\mathbf{y} = [\ldots, y_i, \ldots]$ is $S = \Pi_i s_i^{|x_i - y_i|}$, where s_i is the insignificance of the ith feature.

ELMO assumes that any perfect predictor has a certain insignificance, denoted s_P, and any imperfect predictor has a possibly

Table 2

ELMO Choice Percentages for Each Knowledge State, Using Probe $PC_1.PR_1$ and Four Possible Rules From the Design of Experiment 1 of Kruschke (1996a)

Knowledge state (active rules)	Probability of knowledge state	Response strategy	Choice			
			C_1	R_1	C_2	R_2
1111	$p_C p_R p_C p_R$ = .4124	constr. induct.	50	50	0	0
1110	$p_C p_R p_C \overline{p_R}$ = .2022	constr. induct.	50	50	0	0
1011	$p_C \overline{p_R} p_C p_R$ = .2022	elim. infer.	0	100	0	0
1010	$p_C \overline{p_R} p_C \overline{p_R}$ = .0991	elim. infer.	0	50	0	50
1101	$p_C p_R \overline{p_C} p_R$ = .0185	constr. induct.	50	50	0	0
0111	$\overline{p_C} p_R p_C p_R$ = .0185	elim. infer.	100	0	0	0
1100	$p_C p_R \overline{p_C} \overline{p_R}$ = .0091	constr. induct.	50	50	0	0
0110	$\overline{p_C} p_R p_C \overline{p_R}$ = .0091	elim. infer.	50	0	0	50
1001	$p_C \overline{p_R} \overline{p_C} p_R$ = .0091	elim. infer.	0	50	50	0
0011	$\overline{p_C} \overline{p_R} p_C p_R$ = .0091	elim. infer.	50	50	0	0
1000	$p_C \overline{p_R} \overline{p_C} \overline{p_R}$ = .0045	elim. infer.	0	33	33	33
0010	$\overline{p_C} \overline{p_R} p_C \overline{p_R}$ = .0045	elim. infer.	33	33	0	33
0101	$\overline{p_C} p_R \overline{p_C} p_R$ = .0008	elim. infer.	50	0	50	0
0100	$\overline{p_C} p_R \overline{p_C} \overline{p_R}$ = .0004	elim. infer.	33	0	33	33
0001	$\overline{p_C} \overline{p_R} \overline{p_C} p_R$ = .0004	elim. infer.	33	33	33	0
0000	$\overline{p_C} \overline{p_R} \overline{p_C} \overline{p_R}$ = .0002	elim. infer.	25	25	25	25
Probability-weighted sum			35	59	1	6

Note. PC = perfectly predictive symptom of a common disease; PR = perfectly predictive symptom of a rare disease. A dot between symptoms indicates co-occurrence. C = common disease; R = rare disease; p_C = probability of knowing rule for common disease; $\overline{p_C} = 1 - p_C$; p_R = probability of knowing rule for rare disease; $\overline{p_R} = 1 - p_R$; constr. induct. = constrained induction; elim. infer. = eliminative inference. Knowledge probabilities were computed using $p_C = .957$ and $p_R = .671$. Response strategies were determined from similarities computed using $s_I = .643$ and $s_P = .388$. These values are the best fitting parameter values for Experiment 1 from Kruschke (1996a). Choice probabilities do not total 100% in some rows because of rounding.

different insignificance, denoted s_I. Presumably $s_I > s_P$, to reflect these symptoms' overall diagnosticity, but the parameter values are freely estimated. As an example of similarity computation, suppose the probe is $PC_1.PR_1$, denoted by the vector [0, 1, 1, 0, 0, 0]. Suppose that ELMO knows the rule $I_1.PC_1 \rightarrow C_1$, which has condition denoted by the vector [1, 1, 0, 0, 0, 0]. Then the similarity of the probe to the rule condition is $s_I^1 s_P^0 s_P^1 s_I^0 s_P^0 s_P^0 = s_I s_P$.

After the similarity of the probe to each known rule is computed, the model decides what response strategy to use, either induction (on the basis of the partially matched rules) or eliminative inference. ELMO processes the probe according to three possible situations. First, if there is at least one known rule that deviates by only one symptom (or less) from the probe, then the choice is made on the basis of induction over the known rules that share at least one feature with the probe. This first case is called a situation with high similarity, and the resulting process is called direct induction. Second, if all known rules deviate from the probe by two or more features but there are multiple (two or more) partially matching rules, then the choice is made on the basis of induction over the known rules that share at least one feature with the probe. This second case is called a situation with low similarity but multiple matching rules, and the resulting process is called constrained induction. Third, if all known rules deviate from the probe by two or more features, a situation called low similarity, and there are only one or zero partially matching rules, then the choice is made on the basis of eliminative inference. Concrete examples are provided next.

Suppose that all four rules are known and that the probe I_1 is presented. Then there are two known rules that deviate by only one symptom from the probe, specifically rules $I_1.PC_1 \rightarrow C_1$ and $I_1.PR_1 \rightarrow R_1$. Therefore this is a high-similarity situation, and direct induction is applied over the partially matched rules. The similarity of the probe to the first rule is s_P, and the similarity of the probe to the second rule is also s_P. The probability of applying the first rule (and choosing C_1) is therefore $s_P/(s_P + s_P) = .50$.

The processing for probe $PC_1.PR_1$ is shown in detail in Table 2 because this probe is so critical in the assessment of the inverse base-rate effect. Each row in the table corresponds to a particular knowledge state: Because there are four possible rules, and each rule might be known or not known, there are 16 possible knowledge states. The next column shows the probability of being in this knowledge state, given the probability of knowing the individual rules. The values for p_C and p_R were taken from the best fit of ELMO to data from Experiment 1 of Kruschke (1996a), which is described in full detail later. The value $1 - p$ is denoted $\bar{p}$. The rows are arranged from most to least probable knowledge state.

The first row of Table 2 shows a case of constrained induction. In this first row, the model knows all four rules. The probe $PC_1.PR_1$ deviates from all known rules by two or more features, but it partially matches multiple known rules. The third column indicates this situation with the annotation "constr. induct.," which means there is low similarity but constrained induction is applied. The model therefore applies the partially matched rules according to the proportional similarity of the probe to the rule conditions, as described above for direct induction. The last four columns indicate the probabilities of choosing each disease on the basis of this constrained induction. Because the probe has equal similarity to the conditions for diseases C_1 and R_1, ELMO (in this knowledge state) chooses these two diseases with 50–50 probability.

The third row of Table 2 shows a case of eliminative inference. In this case, the model knows both common disease rules but knows only the second rare disease rule. The probe $PC_1.PR_1$ deviates from all known rules by two or more features, and it partially matches only one known rule. Therefore the model eliminates all the known diseases and chooses from among the unknown diseases. In this case there is only one unknown disease, so it is always chosen, as indicated in the last four columns of the table.

The bottom row of Table 2 shows the probability-weighted sum of all the knowledge states' choice percentages. For these particular parameter values, ELMO chooses the common disease just 35% and the rare disease 59% of the time; that is, it shows a key aspect of the inverse base-rate effect.

In summary, ELMO has free parameters for the probability of knowing the common diseases and the rare diseases, and it has free parameters for the insignificance of mismatching a perfectly predictive symptom or an imperfectly predictive symptom. If the overall similarity of a probe and known rules is high, or if more than one known rule shares at least one feature with the probe, then induction with rules takes place. Otherwise, eliminative inference occurs.

An Attention-Shifting Model (EXIT)

In this section of the article, a connectionist model is described that implements the principles of attentional theory. The model is an extension of the attention to distinctive input (ADIT) model proposed previously by Kruschke (1996a) and was introduced by Kruschke (in press). The model is called EXIT for two reasons: First, the appellation is an acronym for EXtended adIT. Second, an exit sometimes follows an adit.

EXIT expresses several psychological principles in connectionist formalisms. Figure 2 shows the architecture of the model. The theory assumes that people learn to associate cues with outcomes, and these associations are implemented in the model as connections from nodes that represent cues to nodes that represent outcomes. In Figure 2, these learned associations are represented by the thick arrows impinging on the outcome nodes near the top of the diagram.

The theory assumes that people can differentially allocate attention to the cues and that the attention modulates the influence of the cues on both responding and learning. This is implemented in EXIT by multiplicative factors on each cue, indicated in Figure 2 by the Xs in boxes directly above the cue nodes. (The Xs are meant to suggest the multiplicative action of attention.) The theory also assumes that attention is limited in its overall capacity; that is, increasing attention to one cue entails decreasing attention to other cues. This capacity constraint, also known as attention normalization, is indicated in Figure 2 by the crisscrossing connections to the attention nodes from the gain nodes. The gain nodes express the underlying attentional strength allocated to each cue before the attention is normalized. By default, if a cue is present, it is allocated some attentional gain. This default allocation of attention is indicated in Figure 2 by the one-to-one connections from cue nodes to gain nodes.

