Illusory Correlation and the Inverse Base Rate Effect: Different Underlying Mechanisms

Kruschke, Sherman, Conrey & Sherman

Contents

Draft of April 5, 2006.

Submitted to, and rejected from, *JEP:Learning, Memory & Cognition* in spring 2006.

Illusory Correlation and the Inverse Base Rate Effect: Different Underlying Mechanisms

John K. Kruschke
Indiana University, Bloomington

Jeffrey W. Sherman
University of California at Davis

Frederica R. Conrey
Indiana University, Bloomington

Steven J. Sherman
Indiana University, Bloomington

The two phenomena of illusory correlation and the inverse base rate effect are both misperceptions of true contingencies and are evoked by training procedures with analogous structures. We therefore explored the possibility that they share a common underlying mechanism. In two experiments using social stimuli (traits and groups), we measured illusory correlation and the inverse base rate effect in a within-subjects design. We found no hint of correlation in their magnitudes, which suggests that the two effects are generated by different mechanisms. In Experiment 2, we gave no explicit feedback for the rare cue in illusory correlation, yet robust illusory correlation was obtained. This suggests that illusory correlation is a response strategy, not a learning effect. Previously established mathematical models of attentional learning and eliminative inference (a response strategy) were fit to the data via computer simulations. The attentional learning model could accurately fit the inverse base rate effect, but fared less well for illusory correlation. The eliminative inference model fit the illusory correlation results better than the inverse base rate effect. We conclude that while the two phenomena are dominated by different underlying mechanisms, both mechanisms may be at work in social situations.

People do not always accurately learn the true correlations of experienced cues and outcomes. These inaccuracies are interesting because they can have real-world consequences and because they can reveal underlying cognitive mechanisms of learning and judgment. In social psychology, illusory correlation has been investigated because of its possible relationship with real-world negative stereotyping. In the psychology of learning, the inverse base rate effect has been studied as a touchstone for theories of associative learning and choice. In this article we point out structural analogies between the two phenomena, and explore whether or not common mechanisms might underlie them. We conclude that the two effects are dominated by different mechanisms, but that both may be at work in real-world situations.

We are interested in how people learn from experience with cues and outcomes. For example, the learner experiences another person who performs certain behaviors (the cues), and subsequently the learner is informed that the person belongs to a particular group (the outcome). As another example, the learner may be a clinician who sees a patient with certain symptoms (the cues), and subsequently the learner is informed by a lab result that the patient has a particular disease (the outcome).

In typical associative learning experiments, the participant experiences numerous trials in which cues are presented on a computer screen, the participant guesses the outcome, and then the correct outcome is provided. Learning is assessed by accuracy on the training items and also by generalization to novel combinations of cues. Because one of the phenomena we investigate has been extensively studied in the context of social cognition, we will often refer to cues as traits, and outcomes as groups.

Two learning phenomena related to base rates

The inverse base rate effect

In the procedure for the inverse base rate effect, one outcome occurs three times as often as the other. The "base rates" of the outcomes are, therefore, three to one. What happens for one of the ambiguous test items is that people tend to choose the infrequent outcome, even when the cues in the test item are equally diagnostic of the outcomes. This behavior has been called an inverse base rate effect (Medin & Edelson, 1988).

Participants experience many cases in which two cues, denoted PC and I, are displayed together and followed with the

Supported in part by Grant BCS-9910720 from the National Science Foundation to John K. Kruschke. The use of valenced traits was adapted from experiments developed in collaboration between Kruschke and Dr. Douglas Wedell. For help administering the experiments, we thank Jennifer Armstrong, Lauren Bulakowski and Amarachi Igboegwu. Jared Dixon helped with preliminary data analyses. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. The first author's world wide web page is at http://www.indiana.edu/~kruschke/

common, i.e., frequent, outcome Ci. This training case is denoted PC.I→Ci. Intermixed with those cases are less frequent trials in which cues PR and I are displayed together and followed with the rare outcome Ri (PR.I→Ri). Cue PC is therefore a Perfect predictor of outcome Ci, and cue PR is a Perfect predictor of outcome Ri, while cue I is an Imperfect predictor because it is always present for both outcomes.

After experiencing many such trials, people are asked to guess the most likely outcome for cue I by itself, or for the novel combination PC and PR. If people learned the true structure of the cues and outcomes, then they should simply respond consistently with the base rates, and tend to choose outcome Ci more than outcome Ri. It turns out, however, that while people do tend to choose Ci in response to cue I by itself, people instead tend to choose Ri in response to the combination PC with PR.

The inverse base rate effect, and its close cousin "highlighting" (Kruschke, 2003b), is quite robust and has been obtained in many different scenarios. It was first reported in a fictitious disease diagnosis procedure by Medin and Edelson (1988), and several variations involving changes in base rates were reported by Medin and Bettger (1991). Several researchers have used different relative base rates (Juslin, Wennerholm, & Winman, 2001; Shanks, 1992; Winman, Wennerholm, Juslin, & Shanks, 2005), and there have been many other variations in the particulars of the cues and test items (e.g., Kalish, 2001; Kruschke, 1996, 2001a). The effect does not depend on the participants thinking that they are diagnosing diseases, as the effect was found quite strongly in a word memory paradigm by Dennis and Kruschke (1998), in a simple geometric-figure association task by Fagot, Kruschke, Dépy, and Vauclair (1998), and in a simple word association task by Kruschke, Kappenman, and Hetrick (2005).

Theories of the inverse base rate effect. The inverse base rate effect is quite challenging to models of associative learning. Simple co-occurrence counting obviously cannot predict it. The classic error-driven associative learning model by Rescorla and Wagner (1972) cannot predict it. The rational model of J. R. Anderson (1991) can weakly show the effect, but not its generalization known as highlighting (Kruschke, submitted). Exemplar-based models struggle with the effect (Medin & Edelson, 1988; Kruschke, 1992), the Kalman filter model cannot show it (Kruschke, submitted), and many memory models cannot accommodate it (Dennis & Kruschke, 1998).

Some attempts have been made to modify the Rescorla-Wagner model to account for the inverse base rate effect. Markman (1989) suggested that an absent-but-expected cue could be encoded as a negative cue activation, but he offered no model of how to learn cue-to-cue expectations. Shanks (1992) suggested a model in which rare cues have higher learning rates than frequent cues. In particular, the rare cue, PR, has a higher learning rate than the frequent cue, I. On I.PR→Ri trials, the fast-learning PR cue acquires a stronger association with Ri than cue I does. Once Ri is well predicted, no further changes in associations takes place, so the previously learned I-to-Ci connection remains stronger than the I-to-Ri connection, and the PR-to-Ri connection is stronger than the PC-to-Ci connection. This intriguing approach was shown not to generalize to closely related experiment designs, however (Kruschke, 1996). We will see below that this approach is analogous to a prominent explanation of illusory correlation. Although it might not be the dominant cause of the inverse base rate effect, it is reasonable to think that this sort of mechanism is at work in human learning, along with other mechanisms.

A rather different approach was suggested by Juslin et al. (2001), who proposed that the effect is caused not by biases during learning but instead by a decision strategy invoked at test. The idea is that when tested with the cue configuration PC and PR, people think that it is very dissimilar from any items they learned in training, and therefore the correct response cannot be any of the responses they learned in training. Because the frequent outcome is more likely to have been successfully learned than the rare outcome, it is that frequent outcome that will (more likely) be eliminated by this reasoning, leaving the unlearned rare outcome as the only available candidate response. This type of processing is called eliminative inference, and a mathematical model that implements it is called ELMO. Juslin et al. (2001) showed that some of the basic trends in the inverse base rate effect are captured by ELMO, but Kruschke (2001a) showed that ELMO fails to capture many aspects of the data. Despite the fact that ELMO was shown not to be an accurate model of the inverse base rate effect, some mechanism like eliminative inference is indisputably operating in human cognition (Juslin et al., 2001; Kruschke & Bradley, 1995), and we will re-visit the ELMO model in the context of illusory correlation.

To date, the most successful account of the inverse base rate effect suggests that people rapidly shift their attention among cues when learning. Elements of this idea were proposed by Medin and Edelson (1988) and formalized in mathematical models by Kruschke (1996, 2001a, 2001b, 2003a). The explanation goes as follows. The frequent cases of PC.I→Ci are learned first because they occur so often. Both cues are attended to, and moderate-strength associations are built from both cues to the outcome. Later, the rare cases, PR.I→Ri are learned. When such a trial occurs, attention is shifted away from cue I to cue PR, because attention to cue I generates the wrong response. That is, attention is rapidly shifted to reduce error. Then an association is learned from cue PR to outcome Ri. Because this single association bears all the weight for predicting the outcome, the PR-Ri association tends to be stronger than the PC-Ci connection. The shift of attention is also learned, so that people remember to shift attention toward PR, especially when it is accompanied by I. A mathematical model called EXIT implements this idea of error-driven attentional shifts. The model has been shown to accurately fit human performance in the inverse base rate effect and related designs (e.g., Kruschke, 2005; Kruschke et al., 2005).

This attentional account says that the only role of the base rates is for people to learn the common outcome first and the rare outcome later, and therefore emphasizing the base rate in the moniker for the phenomenon is misleading. A better

name is "highlighting" because the perfectly predictive cue of the later learned item is attentionally highlighted.

Illusory correlation

An illusory correlation occurs when a person believes that two factors covary when in fact no correlation exists. For example, a person might believe that a target group has relatively more members with a certain trait than a comparison group, but in fact both groups have the same proportion of members with that trait. Chapman and Chapman (1967) documented illusory correlations in psychiatric diagnosis. Illusory correlations have been extensively studied in social psychology, especially since Hamilton and Gifford (1976) documented the effect and proposed a theory as a potential explanation of negative stereotyping of minority groups. Illusory correlations have been documented in numerous reports. Stroessner and Plaks (2001) provide a recent review.

In a typical illusory correlation experiment, cue A indicates outcome Cx twice as often as it indicates outcome Rx, and cue B also indicates outcome Cx twice as often as outcome Rx, but cases of outcome Cx occur twice as often as cases of outcome Rx. Therefore, no actual correlation of the cues with the outcomes exists. Nevertheless, people tend to associate A with Cx more strongly than they associate A with Rx. That preference for Cx is not as strong for cue B. In fact, despite the greater occurrence of B with Cx than with Rx, people will sometimes associate B more strongly with Rx than with Cx.

