

## **A Feature-Salience Analogue of the Inverse Base-rate Effect**

Corey J. Bohil

*Perceptive Sciences Corporation, USA*

Arthur B. Markman and W. Todd Maddox

*University of Texas at Austin, USA*

Classification learning requires integrating many properties of the items being learned including the base-rate probability that a category will occur as well as the salience of features. Previous research has demonstrated an inverse base-rate effect, in which people classify an item that has features predictive of both a high base-rate and low base-rate category into the rarer category. We suggest that this finding reflects that feature salience plays a greater role in classification than does base-rates. We tested this hypothesis by demonstrating that manipulations of feature salience determine the classification of ambiguous stimuli regardless of the underlying base-rates of the categories.

A host of studies have demonstrated that human classification performance generally conforms to base-rate (i.e., relative frequency) ratios when learned from direct experience (i.e., trial-by-trial exposure to base-rates rather than summary statistics – Christensen-Szalanski & Beach, 1982; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Lindeman, Van Den Brink, & Hoogstraten, 1988; Bohil & Maddox, 2001; Manis, Dovalina, Avis, & Cardoze, 1980). For example, when one of two categories is presented more frequently than the other during training, transfer performance usually reflects a tendency to favor the high frequency or common category response over the low-frequency or rare category response. It is important to understand people's sensitivity to base-rates, because they are important in many decision tasks (Cosmides & Tooby, 1996; Koehler, 1996). In particular, base-rate information is useful when stimulus features are of little or no help in classifying an item (e.g., in medical diagnosis, flu-like symptoms could suggest various ailments, with likelihoods shaped by time of year, patient age and gender, recent outbreaks, etc.).

A striking failure to use base-rate information occurs in the inverse base-rate effect (IBRE). This effect occurs under a special set of conditions (which we describe in the next section) in which participants gain direct exposure to base-rates while learning to

---

Correspondence concerning this article should be addressed to W. Todd Maddox, Institute for Neuroscience, University of Texas, 1 University Station A8000, Austin, Texas 78712, USA. E-mail: maddox@psyvax.psy.utexas.edu.

The authors would like to thank Brian Ross for many helpful discussions about the contents of this article. This research was supported in part by a Postdoctoral Training Grant from the National Institute of Mental Health, National Research Service Award # MH 14257, to the University of Illinois, and in part by Grant # 5 R01 MH59196 from the National Institute of Mental Health, National Institutes of Health to W. Todd Maddox.

classify a set of stimuli, but tend to treat stimuli with one feature of a common category and one feature of a rare category as members of the rare class. The most prominent account of the IBRE suggests that the effect is an expected outgrowth of the learning scenario (Kruschke 1996, 2001, 2003). This theoretical account leads to an interesting prediction, the confirmation of which is the main objective of the current research.

In the next section, we describe the conditions leading to the IBRE, along with the details of Kruschke's (1996) account and the prediction being tested. The results of the experiment lead to some intriguing follow-up questions about the underlying cause of the IBRE. We consider these questions in the Discussion section.

**The Inverse Base-rate Effect** Medin and Edelson (1988) first described the IBRE in an article exploring the influence of base-rates on classification judgments. Medin & Edelson's study utilized a simulated medical diagnosis task in which participants classified symptom combinations that signaled various hypothetical diseases (see also Kruschke, 1996). Because the current research also uses this scenario, the experimental design is described in terms of a medical diagnosis task. However, the design and results are not confined to this cover story (see Dennis & Kruschke, 1998; Fagot & Kruschke, 1998; Kalish, 2001).

During training, each stimulus is composed of a pair of symptoms (i.e., predictors) and the participant's task is to provide a diagnosis for the presented pair, with each response followed by corrective feedback. As shown in Table 1, there are two general predictor types – perfect and imperfect. Each perfect predictor is associated with a single disease; each imperfect predictor is associated with two diseases. The imperfect predictor (I) indicates that either of two diseases is possible. The perfect predictor unambiguously signals the presence of a single disease, thus alleviating the indeterminacy of the imperfect predictor. Finally, a critical aspect of this study is that the diseases differ in base rate. Typically one disease (the common disease) occurs three times more often than the other (the rare disease). In this paper, we call a perfect predictor associated with a common disease a perfect common (PC) predictor, and a perfect predictor associated with a rare disease we call a perfect rare (PR) predictor.