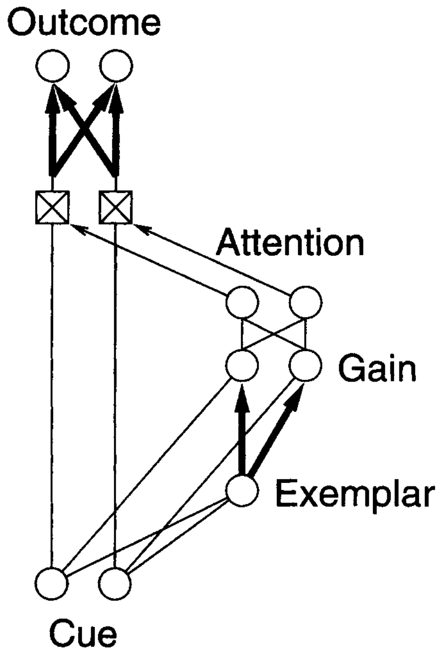

Figure 2. Architecture of the EXIT model. This diagram illustrates a case with two cues, one exemplar, and two outcomes. Thick arrows denote connections with learned weights; thin lines (with or without arrowheads) denote fixed-weight connections.

The theory further assumes that attentional shifts can be learned. That is, if attention has been shifted on a certain trial with a particular stimulus, then this new attentional distribution should be a learned response to the particular stimulus. This is implemented in the model by a set of associations from the cues to the attentional gains. Figure 2 represents the learned attentional associations as thick arrows from exemplar nodes to the attentional gain nodes. The reason that the mapping from cues to gains is mediated by exemplars is that, at least in principle, the model should be capable of learning different allocations of attention for different stimuli (Aha & Goldstone, 1992; Erickson & Kruschke, 1999).

Finally, the theory assumes that people shift attention and learn associations in an efficient manner. The model implements this assumption by using gradient descent on error to drive all aspects of learning. Error is defined as the discrepancy between the correct response and the predicted outcome. Gradient descent on error merely computes the change that would reduce this error most quickly.

The processing steps in EXIT are as follows. First, a stimulus is presented, which activates the corresponding cue nodes. The cue activation spreads to the exemplar nodes, which are activated to the extent that the current stimulus is similar to the stimulus that is represented by the exemplar. The activation from the exemplars then spreads to the attentional gain nodes, which can thereby be selectively activated depending on previous learning about attentional distributions for specific exemplars. Next, the attentional gain is normalized, and the capacity-limited attention strengths are applied (multiplied) to the cue activations. These attentionally gated cue activations then spread to the outcome nodes, where an outcome is selected with probability directly related to the relative activation of the outcomes.

After the model makes a prediction for a given stimulus, corrective feedback is supplied, just as it is for human participants in the experiments. Any error between the prediction and the correct outcome drives a rapid shift in attention. If attention to a certain cue is causing selection of the wrong outcome, then attention to that cue is reduced. If attention to another cue would help generate the correct outcome, then attention to this other cue is increased. After the attentional shift, the associative weights from cues to outcomes are adjusted to reduce the remaining error, and the associative weights from exemplars to gains are adjusted so that this attentional distribution will be better evoked the next time these particular cues are presented.

There are seven free parameters in EXIT, as explained in the formal description provided in the Appendix. In all the fits reported below, the model was trained on the same stimulus–feedback sequences as the human participants who generated the data.

Previous Experiments

The Basic Inverse Base-Rate Effect: Experiment 1 From Kruschke (1996a)

Of all the previous experiments that have shown the inverse base-rate effect, perhaps the simplest design was Experiment 1 of Kruschke (1996a), as shown in Table 1. Table 3 shows people's choices from the test phase of that experiment. A robust inverse base-rate effect is evident for case PC.PR (third row from bottom), for which people chose the rare disease 61.2% and the common disease only 35.3% of the time.

Fit of ELMO. The predictions from the best fit of ELMO are also shown in Table 3. The best fit of ELMO yielded root-mean-square deviation (RMSD) = 0.0740 (i.e., 7.40 percentage points difference between predicted and empirical values, on average), using parameter values of $s_I = 0.643$, $s_P = 0.388$, $p_C = .957$, and $p_R = .671$. These parameter values indicate that ELMO weighted mismatches of perfect predictors more heavily than mismatches of imperfect predictors (because $s_I > s_P$), and ELMO "knew" the rules for the common diseases with high probability ($p_C = .957$) but "knew" the rules for the rare diseases with only a modest probability ($p_R = .671$).

ELMO shows a robust inverse base-rate effect for case PC.PR and matches the data well for this case. ELMO also shows, correctly, a preference for disease C when probed with symptom I and when probed with symptom triplet I.PC.PR. Clearly ELMO successfully exhibits some of the hallmark preferences involved in the inverse base-rate effect.

Discrepancies between ELMO and human choice. The first two rows of Table 3 show the results for probes I and I.PC.PR. Notice that ELMO predicts that the choice probabilities for these two cases should be identical. This prediction of equal choice proportions for these two cases is true regardless of the parameter values, because the predicted equality stems from symmetries in the experimental design. I have not derived an analytical proof that the equality obtains for all parameter values, but the equality has been found for a wide range of parameter values that I have tested numerically. Yet the human data show a significant difference in choice probabilities between these two probes. People prefer disease C over disease R much more strongly for probe I than for probe I.PC.PR. Thus, when PC and PR are added to symptom I, people's preferences shift away from disease C toward disease R (with disease C still preferred, but to a smaller degree). One interpretation of this shift is that PR is more strongly associated with R than PC is associated with C. That is, there is an asymmetry in people's knowledge that is not reflected in ELMO.

The third and fourth rows of Table 3 show results for cases I.PC_o and I.PR_o. Notice that in the human data, $p(R_o|I.PR_o) > p(C_o|I.PC_o)$. ELMO, however, makes the opposite prediction, and quite strongly. This (mis)prediction by ELMO occurs over a wide range of parameter values. The human data can be interpreted as indicating that PR_o more strongly indicates R_o than PC_o indicates C_o. Again, there seems to be an asymmetry in people's knowledge that is not reflected in ELMO.

These systematic discrepancies between ELMO and human data are even more pronounced when the model is fit to data from the

Table 3
Choice Percentages From Test Phase of Experiment 1 From Kruschke (1996a), for Humans, ELMO, and EXIT

	Human choice				ELMO choice				EXIT choice			
Symptom	C	R	C_o	R_o	C	R	C_o	R_o	C	R	C_o	R_o
I	**74.6**	**17.4**	4.9	3.1	**64.2**	**35.6**	0.0	0.2	**74.0**	**15.2**	6.0	4.8
I.PC.PR	**58.0**	**40.2**	1.3	0.4	**64.2**	**35.6**	0.0	0.2	**57.4**	**38.8**	1.9	1.9
I.PC_o	40.6	8.0	**46.9**	4.5	23.0	16.3	**60.3**	0.3	33.5	9.3	**53.4**	3.8
I.PR_o	21.9	8.5	3.1	**66.5**	26.4	25.9	0.2	**47.6**	25.3	6.3	2.7	**65.7**
PC	93.3	3.1	3.1	0.4	98.7	0.6	0.1	0.6	90.1	2.6	3.9	3.3
PR	4.0	91.1	1.8	3.1	0.6	93.5	0.6	5.3	1.7	94.2	2.4	1.8
PC.PR	35.3	61.2	2.2	1.3	35.1	58.5	0.7	5.7	34.5	58.2	3.9	3.4
PC.PR_o	35.3	2.7	5.8	56.3	35.1	5.7	0.7	58.5	36.0	2.3	2.4	59.3
I.PC.PR_o	71.9	3.6	3.6	21.0	72.3	8.0	0.0	19.6	72.8	2.3	1.4	23.6

Note. PC = perfectly predictive symptom of a common disease; PR = perfectly predictive symptom of a rare disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. C = common disease; R = rare disease. A subscript "o" indicates a symptom from the other pair of diseases (e.g., I.PC_o indicates cases of $I_1.PC_2$ and $I_2.PC_1$ collapsed). Bold font indicates data that were particularly challenging for ELMO.

phased-training version of the inverse base-rate effect, as described in the next section.

Fit of EXIT. Table 3 also shows the best fitting predictions of EXIT, which yielded RMSD = 0.0245 (compared with RMSD = 0.0740 for ELMO). The Appendix reports the best fitting parameter values. EXIT reflects the data fairly accurately, including the greater preference of disease C when symptom I is presented than when symptoms I.PC.PR are presented and the fact that $p(\mathrm{R_o}|\mathrm{I.PR_o}) > p(\mathrm{C_o}|\mathrm{I.PC_o})$.

Although EXIT fits the data better than ELMO, the fit is not perfect, of course. For example, in EXIT, $p(\mathrm{C}|\mathrm{PC}) = 90.1 < p(\mathrm{R}|\mathrm{PR}) = 94.2$, contrary to data, in which $p(\mathrm{C}|\mathrm{PC}) = 93.3 > p(\mathrm{R}|\mathrm{PR}) = 91.1$, although this difference in the data is not significant statistically, $\chi^2(1, N = 448) = 0.78$. Moreover, data from comparable experiments do not always show $p(\mathrm{C}|\mathrm{PC}) > p(\mathrm{R}|\mathrm{PR})$. For example, in Experiment 3 of Kruschke (1996a), $p(\mathrm{C_i}|\mathrm{PC_i}) = 89.3 < p(\mathrm{R_i}|\mathrm{PR_i}) = 93.5$. And in Experiment 1 of this article (reported below), $p(\mathrm{C}|\mathrm{PC}) = 83.1 < p(\mathrm{R}|\mathrm{PR}) = 87.2$ (where C corresponds to outcome E and R corresponds to outcome L).