There are some similarities between illusory correlation and the inverse base rate effect. Both involve outcomes with different base rates. Both suggest misperceptions of cue-outcome correlations. Both appear to involve an overemphasis of the distinctive cue of the rare outcome.

Theories of illusory correlation. Hamilton and Gifford (1976) emphasized distinctiveness based illusory correlations that might be explained by people associating rare (i.e., distinctive) traits with rare (distinctive) groups. The theory proposes that distinctive traits and groups garner more attention during encoding. This idea is similar to the model of Shanks (1992) mentioned earlier regarding the inverse base rate effect.

Other explanations of the illusory correlation effect suggest that group judgments are sensitive not to the ratio of common to rare cues, but rather to the difference in the numbers of common and rare cues. For Group Cx, there are typically 16 common and 8 rare cues, yielding a difference of 8. For Group Rx there are typically 8 common and 4 rare cues, yielding a difference of 4. Because the difference in common-rare cues is greater for Group Cx, people perceive that group more in terms of the common cue. Two different models rely on this "difference of the differences" analysis. One, by Smith (1991), is an exemplar model that proposes that memory traces of the different cues yield the key perception that there are differences between the differences in group/cue associations. The second model, proposed by McGarty and his colleagues (e.g., McGarty, Haslam, Turner, & Oakes, 1993), suggests that participants are motivated to differentiate between Groups Cx and Rx, and seek any possible means to do so. One way to perceive meaningful differences between the two groups is to focus on the difference of the differences in group/cue associations, rather than on the ratios of common to rare cues describing the two groups.

Finally, Fiedler (1991) has argued that the illusory correlation effect can arise simply because different amounts of information describe Group Cx and Group Rx. Because people do not process the information perfectly, estimates of the group/cue associations will regress to the mean. This regression should occur for both groups, but particularly for Group Rx rare cues because the sample size for that group/cue association is particularly small. As a result, the Group Rx rare cues are particularly over-estimated (toward the overall group/cue mean of 8), yielding smaller perceived differences in both the numbers and ratios of positive to negative cues for Group Rx than for Group Cx.

These various theories of illusory correlation will be revisited in the final discussion in light of our new empirical results.

Table 1
Frequencies of trait-group combinations in Experiment 1.

Inverse Base Rate Effect			Illusory Correlation		
	Group			Group	
Traits	Ci	Ri	Traits	Cx	Rx
PC.I	27	0	A.X	16	8
PR.I	0	9	B.X	8	4

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., "PC.I" means that trait PC and trait I were presented together.

Structural analogy and shared mechanism?

There has been no previous attempt to explain illusory correlation in terms of rapidly shifting selective attention to cues, despite the success of the approach in explaining the inverse base rate effect. Attentional shifting might account for illusory correlation as follows. When learning the common items, people associate with the common outcome not only cue A but also "context" cues that co-occur with cue A. For example, when cue A is a trait of a person, there are typically other personal attributes presented along with cue A. These contextual traits are often shared with most other members of both the common and rare groups. When subsequently learning the rare items, people shift attention away from these shared context cues toward cue B. This enhanced attention to B causes the illusory correlation.

Table 1 shows the structural analogy between the inverse base rate effect and illusory correlation. The left side of the table shows the cue-outcome pairings for the inverse base rate effect. Cues are listed as traits, and outcomes as groups, because that is how they are instantiated in Experiment 1. The table indicates that item PC.I→Ci occurs 27 times for

every 9 times that PR.I→Ri occurs. This 3-to-1 base rate is typical for experiments investigating the inverse base rate effect.

The right side of Table 1 shows our new version of the illusory correlation procedure. Traits A and B always occur in training with another trait X. Item A.X→Cx occurs 16 times, which is twice as often as item B.X→Cx occurs. That same ratio of cues is true for Group Rx, which occurs only half as often as Group Cx. These relative frequencies are typical of experiments investigating illusory correlation. This design, using a shared trait, has never been used before in illusory correlation experiments, and so it is an empirical question as to whether or not the effect will occur. The results of Experiment 1 confirm that it does.

If there is a shared mechanism in the inverse base rate effect and illusory correlation, then the magnitude of the inverse base rate effect and illusory correlation ought to covary across individuals. This covariation of the effects is implied by the premise that the shared mechanism (regardless of exactly what it is) will vary in its strength from one individual to another. If the mechanism is relatively strong in a particular person, then the inverse base rate effect and illusory correlation will both be large. If in another person the mechanism is relatively weak, then both effects will be weaker. This strategy of looking for covarying effects was used by Kruschke et al. (2005), who found that the inverse base rate effect covaried with associative blocking. They explained the covariation of the effects as individual differences in attentional shifting and learning.

Experiment 1: The inverse base rate effect and illusory correlation in a within-subjects design

In our first experiment, people learned the items for the two effects concurrently. All the items listed in Table 1 were randomly intermixed during training. In this way we can simultaneously assess the magnitudes of illusory correlation and the inverse base rate effect in each individual.

We assayed the effects in two ways. First, we measured choice preferences for assigning traits to groups. This has been the standard means for assessing the inverse base rate effect, and this type of measure has also been used for assessing illusory correlation. Second, we measured subjective ratings of the strength of each trait in each group. This procedure is typical of experiments in illusory correlation, but is novel for experiments investigating the inverse base rate effect.

The experiment used *valenced* traits. The shared traits I and X were positive traits, while all other traits (PC, PR, A, and B) were negative. Use of valenced traits is typical in social psychological research because such valences are central to the social phenomena of interest. But cues of differing valence are often specifically avoided in associative learning research in order to control extraneous sources of variance. The use of valenced traits allows us to look for asymmetries in ratings that would be consistent with selective attention to cues. For example, if positive trait I is associated with group Ci but not as strongly with group Ri, then group Ci should be rated more positively than group Ri.

Method

Participants. Participants volunteered for partial credit in an introductory psychology class at Indiana University in Fall 2002. A total of 84 students participated (71 female, 13 male, age 18-22 years).

Stimuli and procedure. Each training trial proceeded as follows. The initial display had a unique name (e.g., "Jalen") at the top of the screen and two traits (e.g., "nosey" and "sincere"), one above the other, in the middle of the screen, and a response prompt in the lower part of the screen, which read, "This person is a member of which group? (Press F, G, H, or J)". The unique name was chosen on each trial randomly without replacement from a list of 250 male and 250 female distinct names gathered from the Social Security Administration baby name database http://www.ssa.gov/OACT/babynames/. After the participant made a response, corrective feedback was supplied. On the feedback display the two traits continued to be displayed, but the name was removed and the response prompt was replaced by feedback which read, "[Correct!/Wrong!] This is a member of group [F/G/H/J]." A wrong response was also followed by a brief buzzing sound. At the bottom of the feedback screen was a prompt to press the space bar to continue to the next item.

The traits were selected to have positive or negative valence. Traits PC, PR, A and B were always negative traits, randomly chosen for each participant from nosey, unfair, crude, jealous, and intolerant. Traits I and X were always positive traits, randomly chosen for each participant from honest, thoughtful, dependable, sincere, and considerate. These traits are very negative or very positive according to the norms reported by N. H. Anderson (1968). The selection of valenced traits follows previous work by Wedell and Kruschke (2006). The four abstract groups (Ci, Ri, Cx, and Rx) were randomly assigned to response keys (F, G, H, and J) independently for each participant.

There were three types of questions asked in the test phase. One type was a classification question, which was presented in exactly the format of the training trials. The only difference was that the feedback screen stated simply that the participant's response had been recorded; no corrective information was provided. Classification trials in the test phase included trait combinations that were seen in training and novel trait combinations that had not been seen in training. The repeated items from the training phase were not given feedback in the test phase.

A second type of test-phase question asked for ratings of traits for groups. The question was always in the form, "How [trait] are members of group [F/G/H/J]?" For example, the screen could display the question, "How thoughtful are members of group J?" The lower part of the screen displayed the numerals 1-9 on a ruler-like scale, with the ends labeled "Not

At All" and "Very Much", and a prompt that said "Press a top row number key from 1 to 9."

A third type of test-phase questions was group likability ratings. In these cases, the screen displayed a group label in the middle of the screen, "Group [F/G/H/J]", beneath which was a response prompt, "How likable is this group?" Below the prompt was a ruler-like scale with the numerals 1-9, marked at its ends, "Not Likable" and "Very Likable." Below the scale was the instruction, "Press a top row number key from 1 to 9."

Design. The training phase consisted of 3 blocks of the 72 items shown in Table 1. Within each block, the 72 items were randomly ordered.

Each block of the test phase had 12 classification items, 12 ratings for traits of groups, and 4 group likability ratings. Within each block, the 28 questions were randomly intermixed. There were two test blocks. A complete list of classification and rating questions appears in the Results section; the 12 classification items are listed in Table 2, and the 12 trait-group rating combinations appear in Table 3.

Results and discussion

We are interested in the behavior of participants who actually learn about the items seen in training. Therefore we set accuracy criteria on the items for which there were deterministically correct answers, namely items I.PC and I.PR. We decided to include participants who, in the last (i.e., third) block of training, had accuracies significantly better than merely matching the background probabilities of groups Ci and Ri. This implies specific accuracy levels as follows. The last block contained 27 trials of I.PC and 9 trials of I.PR. The background probabilities of relevant groups for these items are 75% Ci and 25% Ri. We want Ci responding on I.PC to significantly exceed 75% of 27, and we want Ri responding on I.PR to significantly exceed 25% of 9. We arbitrarily set significance at .05 one-tailed (but all tests in the results section below are two-tailed). Therefore, according to a binomial distribution, we included for further analysis only those participants who responded correctly on 25 or more of 27 I.PC trials and on 5 or more of 9 I.PR trials in the last training phase. These criteria left 62 (of 84) participants in the pool for further data analysis. It turns out that setting the criteria differently makes virtually no change in the statistical conclusions below. Therefore the accuracy criteria can safely be viewed as merely a precautionary measure to exclude unmotivated participants.

In the statistical analyses reported below, all χ^2 tests assume that multiple responses from individual participants are independent. This is standard procedure in much perceptual and psychophysical research, where the experimenter measures many repeated responses to the same stimulus. In the design of Experiment 1, each test item occurred only twice and was separated (on average) by many trials in the two test blocks, so an assumption of independence is not unreasonable. Even if the two repetitions are not considered independent, the conservative correction is to divide all χ^2 values by 2; when this is done, all results reported below remain significant.

Table 2
Percentage of group classifications in Experiment 1.