A concrete example helps to clarify the design. Suppose that “headache” and “nausea” are symptoms of common Disease C, while “headache” and “fever” are symptoms of rare Disease R. In this example, headache is an imperfect predictor that indicates the presence of either Disease C or R. Nausea is a PC predictor signifying (with certainty) that the common Disease C is present. Fever is a PR predictor signifying (with certainty) that the rare Disease R is present.

During the test phase, participants are shown new combinations of the features. When presented with an ambiguous stimulus such as predictor I alone, or the combination I+PC+PR, participants generally respond that the common disease is present. In contrast, when the equally ambiguous PC+PR combination is presented, people tend to respond with the rare disease. This pattern of responses has come to be known as the inverse base-rate effect.

**Explaining the Inverse Base-rate Effect** Kruschke (1996) provided a compelling account of the IBRE by positing differential learning characteristics for common and rare training combinations. For the Common Stimuli, both the imperfect and perfect predictor become associated with the common category. For example, both headache (I) and nausea (PC) are learned as good predictors of Disease C. Because the imper-

fect predictor becomes associated with the common disease, only the perfect rare predictor gets a strong association to the rare disease. This effect is particularly pronounced early in training. For example, when headache (I) and fever (PR) are presented, the participant may make an error based on the headache symptom (i.e., respond that common Disease C is present) due to the already-formed association between symptom I and Disease C. When the feedback reveals the error, attention shifts rapidly to the distinguishing feature of the stimulus – to the fever (PR) symptom. Shifting attention in this way serves to reduce the error rate over trials. The eventual association between fever (PR) and Disease R is stronger than the association between nausea (PC) and Disease C. This is because Disease C is strongly predicted by symptom I as well as by PC, but the total capacity for predictor-association is limited (and must be divided among the two strong predictors – Kruschke, 1996). When fever and nausea are presented together during the test phase (i.e., PC+PR), the stronger association between symptom PR and Disease R governs the response.

Some of the notions involved in Kruschke's (1996) explanation are embodied in findings from the animal learning literature, and have been suggested by previous researchers beginning with Medin and Edelson (1988) to account for the IBRE (see also Shanks, 1992; Gluck & Bower, 1988; Markman, 1989; Binder & Estes, 1966, Rescorla & Wagner, 1972). However, Kruschke was the first to successfully model the IBRE. His connectionist model, called ADIT (for Attention to Distinctive Input), specifically incorporated the rapid attention shifting mechanism required to develop association weights capable of mimicking the full range of results from inverse base-rate studies (Kruschke, 1996, 2001). (A more recent version of the model is discussed below.)

**The Current Research** The current research extends the inverse base-rate paradigm. An implication of Kruschke's (1996) explanation is that appropriate attention shifts during learning – *attributable to base-rates or otherwise* – can produce the predictor-to-class associations that account for the IBRE.<sup>1</sup> Features that evoke rapid attention shifts during learning (for whatever reason) should eventually dominate classification performance. The research reported here accepts Kruschke's attention shifting explanation, and follows-up on this line of reasoning to examine what makes attention shift in the first place. We wanted to test the prediction that learning-effects similar to the IBRE can be created using the IBRE design but with a different attentional manipulations. More specifically, our study examines the effect of feature salience on classification within the IBRE design.

Operational definitions of salience center on the ability of an item to dominate attention (for example, see Bock & Clifton, 2000; Gibson & Yuhong, 2001; Yantis & Egeth, 1999), but there are numerous ways in which salience can be manifested. It can be endogenous to the features themselves (e.g., intensity attributes like loudness or brightness may influence attention during learning). In medical diagnosis, symptom severity could determine salience. Salience could also be rooted in some functional attribute that pertains to a goal, or it could be a function of surprise; both seem to be exogenous to the feature itself. The study reported here demonstrates the strong

<sup>1</sup> Order of training-stimulus presentation has been shown to be a critical factor in generating the IBRE, rather than unequal base-rates per se (Kruschke, 1996, Experiment 2). In the context of the learning task, however, manipulating the order of presentation is essentially equivalent – from the perspective of the learner – to having unequal base-rates.

effect that salient stimulus features can have over classification learning and subsequent performance.

