# Mirroring the Inverse Base-Rate Effect: The Novel Symptom Phenomenon

**Peter Juslin** (Peter.Juslin@psyk.uu.se)
Department of Psychology, Uppsala University
Box 1225, S-751 42, Uppsala, Sweden

**Pia Wennerholm** (Pia.Wennerholm@psyk.uu.se)
Department of Psychology, Uppsala University
Box 1225, S-751 42, Uppsala, Sweden

**Anders Winman** (Anders.Winman@psyk.uu.se)
Department of Psychology, Uppsala University
Box 1225, S-751 42, Uppsala, Sweden

**Abstract**

The *elimination model* is proposed as an account of the *inverse base-rate effect* (D. L. Medin & S. M. Edelson, 1988). A key-assumption is that participants sometimes rely on *eliminative inference* to decide among candidate categories. A new prediction is that there will be an inverse base-rate effect also for an entirely novel symptom presented in the transfer phase—a prediction that contrasts with that by *ADIT* (J. K. Kruschke, 1996). This was tested and confirmed in 2 experiments.

## Introduction[1]

In 1988, Medin and Edelson reported an interesting but complex pattern of findings regarding how people utilize base-rates. In their experiments, participants were asked to decide whether patients with ambiguous symptom patterns were suffering from previously learned common or rare diseases. Surprisingly, in some cases participants chose the less frequent of the diseases. A number of explanations of this base-rate inverse (BRI) effect have been proposed (Kruschke, 1996; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992).

In this paper, we propose a further mechanism that may contribute to both the BRI effect and the unspecified guessing strategy reported by Kruschke (1996). Basically, the elimination model suggests that the participants eliminate options that are inconsistent with well-supported inference rules, leading to the prediction of an intricate pattern of responses in which participants sometimes favor the common diseases, and sometimes the rare ones. A presentation of the details, and the fit of a quantitative implementation of the elimination model, is provided in Juslin, Wennerholm, and Winman (1999). In this paper we will focus on one prediction by the elimination model that goes beyond what previous models can predict or account for, the prediction of a novel symptom phenomenon.

[1]The research reported here was supported by the Swedish council for Research in the Humanities and Social Sciences.

## The Experimental Paradigm

The basic task introduced by Medin and Edelson (1988) involves a training- and a transfer phase. On each training trial a pair of symptoms is presented, and participants are requested to choose which of six fictitious diseases the hypothetical patient is suffering from. After each choice the participant is informed about the proper diagnosis (disease), after which another training trial is presented. The critical manipulation concerns the base-rate of each disease, with the common diseases occurring three times more often than the remaining rare ones (see Table 1).

Table 1: The basic design of the training phase in the Medin and Edelson Experiment 1 (1988).

| Base-rate | Symptoms | Disease | Inference rule |
|---|---|---|---|
| 3 | $I_1+PC_1$ | $C_1$ | $I_1+PC_1 \rightarrow C_1$ |
| 1 | $I_1+PR_1$ | $R_1$ | $I_1+PR_1 \rightarrow R_1$ |
| 3 | $I_2+PC_2$ | $C_2$ | $I_2+PC_2 \rightarrow C_2$ |
| 1 | $I_2+PR_2$ | $R_2$ | $I_2+PR_2 \rightarrow R_2$ |
| 3 | $I_3+PC_3$ | $C_3$ | $I_3+PC_3 \rightarrow C_3$ |
| 1 | $I_3+PR_3$ | $R_3$ | $I_3+PR_3 \rightarrow R_3$ |

During training every instance of a common disease, C, occurs in the presence of two symptoms: One imperfect, I, and one perfect, PC. Similarly, every instance of a rare disease, R, has two symptoms: One imperfect, I, and one perfect, PR. Thus, each imperfect predictor is associated with both a common and a rare disease, and each perfect predictor is uniquely associated with only one disease.

In a succeeding transfer phase participants are tested with previously uncombined symptoms. Medin and Edelson (1988) found that when tested with the imperfect symptom, I, the majority of participants chose the common disease.

When tested with the ambiguous combination, I+PC+PR (the combined probe), the participants again tended to choose the common disease. However, when tested with two perfect predictors, PC+PR (the conflicting probe), the majority of participants chose the rare disease in contrast to the base-rate—the inverse base-rate effect (Figure 2C below).