	Group Chosen			
Traits	Ci	Ri	Cx	Rx
I.PC	97	3	0	0
I.PR	9	85	2	4
I	82	11	4	2
PC.PR	29	62	4	5
I.PC.PR	47	52	0	1
X.A	2	4	69	26
X.B	2	7	48	44
A	1	2	64	34
B	1	3	34	62
X	6	5	62	27
A.B	1	6	44	49
X.A.B	1	4	61	34

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., "PC.I" means that trait PC and trait I were presented together.

For all t-tests or tests of correlation, outliers were first removed as a precautionary step to better respect the assumptions of the tests. Outliers were defined as scores that were more than 1.0 interquartile range (IQR) above the 75th percentile or more than 1.0 IQR below the 25th percentile (and where the IQR is defined as the distance between the 25th and 75th percentiles). In all cases, this resulted in only a few, if any, points being removed. For all t-tests, effect size *d* is also reported, with *d* defined as the mean difference divided by the estimated population standard deviation.

Classification. Table 2 shows the percentage with which each group was chosen for each combination of traits in the test phase. Recall that each test item was presented twice in the test phase; therefore each row of Table 2 is based on $2 \times 62 = 124$ responses.

The first two rows of Table 2 reveal that accuracy on item I.PC was 97% and on I.PR was 85%. There was no corrective feedback on these items in the test phase, so this high accuracy indicates good retention of learning in the midst of intermixed trials of novel trait combinations and rating tasks.

The inverse base rate effect is strongly exhibited. Participants classified trait I by itself as group Ci 82% of the time, and as group Ri only 11% of the time. This preference is reliability different from 50-50, $\chi^2(1, N{=}116) = 66.76$, $p < 0.001$. On the other hand, for trait pair PC.PR, participants strongly preferred group Ri (62%) over group Ci (29%). This preference is also reliably different from 50-50, $\chi^2(1, N{=}113) = 14.88$, $p < 0.001$. Thus, the inverse base rate effect is strongly in evidence for these socially based stimuli (and when each trait item is marked by a unique name). These results further expand the conditions under which the inverse base rate effect is observed.

Table 3
Mean ratings of traits for groups in Experiment 1.

	Group			
Traits	Ci	Ri	Cx	Rx
I	7.69	5.77	–	–
PC	7.53	3.52	–	–
PR	3.14	7.81	–	–
X	–	–	6.54	5.81
A	–	–	6.75	6.03
B	–	–	5.66	6.46

Note: Cells with "–" indicate trait-group combinations that were not included.

For the items that probe illusory correlation, notice that responding to trait pair X.A was close to probability matching, with 69% Cx choices and 26% Rx choices (the base rates of Cx and Rx were 67% and 33%, respectively). On the other hand, responding to trait pair X.B did not show a strong preference for Cx, with just 48% Cx choices and 44% Rx choices.

The crucial assay of illusory correlation is a comparison of trait A by itself with trait B by itself. Recall that in training, trait A indicated group Cx the same proportion of times that trait B indicated Cx. Therefore an illusory correlation is found if the choice preference observed for A is different from the choice preference observed for B. Such a difference is exhibited: For trait A, participants preferred Cx (64%) over Rx (34%), but for trait B that preference was not as strong; indeed, it was reversed, as participants preferred Rx (62%) over Cx (34%). This difference of preferences is reliable, $\chi^2(1, N{=}240) = 21.59$, $p < 0.001$. These results reveal that illusory correlation is robustly observed even in this novel training task.

Trait in group ratings. Table 3 shows mean ratings of traits for groups. There were 12 trait-group combinations probed during the test phase, as indicated by the 12 cells of Table 3 that have numerical entries. For example, the cell in row I and column Ci indicates ratings from trials in which the participant was asked, How I-ish are members of group Ci? (with "I-ish" replaced by an actual trait such as "thoughtful" and "Ci" replaced by an actual group label such as "G"). Each cell's mean rating is based on a total of 124 ratings (two ratings per participant). In the following discussion, the rating of the strength of trait T for group G is denoted $r(T{\in}G)$.

The first row of Table 3 reveals that group Ci was rated as more "I-ish" than group Ri, with means ratings of 7.69 vs. 5.77, $t(56) = 5.31$, $p < 0.001$, $d = 0.70$. This stronger rating of trait I in group Ci than group Ri echoes the results seen for classification of trait I, which also strongly favored group Ci.

The next two rows of Table 3 show that group Ci was rated as more "PC-ish" than group Ri, but group Ri was more "PR-ish" than group Ci. These differences simply indicate that ratings reflect the correct classifications learned during training. There is a hint in the ratings that PR is rated more strongly for Ri than PC is rated for Ci. That is, $r(PR{\in}Ri){-}r(PR{\in}Ci)$ is larger than $r(PC{\in}Ci){-}r(PC{\in}Ri)$. To test the reliability of this interaction, an interaction contrast was computed for each individual participant, $[r(PR{\in}Ri){-}r(PR{\in}Ci)]{-}[r(PC{\in}Ci){-}r(PC{\in}Ri)]$, and the mean of this contrast was tested for being greater than zero. It was not significant, $t(57) = 0.88$, $p = 0.38$, $d = 0.12$. The analogous contrast in Experiment 2 will turn out to be reliably greater than zero, however, with $d = 0.29$.

The ratings show a strong illusory correlation for traits A and B. Group Cx is rated as more A-ish than group Rx, yet group Rx is rated as more B-ish than group Cx. The interaction is reliable, $t(59) = 3.82$, $p < 0.001$, $d = 0.49$. As was found for the classification preferences, we see again for the trait ratings that a robust illusory correlation is obtained for this novel training paradigm.

Correlations of effects. If illusory correlation and the inverse base rate effect are caused, at least in part, by some common underlying mechanism, and if that mechanism varies in strength across individuals, then the magnitude of illusory correlation and inverse base rate effects should covary across individuals. This logic was used by Kruschke et al. (2005) to demonstrate covariation of highlighting (which is closely related to the inverse base rate effect) and associative blocking. To discover such covariation, we must first define the magnitude of illusory correlation and inverse base rate effect in each individual. Following the procedure of Kruschke et al. (2005), the magnitude of the inverse base rate effect for choices was defined as a sum of response choices consistent with the effect, minus the choices inconsistent with the effect. Formally, letting $fr(G|T)$ denote the frequency of responding group G given traits T, the magnitude of the inverse base rate effect for an individual was defined as $fr(Ri|PC.PR) - fr(Ci|PC.PR) + fr(Ci|I) - fr(Ri|I)$. Because there were two trials of trait PC.PR and two trials of trait I, that magnitude could range between -4 and $+4$. The magnitude of illusory correlation for choices was defined analogously as $fr(Cx|A) - fr(Rx|A) + fr(Rx|B) - fr(Cx|B)$, which also could range between -4 and $+4$.

The correlation of the inverse base rate effect and the illusory correlation effect for choices was not reliably different from zero, and only very weakly positive: $r(df{=}56) = 0.072$, $p = 0.591$ (two-tailed). Kruschke et al. (2005) found a correlation between highlighting and blocking of $r = .38$; if the correlation were that large here, our sample size yields a statistical power of .90 two-tailed. Admittedly, the correlation might not be as strong here because we are using fewer trials per subject to assess each effect. Bolstering our claim to weak if any correlation are the results from Experiment 2, where we will see that the correlation of the two effects is slightly negative.

We can define analogous measures of the inverse base rate effect and illusory correlation for ratings, and ascertain whether the effects as measured by ratings are correlated across individuals. The inverse base rate effect for ratings was computed for each individual as $r(I{\in}Ci){-}r(I{\in}Ri){+}r(PR{\in}Ri){-}r(PR{\in}Ci){-}r(PC{\in}Ci){+}r(PC{\in}Ri)$. This

summarizes in a single mean rating, for each individual, the extent to which there was an asymmetry in ratings of I, PC, and PR. The degree of illusory correlation for ratings was computed as $r(A{\in}Cx){-}r(A{\in}Rx){+}r(B{\in}Rx){-}r(B{\in}Cx)$. This value was used earlier in the test of illusory correlation for ratings. The correlation of the inverse base rate effect and the illusory correlation effect for ratings was not reliably different from zero, and was actually negative: $r(df{=}50) = -0.116$, $p = 0.413$ (two-tailed).

These failures to obtain significant correlations between the inverse base rate effect and illusory correlation might be taken to imply merely that the measures were very noisy and would correlate with nothing, and therefore the measures can be discounted as uninformative. But this is not true. The choice and rating measures *did* significantly correlate *within* effects: For the inverse base rate effect, choices and ratings had a strong correlation, $r(df{=}52) = 0.561$, $p < 0.001$ (two-tailed). For illusory correlation, choices and ratings again correlated significantly, $r(df{=}54) = 0.313$, $p = 0.019$ (two-tailed).

In summary, the ratings and choices correlated significantly within effect (inverse base rate or illusory correlation), but the two effects did not correlate across effects, either for choice or for ratings. This lack of correlation between the effects suggests that the two phenomena are not caused by the same underlying mechanisms; or, more precisely, whatever common mechanisms they share are not very strong compared to other sources of variation.

Group likability ratings. Participants were also asked how likable each group was. Mean likabilities were as follows: 5.92 for group Ci, 4.97 for Ri, 5.06 for Cx, and 4.94 for Rx. The ratings for Ci and Ri were reliably different, $t(61) = 3.31$, $p = 0.002$, $d = 0.42$. This difference is consistent with the positively valenced trait I being more strongly associated with group Ci than group Ri. Evidently, however, trait X was not strongly enough associated with group Cx to make it much more likable than Rx. This relative weakness of X compared with I is consistent with the classification preferences and trait ratings.

Summary of results. Experiment 1 yielded several new findings regarding the inverse base rate effect. The experiment showed that the effect occurs for social (trait-group) stimuli. Previous research on the inverse base rate effect, except for Wedell and Kruschke (2006), used non-social stimuli. Experiment 1 also showed that the inverse base rate effect is manifest in trait-group ratings and likability ratings; all previous research used choice proportion as the only measure.

Experiment 1 also produced new findings regarding illusory correlation. The experiment showed that illusory correlation occurs in a trait-group predictive training procedure; most previous research used just exposure to cases. The experiment design was novel: All training instances had a shared trait; no previous research has used this structure. The design also used equally positively valenced traits in both groups.

Importantly, the inverse base rate effect and illusory correlation were tested in a within subjects design, and the magnitudes of the inverse base rate effect and illusory correlation did not covary across individuals (unlike highlighting and blocking Kruschke et al., 2005). The lack of covariation suggests that different mechanisms underlie the phenomena.