As will be discussed below, the results of the experiment suggest an interesting possibility for the IBRE that remains to be tested directly – attention shifting may be due to greater salience of rare features over common features. In the ADIT model, error reduction was the mechanism for learning the IBRE trends. Attention was shifted in the model in order to achieve this goal. More recent models developed by Kruschke incorporate parameters specifically intended to account for salience effects (Kruschke & Johansen, 1999; Kruschke, 2001). Although salience effects have been observed and accounted for in other experimental paradigms, we are not aware of any previous attempts to manipulate salience in order to mimic the IBRE trends. It is interesting to note that the ADIT model didn't require a salience parameter to account for the IBRE, even though its successor (called EXIT, Kruschke, 2001, 2003) includes such a parameter. In the Discussion section, we consider the possibility of empirically distinguishing between these two mechanisms.

## EXPERIMENT

Three experimental conditions were designed to assess the influence of feature salience on classification learning. Table 1 displays the abstract design for each condition. The 3:1 BR, Non-salience condition employed the standard inverse base-rate design and replicated Kruschke's (1996) Experiment 1. There were four diseases composed of six symptoms, a 3:1 base-rate (BR) ratio, and no feature salience (Non-salience) manipulation. The same design was used in the 3:1 BR, Salience condition, except that the PC predictors were associated with more serious symptoms than were the other predictors. This condition tested whether salient PC features could weaken or reverse the IBRE by generating stronger associations between PC and Disease C than between PR and Disease R. Finally, in the 1:1 BR, Salience condition, the disease base-rates were unbiased. Each of the training diseases was presented with equal frequency. The feature salience manipulation was applied to two of the perfect predictors (i.e., two perfect predictors were associated with serious symptoms). For the 1:1 BR, Salience condition, the salient features are arranged in Table 1 so that they correspond to the PR predictors in the 3:1 BR, Non-salience condition. This was done to facilitate comparison of the results for these two conditions in the Results section (see Table 2). Due to the increased salience of the two perfect predictors, the pattern of results in the 1:1 BR, Salience condition should mirror those from the 3:1 BR, Non-salience condition. The 1:1 BR, Salience condition tested whether feature salience can influence responding via the mechanism of rapid attention shifting during learning and lead to the same response pattern as base-rates.

## METHOD

**Participants** Thirty University of Texas undergraduate students participated in the 3:1 BR, Non-Salience condition in exchange for course credit. The remaining participants were students and staff recruited from the University of Texas community who were paid for their time. There were twenty-seven participants in the 3:1 BR, Salience condition and twenty participants in the 1:1 BR, Salience condition.

**Table 1**  
*Abstract Design of Experimental Conditions*

Disease	Training Trials		
	3:1 BR, Non-Salience (Frequency) Symptoms	3:1 BR, Salience (Frequency) Symptoms	1:1 BR, Salience (Frequency) Symptoms
1 (C <sub>1</sub> )	(3) I <sub>1</sub> + PC <sub>1</sub>	(3) I <sub>1</sub> + PC <sub>1SAL</sub>	(2) I <sub>1</sub> + PE <sub>1</sub>
2 (R <sub>1</sub> )	(1) I <sub>1</sub> + PR <sub>1</sub>	(1) I <sub>1</sub> + PR <sub>1</sub>	(2) I <sub>1</sub> + PE <sub>1SAL</sub>
3 (C <sub>2</sub> )	(3) I <sub>2</sub> + PC <sub>2</sub>	(3) I <sub>2</sub> + PC <sub>2SAL</sub>	(2) I <sub>2</sub> + PE <sub>2</sub>
4 (R <sub>2</sub> )	(1) I <sub>2</sub> + PR <sub>2</sub>	(1) I <sub>2</sub> + PR <sub>2</sub>	(2) I <sub>2</sub> + PE <sub>2SAL</sub>
Transfer Trials			
	3:1 BR, Non-Salience	3:1 BR, Salience	1:1 BR, Salience
Single	I, PC, PR	I, PC <sub>SAL</sub> , PR	I, PE, PE <sub>SAL</sub>
Conflicting	PC + PR (or PR <sub>0</sub> )	PC <sub>SAL</sub> + PR (or PR <sub>0</sub> )	PE + PE <sub>SAL</sub> (or PE <sub>0SAL</sub> )
Combined	I + PC + PR (or PR <sub>0</sub> )	I + PC <sub>SAL</sub> + PR (or PR <sub>0</sub> )	I + PE + PE <sub>SAL</sub> (or PE <sub>0SAL</sub> )
Other	I + PC <sub>0</sub> , I + PR <sub>0</sub>	I + PC <sub>0SAL</sub> , I + PR <sub>0</sub>	I + PE <sub>0</sub> , I + PE <sub>0SAL</sub>