### Accounts of The Inverse Base-rate Effect

Most previous accounts of the BRI effect revolve around a common theme: Because of cue-competition, symptom PR becomes more strongly associated with disease R than symptom PC does with disease C (Gluck & Bower, 1988; Kruschke, 1996; Shanks, 1992).

Kruschke (1996) suggested that ADIT can explain both the inverse base-rate effect and apparent base-rate neglect (Gluck & Bower, 1988). By the application of two separate mechanisms: (a) A base-rate bias, that participants apply consistently on all training trials, and (b) an attention-shifting mechanism that rapidly shifts attention from typical to distinctive features, ADIT provided a good fit to the transfer data. Specifically, because the common disease C, occurs more often than the rare one, R, participants first learn to associate both the imperfect symptom I, and the perfect symptom PC with the common disease. Later in training when they are presented with the symptoms that are associated with the rare disease, R, they focus on the symptom that is perfectly predictive of that disease, PR, and thereby encode it by this single symptom. This explains why participants choose the rare disease on the PC+PR (conflicting) test case. When confronted with the remaining two ambiguous test cases, I and I+PC+PR, people apply both their base-rate knowledge and their associative knowledge, where the base-rate knowledge dominates the responses.

Although ADIT provides a good quantitative fit to transfer data, Kruschke (1996) reported that his participants responded better-than-chance for the rare categories—an effect he attributed to an unspecified non-random guessing strategy (Kruschke, 1996). Likewise, when ADIT was fitted to the training data it performed much worse than human learners on early training trials. Thus, although appealing, ADIT fails to fully account for the complete pattern of data observed with the Medin and Edelson (1988) design.

## The Elimination Model

To illustrate the inferential mechanisms of the elimination model, consider the following example: You are told that a friend of yours has bought a pet animal called George who is either a goldfish or a Psittaciformes. Not being a zoologist, you have a pretty good idea of what a goldfish is, but you have no notion whatsoever of what a Psittaciformes is. Your task is to guess what kind of pet animal George is. First, you receive the cue George lives in water. George is thus similar to a goldfish in the sense that he lives in water. In the absence of knowledge about what a Psittaciformes is you might be tempted to guess that George is a goldfish. This illustrates one (weak) form of induction.

Now consider the situation where you instead are given the cue George can fly. In this case you would probably guess that George is a Psittaciformes—he is certainly not a goldfish (in fact a Psittaciformes is a parrot). You would use your knowledge about the category goldfish to eliminate the possibility that George is a goldfish. The elimination model takes this latter kind of inference into account.

A reasonable assumption in the Medin and Edelson (1988) design is that the participants perceive the task as involving a set of perfectly valid inference rules (see Nosofsky, Palmeri, & McKinley, 1994, for similar approaches). If the participants succeed to learn these rules, they will make 100 percent correct classifications at the end of training (see Table 1).

In the quantitative implementation in Juslin et al. (1999), we assume that at each trial the inference rule appropriate for the presented training probe is formed with a rule-activation probability. This probability, that is higher for the early training trials (implementing "freezing" at the initial stages of learning, cf. Medin & Bettger, 1991), is controlled by a single parameter. From these rule-activation probabilities, we can compute the probability c that the rule appropriate for a common disease is active and accessible at the transfer phase, and the corresponding probability r that the rule appropriate for a rare disease is accessible.

In the training phase, every probe precisely matches one of the six inference rules (see Table 1). In the transfer phase, however, the participants' inferences will have to be based on the similarity between the new symptom combinations and the conditions of the inference rules. The elimination model consists of two decision mechanisms that determine how an inference rule is applied to a probe:

(1). The induction mechanism applies when the probe has exactly the symptoms in the condition-part of the inference rule, or when the symptoms of the probe are perceived to be sufficiently similar to the rule conditions. Whenever similarity is larger than a similarity criterion the induction mechanism applies, and the probe is assigned to the category with the most similar rule. If the probe is equally similar to several rules, the participants will decide randomly among the set of equally similar rules.