One theoretical implication of Experiment 1 is that the inverse base rate effect could be a contributor to social stereotyping, because the effect is now known to occur with social stimuli. Wedell and Kruschke (2006) found that highlighting (i.e., the inverse base rate effect) is strongest when the rare trait is negative and the frequent trait is positive, making the phenomenon even more intriguing as a contributor to social stereotyping. Attentional shifting, as an explanation of the inverse base rate effect, is a new theoretical approach to mechanisms of stereotyping. Neither distinctiveness nor eliminative inference (which have analogous theories in social psychology) account for highlighting, but attentional shifting does. To the extent that both the inverse base rate effect and illusory correlation contribute to stereotyping, the lack of covariation between the inverse base rate effect and illusory correlation suggests that there are different underlying mechanisms contributing to the creation of stereotypes.

Modeling the results of Experiment 1

Intuitively, either the attentional shifting theory or the eliminative inference theory could account for both the inverse base rate effect and illusory correlation. One benefit of formalizing the theories is that specific quantitative predictions can be derived. The quantitative fits can indicate specific merits or shortcomings of each model in far greater detail than armchair theorizing. As will be seen, despite the intuitive appeal of each theory, the models do not fit all aspects of the data equally well.

The attention shifting theory has been formalized in the EXIT model (Kruschke, 2001a, 2001b), and the eliminative inference theory has been formalized in the ELMO model (Juslin et al., 2001). The reader is referred to the original articles for full details of the models. The following sections provide an overview.

Attention shifting in the EXIT model

The EXIT model is a connectionist network with two layers of associative weights. The first layer connects input cues to attentional gates on those cues. The attentional gates are simply multipliers, so that a cue can be ignored by multiplying it by a value near zero, and a cue can be strongly attended to by multiplying it by a value near one. The connections in the first layer learn to associate particular cue combinations with appropriate distributions of attention across the gain nodes. For example, the first layer can learn that for input I.PC, attention should be partially allocated to both cue I and cue PC, but for input I.PR, attention should be strongly allocated to cue PR but not to cue I. The second layer of associations connects the attentionally-gated cues to response

choices. For example, the second layer can learn to associate the attended cue PC to the outcome Ci.

On any single training trial, EXIT operates as follows. When the cues are presented, the input nodes are activated, and activation propagates to the attention gates via previously learned associations. The attentionally gated input activations are then propagated to the output nodes, which represent response options. The relative activations of the output nodes are converted to choice probabilities such that more highly activated nodes have a higher choice probability. When corrective feedback is provided, the network determines the discrepancy between the correct response and its output activations. This discrepancy is the error that the network reduces by shifting its attention and adjusting its associative weights. Before weights are adjusted, the network rapidly shifts its attention to the input cues to reduce the error. If attention to a cue causes error because the cue has been previously associated with a now-incorrect response, then attention is shifted away from that cue. If attention to a cue reduces error because it is already associated with the currently correct response, then attention is shifted toward that cue. Once attention has been shifted, then the weights from the input nodes to the attention gates are adjusted so that the distribution of attention will be better evoked by the current inputs on future trials, and the weights to the response nodes are adjusted so that the correct response will be better evoked on future trials. In summary, on any given trial, the model learns both a (covert) attentional response and a (overt) category/group response.

The model has several parameters, all of which have direct psychological interpretation. Three parameters involve attention. One parameter governs how large the attentional shift is. Another parameter governs how quickly the associative weights can learn to reproduce that shift. Another parameter determines how much a learned attentional allocation generalizes from one cue combination to another. Other parameters involve learning of output associations. One of these parameters determines how quickly the associative weights learn to reproduce correct responses. Another determines the overall capacity of the network to learn about several cues simultaneously. A final parameter has nothing to do with learning, but instead acts as a scaling parameter for the mapping from output node activations to response probabilities. Despite having several freely estimated parameters, the model is not infinitely flexible. As will be seen below, the model can nicely fit some aspects of our data but not others.

Previous research has shown that the EXIT model can accurately fit human behavior in the inverse base rate effect. It produces the effect by learning to attend to both cues I and PC when I.PC appears, but learning to ignore cue I when I.PR appears. The model therefore has a moderately strong association from I to Ci but not from I to Ri, and the model builds a strong association from PR to Ri. Therefore, when tested with cue I alone, the model tends to respond Ci, but when tested with PC.PR, the model tends to respond Ri.

Intuitively, it is plausible that EXIT could generate analogous learning in the illusory correlation structure that we used in Experiment 1. The notion is that EXIT would initially learn to associate both A and X with Cx, because $A.X \rightarrow Cx$ occurs most frequently. Then when cases of $B.X \rightarrow Rx$ occur, the model will shift attention away from cue X to cue B, and associate B with Rx. As we will see, it turns out that this armchair simulation of EXIT is only partly correct.

Eliminative inference in the ELMO model

The ELMO model is a formal implementation of the eliminative inference theory (Juslin et al., 2001). Although ELMO was shown to have serious shortcomings as a model of the inverse base rate effect (Kruschke, 2001a), there is no doubt that people will use eliminative inference in some situations (Kruschke & Bradley, 1995; Juslin et al., 2001). One such situation might be the illusory correlation procedure.

In the ELMO model, each training item has a certain probability of being learned and available in memory at the time of test. The probability is assumed to be monotonically related to the frequency with which the item occurred in training. For Experiment 1, for example, there are six possible rules that ELMO might have available at time of test, corresponding to the six possible instances that occurred during training. The rule $A.X \rightarrow Cx$ will have a higher probability of being available in memory than the rule $B.X \rightarrow Rx$. Notice that if a training item is in memory, then all of it is in memory; there are no fragments or selected components. This lack of cue selection contrasts with EXIT, which can selectively attend to cues within instances.

The emphasis of ELMO is not the learning but the decision process used at test. When a test item (a.k.a. the probe) appears, its cues are compared with all the known rule conditions. If the probe is similar enough to known rules, then the response is selected from among the outcomes associated with those rules. But if the probe is sufficiently different from all known rules, then all the outcomes associated with known rules are eliminated, and the response is selected from among the remaining outcomes for which no rules are known.

ELMO computes the overall probability of each response to a probe by averaging over all possible knowledge states, weighted by the probability of being in each knowledge state. For example, denote the probability of knowing rule R_1 by p_1 and the probability of knowing rule R_2 by p_2. Denote the state of knowing both rules as R_1R_2. Its probability_is p_1p_2. Denote the state of knowing R_1 but not R_2 as $R_1\overline{R}_2$. Its probability is $p_1(1-p_2)$, and so forth. ELMO computes the response probabilities for each knowledge state, and averages them together, weighted by the probability of being in each state.

ELMO has parameters for the probabilities of knowing the rules, and it also has parameters that govern how to compute the similarity of a probe to the rule conditions. The similarity parameters indicate how much each cue is weighted if it mismatches. The similarity parameter for a cue is assumed to indicate how much the cue covaries with the outcome; therefore cues with the same outcome correlations are given the same similarity weighting.

ELMO can generate some aspects of the inverse base rate

effect. In particular, consider what happens when ELMO is probed in the test phase with PC.PR. Because $PR.I \rightarrow Ri$ occurred rarely in training, the probability of knowing the rule for outcome *Ri* is relatively low. It turns out that PC.PR is not sufficiently similar to other rules for them to be used, so the known outcomes are eliminated, and outcome *Ri* is guessed instead. Thus, probe PC.PR tends to evoke a response of *Ri* because ELMO does *not* know an association to *Ri*.

A response mechanism like eliminative inference seems intuitively to apply to the situation of illusory correlation. If rules involving *Rx* are not known, then when probe *B* appears, *Rx* will be chosen because *Cx* has been eliminated.

Fits of the models

To determine the extent to which the models can fit the inverse base rate effect or illusory correlation, we weighted components of the data differently in different fits. The root mean squared error (RMSE) for the five probes involving traits from the inverse base rate effect is denoted E_I, and the RMSE for the seven probes involving illusory correlation is denoted E_X. The fitting algorithm found parameter values that minimized a weighted combination of the two errors: $E = w_I E_I + w_X E_X$. For the fit that emphasized the inverse base rate effect, we arbitrarily set $w_I = 0.99$ and $w_X = 0.01$. For the fit that emphasized illusory correlation, we set $w_I = 0.01$ and $w_X = 0.99$. [1]

The middle column of Figure 1 graphs the best fits of EXIT to the data from Experiment 1. (The human data were reported numerically in Table 2. EXIT's predicted choice percentages for all probe items are reported in Appendix Table A5, with the corresponding parameter values and the minimized errors in Table A1.) The upper middle panel of Figure 1 shows the best fit of EXIT when emphasizing the inverse base rate effect. The fit is superb, with an RMSE of only 1.40. Clearly EXIT continues to be an excellent model of the inverse base rate effect even when using social stimuli.

The lower middle panel of Figure 1 shows the best fit of EXIT when emphasizing illusory correlation. The fit here is notably poor. EXIT does not generate the crucial cross-over of the A and B test items.[2]

One might wonder if EXIT's poor fit to the illusory correlation results is due merely to the fact that illusory correlation uses a probabilistic mapping. Perhaps any probabilistic mapping is difficult for EXIT to accommodate. No, this is not the case, because some previous applications to probabilistic mappings have shown good fits. In particular, Kruschke (1996) emphasized that the predecessor of EXIT (called ADIT) showed good fits to data from probabilistic designs in which people showed apparent base rate neglect. Thus, the poor fit of EXIT to the illusory correlation results is quite notable when juxtaposed to good fits in other probabilistic designs.

The right column of Figure 1 graphs the best fits of ELMO to the data from Experiment 1. (The human data were reported numerically in Table 2. ELMO's predicted choice percentages for all probe items are reported in Appendix Table A6, with the corresponding parameter values and the minimized errors in Table A2.) The upper right panel of Figure 1 shows the best fit of ELMO when emphasizing the inverse base rate effect. The fit is poor. In particular, ELMO cannot show a difference between the I-alone and I.PC.PR items (and so the points for these items are superimposed in the graph). This problem and others were previously pointed out by Kruschke (2001a).

The lower right panel of Figure 1 shows the best fit of ELMO when emphasizing illusory correlation. The fit here is notably better than EXIT. In particular, ELMO generates the crucial cross-over of the A and B test items: For A alone, ELMO correctly predicts more Cx responses than Rx, but for B alone, ELMO correctly predicts more Rx responses than Cx.