**Note:** C = Common Disease, R = Rare Disease, I = Imperfect predictor, PC = Perfect Common predictor, PR = Perfect Rare predictor, PE = Perfect Equal-base-rate predictor. Subscript Sal denotes salient feature, subscript 1, 2, or 0 denotes grouping, and (Frequency) shows the number of presentations in each block of 8 training trials (see text for details).

**Stimuli** As Table 1 shows, the design for all three conditions was identical to that used by Kruschke (1996, Experiment 1), with the exception of the base-rates used in the 1:1 BR, Salience condition. In each condition, six symptom labels were associated with four diseases. Table 1 also displays the difference between conditions in terms of seriousness of the symptoms assigned to each predictor type (the subscript Sal indicates a serious symptom). A total of eight symptoms were used to generate the training and transfer stimuli for the three experimental conditions. Six of the symptoms were considered mild. These included "ear aches", "skin rash", "back pain", "dizziness", "sore muscles", and "stuffy nose". The two remaining symptoms, "paralysis" and "bleeding ulcers", were considered to be of a more serious nature, although this was not independently confirmed prior to the experiment.

In the 3:1 BR, Non-salience condition, mild symptoms were randomly assigned to all three abstract predictor types. There were two imperfect predictors (I), two perfect predictors of common diseases (PC), and two perfect predictors of rare diseases (PR). During training, each common disease was signaled by an I and a PC predictor (I+PC); each rare disease by an I and a PR predictor (I+PR). There were two distinct disease groupings (i.e., sets of common and rare diseases sharing an imperfect predictor – C<sub>1</sub> & R<sub>1</sub> and C<sub>2</sub> & R<sub>2</sub> – see Table 1). In the 3:1 BR, Salience condition, I and PR predictors were randomly assigned to four of the six possible benign symptoms, while the PC predictors were assigned to the more serious symptoms. In the 1:1 BR, Salience condition, the serious symptoms were associated with the PE<sub>SAL</sub> predictors (see Table 1), while I and PE predictors were randomly chosen from among the six benign symptoms (PE denotes a perfect predictor with equal base-rates).

During transfer, participants were presented with novel stimulus configurations. The configurations for the 3:1 BR, Non-salience condition are described here. Table 1 displays the analogous combinations for the 3:1 BR, Salience and 1:1 BR, Salience conditions. There were nine unique transfer configurations, including the three single predictor test items (I, PC, or PR presented alone), conflicting tests (PC+PR), and

combined tests (I+PC+PR) as used in previous studies. Additional tests resulted from combining predictors across groupings, (e.g., PC<sub>1</sub>+PR<sub>2</sub>). This type of cross-group pairing is denoted by the more general term PC+PR<sub>o</sub> (the subscript o denotes a pairing of symptoms from "other" disease groupings). Additional cross-group test items included I+PC+PR<sub>o</sub>, I+PC<sub>o</sub>, and I+PR<sub>o</sub>.

In a dimly lit room, stimuli were presented on a desktop computer monitor and responses were made on a standard keyboard. The predictor pairs shown during training were randomly assigned to four disease labels: "F", "G", "H", and "J". Participants pressed the corresponding keys to respond. On each trial, symptoms were arranged vertically on the screen, and the location of each symptom's appearance varied from trial to trial.