Nevertheless, if we suppose that there is a tendency in the data for $p(\mathrm{C}|\mathrm{PC}) > p(\mathrm{R}|\mathrm{PR})$, then EXIT's failure to produce the common disease more strongly than the rare disease in this case might be attributed to many different possible sources. Perhaps the implementation of bias (base rate) learning is a cause (see the Appendix for discussion). Perhaps direct connections from cues to gain nodes are needed in addition to exemplar-mediated connections. Or perhaps there need to be exemplar-mediated connections from cues to response nodes, in addition to direct connections. Any of these alterations could change the detailed response rates without violating the model's motivating principles, but any of these alterations would also involve additional free parameters without dramatically improving the RMSD overall. Therefore exploration of these modeling options awaits future research.

The Phased Training Analogue: Experiment 2 From Kruschke (1996a)

According to the attentional theory of the inverse base-rate effect, the main role of differential base rates is to encourage the common diseases to be learned before the rare diseases. Instead of relying on base rates to probabilistically encourage learning some diseases before others, Experiment 2 of Kruschke (1996a) simply trained people on some diseases before others. The design is shown in Table 4. A comparison of Tables 1 and 4 indicates that what was a common disease or symptom in Experiment 1 is an early disease or symptom in Experiment 2. What was a rare disease or symptom in Experiment 1 is a late disease or symptom in Experiment 2. The late training phase contained equal numbers of trials of each of the four diseases; that is, the base rates of the four diseases were equal in the late training phase.

Table 5 shows people's choices from the test phase. A robust analogue of the inverse base-rate effect is evident for case PE.PL (third row from bottom), for which people chose the late disease 61.2% and the early disease only 35.3% of the time.

Fit of ELMO. The predictions from the best fit of ELMO to these data are also shown in Table 5. The best fit yielded RMSD = 0.0817 (i.e., 8.17 percentage points deviation on average), using parameter values of $s_\mathrm{I} = 0.702$, $s_\mathrm{P} = 0.405$, $p_\mathrm{E} = .970$, and $p_\mathrm{L} = .677$. (These parameter values are close to those that best fit the data of Experiment 1.) ELMO correctly shows a preference for disease L when probed with PE.PL and a preference for disease E when probed with symptom I or with symptom triplet I.PE.PL.

Table 4
Design of Experiment 2 From Kruschke (1996a)

Phase	Symptoms → Disease
Early training	$\mathrm{I_1.PE_1} \rightarrow \mathrm{E_1}$ $\mathrm{I_2.PE_2} \rightarrow \mathrm{E_2}$
Later training	$\mathrm{I_1.PE_1} \rightarrow \mathrm{E_1}$ $\mathrm{I_2.PE_2} \rightarrow \mathrm{E_2}$ $\mathrm{I_1.PL_1} \rightarrow \mathrm{L_1}$ $\mathrm{I_2.PL_2} \rightarrow \mathrm{L_2}$
Testing	I PE PL PE.PL I.PE.PL $\mathrm{I.PE_o}$ $\mathrm{I.PL_o}$ $\mathrm{PE.PL_o}$ $\mathrm{I.PE.PL_o}$

Note. E = early-trained disease; L = late-trained disease; PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. A subscript "o" indicates a symptom from the other pair of diseases (e.g., $\mathrm{I.PE_o}$ indicates cases of $\mathrm{I_1.PE_2}$ and $\mathrm{I_2.PE_1}$ collapsed).

Discrepancies between ELMO and human choice. The first two rows of Table 5 show the data and predictions for probes I and I.PE.PL. Notice that ELMO predicts that the choice probabilities for these two cases should be identical. This prediction of equal choice proportions of these two cases is true regardless of the parameter values, because the predicted equality stems from symmetries in the experimental design. Yet the human data show a significant difference in choice probabilities between these two probes. People prefer disease E over disease L much more strongly for probe I than for probe I.PE.PL. Thus, when PE and PL are added to symptom I, people's preferences shift away from disease E toward disease L (with disease E still slightly preferred, but to a much smaller degree). One interpretation of this shift is that PL is more strongly associated with L than PE is associated with E. That is, there is an asymmetry in people's knowledge that is not reflected in ELMO.

The third and fourth rows of Table 5 show results for cases $\mathrm{I.PE_o}$ and $\mathrm{I.PL_o}$. Notice that in the human data, $p(\mathrm{L_o}|\mathrm{I.PL_o}) > p(\mathrm{E_o}|\mathrm{I.PE_o})$. ELMO, however, makes the opposite prediction, and quite strongly. The human data can be interpreted as indicating that $\mathrm{PL_o}$ more strongly indicates $\mathrm{L_o}$ than $\mathrm{PE_o}$ indicates $\mathrm{E_o}$. Again, there seems to be an asymmetry in people's knowledge that is not reflected in ELMO.

Fit of EXIT. The phased training analogue of the inverse base-rate effect is also very well fit by EXIT. Best fitting predictions to Experiment 2 from Kruschke (1996a) are shown in Table 5, which yielded RMSD = 0.0244 (compared with RMSD = 0.0817 for ELMO). EXIT shows the effects in the data fairly accurately, including the greater preference of disease E when symptom I is presented than when symptoms I.PE.PL are presented and the fact that $p(\mathrm{L_o}|\mathrm{I.PL_o}) > p(\mathrm{E_o}|\mathrm{I.PE_o})$.

Although EXIT fits the data better than ELMO, the fit is not perfect, of course. For example, in EXIT, $p(\mathrm{E}|\mathrm{PE}) = 90.0 < p(\mathrm{L}|\mathrm{PL}) = 94.9$, contrary to data, in which $p(\mathrm{E}|\mathrm{PE}) = 92.5 > p(\mathrm{L}|\mathrm{PL}) = 91.5$, although this difference in the data is not significant statistically, $\chi^2(1, N = 424) = 0.13$. In Experiment 1 (reported below), however, the data show $p(\mathrm{E}|\mathrm{PE}) = 83.1 < p(\mathrm{L}|\mathrm{PL}) = 87.2$, consistent with EXIT's prediction. As suggested

Table 5
Choice Percentages From Test Phase of Experiment 2 From Kruschke (1996a), With Best Fitting Predictions of ELMO and EXIT

	Human choice				ELMO choice				EXIT choice			
Symptom	E	L	E_o	L_o	E	L	E_o	L_o	E	L	E_o	L_o
I	**80.2**	**11.8**	3.3	4.7	**64.6**	**35.3**	0.0	0.1	**78.4**	**11.3**	5.4	4.9
I.PE.PL	**51.4**	**46.7**	1.4	0.4	**64.6**	**35.3**	0.0	0.1	**50.0**	**45.5**	2.2	2.3
$I.PE_o$	37.9	7.6	**50.2**	4.3	21.8	15.1	**60.3**	0.2	36.4	9.0	**51.3**	3.3
$I.PL_o$	20.8	10.0	1.4	**67.8**	25.5	25.3	0.1	**49.1**	24.9	6.1	2.1	**66.9**
PE	92.5	1.4	3.8	2.4	99.1	0.4	0.0	0.4	90.0	2.3	4.1	3.6
PL	2.8	91.5	1.9	3.8	0.4	94.0	0.4	5.1	1.2	94.9	2.1	1.8
PE.PL	31.6	64.6	1.9	1.9	34.9	59.2	0.5	5.4	30.5	60.8	4.5	4.2
$PE.PL_o$	37.4	4.7	0.9	56.9	34.9	5.4	0.5	59.2	34.0	2.2	2.1	61.6
$I.PE.PL_o$	69.8	4.7	2.8	22.6	70.1	8.2	0.0	21.7	74.3	2.2	1.2	22.3

Note. PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. E = early-trained disease; L = late-trained disease. A subscript "o" indicates a symptom from the other pair of diseases (e.g., $I.PE_o$ indicates cases of $I_1.PE_2$ and $I_2.PE_1$ collapsed). Bold font indicates data that were particularly challenging for ELMO.

earlier, if the data of future experiments were to reliably go against EXIT for this case, there are several implementation details that might be altered in EXIT to address this discrepancy.

Summary of Fits to Previous Experiments

When ELMO is fit to Experiments 1 and 2 of Kruschke (1996a), it robustly shows some crucial aspects of the inverse base-rate effect and its phased-training analogue. Unfortunately, ELMO also shows large and systematic discrepancies from the data. The discrepancies were interpreted as indicating that human learners have a stronger association of PR with R than of PC with C or, more appropriately, a stronger association of PL with L than of PE with E, but this asymmetry is not possible in ELMO. The attention-shifting model, EXIT, fits the data much better. The next sections of the article describe two new experiments that reveal additional mispredictions of ELMO. In the second new experiment, ELMO produces no inverse base-rate effect, but people do.

Experiment 1: No Shared Symptom During Training

The attentional theory predicts a shift of attention during learning the rare disease because the shared feature is already associated with the earlier-learned common disease. If there is no shared feature, there should be no inverse base-rate effect. The eliminative inference theory, however, predicts that the inverse base-rate effect should persist. This is because the inverse base-rate effect is caused entirely by lack of learning the rare disease, not by any structural overlap with the common disease.