In summary, both models have difficulties accurately fitting the data from both effects simultaneously. EXIT accurately fits the data from the inverse base rate effect while it does not fit the data from illusory correlation. ELMO has the opposite qualities, fitting illusory correlation better than the inverse base rate effect. Because the models implement different mechanisms, the fits suggest that the two effects are caused (at least in part) by different mechanisms. Thus, our conclusion from the modeling corroborates our interpretation of the lack of correlation between the inverse base rate effect and illusory correlation.

Experiment 2: Illusory correlation with no corrective feedback for the rare trait

The design of Experiment 1 produced robust illusory correlation and inverse base rate effect, but no notable covariation between them. Fits by models also suggested that attentional shifting nicely addresses the inverse base rate effect but does not fare as well with illusory correlation, whereas

[1] We used RMSE as a measure of fit, rather than likelihood and its variants such as AIC or BIC, for two reasons. First, likelihood strongly emphasizes aspects of the data in which we are least interested, namely, predicted choice probabilities near zero (e.g., the probability of choosing group Cx when presented with trait I). Second, the predictions based on minimized RMSE turn out to be dramatically and qualitatively different across models, so no statistical hypothesis testing is needed to be confident that the models are reliably different in their behaviors.

[2] The associative weights that EXIT learns for illusory correlation tend to fall into two distinct clusters, rather than in one tight cluster as is the case for the inverse base rate effect. In the more frequent of the clusters, the weight from X to Cx is strongly positive, and both A and B are more positively associated with Rx than with Cx. For these weights, the network tends to respond Rx for both B by itself and for A by itself. In other words, simulated subjects in this cluster do not show the A-by-B cross-over shown in the average human data. In a second cluster of weights, X is more strongly associated with Rx than Cx, A is more strongly associated with Cx than Rx, and B is a little more strongly associated with Cx than Rx. For these weights, the network tends to respond Cx for both A by itself and for B by itself. Again, simulated subjects in this cluster do not show the A-by-B cross-over shown in the average human data.

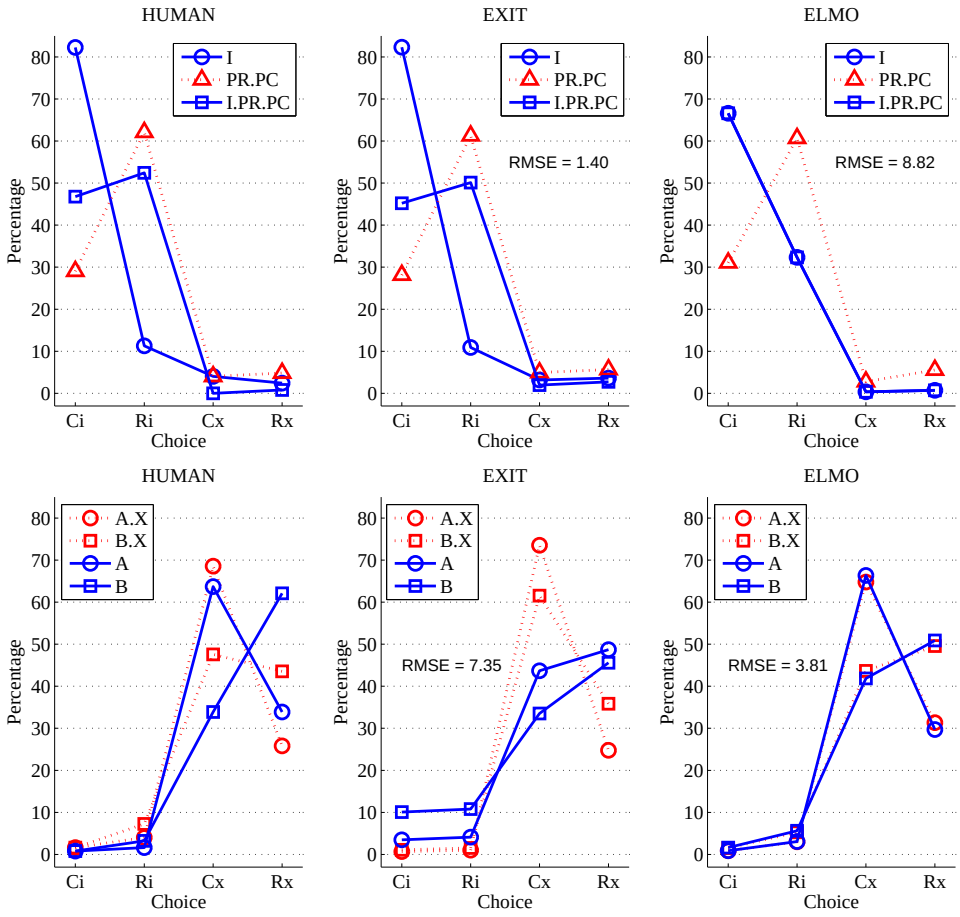

Figure 1. Best fits of models. Left column shows selected human data. Middle column shows behavior of the EXIT model when fit primarily to the data from the inverse base rate effect (upper row) or fit primarily to the data from illusory correlation (lower row). Right column shows behavior of the ELMO model when fit primarily to the data from the inverse base rate effect (upper row) or fit primarily to the data from illusory correlation (lower row). RMSE = root mean squared error. The RMSE's in the upper panels refer to the fit to all five test items involving cues from the inverse base rate effect. The RMSE's in the lower panels refer to the fit to all seven test items involving cues from illusory correlation.

the response strategy of eliminative inference better accommodates illusory correlation than the inverse base effect.

The explanation of illusory correlation in terms of eliminative inference goes as follows. People first learn an association of trait A with group Cx, because that group occurs so often. But people do not learn as well about trait B (or group Rx) because it occurs more rarely. Therefore, when tested with trait B, people reason that because trait B is something they do not know, it must correspond to a group they do not know, namely, the rare group Rx. This response strategy does not depend on any learned association between B and Rx; instead, it uses only a process of eliminating other groups that do have associations.

An implication of the explanation by eliminative inference is that we should be able to observe illusory correlation even when there is no corrective feedback for trait B. That is, any time trait B is presented, we simply withhold the feedback, as if the trial were a test trial. The design is identical to Experiment 1 (Table 1) except that for the trials in which B.X is presented in Experiment 1, participants in Experiment 2 get no corrective feedback and merely proceed to the next trial. We predict that people should exhibit illusory correlation in the test phase, despite the fact that they have never received any explicit training for trait B.

Method

Participants. 95 students from introductory psychology courses at Indiana University in Fall 2002 volunteered for partial course credit.

Design and procedure. All details of Experiment 2 were identical to Experiment 1 except that for B.X training trials there was no corrective feedback; the participant was shown

Table 4
Percentage of group classifications in Experiment 2.

	Group Chosen			
Traits	Ci	Ri	Cx	Rx
I.PC	91	3	2	4
I.PR	20	71	6	4
I	81	15	3	1
PC.PR	39	57	2	2
I.PC.PR	63	32	4	1
X.A	3	4	69	24
X.B	6	21	27	46
A	4	7	58	31
B	4	26	19	50
X	5	4	63	28
A.B	4	21	26	48
X.A.B	2	8	51	39

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., "PC.I" means that trait PC and trait I were presented together.

merely the feedback for a testing trial, i.e., "Your response has been recorded. Press the space bar to continue."

Results and discussion

The procedure for data analysis used for Experiment 1 was applied exactly to Experiment 2. 68 participants met the learning criterion on the training items of the inverse base rate effect.

Classification. Table 4 shows the percentage with which each group was chosen for each combination of traits in the test phase. Recall that each test item was presented twice in the test phase; therefore each row of Table 4 is based on $2 \times 68 = 136$ responses.

The first two rows of Table 4 reveal that accuracy on item I.PC was 91% and on I.PR was 71%. The accuracy on I.PR is lower than in Experiment 1, and lower than many results in the literature regarding the inverse base rate effect. Nevertheless, the data are revealing, as will be shown.

The inverse base rate effect is strongly exhibited. Participants classified trait I by itself as group Ci 81% of the time, and as group Ri only 15% of the time. This preference is reliability different from 50-50, $\chi^2(1, N{=}130) = 62.31$, $p < 0.001$. On the other hand, for trait pair PC.PR, participants strongly preferred group Ri (57%) over group Ci (39%). This preference, though weaker than in Experiment 1, is also reliably different from 50-50, $\chi^2(1, N{=}130) = 4.43$, $p = 0.035$. Thus, the inverse base rate effect is clearly in evidence. Why it is a bit weaker than Experiment 1 could be due to random variation from one set of subjects to the next (this experiment was run near the end of the semester, whereas Experiment 1 was run near the beginning of the semester), or due to possibly more confusion caused by lack of feedback on some training trials.

Table 5
Mean ratings of traits for groups in Experiment 2.

	Group			
Traits	Ci	Ri	Cx	Rx
I	7.34	5.40	–	–
PC	7.64	3.96	–	–
PR	2.97	7.39	–	–
X	–	–	6.78	6.01
A	–	–	6.59	5.96
B	–	–	4.65	5.89

Note: Cells with "–" indicate trait-group combinations that were not included.

For the items that probe illusory correlation, notice that responding to trait pair X.A was close to probability matching, with 69% Cx choices and 24% Rx choices. Interestingly, responding to traits X.B is not uniform. There is a clear preference (46%) for group Rx over group Cx (27%), despite the fact that participants were never given any feedback for cases involving trait B. The preference for Rx over Cx is reliable, $\chi^2(1, N{=}100) = 6.76$, $p = 0.009$. Also interesting is that if participants did not respond with Rx or Cx, then the next most prevalent response was Ri (21%), much more strongly than Ci (6%). The preference for Ri over Ci is also reliable, $\chi^2(1, N{=}36) = 11.11$, $p = 0.001$. These preferences for the rare groups over the frequent groups are predicted by the eliminative inference theory: Because trait B is uncertain, the known groups, Cx and Ci, tend to be eliminated, leaving the rare groups, Rx and Ri, as candidate responses. As Ri is learned somewhat, it is also eliminated sometimes, leaving Rx as the most frequently selected group.

The crucial assay of illusory correlation is a comparison of trait A by itself with trait B by itself. For trait A, participants preferred Cx (58%) over Rx (31%), but for trait B that preference was reversed, as participants preferred Rx (50%) over Cx (19%). This difference of preferences is reliable, $\chi^2(1, N{=}215) = 29.98$, $p < 0.001$. Thus, illusory correlation is robust, even when trait B was never given explicit group membership.