(2). When a probe is dissimilar to the rule-conditions, as indicated by a similarity smaller than the similarity criterion, the elimination mechanism is used to eliminate the possibility that the probe belongs to the category and the probe is assigned randomly to any category but the dissimilar one. For example, if there is no basis for induction and the probe eliminates one or several of the categories, the participant will have to decide randomly among the still admissible categories—the diseases that are not inconsistent with the symptoms of the probe.

When the elimination model is applied to the Medin and Edelson design, we need to impose a similarity structure on the probes presented in the transfer phase. There are two crucial assumptions: (a) The conflicting probe, PC+PR, is less similar to the inference rules formed in the training phase for C and R than the combined probe, I+PC+PR and the imperfect probe, I. While the combined probe is ambiguous in the sense of being consistent with two inference rules, the conflicting probe actually contradicts both. (b) The similarity criterion for induction versus elimination is located between the similarities of the

combined and the conflicting probes implying that the combined probe elicits induction and the conflicting probe elimination. The example of such a similarity structure, derived from the multiplicative similarity rule of the original context model (Medin & Schaffer, 1978), is provided in Juslin et al. (1999).

If we refer to the common-rare disease-pairs relevant to a particular probe (e.g., $C_1$ and $R_1$ in Table 1) as the focal disease-pair, a participant may be in one of four knowledge states when entering the transfer phase. State 1: With probability (1-c) (1-r) neither the inference rule for the focal common $C_1$ nor the focal rare disease $R_1$ is accessible. State 2: With probability (c-cr) only the inference rule for the focal common disease $C_1$ is accessible. State 3: With probability (r-cr) only the inference rule for the focal rare disease $R_1$ is accessible. State 4: With probability (cr) both of the focal inference rules, $C_1$ and $R_1$, are accessible. Table 2 shows an example of Knowledge State 1, the state in which neither of the rules for the focal diseases are accessible.

The predicted response patterns may be exemplified by reference to the combined and the conflicting probes. For example, imagine that you are presented with the symptom combination $I_1+PC_1+PR_1$. These three symptoms are consistent with two of the six inference rules, the first and the second in the right-most column of Table 1. If you are in Knowledge state 2 you will only know the rule $I_1+PC_1 \rightarrow C_1$ which is executed by the induction mechanism. If you are in Knowledge state 3 you will only know the rule $I_1+PR_1 \rightarrow R_1$ which is executed by the induction mechanism. Because of the base-rate manipulation the probability of Knowledge state 2 is higher and most responses will thus favor $C_1$. Knowledge states 1 and 4 are assumed to elicit random decisions favoring neither common nor rare categories.

Now you are presented with the conflicting probe, $PC_1+PR_1$ that is dissimilar to both the first and the second inference rule.

Table 2: An example of a hypothetical Knowledge State.

| Ratio | Symptoms | Disease | Rule |
|---|---|---|---|
| 3 | *$I_1$: Stomach pain*<br>*$PC_1$: Loss of hair* | *$C_1$: Coralgia* | Unknown |
| 1 | *$I_1$: Stomach pain*<br>*$PR_1$: Impaired hearing* | *$R_1$: Buragamo* | Unknown |
| 3 | *$I_2$: Epidermophytosis*<br>*$PC_2$: Back pain* | *$C_2$: Midosis* | Known or Unknown |
| 1 | *$I_2$: Epidermophytosis*<br>*$PR_2$: Loosening of the teeth* | *$R_2$: Namitis* | Known or Unknown |
| 3 | *$I_3$:Visual defect*<br>*$PC_3$: Impaired short-term memory* | *$C_3$: Terrigitis* | Known or Unknown |
| 1 | *$I_3$: Visual defect*<br>*$PR_3$: Swollen arms* | *$R_3$: Althrax* | Known or Unknown |

If you are in Knowledge state 2 you will eliminate the rule $I_1+PC_1 \rightarrow C_1$; If you are in Knowledge state 3 you will eliminate the rule $I_1+Pr_1 \rightarrow R_1$. Again, because of the base-rate manipulation Knowledge state 2 is more probable, and most eliminations will concern $C_1$ and thus favor the choice of a rare disease category. Given the particular knowledge-state and the similarity between the inference rules formed and the presented transfer probe, the decision mechanisms of induction and elimination can be applied in the manner specified above. Although this is straightforward in principle, application to the Medin and Edelson design is complicated by the fact that the number of unknown diseases is a random variable.