The design of Experiment 1 is shown in Table 6. One pair of early trained (common) and later trained (rare) diseases is the standard design for the inverse base-rate effect: There is a single shared, imperfectly predictive symptom (I), and each disease has a single perfectly predictive symptom (PE or PL), as shown in the right column of Table 6. Another pair of early and later trained diseases has no shared symptom. These disjoined diseases are therefore labeled E_D and L_D. The earlier trained disease has two perfectly predictive symptoms, labeled PE'_D and PE''_D, and the later trained disease has two perfectly predictive symptoms, labeled PL'_D and PL''_D, as shown in the middle column of Table 6. The design is a simplified version of Experiment 2 from Medin and Edelson (1988), who also examined the influence of shared cues during learning.

The testing trials included the usual tests of the inverse base-rate effect, including case PE.PL for the diseases with the shared symptom and cases of type $PE^*_D.PL^*_D$ for the disjoined diseases. The notation PE^*_D indicates cases of PE'_D and PE''_D collapsed because they are structurally equivalent. Thus, there were four cases of type $PE^*_D.PL^*_D$: $PE'_D.PL'_D$, $PE''_D.PL'_D$, $PE'_D.PL''_D$, and $PE''_D.PL''_D$. The attentional theory predicts no inverse base-rate effect for type $PE^*_D.PL^*_D$, but the eliminative inference theory does,

Table 6
Design of Experiment 1

Phase	Symptoms → Disease	
Initial training	$PE'_D.PE''_D \to E_D$	$I.PE \to E$
3:1 base rate training	$(3\times)\ PE'_D.PE''_D \to E_D$ $(1\times)\ PL'_D.PL''_D \to L_D$	$(3\times)\ I.PE \to E$ $(1\times)\ I.PL \to L$
Testing	$PE'_D.PE''_D$ $PL'_D.PL''_D$ — PE^*_D PL^*_D $PE^*_D.PL^*_D$ $PE'_D.PE''_D.PL'_D.PL''_D$ $PE^*_D.PE.PL$ $PL^*_D.PE.PL$	I.PE I.PL I PE PL PE.PL I.PE.PL

Note. E = early-trained disease; L = late-trained disease. Subscript "D" indicates symptoms or diseases from the disjoined pair, that is, the pair of diseases with no shared symptom. I = imperfectly predictive symptom; PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease. Superscript single and double primes indicate two different symptoms for the disjoined diseases. PE^*_D = cases of PE'_D and PE''_D collapsed; PL^*_D = cases of PL'_D and PL''_D collapsed. A dot between symptoms indicates co-occurrence.

assuming that later trained, rare disease L_D is known with a lesser probability than the earlier trained, common disease E_D.

The testing trials also included two new types, PE_D^*.PE.PL and PL_D^*.PE.PL. These new types are also diagnostic between the two theories. Attentional theory predicts that of the responses for E or L, an inverse base-rate effect should persist. That is, there should be a greater preference for L than for E. ELMO, to the contrary, predicts no inverse base-rate effect for this case because multiple rules are matched and constrained induction dominates eliminative inference.

The training included a brief (16 trial) initial phase in which only the earlier trained diseases were shown, before the 3:1 base-rate phase, to be sure that people experienced the common diseases before the rare diseases. This is theoretically motivated by the attentional theory, for which order of learning is important. This does not put ELMO at any disadvantage, because ELMO is provided with free parameters for the probabilities of knowing each disease.

Method

Participants. Participants were 74 students (49 female, 25 male, median age = 19, range = 18–26) who volunteered for partial credit in an introductory psychology class at Indiana University.

Stimuli. Because prior knowledge about real symptoms and diseases might affect learning, random words were used in place of actual symptoms. The Medical Research Council (MRC) psycholinguistic database (Coltheart, 1981; the MRC psycholinguistic database is available online at http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm) was searched for five-letter nouns with a mean imagability at least 1.5 *SD* above the mean. This search yielded 48 candidate words, from which 7 were selected that were not obviously suggestive of anything medical. These 7 words were the following: *snake, mouse, robin, whale, puppy, skunk,* and *trout.* For each participant, the 7 words were randomly assigned to abstract symptoms.

Procedure. Participants sat individually before a standard desktop computer, in a dimly lit, ventilated, sound-dampened booth. Participants responded by pressing one of the keys *F, G, H,* or *J* on the standard keyboard. The assignment of response keys to abstract disease roles was randomized separately for each participant.

Written instructions were presented on the computer screen, which participants read at their own pace. The instructions indicated that the task was to learn which symptoms tended to go with which diseases and that the goal was to be as accurate as possible. Participants were told that random words would be used as place holders for symptoms. The instructions also indicated before the testing phase that the test trials would not be supplied with correct diagnoses but that the participant should give his or her best guess on the basis of previous learning.

On a training trial, the symptom words were presented in a vertical list, separately randomized on each trial. A response prompt appeared below the words. After the learner's response, on training trials, corrective feedback was supplied. If the response was wrong, a tone sounded and the word *WRONG!* appeared. If the response was correct, the word *CORRECT!* appeared. In either event, the correct disease label was also supplied. The learner then had up to 30 s to study the symptoms and correct diagnosis, and he or she initiated the next trial by pressing the space bar. On testing trials, the feedback after a response said only "No official diagnosis is available. Your response has been recorded."

The initial training consisted of 16 trials, 8 trials of each of the common diseases. The 3:1 base-rate training consisted of either 20 or 26 blocks of 8 trials. Of the participants, 39 did the 20-block version, and 35 did the 26-block version. The testing phase results were statistically indistinguishable, and so results of the two versions are collapsed. Within each block, there were three occurrences of each of the common disease cases shown in Table 6 and one occurrence of each of the rare disease cases. The 8 trials were separately permuted for every block and every participant.

The testing block consisted of 2 blocks of the test items shown in Table 6. The cases were separately permuted for every participant and block. The entire experiment lasted approximately 40 min.

Results

Table 7 shows the results of the test phase. The first two rows show choice percentages for cases I and I.PE.PL. The results replicate the trends emphasized from previously published experiments (see Tables 3 and 5): For symptom I, people strongly preferred disease E over disease L, but for symptom triplet I.PE.PL there was only a very weak preference for disease E over disease L. This interaction is reliable, $\chi^2(1, N = 79) = 23.43, p < .001$.

Rows three and four of Table 7 show results from the conflicting cue pairs, PE.PL and $PE_D^*.PL_D^*$. People showed an inverse base-rate effect for the cues that were trained with the shared symptom, 50.9% versus 39.0%. This difference was only marginally significant, however, even using the raw frequencies, $\chi^2(1, N = 196) = 3.45, p \approx .06$. (When poor learners are excluded from the analysis, the effect is stronger.) For the cues that were trained with no shared symptom, however, people showed base-rate consistency, 50.3% versus 42.5%. This difference was only marginally significant, $\chi^2(1, N = 548) = 3.86, p = .05$. However, the interaction between the inverse base-rate effect for probe PE.PL and the base-rate consistency for probe $PE_D^*.PL_D^*$ was reliable, $\chi^2(1, N = 744) = 6.74, p < .01$. This interaction replicates the interaction reported in Experiment 2 of Medin and Edelson (1988).

Rows five and six of Table 7 show results from the new probe types, PE_D^*.PE.PL and PL_D^*.PE.PL. In both cases, an inverse base-rate effect was evident, with the later trained (rare) disease L being preferred over the earlier trained (common) disease E, $\chi^2(1, N = 196) = 10.80, p = .001$, and $\chi^2(1, N = 219) = 8.44, p < .005$, respectively.

The remaining rows of Table 7 show test performance on the training items and some other probe combinations. Notice in particular that test performance was fairly good (around 90% correct) on all trained items except I.PL. This fact will be relevant when ELMO is fit to these data.

Fit of ELMO

If ELMO is constrained to use the same knowledge probabilities for both common and both rare diseases, that is, if we constrain $p_{E_D} = p_E$ and $p_{L_D} = p_L$, then ELMO's fit is poor, with RMSD = 0.0729 (although this RMSD is less than the ELMO's RMSDs for the previous experiments). In particular, with this constraint, ELMO fails to show base-rate consistency for $PE_D^*.PL_D^*$. Instead, it has a preference for the later trained (rare) disease, 48.0% L_D to 38.9% E_D, precisely the same as its preference for the later trained (rare) disease (L) when probed with PE.PL. This equality of later trained (rare) preference makes sense because ELMO generates the inverse base-rate effect from the lower probability of knowing the rules for the later trained (rare) diseases.