Trait in group ratings. Table 5 shows mean ratings of traits for groups. The first row reveals that group Ci was rated as more "I-ish" than group Ri, with mean ratings of 7.34 vs. 5.44, $t(65) = 7.21$, $p < 0.001$, $d = 0.89$. This stronger rating of trait I in group Ci than group Ri echoes the results seen for classification of trait I, which also strongly favored group Ci.

The next two rows of Table 5 show that group Ci was rated as more "PC-ish" than group Ri, but group Ri was more "PR-ish" than group Ci. These differences simply indicate that ratings reflect the correct classifications learned during training. The ratings show that PR is rated more strongly for Ri than PC is rated for Ci. That is, $r(PR{\in}Ri){-}r(PR{\in}Ci)$ is larger than $r(PC{\in}Ci){-}r(PC{\in}Ri)$. An interaction contrast was computed for each individual participant, $[r(PR{\in}Ri){-}r(PR{\in}Ci)]{-}[r(PC{\in}Ci){-}r(PC{\in}Ri)]$, and the mean of

this contrast was tested for being greater than zero. It was significant, $t(61) = 0.77$, $p = 0.024$, $d = 0.29$.

This significantly stronger rating for the rare trait has interesting ramifications for social cognition: Associations of rare traits with rare groups can be stronger than associations of common traits with common groups. Thus, if social stereotypes are created (at least in part) by the mechanism of attention shifting, rare traits might be especially strongly associated with their groups.

The ratings show a strong illusory correlation for traits A and B. Group Cx is rated as more A-ish than group Rx, yet group Rx is rated as more B-ish than group Cx. The interaction contrast $(r(A{\in}Cx)-r(A{\in}Rx))-(r(B{\in}CX)-r(B{\in}Rx))$ is reliably greater than zero, $t(64) = 3.27$, $p < 0.002$, $d = 0.41$. As was found for the classification preferences, we see again for the trait ratings that a robust illusory correlation is obtained for this novel training paradigm, even when trait B is never explicitly trained as a member of any group.

Correlations of effects. As explained for Experiment 1, if illusory correlation and the inverse base rate effect are caused, at least in part, by some common underlying mechanism, and if that mechanism varies in strength across individuals, then the magnitude of illusory correlation and inverse base rate effects should covary across individuals. The correlation of the inverse base rate effect and the illusory correlation effect, as measured by choices, was not reliably different from zero, and even weakly negative: $r(df{=}61) = -0.190$, $p = 0.136$ (two-tailed). The correlation of the inverse base rate effect and the illusory correlation effect, as measured by ratings, was also not reliably different from zero and weakly negative: $r(df{=}56) = -0.212$, $p = 0.109$ (two-tailed). Although there were not significant correlations across effects, there were significant correlations within effects: The choice and rating measures of the inverse base rate effect were significantly correlated, $r(df{=}58) = 0.268$, $p = 0.039$ (two-tailed). The choice and rating measures of illusory correlation were strongly correlated, $r(df{=}58) = 0.538$, $p < 0.001$ (two-tailed).

These correlations of choice and rating measures within effects, and lack of correlations across effects, echo the results of Experiment 1. The presence of significant correlations within effects indicates that the measures are robust enough to detect individual differences. The lack of correlation across effects suggests that individual variations in the two effects do not have a strong common component, because if individual variation in the inverse base rate effect and illusory correlation had a common source, there should detectable covariation.

Group likability ratings. Participants were also asked how likable each group was. Mean likabilities were as follows: 5.82 for group Ci, 4.99 for Ri, 5.23 for Cx, and 5.00 for Rx. The ratings for Ci and Ri were reliably different, $t(63) = 3.67$, $p < 0.001$, $d = 0.46$. This difference is consistent with the positively valanced trait I being more strongly associated with group Ci than group Ri. Evidently, however, trait X was not strongly enough associated with group Cx to make it much more likable than Rx. This relative weakness of X compared with I is consistent with the classification preferences and trait ratings.

Modeling the results of Experiment 2

The same modeling procedure was used to fit the data from Experiment 2 as for Experiment 1. In particular, parameter searches were conducted that emphasized fit to the data from the inverse base rate effect or from illusory correlation.

The middle column of Figure 2 graphs the best fits of EXIT to the data from Experiment 2. (The human data were reported numerically in Table 4. EXIT's predicted choice percentages for all probe items are reported in Appendix Table A7, with the corresponding parameter values and the minimized errors in Table A3.) The upper middle panel of Figure 2 shows the best fit of EXIT when emphasizing the inverse base rate effect. The fit is fairly good, with an RMSE of 3.45.

The lower middle panel of Figure 2 shows the best fit of EXIT when emphasizing illusory correlation. The fit here is notably poor. EXIT does not generate the crucial cross-over of the A and B test items. Indeed, EXIT shows essentially uniform responding when trait B is presented by itself, because there has been no learned association between B and any group.[3]

The right column of Figure 2 graphs the best fits of ELMO to the data from Experiment 2. (The human data were reported numerically in Table 4. ELMO's predicted choice percentages for all probe items are reported in Appendix Table A8, with the corresponding parameter values and the minimized errors in Table A4.) The upper right panel of Figure 2 shows the best fit of ELMO when emphasizing the inverse base rate effect. The fit is poor. Again, ELMO cannot show any difference between the I-alone and I.PC.PR items. This problem and others were previously pointed out by Kruschke (2001a).

The lower right panel of Figure 2 shows the best fit of ELMO when emphasizing illusory correlation. The fit here is remarkably better than EXIT. In particular, ELMO generates the crucial cross-over of the A and B test items: For A alone, ELMO correctly predicts more Cx responses than Rx, but for B alone, ELMO correctly predicts more Rx responses than Cx. ELMO also captures for trait B the larger choice of Ri relative to Ci. ELMO is able to show these trends despite the lack of explicit training on trait B because the model is eliminating the better-known Cx and Ci groups, leaving the Rx and Ri groups as the preferred choices.

In summary, the fits of the models to data from Experiment 2 show clearly that attentional learning alone cannot account for the illusory correlation exhibited by people. The fits also suggest that a response strategy such as eliminative

[3] The tiny variation from uniform responding to trait B is caused by a very weak influence of associations from the bias node in EXIT. The best fitting parameter values gave the bias node a very small salience; see Table A3.

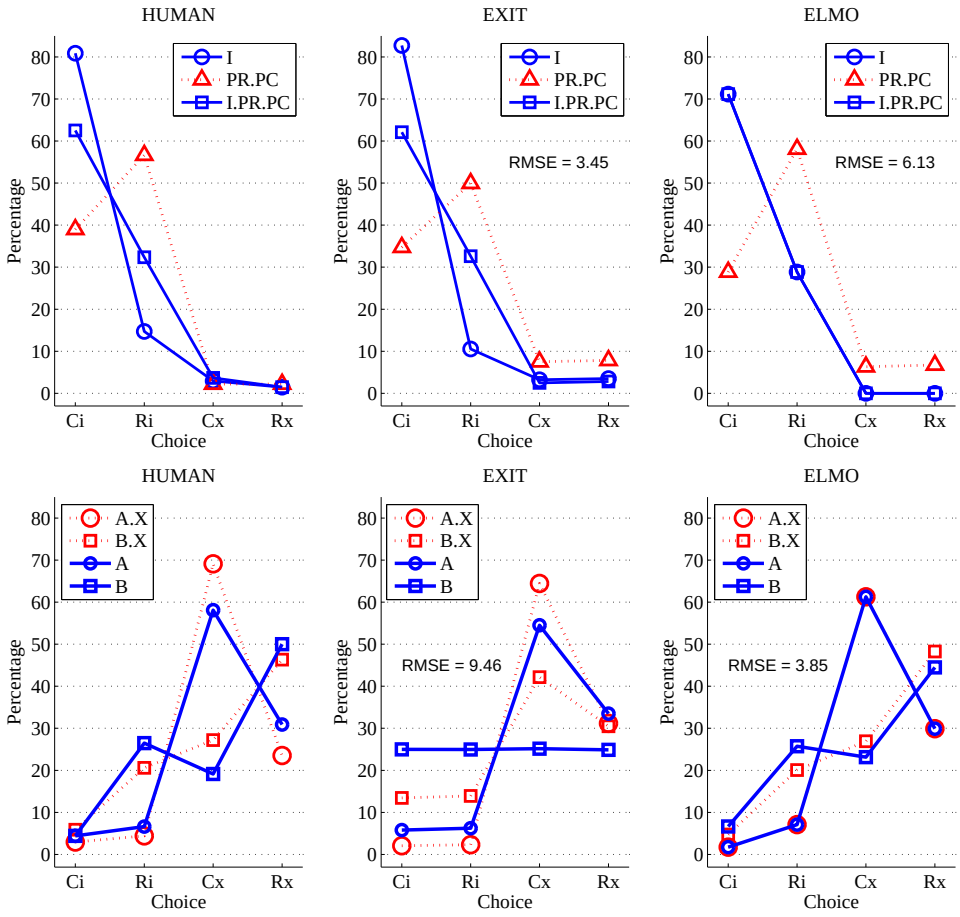

Figure 2. Best fits of models. Left column shows selected human data. Middle column shows behavior of the EXIT model when fit primarily to the data from the inverse base rate effect (upper row) or fit primarily to the data from illusory correlation (lower row). Right column shows behavior of the ELMO model when fit primarily to the data from the inverse base rate effect (upper row) or fit primarily to the data from illusory correlation (lower row). RMSE = root mean squared error. The RMSE's in the upper panels refer to the fit to all five test items involving cues from the inverse base rate effect. The RMSE's in the lower panels refer to the fit to all seven test items involving cues from illusory correlation.

inference is a viable candidate for explaining illusory correlation in this task procedure. Finally, the fits recapitulate arguments made by Kruschke (2001a) that attentional learning is a crucial component of the inverse base rate effect, but eliminative inference is not the central mechanism underlying that effect.