This can be illustrated by reference to the example in Table 2. You may be faced with the conflicting probe *loss of hair + impaired hearing*. Because you are in Knowledge state 1 and know neither of the focal rules, there is no possibility for an inductive inference with the focal rules. The choice of an unknown category in this case amounts to an elimination of any of the four non-focal diseases that possibly are known. The number of unknown diseases is a random variable controlled by the same parameter that defines c and r; that is, one to four of the non-focal rules may have been activated the training. Thus, in knowledge State 1, the guessing rate of responses in the focal common category ($C_1$) is anything between 1/6 (no disease has been learned) to 1/2 (all the four non-focal diseases have been learned).

The predicted response proportions for each probe therefore equal the proportions of inductive inferences that fall in this category plus the expected value of the guessing rates when the participant eliminates, where the probability of induction and elimination is jointly determined by the knowledge-state and the similarity between the probe and the known inference rules. Since both the probabilities of the knowledge states and the expected guessing-rates are controlled by the parameter that defines the rule-activation probability, predicted response proportions are controlled by a single parameter. These computations are detailed in Juslin et al. (1999).

In Juslin et al. (1999) the quantitative predictions were fitted to the data from Experiment 1 by Medin and Edelson (1988), to Kruschke (1996, Exp. 1), and to the data from the two experiments reported below. In all of these data sets, the model reproduced the observed pattern of base-rate findings, in general with an impressive quantitative fit given the use of one single free parameter. Figure 1 illustrates the fit of the model to the data from Kruschke' s (1996) Experiment 1. Although factors such as cue competition probably play an important role in the Medin and Edelson paradigm, the quantitative predictions presented above demonstrate that the elimination mechanism alone has the potential to reproduce the BRI effect.

## The Novel Symptom Phenomenon

A straightforward prediction by the model is that the presentation of a novel symptom in the transfer phase will lead to a preponderance of *rare-disease* responses, mirroring the BRI effect for the conflicting symptoms. Participants will notice that the novel symptom is dissimilar to the symptoms of known categories and as a result they will guess on some of the unknown diseases.

By virtue of the base-rate manipulation these *unknown* categories are likely to be rare rather than common ones.

This prediction is important for two reasons: First, it is the most critical test of the presence of eliminative inferences. To the extent that this type of inferences underlies both the responses for the conflicting and novel transfer probes, the response patterns observed for these probes should be similar. Second, this prediction amounts to a response pattern contrary to that by ADIT (Kruschke, 1996). The novel transfer probe has not been affected by any shift of attention during training, thus the only factor at work is the base-rate bias toward common-disease responses. As we have seen, the elimination model predicts rare-disease responses that mirror those observed for the conflicting test probe. Next, we present results from two experiments that confirm this prediction.