The constraint of equal L_D and L knowledge probabilities is unfair, however, because it is clear from the test data that people learned the later trained (rare) disease (L_D) of the disjoined pair

Table 7
Choice Percentages From Test Phase of Experiment 1 and Best Fits of ELMO (With $p_{E_D} < p_{L_D}$) and of EXIT

	Human choice				ELMO choice				EXIT choice			
Symptoms	E_D	L_D	E	L	E_D	L_D	E	L	E_D	L_D	E	L
I	4.7	4.1	**80.4**	**10.8**	0.7	0.4	**61.7**	**37.2**	5.8	5.6	**82.4**	**6.1**
I.PE.PL	2.0	3.4	**48.6**	**45.9**	0.7	0.4	**61.7**	**37.2**	6.2	6.2	**50.2**	**37.4**
PE.PL	8.3	1.8	**39.0**	**50.9**	8.3	5.2	**34.6**	**51.9**	6.5	6.5	**31.8**	**55.2**
$PE^*_D.PL^*_D$	**50.3**	**42.5**	2.5	4.7	**48.5**	**38.6**	3.3	9.6	**48.0**	**37.0**	7.5	7.5
PE^*_D.PE.PL	31.2	2.4	**25.4**	**41.0**	28.0	1.9	**37.1**	**33.0**	26.0	8.0	**26.5**	**39.4**
PL^*_D.PE.PL	3.1	22.4	**29.9**	**44.6**	2.1	29.2	**38.1**	**30.6**	8.2	22.9	**27.6**	**41.3**
$PE'_D.PE''_D$	91.9	2.0	0.7	5.4	86.8	4.3	2.3	6.7	92.7	2.4	2.4	2.4
$PL'_D.PL''_D$	5.4	90.5	1.4	2.7	4.3	90.0	1.5	4.2	3.4	89.8	3.4	3.4
I.PE	4.7	1.4	89.9	4.1	2.3	1.5	83.9	12.4	2.8	2.8	91.6	2.8
I.PL	5.4	2.7	20.9	70.9	6.7	4.2	12.4	76.7	5.6	5.6	16.3	72.5
PE	6.8	6.1	83.1	4.1	2.3	1.5	94.0	2.3	4.0	3.9	88.3	3.9
PL	2.7	4.7	5.4	87.2	6.7	4.2	2.3	86.8	3.3	3.3	3.3	90.1
PE^*_D	87.5	4.4	4.4	3.7	86.8	4.3	2.3	6.7	88.1	3.9	4.0	4.0
PL^*_D	3.4	88.5	2.4	5.7	4.3	90.0	1.5	4.2	5.3	84.0	5.3	5.4
$PE'_D.PE''_D.PL'_D.PL''_D$	54.4	36.7	4.1	4.8	48.5	38.6	3.3	9.6	50.2	37.6	6.1	6.1

Note. E = early-trained disease; L = late-trained disease. Subscript "D" indicates symptoms or diseases from the disjoined pair; that is, the pair of diseases with no shared symptom. I = imperfectly predictive symptom; PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease. Superscript single and double primes indicate two different symptoms for the disjoined diseases. PE^*_D = cases of PE'_D and PE''_D collapsed; PL^*_D = cases of PL'_D and PL''_D collapsed. A dot between symptoms indicates co-occurrence. Bold font indicates data that were particularly challenging for ELMO.

much better than the later trained (rare) disease (L) of the pair with a shared symptom. Therefore ELMO was allowed four different knowledge probabilities, one for each of the four rules that specify the four training cases.

As in the fits of ELMO to previous experiments, ELMO was provided with a free similarity parameter for the shared symptom I and with another free similarity parameter for the perfect predictors PE and PL. Although the symptoms of the diseases E_D and L_D are also perfect predictors and might therefore be given the same similarity parameter as the perfect predictors of diseases E and L, ELMO was instead provided with another free parameter for the similarity calculations involving the symptoms of E_D and L_D. ELMO therefore had seven free parameters.

The best fit yielded RMSD = 0.0641, with parameter values of $s_I = 0.648$, $s_P = 0.491$, $s_{P_D} = 0.692$, $p_{E_D} = .607$, $p_{L_D} = .738$, $p_E = .855$, and $p_L = .610$. Notice that $p_{E_D} < p_{L_D}$, meaning that the probability of knowing the early trained, common disease is less than the probability of knowing the later trained, rare disease; this is highly implausible pyschologically.

With this enhanced freedom in parameters, ELMO was able to show base-rate consistency for $PE^*_D.PL^*_D$ (see the fourth row of Table 7), but only because $p_{E_D} < p_{L_D}$ and consequently eliminative inference paradoxically favored the common disease. An inspection of all the possible knowledge states revealed that when ELMO knows the three rules $PL^*_D.PL''_D \rightarrow L_D$, I.PE → E, and I.PL → L but does not know the rule $PE'_D.PE''_D \rightarrow E_D$, then eliminative inference chooses disease E_D with 100% probability. This knowledge state occurs with 15.1% probability, which is enough to sway the overall choice probability toward disease E_D. To reiterate, ELMO shows base-rate consistency for this probe because it knows the common disease with lower probability than it knows the rare disease, and eliminative inference then leads to choosing the common disease.

ELMO could not show the difference between cases I and I.PE.PL. As pointed out for the previous experiments, ELMO makes precisely the same predictions for cases I and I.PE.PL, but the data are very different.

ELMO also could not show the inverse base-rate effect for the new test cases PE^*_D.PE.PL and PL^*_D.PE.PL. ELMO instead predicts base-rate consistency for these cases, because several fairly high probability knowledge states entail constrained induction rather than eliminative inference and because these particular knowledge states did not include the rule for L.

Fit of EXIT

The data from Experiment 1, which compared training of diseases with no shared symptom with training of diseases with a shared symptom, are fit fairly well by EXIT. The best fitting predictions are shown in Table 7. EXIT yielded RMSD = 0.0309 (compared with RMSD = 0.0641 for ELMO). EXIT accurately shows all the effects seen in the human data, including the larger base-rate consistency for I than for I.PE.PL, the base-rate consistency for $PE^*_D.PL^*_D$ but inverse base-rate effect for PE.PL, and the inverse base-rate effect for PE^*_D.PE.PL and for PL^*_D.PE.PL.

Experiment 2: Reversing the Base Rates During Training

ELMO predicts that if people learn the rare diseases very well, then the inverse base-rate effect should disappear. That is, if the rare diseases are known as well as the common diseases, then they are not differentially eliminated and no inverse base-rate effect occurs. One way to encourage people to learn the rare diseases is to increase their frequency. In Experiment 2, the rare diseases only began training with relatively low frequency. By the end of training, these disease cases were the higher frequency disease.

Attentional theory predicts that the inverse base-rate effect should persist in this situation, because the earlier phase of training, during which attentional shifts occur, is much like the standard paradigm that produces the inverse base-rate effect. The attentional theory allows for changed knowledge of base rates but does not change the asymmetric representation of the diseases unless predictive error demands it.

The design of Experiment 2 is shown in Table 8. Training began with just the two early diseases, then progressed to a phase in which the early diseases were more common than the later trained diseases. The relative base rates were 3:1, indicated in Table 8 by the multiplicative factors, 3× and 1×. Training continued with the base rates reversed, such that the early trained disease was relatively rare and the later trained disease occurred three times as often. The testing phase then probed some of the standard cases for assessing the inverse base-rate effect, plus the new probe $\mathrm{I.PE_o.PL_o}$, which was introduced in Experiment 2 and proved to be especially challenging for ELMO.

Previous experiments by Medin and Bettger (1991) also examined effects of changed base rates during training, but none had the specific design used here or the new test probe.

Method

Participants. Participants were 56 students (37 female, 19 male, median age = 19, range = 18–23) who volunteered for partial credit in an introductory psychology class at Indiana University.

Stimuli. As in Experiment 1, high-imagability words were used as placeholders for symptoms. The six words were the following: *child, mouse, ocean, tulip, piano,* and *arrow.* For each participant, the six words were randomly assigned to abstract symptoms.

Procedure. The instructions and procedure were the same as in Experiment 1. The initial training phase (see Table 8) had 3 blocks of 8 trials, each block having 4 trials of E_1 and 4 trials of E_2. The 3:1 base-rate training phase had 15 blocks of 8 trials, and the 1:3 base-rate training phase had 5 blocks of 8 trials. The testing phase had 2 blocks of 12 trials. For each participant, trials were separately permuted within each block. There were a total of 208 trials. The experiment lasted approximately 35 min.

Table 8
Design of Experiment 2

Phase	Symptoms → Disease	
Initial training	$\mathrm{I_1.PE_1} \rightarrow \mathrm{E_1}$	$\mathrm{I_2.PE_2} \rightarrow \mathrm{E_2}$
3:1 base-rate training	(3×) $\mathrm{I_1.PE_1} \rightarrow \mathrm{E_1}$ (1×) $\mathrm{I_1.PL_1} \rightarrow \mathrm{L_1}$	(3×) $\mathrm{I_2.PE_2} \rightarrow \mathrm{E_2}$ (1×) $\mathrm{I_2.PL_2} \rightarrow \mathrm{L_2}$
1:3 base-rate training	(1×) $\mathrm{I_1.PE_1} \rightarrow \mathrm{E_1}$ (3×) $\mathrm{I_1.PL_1} \rightarrow \mathrm{L_1}$	(1×) $\mathrm{I_2.PE_2} \rightarrow \mathrm{E_2}$ (3×) $\mathrm{I_2.PL_2} \rightarrow \mathrm{L_2}$
Testing	I.PE I I.PE.PL	I.PL PE.PL $\mathrm{I.PE_o.PL_o}$

Note. E = early-trained disease; L = late-trained disease; PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. A subscript "o" indicates a symptom from the other pair of diseases (i.e., $\mathrm{I.PE_o.PL_o}$ indicates cases of $\mathrm{I_1.PE_2.PL_2}$ and $\mathrm{I_2.PE_1.PL_1}$ collapsed).