**Procedure** A training phase consisted of fifteen 8-trial blocks totaling 120 trials. The stimulus presentation order in each block was randomized. On each training trial, a pair of symptoms appeared onscreen, along with the statement "Diagnose either as F, G, H, or J". The time allowed to make a response was unlimited, although if the participant failed to respond within 15 seconds of stimulus onset a prompt to go faster would appear. Following a response, the stimulus remained onscreen while feedback was presented. After a correct response, the word "correct" appeared, while incorrect responses prompted the word "incorrect". Regardless of the response, the correct diagnosis was also presented. Participants could examine the feedback as long as they wished. Presented along with the corrective feedback was the statement "after you have studied the feedback, press a key to start the next trial". During training, the participant received a break after every 40 trials.

Stimulus presentation during the test phase was the same as during training, although instead of receiving corrective feedback, each trial was followed by the statement "your diagnosis has been recorded". This statement remained onscreen for one second, followed by the next stimulus. Because there were two disease groupings (i.e., sets of common and rare diseases sharing a single imperfect predictor), and nine abstract transfer combinations (described above), there were a total of 18 unique transfer combinations. Each of these transfer stimuli was presented twice, for a total of 36 transfer trials in each condition.

## RESULTS

**Training Phase** Following Kruschke (1996), learning was evaluated based on the final third (40 trials) of training. Because of the precedent for high training accuracy set by previous research, some data sets exhibiting very low accuracy were omitted from further analysis. In the 3:1 BR, Salience condition, five data sets were removed. Overall accuracy for this condition (over the last 40 training trials) after removing these outliers was 95% correct. Average accuracy for the outliers was 45% correct. In the 1:1 BR, Salience condition a single outlier was removed. Average accuracy for this condition after removal was 97% correct. The lone outlier's accuracy rate was 15% correct.

With the removal of these poor performers, average accuracy rates were similar to those reported in previous studies (e.g., Kruschke, 1996, Experiment 1). In all three experimental conditions, average accuracy for common diseases was 98% correct. Average accuracy for rare diseases was 96%, 92%, and 96% correct for the 3:1 BR, Non-salience; 3:1 BR, Salience; and 1:1 BR, Salience conditions, respectively. Clear-

ly, with the exception of the outliers, participants learned to classify the training stimuli at a high level.

### Test Phase

*3:1 BR, Non-Salience condition* Table 2 displays transfer response proportions for all three conditions. Based on the corresponding response frequencies, the significance of each trend was determined by  $\chi^2$  test. In the 3:1 BR, Non-salience condition, the major inverse base-rate trends were replicated. Responding favored the "Rare" response for the conflicting test combinations [PC+PR:  $\chi^2(1, N=120) = 10.862, p < .005$ ; PC+PR<sub>o</sub>:  $\chi^2(1, N=120) = 3.461, .05 < p < .1$ ]. Responding favored the "Common" response for the I alone test probes,  $\chi^2(1, N=120) = 16.938, p < .005$ , as well as for the I+PC+PR<sub>o</sub> test probes,  $\chi^2(1, N=120) = 49.814, p < .005$ . Although the trend was toward the Rare response for the I+PC+PR probe, this trend was non-significant,  $\chi^2(1, N=120) = 1.242, p > .1$ . Kruschke (1996, Experiment 1) found a significant trend toward the Common response for the probe I+PC+PR, although the trend was much weaker than for the probe I+PC+PR<sub>o</sub>. This may or may not have been the case for Medin and Edelson's (1988, Experiment 1) data (their presented response proportions were averaged across these two test types). In discussing the difference in response proportions for these two test cases, Kruschke points out that "they were evidently construed differently by the participants" (Kruschke, 1996, page 6).

*3:1 BR, Salience condition* The trends found in the 3:1 BR, Salience condition were similar to those found in the 3:1 BR, Non-salience condition with the notable exception of the conflicting test pairs (e.g., PC+PR). Due to the feature-salience manipulation, responding now favored the Common disease for both the PC+PR case,  $\chi^2(1, N=88) = 5.762, p < .025$ , and the PC+PR<sub>o</sub> case,  $\chi^2(1, N=88) = 4.651, p < .05$ . Responding again favored the Common disease for the I alone test case,  $\chi^2(1, N=88) = 14.098, p < .005$ , as well as for both combination probes [I+PC+PR:  $\chi^2(1, N=88) = 3.200, .05 < p < .1$ ; I+PC+PR<sub>o</sub>:  $\chi^2(1, N=88) = 46.694, p < .005$ ]. Apparently, making the PC predictors more salient than the other predictor types had a strong enough impact to disrupt the normally robust IBRE.