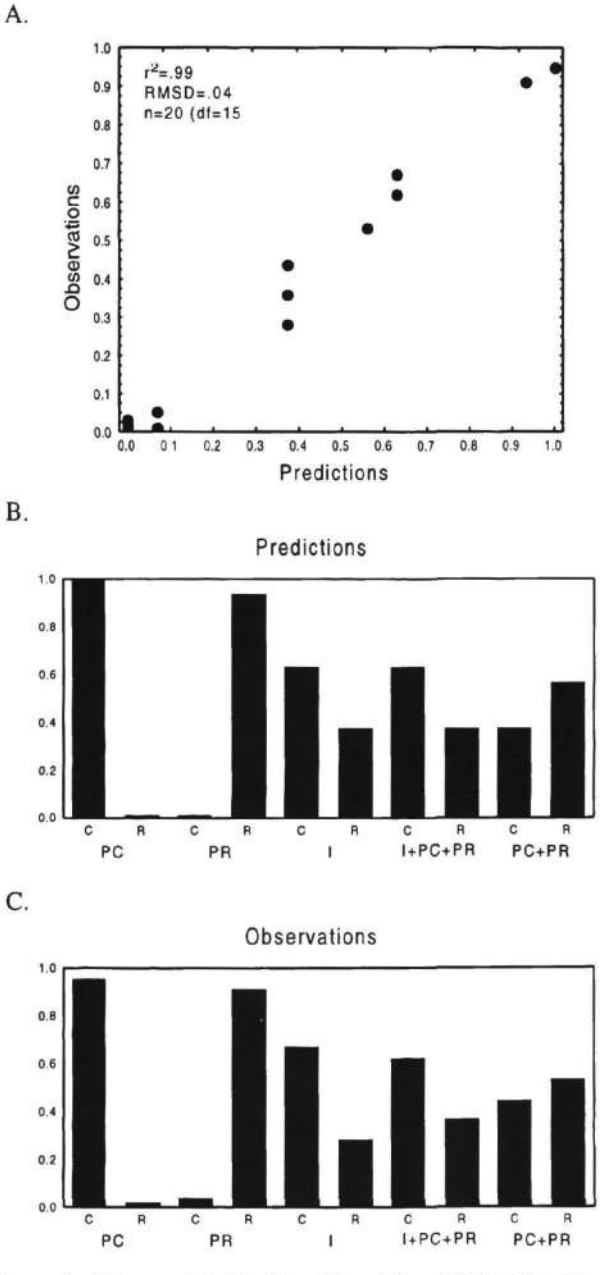


Figure 1: The model fitted to Kruschke (1996, Exp.1). A) Predicted/Observed response proportions. B) The pattern of data predicted for the transfer probes, and C) the corresponding observed pattern.

## Experiment 1: A Test of the Novel Symptom Phenomenon

The main purpose of Experiment 1 was to test the prediction of a BRI effect for novel symptoms. One hundred and nine participants were divided into a 3:1 base-rate ratio group and a 7:1 base-rate ratio group (cf. Shanks, 1992). The material, stimuli and procedure were more or less identical to those used by previous researchers (see e.g., Medin & Edelson, 1988; Kruschke, 1996; Shanks, 1992). Participants were told that they would be allowed to practice on 168 patients with feedback informing them if they had made a correct or incorrect diagnosis. They were instructed to apply the knowledge they had acquired during the training phase to 24 patients in a transfer phase with no feedback. For each participant nine of the twelve symptoms were randomly selected and matched with the six different diseases (see Tables 1 and 2). The three remaining "novel" symptoms were used in the transfer phase in order to test the novel symptom phenomenon.

On the transfer trials participants were required to make responses to six perfect predictors, PC and PR, three imperfect predictors, I, three combined probes, I+PC+PR, three conflicting probes, PC+PR, and three novel probes, N. The remaining six test trials in which the imperfect predictors were paired with perfect predictors were mainly used to disguise the purpose of the transfer phase. The transfer trials were followed by a short break after which another training - and transfer phase followed. The purpose of this manipulation was to see whether additional training would decrease the BRI effect. The experiment including the break lasted approximately one hour.

### Training Results

As in most previous studies, only those participants who had reached asymptotic learning were used in the subsequent analysis. Participants with more than one incorrect answer in the last 24 trials of the first training session were excluded. In the 3:1 ratio group, 40 participants out of 55 (73%) met this criterion. In the 7:1 ratio group, 41 out of 54 (76%) participants met the criterion. Whereas the 7:1 group showed evidence of slightly faster learning, both groups converged on asymptotic learning at the end of the first session and these levels of performance were maintained throughout the second training session.

### Transfer Results for the 3:1 Group

For the perfect predictors, PC and PR, in the 3:1 condition, most responses were assigned to the disease that the cue had been a perfect predictor of, both after training session 1 and 2 (all $t(39) = 15$ with $p < .05$, given a null-hypothesis of .5). The use of base-rate information is evident for the imperfect transfer probes, I, both after training sessions 1 and 2 (.67, $t(39) = 5.77$, $p <$ .05, and .66, $t(39) = 4.03$, $p < .05$, respectively), and for the combined transfer probe, I+PC+PR, both after training sessions 1 and 2 (.62, $t(39) = 3.60$, $p < .05$, and .62, $t(39) = 2.07$, $p < .05$, respectively).