Results

Table 9 shows the choice percentages for the six types of test items. The first two rows show that memory for the test items was very good, with performance on the later trained items at 86.6% correct.

The third row shows that for the ambiguous probe symptom I, people preferred to choose disease E rather than disease L, 65.6% to 23.2%, $\chi^2(1, N = 199) = 45.35, p < .001$. This preference for disease E goes against the base rates in the final phase of training. When the perfect predictors were added to the ambiguous symptom, in triplet I.PE.PL, then people preferred to choose disease L rather than disease E, 36.2% versus 53.6%, $\chi^2(1, N = 201) = 7.57, p < .01$.

For the conflicting symptoms, PE.PL, people strongly preferred the later trained disease L over the early trained disease E, 63.8% to 26.3%, $\chi^2(1, N = 202) = 34.93, p < .001$. Moreover, when the ambiguous symptom from the other pair was added to the conflicting symptoms (i.e., $\mathrm{I.PE_o.PL_o}$), people continued to prefer the later trained disease over the earlier trained disease, 53.1% to 21.4%, $\chi^2(1, N = 167) = 30.19, p < .001$.

Fit of ELMO

ELMO was supplied with the same four parameters as it had for fits to Experiments 1 and 2 of Kruschke (1996a). The best fit was obtained with $s_I = 0.485$, $s_P = 0.238$, $p_C = .707$, and $p_R = .712$, yielding RMSD = 0.1013. Notice that the probability of knowing a rare disease rule is about equal to the probability of knowing a common disease rule. The predictions given by these parameter values are displayed in Table 9.

ELMO is unable to exhibit the effects seen in the data. ELMO cannot produce a preference for E for probe I. ELMO cannot produce a preference for L for probe I.PE.PL. ELMO cannot produce a preference for L with probe PE.PL, nor can ELMO produce a preference for $\mathrm{L_o}$ with probe $\mathrm{I.PE_o.PL_o}$.

ELMO fails in this case because the later trained (originally rare) disease is well learned and because base rates at the end of training actually favor the later trained (originally rare) disease. ELMO cannot favor the later trained disease while not knowing its rule (which is necessary for ELMO to generate the inverse base-rate effect).

Fit of EXIT

The data from Experiment 2 are fit well by EXIT. The best fitting predictions are shown in Table 9. EXIT yielded RMSD = 0.0196 (compared with RMSD = 0.1013 for ELMO). EXIT again accurately shows all the effects seen in the human data, including the inverse base-rate effect for both I.PE.PL and $\mathrm{I.PE_o.PL_o}$.

Summary and Discussion

Juslin et al. (2001) showed that people can use eliminative inference. This fact is not disputed. Indeed, Kruschke and Bradley (1995) previously reported essentially the same phenomenon under the name of "strategic guessing." Juslin et al. (2001) argued, however, that eliminative inference might be enough to account entirely for the inverse base-rate effect, without recourse to atten-

Table 9
Choice Percentages From Test Phase of Experiment 2 and Best Fits of ELMO and EXIT

	Human choice				ELMO choice				EXIT choice			
Symptom	E	L	E_o	L_o	E	L	E_o	L_o	E	L	E_o	L_o
I.PE	88.8	6.2	2.7	2.2	86.9	6.2	3.5	3.5	92.4	3.0	2.3	2.3
I.PL	7.6	86.6	4.0	1.8	6.2	87.0	3.5	3.4	5.7	86.6	3.8	3.8
I	**65.6**	**23.2**	8.0	3.1	**49.0**	**49.5**	0.8	0.8	**65.7**	**20.3**	7.0	7.0
I.PE.PL	**36.2**	**53.6**	4.9	5.4	**44.0**	**43.6**	6.2	6.1	**35.5**	**54.9**	4.8	4.8
PE.PL	**26.3**	**63.8**	4.5	5.4	**49.0**	**49.5**	0.8	0.8	**23.4**	**61.7**	7.5	7.5
I.PE_o.PL_o	15.6	9.8	**21.4**	**53.1**	12.6	12.7	**37.2**	**37.5**	17.4	10.8	**20.5**	**51.3**

Note. E = early-trained disease; L = late-trained disease; PE = perfectly predictive symptom of an early-trained disease; PL = perfectly predictive symptom of a late-trained disease; I = imperfectly predictive symptom. A dot between symptoms indicates co-occurrence. A subscript "o" indicates a symptom from the other pair of diseases (i.e., I.PE_o.PL_o indicates cases of I_1.PE_2.PL_2 and I_2.PE_1.PL_1 collapsed). Bold font indicates data that were particularly challenging for ELMO.

tional shifting or other forms of cue competition. To bolster their argument, Juslin et al. (2001) showed that a model of eliminative inference called ELMO can produce many of the hallmark preferences in the inverse base-rate effect.

In the present article I have shown that ELMO fails to fit the data in a number of fundamental ways. ELMO predicts identical preferences for test cases I and I.PC.PR, but people do not. People's preferences suggest that PR more strongly indicates R than PC indicates C. ELMO predicts that $p(C_o|I.PC_o) > p(R_o|I.PR_o)$, but people strongly show the opposite. Again, this suggests for human learners that PR more strongly indicates R than PC indicates C. When there is no shared symptom during training, people show base-rate consistency instead of an inverse base-rate effect, but ELMO can do this only by assuming that the rare disease is known better than the common disease; this assumption is rather implausible. The dependency of the inverse base-rate effect on a shared symptom during training suggests that human learners shift attention away from the shared symptom when learning the rare disease. Finally, people show a strong inverse base-rate effect when presented with I.PC_o.PR_o, but ELMO does not. Again, this suggests for human learners that PR more strongly indicates R than PC indicates C, whether or not a feature from another set of outcomes is present.

Perhaps most dramatically, Experiment 2 showed that ELMO cannot produce an inverse base-rate effect when the initially rare disease is well learned, whereas people produce a strong inverse base-rate effect. ELMO fails in this case presumably because it relies on lack of knowing an outcome to produce the effect, but people produce the effect even when they know the outcome well.

The attention-shifting theory proposed by Kruschke (1996a) was shown to account for the data much better. A connectionist implementation of the theory, called EXIT, fit the data well. The model shifts attention to rapidly reduce error. The main claim of the theory is that people's knowledge is asymmetric, unlike the symmetric structure of the actual training cases. For people, it is claimed, symptom I more strongly indicates disease C than disease R and symptom PR more strongly indicates disease R than symptom PC indicates disease C. The connectionist model generates these asymmetries as a natural consequence of error reduction.

Model Comparison: Parameters and Flexibility

The best fitting parameter values for EXIT apparently vary extensively from one experiment to another (see the Appendix). This variation is cause for concern, but it is not necessarily a major liability of the model. There is no reason to presume that the parameter values should be universal constants instead of situation specific. That is, by analogy to a well-known theory of gravity, the parameters in EXIT might correspond to parameters like mass instead of parameters like the gravitational constant. Unfortunately EXIT does not yet have any independent operational measurement of the situation-specific parameter values unlike mass in the theory of gravity.

For theorists who would prefer the parameter values to be constant across experiments, I fit EXIT simultaneously to all four data sets (Tables 3–9). The four data sets comprise 135 degrees of freedom, and EXIT has 7 free parameters. The resulting quantitative fit was, of course, worse than the four separate fits, but all the qualitative effects emphasized in this article were exhibited by EXIT in every experiment. The RMSDs for the four experiments, when fit simultaneously, were 0.0457, 0.0295, 0.0415, and 0.0508, respectively. These RMSDs are still much better in every case than are the RMSDs for ELMO, using separate parameter values for each experiment. ELMO cannot be fit simultaneously to all four experiments because, unlike EXIT, the parameters are structurally linked to the experiment design. For example, Experiment 1 needs 7 parameters in ELMO, but Experiment 2 needs 4 parameters in ELMO. Thus EXIT with 7 parameters fits the 135-*df* data much better than ELMO with 19 parameters. Even for Experiment 1 fit by itself, for which both models have 7 parameters, EXIT fits much better than ELMO.

Of course, model comparison involves more than just equating numbers of parameters and then comparing some measure of fit to the data. Different models with the same number of parameters may have different inherent flexibility (e.g., Myung & Pitt, 2000). Assuming that complexity can be formally measured in a way that is agreeable to proponents of competing theories, then when two models have the same number of parameters and are equally complex, the models can be more or less directly compared for fit to the data. The winning model wins because it flexes in just the right places, rather than merely being more flexible everywhere.

In the present case, however, this issue of model flexibility can be addressed informally. Juslin et al. (2001) designed experiments that highlight the influence of eliminative inference. For those situations, ELMO fits the data better than EXIT can, because EXIT implements no mechanism for eliminative inference. The present article, on the other hand, reports experiments that highlight the influence of attention shifting during learning. For these situations, EXIT fits the data better than ELMO can, because ELMO implements no mechanism for attention shifting. The two models flex in different ways, by design. The main points of the present article are that (a) in previous experiments demonstrating the inverse base-rate effect there are strong asymmetries that eliminative inference cannot address but that attentional shifting can, and (b) in new experiments there are strong inverse base-rate effects that eliminative inference cannot produce but that attentional shifting can. The present article argues that the inverse base-rate effect is different than eliminative inference and that ELMO cannot flex appropriately to fit the inverse base-rate effect. Thus, it is not the case that EXIT is merely more flexible than ELMO and could fit any data better than ELMO. On the contrary, EXIT fits the inverse base-rate effect better than ELMO, whereas ELMO fits eliminative inference better than EXIT.