*1:1 BR, Salience condition* In the 1:1 BR, Salience condition, performance trends mirrored those found in the 3:1 BR, Non-salience condition (the standard inverse base-rate task) despite the absence of a base-rate manipulation during training. As predicted, conflicting test pairs elicited a strong response trend toward diseases corresponding to salient (PE<sub>SAL</sub>) predictors (analogously Rare categories in the 3:1 BR, Non-Salience condition – see Table 1), [PE+PE<sub>SAL</sub>:  $\chi^2(1, N=76) = 8.000, p < .005$ ; PE+PE<sub>OSAL</sub>:  $\chi^2(1, N=76) = 11.845, p < .005$ ]. Presenting I probes alone elicited a significant trend toward diseases corresponding to non-salient (PE) predictors (analogously Common categories in the 3:1 BR, Non-Salience condition),  $\chi^2(1, N=76) = 6.667, p < .01$ . As expected, there was a (non-significant) trend toward the non-salient-category (analogously Common) response for the I+ PE+PE<sub>OSAL</sub> test case,  $\chi^2(1, N=76) = .352, p > .05$ . However, there was a strong trend toward the salient-category (analogously Rare) response for the I+ PE+PE<sub>SAL</sub> case,  $\chi^2(1, N=76) = 22.222, p < .005$ .

With unbiased base-rates in the 1:1 BR, Salience condition, the trends toward the Salient-Disease (analogously Rare) response for conflicting test cases (PE+PE<sub>SAL</sub> & PE+PE<sub>OSAL</sub>) are apparently due to higher salience of the PE<sub>SAL</sub> predictors. The effect

on attention during training may have been so large, however, that the  $PE_{SAL}$  symptoms dominated responding every time they were presented. This could account for the trend toward the Salient-Disease response when shown test probe  $I + PE + PE_{SAL}$ , but the trend toward the Non-Salient-Disease (analogously Common) response to probe  $I$  when presented alone. According to the attention-shifting theory, the  $I \rightarrow$  Non-Salient-Disease (Common) association should be stronger than the  $I \rightarrow$  Salient-Disease (Rare) association, owing to the greater attention to  $PE_{SAL}$  than to  $PE$  during training. The strong association between  $PE_{SAL}$  and the Salient-Disease response erodes the association strength between  $I$  and the salient disease, leaving the  $I \rightarrow$  Non-Salient-Disease association relatively stronger.

Taken together, the experimental results support the notion that attention-shifts in response to high feature-salience can strongly influence learning. Not only did the IBRE disappear when attention shifted toward PC predictors in the standard IBRE task (the 3:1 BR, Salience condition), but analogous trends appeared when attention was shifted appropriately in complete absence of unequal base-rates (the 1:1 BR, Salience condition). This support for our prediction, based on Kruschke's (1996) attention-shifting theory, leads to some interesting questions about the IBRE which are considered in the Discussion section.

## DISCUSSION

The results reported here demonstrate a pronounced influence of feature-salience on classification learning and subsequent judgments. As predicted by Kruschke's attention-shifting theory (Kruschke 1996, 2001, 2003), salient features clearly dominated responding and led to an effect analogous to the IBRE.

**Rare-feature salience** Given the similarity in design between the current study and past inverse base-rate experiments, it is tempting to speculate that similar processes underlie both the feature-salience and base-rate effects. The logical conclusion of this speculation is that, the IBRE reflects that rare features are more salient than common features. If the rare features are more salient (i.e., garner more attention from the learner), then a reasonable expectation would be that rare (i.e., more salient) features are better learned, and perhaps more easily brought to mind, than are common ones. Although the data from the current study do not provide a direct test of this possibility, there is evidence to suggest the plausibility of this idea. In Kruschke's (1996, Experiment 1) study, participants learned the rare items in far fewer presentations than the common items (about 10 and 21 presentations, respectively).