General Discussion

Summary

A summary of this project's empirical and theoretical contributions:

- The inverse base rate effect occurs for social (trait-group) stimuli. Previous research on the inverse base rate effect (except for Wedell & Kruschke, 2006) used non-social stimuli such as symptoms and diseases, random words, or simple geometric figures.
- The inverse base rate effect is manifest in trait-group ratings and likability ratings. Previous research used choice proportion as the only measure of the effect.
- Because the inverse base rate effect occurs with social stimuli, it could be an effect contributing to social stereotyping. Indeed, Wedell and Kruschke (2006) found that the inverse base rate effect is strongest when the rare trait is negative and the frequent trait is positive.
- Attentional shifting is a favored explanation of the inverse base rate effect, and is a new theoretical approach to mechanisms of stereotyping. Neither distinctiveness (i.e., attention to rare items) nor eliminative inference adequately account for the inverse base rate effect, but attentional shifting does. Attentional shifting also suggests that the strength of association between the rare trait and the rare group, i.e., the stereotyping, can be stronger than the strength of association between the common trait and the common group.

• Illusory correlation occurs in a trait-group predictive training procedure. Previous research used just exposure to cases.

• Illusory correlation occurs when there is a shared trait in all stimuli. Previous research had no explicit shared trait.

• Illusory correlation occurs when both of the differentiating traits are of positive valence and are from distinct dimensions. Most previous research has used a frequent positive trait with an infrequent negative trait, and typically the two traits are polar opposites from a single dimension.

• Illusory correlation occurs even when there is no explicit group label on the rare trait.

• Importantly, the inverse base rate effect and illusory correlation were tested in a within subjects design, and the magnitudes of the inverse base rate effect and illusory correlation did not covary across individuals, either for choice preferences or ratings. This lack of covariation contrasts with the reliable covariations found between the inverse base rate effect (a.k.a. highlighting) and associative blocking (Kruschke et al., 2005). The lack of covariation between the inverse base rate effect and illusory correlation suggests that different mechanisms underlie the phenomena.

• Two mathematical models were quantitatively fit to the choice data. Few if any previous studies have performed quantitative model fitting.

• The attentional shifting model (Kruschke, 2001a) fit the the inverse base rate effect data well but did not fit the illusory correlation data. The eliminative inference model (Juslin et al., 2001) fit the illusory correlation data fairly well but did not fit the the inverse base rate effect data. These fits also suggest that the two phenomena are produced, at least in part, by different mechanisms.

Relation to previous theories

Smith (1991) used an exemplar-based model (Hintzman, 1986) to address aspects of illusory correlation. Unfortunately, the model cannot account for the results of our Experiment 2. The exemplar memory model needs exemplars of trait-group co-occurrences to remember, and the procedure of Experiment 2 provided no cases of the rare trait explicitly co-occurring with any group. Exemplar models also cannot produce the inverse base rate effect (Kruschke, 1996). Fiedler (2000) presented an associative learning model that can address many findings in the literature on illusory correlation, but that model also requires co-occurrences of traits and groups to associate, and therefore cannot address the results of our Experiment 2. The mechanisms proposed by these learning models might well be at work in producing some aspects of human behavior, but these mechanisms cannot be the primary causes of the inverse base rate effect or illusory correlation in our experiments.

In another article, Fiedler (1991) argued that much of illusory correlation can be explained by loss of information about the rare group. This alternative, lack-of-learning approach can explain how the rare group can be judged to be less A-ish than the frequent group, but the approach cannot explain how the rare group can be judged to be significantly B-ish, as we found in our experiments. That is, information loss predicts that the rare group will be judged more neutrally than the common group, but does not predict that the rare group will be judged opposite from the common group.

Hamilton and Gifford (1976) proposed that illusory correlation was caused by distinctive, i.e., rare, traits being strongly attended to and associated with distinctive, i.e., rare, groups. A similar idea was formalized in the "attention" model of Shanks (1992), which gives higher learning rates to infrequent cues. In that model, on a given training trial, the change in an association's strength is proportional to the degree to which the group is mis-predicted, as in the classic Rescorla-Wagner (1972) model. The constant of proportionality is the learning rate. The learning rate is cue-specific and set higher for less frequently occurring cues. In particular, the learning rate for cue PR is higher than the learning rate for cue I. The model addresses some aspects of the inverse base rate effect, but was shown not to account very well for results from closely related experiments (Kruschke, 1996). The model also fails to account for the results from our Experiment 2 because, like EXIT, the model relies on explicit learning trials in which the traits and groups are presented together. The mechanism of enhanced attention to rare cues is probably a real influence in human performance, but our experiments and modeling suggest that it is not enough to explain the inverse base rate effect and illusory correlation here.

McGarty et al. (1993) presented experiments, akin to our Experiment 2, in which no explicit pairings of traits with groups were presented at all. In their Experiment 2, participants were told merely that group B had fewer members than group A, and participants were then shown a series of positive or negative behaviors in which there were more positive than negative cases. In subsequent tests, participants indicated that the larger group A had more positive traits and the smaller group B had more negative traits. McGarty et al. (1993) argued that the framing of the task induces participants to "seek meaning" or "make sense" of the situation by conceiving two competing hypotheses (H1 and H2): Either (H1) large group A is good and small group B is bad, or (H2) large group A is bad and small group B is good. The probability of the data (i.e., more good statements than bad) is higher given H1 than given H2. Therefore, from a Bayesian perspective, the probability of H1 given the data is higher than the probability of H2 (assuming that the priors do not favor H2). This approach might also account for the results of our Experiment 2, if we assume that the participants have generated hypotheses that assign different traits (such as A and B) to different groups (such as Cx and Rx).

What is missing from their theory is exactly how the framing of the task induces the participants to set up specific competing hypotheses. The hypotheses needed to account for our Experiment 2 could be (H3) Cx is A-ish and Rx is B-ish, or (H4) Cx is B-ish and Rx is A-ish, but the genesis of these hypotheses is unclear. Study 2 of Berndsen, McGarty, van der Pligt, and Spears (2001) had participants think aloud as they worked through a standard illusory correlation task, and found that subjects do spontaneously set up hypothe-

ses about types of traits and about assignments of traits to groups. Therefore, hypothesis induction certainly can take place; the specific mechanisms of hypothesis induction are yet to be specified.

As further confirmation of their meaning-seeking explanation, Berndsen et al. (2001) conducted an experiment analogous to our Experiment 2. In the "behavior-constrained" condition of their Study 1, they showed participants only the 36 cases of the frequent positive trait, with no cases of the infrequent negative trait. They also told participants that there were 18 cases of negative traits that would not be displayed. Robust illusory correlation occurred in all assessments. The results of our Experiment 2 are entirely compatible with theirs: People produce an illusory correlation even without explicit co-occurrences of the rare trait with the rare group.[4]

An explanation of illusory correlation in terms of eliminative inference has some similarities to McGarty and Berndsen et al.'s explanation in terms of searching for meaningful differences. Just as a participant's search for meaningful differences sets up a mental space of hypotheses, the operation of eliminative inference assumes that the participant sets up an analogous mental hypothesis. This mental hypothesis assigns distinct traits to distinct groups, so that when unknown traits (or trait combinations) are presented, the known trait-group associations are eliminated. Eliminative inference only operates in the context of a constrained hypothesis space of the sort that McGarty et al. suggest. What we have done is shown that a formal model of eliminative inference (Juslin et al., 2001) can quantitatively fit human performance.

The meaning-seeking and eliminative-inference theories are quite different, however, in that one emphasizes learning and the other emphasizes guessing. In meaning seeking, participants are learning which hypotheses to believe in. As cases are presented, participants actively change their beliefs. In illusory correlation, ultimately one of the explicit beliefs is that the rare trait goes with the rare group. In eliminative inference, on the other hand, participants are assumed to learn some cases but are also assumed not to learn other cases. In illusory correlation, it is specifically a *lack* of knowledge about the rare trait and group that generates guesses that the rare trait indicates the rare group.

Ultimately, it is likely that multiple underlying mechanisms conspire to produce overt behaviors such as illusory correlation and the inverse base rate effect. Despite the fact that our Experiment 2 prohibited simple associative mechanisms between the rare trait and the rare group, it is likely that associative learning mechanisms do play a role when they can be used. For example, Johnson and Mullen (1994) found that participants were faster when classifying rare traits (into rare groups) than when classifying frequent traits. Such faster processing of the rare traits is not naturally explained by eliminative inference. Stroessner, Hamilton, and Mackie (1992, Experiment 2) found that (neutral mood) participants who spontaneously studied cases of rare traits and rare cases longer during training also tended to show higher illusory correlations during subsequent frequency judgments. This correlation implies that learning, not just lack of learning, can influence illusory correlation.

Conclusion

We began this research project with the hypothesis that attention shifting and attention learning, which have been shown to play a strong role in the inverse base rate effect, might also play a pivotal role in illusory correlation. We therefore created experiments in which the two effects have analogous structures and are learned simultaneously in a within-subjects design. We found robust effects, but zero correlation between them, which suggested that different mechanisms underlie the effects.

Another candidate mechanism for illusory correlation was suggested, namely eliminative inference. This mechanism does not rely on learning trait-group associations, but instead relies on lack of learning: A trait that is recognized as unlearned is assigned to a group that is known to be unlearned. Experiment 2 provided learners with no group labels regarding the rare trait, yet people exhibited a strong illusory correlation effect.

Modeling revealed that the inverse base rate effect is better fit by attentional shifting and learning than by eliminative inference, whereas illusory correlation is better fit by eliminative inference than by attentional shifting and learning. Both effects are probably caused by multiple mechanisms in humans. Our claim is that, at least in our experiments, the dominant causes of the two effects are different.

Finally, one of the important motivations for studying these effects in the lab is that they might have real-world analogues. McGarty and de la Haye (1997) cautioned that illusory correlation in the laboratory might be distinct from stereotyping in the real world, but concluded that at least some mechanisms that generate laboratory illusory correlation also operate in real-world stereotype formation. We have argued here that attentional shifting and learning is important to account for the inverse base rate effect using traits and groups, and we believe that attentional shifting and learning is also at work in real-world social learning. We have also argued here that other mechanisms are important to account for

[4] Our Experiment 2 differed from Study 1 of Berndsen et al. (2001) in many aspects. Their stimulus cases exhibited just one behavior, which was either positive or negative, whereas our cases exhibited two explicit traits (A and X, or B and X) one of which was negative and one of which was positive. There were also several procedural differences between experiments. Importantly, their participants did not experience any cases of the rare trait, whereas our participants did experience the cases but without feedback regarding group membership. Their participants merely studied the cases, whereas ours predicted group membership for each case before being given corrective feedback. Their stimuli were descriptions of unique behaviors, our stimuli were repeated single-word traits. Their traits had more positive than negative cases, our traits had an equal number of positive and negative cases. Finally, the experiments also differed in how they assessed illusory correlation. In particular, we obtained ratings regarding each of the presented traits with respect to each group.

the illusory correlation effect, and we believe that these other mechanisms are also at work in real-world social learning.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*, 272–279.