Other Evidence

The case against eliminative inference as an explanation of the inverse base-rate effect does not rest on only the fits of ELMO to the four experiment results presented above. For example, Dennis and Kruschke (1998) reported a robust inverse base-rate effect in cued recall. Participants learned random word associations. Participants were tested with cue words, in response to which they typed the associated word that came to mind. Unlike category learning experiments, there was no list of response options displayed at test. Indeed, there were a huge number of possible responses, namely, all the words in the participant's vocabulary. Eliminative inference seems to predict that for test cases like PE.PL (or PC.PR), people should eliminate all the associated words they remember and choose from among the universe of other words in their vocabulary. Yet the data showed very few such responses. Perhaps the eliminative inference approach could be salvaged by assuming that people felt constrained to respond with only words that occurred during study. This assumption implies, however, that people did learn the words, which implies that eliminative inference would eliminate those words. Thus, the theory would have to assume that people learned the words but not their associations. This is plausible, but is also a notable expansion of the representational assumptions in the theory.

Other work in progress in my laboratory presents further challenges to the eliminative inference account of the inverse base-rate effect. I have found that the inverse base-rate effect occurs in *function learning*. In function learning, people learn a continuous-valued outcome, instead of a nominal-valued outcome as in category learning. When presented with the conflicting cues PE.PL at test, the PE cue corresponds with a moderately high outcome (for example), whereas the PL cue corresponds with a moderately low outcome. Eliminative inference predicts that people should eliminate the known, moderately high response and instead choose from among moderately low responses or untrained levels of the scale. But people show relatively few such responses at untrained levels of the scale. Moreover, when tested with cue PL alone, people respond correctly and extrapolate appropriately to untrained levels.

Other experiments in my laboratory have investigated learning *after* the inverse base-rate effect. The motivating idea is an old one, that learned attentional shifts should perseverate into subsequent learning (e.g., Kruschke, 1996b; Lawrence, 1950). Attentional theory suggests that when learning I.PL → L, attention shifts away from cue I toward cue PL. Therefore subsequent learning about cue I, in the context of cue PL, might be slow relative to learning about cue PL in the context of cue I. On the other hand, the analogous result is predicted not to occur for I.PE, because attention was not shifted away from cue I toward cue PE during learning. Our experiments have confirmed these predictions. This difference in subsequent associabilities of I.PL and I.PE is difficult to explain in terms of eliminative inference.

The Status of Eliminative Inference

The ELMO model invented by Juslin et al. (2001) is one implementation of the idea of eliminative inference; other implementations are possible in principle. (Indeed, Kruschke & Bradley, 1995, proposed a connectionist implementation but did not apply it to the inverse base-rate effect.) Therefore the failures of ELMO pointed out in this article do not necessarily imply that all possible implementations of eliminative inference must fail. For example, it might be possible to revamp the processes leading to constrained induction, and thereby salvage the approach.

A complete model of learning and performance should include both attention shifting and eliminative inference. Different experiment designs can influence the relative contributions of (a) eliminative inference as a response strategy and (b) attention shifting during learning. The attention-shifting model cannot account for cases of eliminative inference such as those reported by Juslin et al. (2001) and Kruschke and Bradley (1995). The fact that people can use eliminative inference does not imply that it accounts for the inverse base-rate effect, however. This article has argued that eliminative inference both fails to show fundamental asymmetries in human data and even fails entirely to show the inverse base-rate effect in some cases. On the other hand, the attention-shifting model accounted for the data well. Thus, the claim is that the fundamental cause of the inverse base-rate effect is attention shifting during learning.

References

Aha, D. W., & Goldstone, R. (1992). Concept learning and flexible weighting. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulié & J. Hérault (Eds.), *Neurocomputing: Algorithms, architectures and applications* (pp. 227–236). New York: Springer-Verlag.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497–505.

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology, 50*, 131–138.

Erickson, M. A., & Kruschke, J. K. (1999). *Rule and exemplar representation in rule-defined category structures.* Manuscript submitted for publication. Available on the World Wide Web at http://www.indiana.edu/~kruschke/ek99_abstract.html.

Estes, W. K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language, 27,* 196–212.

Estes, W. K. (1994). *Classification and cognition.* New York: Oxford University Press.

Fagot, J., Kruschke, J. K., Depy, D., & Vauclair, J. (1998). Associative learning in baboons (*Papio papio*) and humans (*Homo sapiens*): Species differences in learned attention to visual features. *Animal Cognition, 1,* 123–133.

Garner, W. R. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Gluck, M. A., & Bower, G. H. (1988a). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

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

Juslin, P., Wennerholm, P., & Winman, A. (1999). Mirroring the inverse base rate effect: The novel symptom phenomenon. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 252–257). Hillsdale, NJ: Erlbaum.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 849–871.

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

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

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science, 8,* 201–223.

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

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7,* 636–645.

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for human associative learning* (Research Report No. 141). Indiana University, Cognitive Science Department.

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

Lawrence, D. H. (1950). Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology, 40,* 175–188.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

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

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

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

Myung, I.-J., & Pitt, M. A. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 194–204.

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

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

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

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology, 53A,* 983–1011.

Appendix

Formal Description of EXIT

Activation Propagation to the Category Nodes

Figure 2 shows the architecture of EXIT. Each cue in the stimulus is represented by a corresponding input node in a connectionist network, and each possible outcome is represented by a corresponding output node. If cue i is present in a stimulus, then node i is activated, with activation value $a_i^{\text{in}} = 1$. When cue i is absent, $a_i^{\text{in}} = 0$. When outcome k is present, then the corresponding output node receives a teacher signal, $t_k = 1$, which indicates that the node should be activated. When the outcome is absent, then $t_k = 0$. Input node i is connected to output node k by a link with an associative weight denoted w_{ki}.

When a stimulus is presented, the corresponding input nodes are activated, and activation spreads to the output nodes via the weighted connections. The attentional strengths modulate the influence of the input activations, such that the output activation is determined by a weighted sum across the attentionally gated input activations. Formally, the activation of output node k is determined as follows:

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

where α_i is the attention strength on input node i. The source of these attention values is described below. The input-to-category association weights are initialized at zero but change with learning, as described later.

Category node activations are mapped to response probabilities using a version of the Luce (1959) choice rule, also known as the softmax rule (Bridle, 1990; Rumelhart, Durbin, Golden, & Chauvin, 1995). Specifically, the probability of choosing category c is given by

$$p(c) = \exp(\phi a_c^{\text{out}}) / \sum_k \exp(\phi a_k^{\text{out}}), \tag{2}$$

where ϕ is a scaling constant. In other words, the probability of classifying the given stimulus into category c is determined by the magnitude of category c's activation relative to the sum of all category activations. The constant, ϕ, determines the decisiveness of the network: A large value of ϕ expresses a highly decisive choice, in that it causes just a small activation advantage for category c to be translated into a large choice preference for category c. A small value of ϕ expresses an indecisive or unconfident network, in that the small ϕ causes large activation differences to be translated into ambivalent choices.

This rule for mapping output activations to choice probabilities has many precedents in the psychological literature (e.g., Estes, 1988, 1994; Gluck & Bower, 1988a; Kruschke, 1992) and in the engineering literature (e.g., Bridle, 1990; Rumelhart et al., 1995). An added computational benefit beyond the psychological plausibility is that exponentiation of the output activations monotonically transforms possibly negative activations into positive values; this is essential if the transformed values are interpreted as probabilities. Wills, Reimers, Stewart, Suret, and McLaren (2000) pointed out potential problems with this sort of ratio rule, however.

Base Rates

The original ADIT model (Kruschke, 1996a, Equation 11, p. 15) used a separate formula for mixing category base rates with the choice probabilities generated from the associative network. Juslin et al. (1999, 2001) emphasized this use of two apparently different explanatory mechanisms in ADIT. It turns out, however, that this separate formula is essentially equivalent to handling base rates as learned associations from a bias cue. A proof is provided by Kruschke (in press, Appendix 2). The bias cue is fully activated on every trial, and attention shifts to and from it like any other cue. In effect, the bias cue encodes the response prompt that appears on every trial during the experiment. It is possible for the bias cue to have a different salience than the other cues, however (cf. Kruschke & Johansen, 1999). The salience of cues is formalized below. Thus, what was presented in the original ADIT model as a separate principle for mixing base rates with other choice probabilities is actually equivalent to the singular attentional and associative learning system applied to a bias cue.

The weights from the bias cue grow to reflect the base rates of the categories when the same attention is paid to the bias cue on every trial. When attention is rapidly shifted, however, the weights need not perfectly reflect the base rates. In practice, the weights are influenced by the base rates early in training but not later in training, because attention shifts away from the bias. In essence, learning about the bias cue is blocked. Future revisions of the model will need to better address base-rate learning.