One explanation for this result might be what is termed "eliminative inference", which constitutes the major opposition to the attentional theory (Juslin, Wennerholm, & Winman, 1999, 2001). The idea behind eliminative inference is that once participants learn the common categories, responses to rare stimuli can be partially inferred (at least the number of possibilities can be narrowed down) by process of elimination since there are a limited number of response alternatives. Kruschke has demonstrated the inadequacy of this account, although he has not ruled out the possibility that it may play some role in the IBRE (Kruschke 2001, 2003). Another possible explanation could be that rare events evoke surprise, which serves to focus attention (Kamin, 1969; Wagner, 1978). Shanks (1992) based a connectionist account of the IBRE on this notion. Shanks' model modified association weights in a manner similar to Kruschke's ADIT model, with attention-shifting based on the degree to which a stimulus

**Table 2**  
*Transfer Response Proportions by Condition*

Symptoms	3:1 BR, Non-Salience			
	C	R	C <sub>o</sub>	R <sub>o</sub>
I	0.633	0.275	0.033	0.058
PC	0.933	0.025	0.008	0.033
PR	0.025	0.908	0.025	0.042
PC+PR	0.325	0.617	0.050	0.008
I+PC+PR	0.433	0.533	0.025	0.008
I+PC <sub>o</sub>	0.308	0.075	0.575	0.042
I+PR <sub>o</sub>	0.183	0.042	0.025	0.750
PC+PR <sub>o</sub>	0.400	0.008	0.025	0.567
I+PC+PR <sub>o</sub>	0.783	0.050	0.008	0.158
Symptoms	3:1 BR, Salience			
	C	R	C <sub>o</sub>	R <sub>o</sub>
I	0.659	0.273	0.034	0.034
PC <sub>SAL</sub>	0.989	0.000	0.000	0.011
PR	0.011	0.932	0.023	0.034
PC <sub>SAL</sub> +PR	0.602	0.352	0.011	0.034
I+ PC <sub>SAL</sub> +PR	0.545	0.364	0.057	0.034
I+ PC <sub>oSAL</sub>	0.148	0.057	0.784	0.011
I+PR <sub>o</sub>	0.182	0.114	0.011	0.693
PC <sub>SAL</sub> +PR <sub>o</sub>	0.602	0.011	0.011	0.375
I+ PC <sub>SAL</sub> +PR <sub>o</sub>	0.841	0.011	0.023	0.125
Symptoms	1:1 BR, Salience			
	NS	S	NS <sub>o</sub>	S <sub>o</sub>
I	0.526	0.263	0.118	0.092
PE	0.908	0.039	0.053	0.000
PE <sub>SAL</sub>	0.079	0.855	0.026	0.039
PE+ PE <sub>SAL</sub>	0.316	0.632	0.026	0.026
I+ PE+ PE <sub>SAL</sub>	0.211	0.737	0.000	0.053
I+ PE <sub>o</sub>	0.184	0.171	0.632	0.013
I+ PE <sub>oSAL</sub>	0.079	0.105	0.026	0.789
PE+ PE <sub>oSAL</sub>	0.276	0.026	0.039	0.658
I+PE+ PE <sub>oSAL</sub>	0.500	0.066	0.000	0.434

**Note.** I = Imperfect predictor, PC = Perfect-common predictor, PR = Perfect-rare predictor, C = Common response, R = Rare response, NS = Non-salient-disease response, S = Salient-disease response. Subscript o refers to the "other" disease grouping (see text for details).

was expected. Although this model was able to fit the IBRE trends qualitatively, it was rejected on the grounds that it was unable to account for other related base-rate phenomena (in addition to providing an inferior qualitative account compared to ADIT, the model failed to account for the effect known as “apparent base-rate neglect” – Kruschke, 1996, pp. 21-22).

A more direct test may be required to determine whether rare features are indeed more salient (garner more attention) than common features. A simple test might employ the standard IBRE training regime (participant learns to classify two-feature stimuli followed by corrective feedback), followed by a new type of testing phase. At test, the stimulus labels would be presented one at a time and the participant’s task would be to name the corresponding features (e.g., for a presentation of “Disease C”, the correct response would be the appropriate I and PC features). A reasonable prediction based on the rare-feature salience hypothesis would be that participants should exhibit better (quicker and/or more accurate) recall for PR features than for PC features, despite having more exposures to the PC features.