The results show a BRI effect for the conflicting transfer probe, PC+PR, although the trend is non-significant. After session 1, the proportion of common category responses, C, was .410, t(39) = 1.42, N. S. After training session 2, the proportion increased to .483, t(39) = .24, N. S. Finally, the results for novel symptoms, N, mirror the result for the conflicting probe, PC+PR (see Figure 2A). Although slightly less pronounced, the participants favor the rare diseases (response proportion .45, t(39) = 1.12, N. S.). After training session 2, the participants have altered into base-rate use (response proportion .58, t(39) = 1.74, N. S.).

### Transfer Results for the 7:1 Group

Results for the 7:1 group parallel those for the 3:1 group, although the effects were larger, and in contrast to the 3:1 condition there was a significant BRI effect for the conflicting transfer probe (see Figure 2B, session 1: proportion .38, t(40) = 2.10, p< .05: session 2; proportion .40, t(40) = 1.65, N. S.). As described above, the elimination model is supported if the BRI effect on the conflicting and novel probes is similar Figure 2 presents the mean response proportions for the novel and conflicting transfer probes in the 3:1 and 7:1 base-rate ratio conditions of Experiment 1, and of the (single) 3:1 condition of Experiment 2

As can be seen, the response patterns are similar for novel and conflicting probes in both conditions, a result predicted by the elimination model but contrary to ADIT (Kruschke, 1996).

## Experiment 2: Does the Inverse Base-Rate Effect Disappear with Additional Training?

Experiment 1 replicated the base-rate effects of the Medin and Edelson (1988) study. Interestingly, however, the BRI effect was diminished after training session 2, and it vanished altogether for the novel transfer probes. This could however be due to the testing between the first and second training phase. When the novel transfer probe has been presented in the transfer phase after training session 1, it will obviously not be "novel" when the retested in the second transfer phase. As a result Experiment 2 was designed to test whether the diminished BRI effect was due to the repeated transfer exposure. Another potential explanation is that the BRI effect disappears after extensive training. Medin and Bettger (1991), for example, discussed the possibility that "with enough experience the BRI effect that we have attributed to competitive learning may be overcome altogether" (p. 328). Thus, the alternative hypothesis was that the diminished effect was due to the higher extent of learning. Experiment 2 involved a prolonged training phase without transfer phases interspersed midways through the training trials. If the BRI effect would persist even with this prolonged training, it would suggest that the diminished BRI effect after training session 2 in Experiment 1 was a consequence of the repeated exposures. On the other hand, if the correct explanation lies in the increased learning the BRI effect should be gone after four times as many trials. Twenty-five students participated in Experiment 2. Procedures were identical to Experiment 1, with the differences that (a) the number of training trials was 672 (168 × 4), and that (b) the base-rate ratio was 3:1 for all participants. Between the first and second half of learning phase the participants were given a one-hour lunch break. At the end of training session 2 they were presented with the transfer phase (identical to the one in Experiment 1).