Activation Propagation in the Attentional System

The activation of an exemplar node corresponds to the psychological *similarity* of the current stimulus to the exemplar represented by the node. Similarity drops off exponentially with distance in psychological space, as suggested by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). Each exemplar node is significantly activated by only a relatively localized region of input space; that is, it has a small receptive field. Formally, the activation value of exemplar x is given by

$$a_x^{\text{ex}} = \exp(-c \sum_i \sigma_i |\psi_{xi} - a_i^{\text{in}}|), \tag{3}$$

where the superscript *ex* indicates that this is an exemplar node, where c is a constant called the *specificity* that determines the overall narrowness of the receptive field, where σ_i is the *salience* of cue i, and where ψ_{xi} represents the presence or absence of cue i in exemplar x, such that $\psi_{xi} = 1$ if cue i is present in the exemplar and $\psi_{xi} = 0$ if cue i is absent. This is essentially the same exemplar-similarity function used in the attentional learning covering map (ALCOVE) model (Kruschke, 1992) and in the generalized context model (Nosofsky, 1986).

The salience of a feature is assumed to be a nonnegative value that reflects the tonic, stable power of the cue to attract attention. Thus the salience of a cue influences portions of the model that compute the amount of attention allocated to the cue. In the simulations in this article, all the word cues were randomized and selected to be of roughly equal salience; therefore, the salience of these cues is arbitrarily fixed at a value of 1.0. The bias (i.e., response-prompt) cue was allowed a separate value for its salience.

Within the attention module, activation propagates from the input nodes to the gain nodes via two paths: along previously described one-to-one connections from input nodes to gain nodes and via exemplar nodes to gain nodes. The activation of gain node i is given by

$$g_i = a_i^{\text{in}} \sigma_i \exp(\sum_x \text{w}_{ix} \text{a}_x^{\text{ex}}), \tag{4}$$

where w_{ix} is the associative weight from exemplar node x to gain node i. The weights in Equation 4 are initialized at zero but change to new values with learning, as described below. Equation 4 gives zero gain to input cues with zero activation and a gain of σ_i (the cue's underlying salience) to input cues about which nothing has yet been learned. Notice also that the gains on all cues are nonnegative.

From the gain nodes, activation propagates to the attention nodes. The capacity constraint is formalized by requiring the length of the attention vector to be equal to 1, with length measured by a Minkowski power

(*Appendix continues*)

metric. Formally, this is denoted as the constraint that $\Sigma_i\ \alpha_i^P = 1$, where $P > 0$ is the value of the power in the Minkowski metric. Then the attention to the *i*th cue is just the normalized gain of the *i*th cue:

$$\alpha_i = g_i/(\sum_j g_j^P)^{1/P}. \qquad (5)$$

The denominator is certain to be greater than zero because the gains computed from Equation 4 are nonnegative and at least one gain is nonzero by design. Increased attentional capacity is reflected by larger values of P. When $P = 1$, the attention strengths must sum to unity and the attention to any one cue is just the proportion of its gain relative to the total of the other gains: $\alpha_i = g_i/\Sigma_j\ g_j$. In this case, any increase of attention to one cue comes at the cost of the same amount of decrease in attention to other cues. When the capacity P approaches infinity, the attention to each cue approaches the proportion of its gain relative to the maximal gain of any cue: $\alpha_i = g_i/\max_j \{g_j\}$. The cue with maximal gain gets an attentional strength of nearly 1, and other cues get attention proportional to the maximal gain. If several cues are tied for maximal gain, they all get attention of nearly 1. When $0 < P < 1$, any increase in attention to a cue causes more than that amount of decrease to other cues; in this case there is severe competition for attention among cues and there is relatively little attention to any cue unless all cues but one have attention strengths close to zero.

Attention Shifting

After activation is propagated to the category nodes and categorization probabilities are determined, corrective feedback is supplied, just as in human learning experiments. The first response to this corrective feedback is a rapid shift of attention to reduce error. Error is measured as the sum squared deviation between the teacher values and the generated activation values, across the output nodes:

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

The coefficient .5 appears in Equation 6 for convenience in subsequent derivations. This definition of error is typical for models that learn by gradient descent on error (Gluck & Bower, 1988b; Kruschke, 1992; Rumelhart, Hinton, & Williams, 1986), but other definitions of error are possible (Rumelhart et al., 1995).

Gradient descent on error with respect to gains yields

$$\Delta g_I = -\lambda_g \frac{\partial E}{\partial g_I} = \lambda_g \sum_k (t_k - a_k^{\text{out}})(w_{kI}a_I^{\text{in}} - \alpha_I^{P-1}a_k^{\text{out}})/(\sum_j g_j^P)^{1/P}, \qquad (7)$$

where λ_g is a positive constant of proportionality called the shift rate for attention. In this and all subsequent formulas, a lowercase subscript denotes an index that can vary, whereas an uppercase subscript denotes an index that has a fixed value.

Psychologically, attention is hypothesized to shift a large extent on a single trial. This large shift cannot be achieved formally with a single large step in the direction of the gradient because attention is a highly nonlinear function of gain; that is, the gradient changes as the attention changes. Therefore, the change specified by the equation for gain change is iterated 10 times (an arbitrary number) on each trial, so that the nonlinearity of the function can be approximated with 10 relatively small steps. After each small attention change, the activation is repropagated to the category nodes to generate a new error and attention is changed a small amount again, for 10 iterations. (On any one of these iterations, if a gain value is driven to a negative value, it is simply reset to zero before the attention values are computed.) The result of these 10 small steps constitutes the single large shift. The same method was applied in the Rapid Attention SHifting 'N' Learning (RASHNL) model (Kruschke & Johansen, 1999). ADIT (Kruschke, 1996a), the predecessor of EXIT, also shifted attention by gradient descent on error but did not incorporate the more sophisticated attention normalization method of EXIT. Instead, ADIT took one step along the gradient, and if this step happened to yield negative attention values, they were simply truncated.

Learning of Associations

After the attention is shifted, the association weights are adjusted, also by gradient descent on error:

$$\Delta w_{KI} = -\lambda_w \frac{\partial E}{\partial w_{KI}} = \lambda_w(t_K - a_K^{\text{out}})\alpha_I a_I^{\text{in}}, \qquad (8)$$

where λ_w is a constant of proportionality called the *learning rate for output weights.*

The ADIT model (Kruschke, 1996a) had attentional multipliers on the cue nodes, which were shifted by gradient descent on error, but ADIT did not have any way of learning the shifts and retaining them for subsequent trials. EXIT implements a learning mechanism for the shifted attentional distribution. The associative weights for the gain nodes are also adjusted by gradient descent on error, where error is defined as the sum of squared differences between the shifted value and the initial, preshift value. That is, the shifted values act as the teachers for the gain node activations. Formally, this yields

$$\Delta w_{IX}^g = \lambda_x(g_I^{\text{shift}} - g_I^{\text{init}})g_I^{\text{init}}a_X^{\text{ex}}, \qquad (9)$$

where λ_x is the learning rate for the associative weights from exemplar to gain nodes.

List of Free Parameters in EXIT

The free parameters of the EXIT model are the following:

1. The response probability scaling constant, ϕ, used for converting output activation to response probability, in Equation 2.
2. The salience, σ, of the bias (i.e., response prompt) cue, used in Equations 3 and 4. (All other cue saliences were fixed at 1.0.)
3. The specificity, c, of the exemplar nodes in the attention module, in Equation 3.
4. The attention normalization power, P, that is, the attentional capacity, in Equation 5.
5. The attention shift rate, λ_g, in Equation 7.
6. The associative weight learning rate, λ_w, for categorization module, in Equation 8.
7. The learning rate, λ_x, for the associative weights from exemplar nodes to gain nodes, in Equation 9.

Best Fitting Parameter Values For EXIT

The best fit of EXIT to Experiment 1 of Kruschke (1996a) was obtained with parameter values of $c = 2.87$, $P = 2.48$, $\phi = 4.42$, $\lambda_g = 4.42$, $\lambda_w = 0.212$, $\lambda_x = 1.13$, and $\sigma = 0.401$.

The best fit of EXIT to Experiment 2 of Kruschke (1996a) was obtained with parameter values of $c = 1.86$, $P = 2.36$, $\phi = 4.35$, $\lambda_g = 0.070$, $\lambda_w = 0.182$, $\lambda_x = 10.0$ (the highest value searched), and $\sigma = 0.423$.

The best fit of EXIT to Experiment 1 was obtained with parameter values of $c = 0.001$, $P = 1.128$, $\phi = 3.72$, $\lambda_g = 1.746$, $\lambda_w = 0.256$, $\lambda_x = 0.0092$, and $\sigma = 0.568$.

The best fit of EXIT to Experiment 2 was obtained with parameter values of $c = 0.001$, $P = 2.39$, $\phi = 3.936$, $\lambda_g = 0.360$, $\lambda_w = 0.050$, $\lambda_x = 0.017$, and $\sigma = 0.00$.

When fitting all four experiment data sets simultaneously, the best parameter values were $c = 0.204$, $P = 2.00$, $\phi = 3.95$, $\lambda_g = 0.457$, $\lambda_w = 0.159$, $\lambda_x = 0.004$, and $\sigma = 0.336$.

Received July 24, 2000
Revision received March 26, 2001
Accepted April 13, 2001 ■