**Salience vs. error minimization** As stated earlier, an important mechanism in the ADIT account is error reduction. The model implies that the participant shifts attention to minimize error, which ultimately affects the stimulus-response associations in a way that leads to the IBRE. The ADIT model cannot account for the salience effect, but EXIT (a recent extension of ADIT – Kruschke 2001, 2003) can because it has a salience parameter. This is interesting because this parameter isn’t required to account for the IBRE trends. The two accounts, salience and error reduction, are confounded in the IBRE task – it is difficult to know whether there is a salience mechanism at work along with the error-reduction mechanism. However, the two mechanisms could possibly be de-confounded. If the IBRE can be obtained in absence of error-driven learning, then the salience parameter would be the only way to account for this effect. Such an experiment might utilize an observational-learning task in order to eliminate overt error-driven learning. Of course, learning may still be error driven to some extent in the mind of the participant.

In sum, the current work supports the prediction, based on Kruschke’s attention-shifting account, that any stimulus element that can cause attention to shift appropriately can lead to a learning effect analogous to the IBRE. It also demonstrates the strong influence of feature salience on classification learning. Finally, it lays the groundwork to pursue interesting follow-up questions pertaining to the possibility that something like the salience effect is at work in the IBRE. On this view, the IBRE is a marker of the differential salience of the common and rare predictors. It is important to note that this salience account is not at odds with Kruschke’s theory of attention shifting. Although the salience view conflicts with the ADIT model instantiation of the attention-shifting theory, it is in concordance (at least to some degree) with the EXIT model instantiation. Our goal is simply to capitalize on what the various theory instantiations can tell us about the underlying reasons for the attention shifting phenomenon.

## REFERENCES

- Binder, A., & Estes, W. K. (1966). Transfer of response in visual recognition situations as a function of frequency variables. *Psychological Monographs*, *80* (23, Whole No. 631).
- Bock, J. S., & Clifton, C., Jr. (2000). The role of salience in conceptual combination. *Memory & Cognition*, *28*, 1378-1386.
- Bohil, C. J., & Maddox, W. T. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception and Psychophysics*, *63*, 361-376.
- Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance*, *29*, 270-278.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1-73.
- Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology*, *50*, 131-138.
- Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 556-576.
- 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.
- Gibson, B. S., & Yuhong, J. (2001). Visual marking and the perception of salience in visual search. *Perception & Psychophysics*, *63*, 59-73.
- Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning. *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 21<sup>st</sup> 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.
- Kalish, M. L. (2001). An inverse base rate effect with continuously valued stimuli. *Memory and Cognition*, *29*, 587-597.
- Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.
- Koehler, J. J. (1996). The base-rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1-53.
- Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 3-26.
- Kruschke, J. K. (2001). The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1385-1400.

- Kruschke, J. K. (2003). Attentional theory is a viable explanation of the inverse base-rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1396-1400.
- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083-1119.
- Lindeman, S. T., Van Den Brink, W. P., & Hoogstraten, J. (1988). Effect of feedback on base-rate utilization. *Perceptual and Motor Skills*, *67*, 343-350.
- Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology*, *38*, 231-248.
- Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General*, *118*, 417-421.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68-85.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64-99). New York: Appleton-Century-Crofts.
- Shanks, D. F. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3-18.
- Yantis, S., & Egeth, H. E., (1999). On the distinction between visual salience and stimulus-driven attention capture. *Journal of Experimental Psychology: Human Perception & Performance*, *25*, 661-676.
- Wagner, A. R. (1978). Expectancies and the priming of STM. In Hulse, S. H., Fowler, H., & Honig, W. H. (Eds.), *Cognitive Processes in Animal Behaviour*, pp. 177-210. Hillsdale, NJ: Lawrence Erlbaum Associates.

Received February 17, 2005

Accepted March 24, 2005

Key word: Inverse base rate effect, Feature salience, Attention shifting, Attentional theory, Learning, Cues