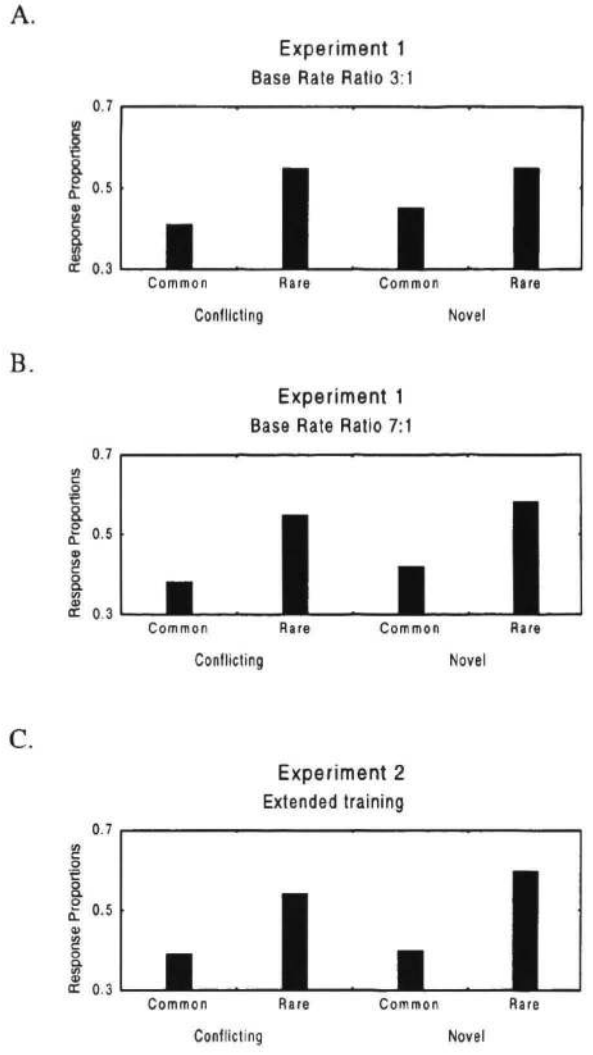


Figure 2: Mean response proportions for the conflicting and novel transfer probes in the 3:1 (Panel A) and the 7:1 (Panel B) base-rate ratio condition of Experiment 1, and in Experiment 2 (Panel C).

### Results

As in Experiment 1, the learning criterion was set at 96% correct responses in the last training block of 24 trials. Again, Experiment 2 replicated the standard pattern found by Medin and Edelson: The common disease was chosen more often than the rare one for the imperfect (common response proportion .66, t(22) = 3.6, p<.05) and the combined probes (common response proportion .46, t(22) = .38, N. S), but the rare disease was chosen for the conflicting probe (common response proportion .39, t(22) =1.1, p<.28, N. S) - the BRI effect. Although not significant, the effect is not diminished in size after ample learning (see Figure 2C). Finally, a BRI effect for the novel symptoms was observed too (common other response proportion .40, t(22) =1.59, p<.125, N. S). We found it hard to obtain statistical

significance due to the very high measurement error. Therefore, it should be noted that if all data in Figure 2 are collapsed, there is no doubt a reliable BRI effect ($t(103) = 2.7$, $p < .01$) as well as a reliable novel symptom phenomenon ($t(103) = 2.6$, $p < .01$).)

In sum, it seems that the diminished BRI effect after training session 2 reported in Experiment 1 is due to the repeated exposures with the transfer probes rather than to more extensive training. Similar to Experiment 1, participants guessed on a rare disease category when presented with a novel symptom, as predicted by the elimination model but in contrast to ADIT (Kruschke, 1996).

## General Discussion

In this paper, a new mechanism has been proposed that may contribute to the base-rate effects observed with the Medin and Edelson design. The main merits of the elimination model are threefold: First, the model has intuitive appeal in the sense that it seems hard to deny that people at least sometimes rely on eliminative inferences. Second, in terms of the psychological mechanisms involved, the model is simple: The participants either make inductive or eliminative inferences depending on the similarity of the probe to the known diseases. Finally, the model provides a good quantitative account of the data given the reliance on one single free parameter (Juslin et al., 1999). ADIT is unable to account for the novel symptom phenomenon, and Kruschke (1996, p. 20) noted that ADIT needs "additional mechanisms not implemented in the model" to account for the non-random guessing strategy observed in the early training trials. The elimination model provides such a mechanism.

Nevertheless, we do not suggest that the mechanism proposed by the elimination model is the only factor that contributes to the BRI effect. The ideas of cue-competition have immense support in the literature on animal learning and Kruschke's (1996) proposal of a rapid attention-shifting mechanism is both reasonable and appealing. We do, however, take the confirmation of the novel symptom phenomenon as fairly strong evidence that eliminative inferences are at work at least to some extent. The quantitative formulation of the elimination model in Juslin et al. (1999) serves to demonstrate that these processes alone have the potential to produce an BRI effect. The relative importance of explanations in terms of attention-shifting mechanisms and eliminative inferences needs to be determined by future research.

## References


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

Juslin, P., Wennerholm, P., & Winman A. (1999). *The elimination model: The inverse base-rate effect as a result of eliminative inferences*. Manuscript submitted for publication. Department of Psychology, Uppsala University, Sweden.

Kruschke, J. K. (1996). Base-rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 1*, 3-26.

Medin D. L., & Bettger, L.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 model of classification learning. *Psychological Review, 85*, 207-238.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

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