Berndsen, M., McGarty, C., van der Pligt, J., & Spears, R. (2001). Meaning-seeking in the illusory correlation paradigm: The active role of participants in the categorization process. *British Journal of Social Psychology*, *40*, 209–233.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic signs. *Journal of Abnormal Psychology*, *72*, 193–204.

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

Fagot, J., Kruschke, J. K., Dépy, 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.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *60*, 24–36.

Fiedler, K. (2000). Illusory correlations: A simple associative algorithm provides a convergent account of seemingly divergent paradigms. *Review of General Psychology*, *4*, 25–58.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in intergroup perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*, 392–407.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Johnson, C., & Mullen, B. (1994). Evidence for the accessibility of paired distinctiveness in distinctiveness-based illusory correlation in stereotyping. *Personality and Social Psychology Bulletin*, *20*(1), 65–70.

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 & Cognition*, *29*(4), 587–597.

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

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

Kruschke, J. K. (2001a). 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. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812-863.

Kruschke, J. K. (2003a). 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. (2003b). Attention in learning. *Current Directions in Psychological Science*, *12*, 171–175.

Kruschke, J. K. (2005). Learning involves attention. In G. Houghton (Ed.), *Connectionist models in cognitive psychology.* London: Psychology Press.

Kruschke, J. K. (submitted). Locally Bayesian learning. *Proceedings of the Annual Conference of the Cognitive Science Society*.

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for associative learning.* (Indiana University Cognitive Science Research Report #141. Available via WWW at http://www.indiana.edu/ ~kruschke/ articles/ KruschkeB1995.pdf)

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 830–845.

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.

McGarty, C., & de la Haye, A.-M. (1997). Stereotype formation: Beyond illusory correlation. In R. Spears (Ed.), *The social psychology of stereotyping and group life* (pp. 144–170). Malden, MA: Blackwell.

McGarty, C., Haslam, S. A., Turner, J. C., & Oakes, P. J. (1993). Illusory correlation as accentuation of actual intercategory difference: Evidence for the effect with minimal stimulus information. *European Journal of Psychlogy*, *23*, 391–410.

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.

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

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

Smith, E. R. (1991). Illusory correlation in a simulated exemplar-based memory. *Journal of Experimental Social Psychology*, *27*, 107–123.

Stroessner, S. J., Hamilton, D. L., & Mackie, D. M. (1992). Affect and stereotyping: The effect of induced mood on distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *62*(4), 564–576.

Stroessner, S. J., & Plaks, J. E. (2001). Illusory correlation and stereotype formation: Tracing the arc of research over a quarter century. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The princeton symposium on the legacy and future of social cognition* (pp. 247–259). Mahwah, NJ: Erlbaum.

Wedell, D. H., & Kruschke, J. K. (2006). Consequences of competitive learning on social preference. **, **, **. (In preparation.)

Winman, A., Wennerholm, P., Juslin, P., & Shanks, D. R. (2005). Evidence for rule-based processes in the inverse base-rate effect. *The Quarterly Journal of Experimental Psychology*, *58A*(5), 789–815.

Appendix
Best fitting parameter values

Best fitting parameter values, with corresponding minimized root-mean squared error (RMSE), are shown in Tables A1, A3, A2 and A4. The root mean squared error

Table A1
Parameter values for best fits by EXIT to Experiment 1, corresponding to predictions in Table A5.

Parameter	IBRE	Illus.Corr	All
c	0.268	0.0277	0.1243
P	4.023	5.9016	5.4212
ϕ	6.272	6.3033	5.7139
λ_g	2.663	1.5305	1.7845
λ_w	0.025	0.0391	0.0363
λ_x	0.101	0.0227	0.0328
σ	0.042	0.041	0.0262
E_I	1.398	4.038	2.896
E_X	10.112	7.353	7.545
Wtd RMSE	1.485	7.320	5.608

(RMSE) for the five probes involving traits from the inverse base rate effect is denoted E_I, and the RMSE for the seven probes involving illusory correlation is denoted E_X. The fitting algorithm found parameter values that minimized a weighted combination of the two errors: $E = w_I E_I + w_X E_X$. For the fit that weighted all probe items equally, $w_I = 5/12$ and $w_X = 7/12$ because E_I is generated by 5 of 12 probes and E_X is generated by 7 of 12 probes. For the fit that emphasized the inverse base rate effect, $w_I = 0.99$ and $w_X = 0.01$. For the fit that emphasized illusory correlation, $w_I = 0.01$ and $w_X = 0.99$.

In the EXIT model there are seven parameters. The notation follows the usage of Kruschke (2001a). The predictions of EXIT were determined by averaging over 50 simulated subjects, each with a different random training order that exactly respected the number of trials in each experiment. This number of simulated subjects yielded stable average model predictions.

For the ELMO model, there are two types of parameters, namely, the insignificance of each feature for similarity computation, and the probability of knowing a trait-to-group rule. The notation s_T indicates the insignificance of trait T. The notation $p_{T\to G}$ indicates the probability of knowing rule $T \to G$. The number of parameters depends on the number of traits and groups in the particular training scenario. Again, this notation follows the usage of Kruschke (2001a). The fits were constrained such that $p_{A.X\to Cx} \geq p_{A.X\to Rx} = p_{B.X\to Cx} \geq p_{B.X\to Rx}$, to respect the relative frequencies of the training cases. ELMO does not use trial-by-trial learning, so it does not have different simulated subjects.

Table A2
Parameter values for best fits by ELMO to Experiment 1, corresponding to predictions in Table A6.

Parameter	IBRE	Illus.Corr	All
$s_{PC} = s_{PR}$	0.008	0.249	0.217
s_I	0.224	0.557	0.580
$s_X = s_A = s_B$	1.000	0.946	0.964
$p_{PC.I\to Ci}$	0.791	0.908	1.000
$p_{PR.I\to Ri}$	0.448	0.695	0.757
$p_{A.X\to Cx}$	1.000	0.463	0.462
$p_{A.X\to Rx} = p_{B.X\to Cx}$	0.903	0.096	0.095
$p_{B.X\to Rx}$	0.903	0.096	0.095
E_I	8.817	11.748	10.501
E_X	9.479	3.811	3.969
Wtd RMSE	8.823	3.891	6.691

Table A3
Parameter values for best fits by EXIT to Experiment 2, corresponding to predictions in Table A7.

Parameter	IBRE	Illus.Corr	All
c	0.1022	0.0003	0.1009
P	5.49	2.7536	5.5073
ϕ	7.5721	5.4771	6.5152
λ_g	1.9459	2.2632	2.0203
λ_w	0.0134	0.0503	0.014
λ_x	0.2446	0.1249	0.2628
σ	0.0209	0.0045	0.0213
E_I	3.446	20.729	4.057
E_X	13.002	9.462	11.729
Wtd RMSE	3.542	9.574	8.532

Table A4
Parameter values for best fits by ELMO to Experiment 2, corresponding to predictions in Table A8.

Parameter	IBRE	Illus.Corr	All
$s_{PC} = s_{PR}$	0.512	0.404	0.443
s_I	0.210	0.996	0.759
$s_X = s_A = s_B$	0.529	0.896	0.932
$p_{PC.I\to Ci}$	1.000	0.852	0.932
$p_{PR.I\to Ri}$	0.577	0.455	0.512
$p_{A.X\to Cx}$	0.661	0.501	0.578
$p_{A.X\to Rx} = p_{B.X\to Cx}$	0.641	0.187	0.276
$p_{B.X\to Rx}$	n/a	n/a	n/a
E_I	6.129	8.896	7.505
E_X	11.166	3.849	4.325
Wtd RMSE	6.179	3.899	5.650

Table A5
Best fits of EXIT to data from Experiment 1 (Table 2). Corresponding parameter values are shown in Table A1.

	Fit Mainly To Inverse Base Rate Probes				Fit Mainly To Illusory Correlation Probes				Fit Equally To All Data			
	Group Chosen				Group Chosen				Group Chosen			
Traits	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx
I.PC	99	1	0	0	99	0	0	0	99	1	0	0
I.PR	7	87	3	3	2	96	1	1	4	92	2	2
I	82	11	3	4	79	15	3	3	76	15	4	5
PC.PR	28	61	5	6	22	70	4	4	26	65	4	5
I.PC.PR	45	50	2	3	43	54	1	1	47	50	2	2
X.A	1	1	81	17	1	1	74	25	1	2	74	23
X.B	1	2	69	27	1	2	62	36	2	2	62	34
A	6	7	50	37	3	4	44	49	5	6	43	46
B	15	15	35	36	10	11	34	46	12	13	31	44
X	2	2	73	23	1	1	65	32	2	2	66	30
A.B	4	5	48	43	2	2	42	54	3	4	41	53
X.A.B	1	1	76	22	0	1	60	39	1	1	63	35

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., “PC.I” means that trait PC and trait I were presented together.

Table A6
Best fits of ELMO to data from Experiment 1 (Table 2). Corresponding parameter values are shown in Table A2.

	Fit Mainly To Inverse Base Rate Probes				Fit Mainly To Illusory Correlation Probes				Fit Equally To All Data			
	Group Chosen				Group Chosen				Group Chosen			
Traits	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx
I.PC	92	5	1	2	89	4	3	4	97	3	0	0
I.PR	5	88	2	5	4	73	10	13	3	78	8	10
I	67	32	0	1	60	38	1	1	62	38	0	0
PC.PR	31	61	3	6	34	39	12	16	38	44	8	10
I.PC.PR	67	32	0	1	60	38	1	1	62	38	0	0
X.A	0	0	52	48	1	3	65	31	0	2	65	32
X.B	0	0	52	48	2	5	44	50	0	4	45	51
A	0	0	55	45	1	3	66	30	0	2	67	30
B	0	0	50	50	2	6	42	51	0	5	43	52
X	0	0	52	48	1	2	65	32	0	2	66	32
A.B	0	0	52	48	2	6	43	49	0	5	45	51
X.A.B	0	0	52	48	1	2	65	32	0	2	66	32

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., “PC.I” means that trait PC and trait I were presented together.

Table A7
Best fits of EXIT to data from Experiment 2 (Table 4). Corresponding parameter values are shown in Table A3.

	Fit Mainly To Inverse Base Rate Probes				Fit Mainly To Illusory Correlation Probes				Fit Equally To All Data			
	Group Chosen				Group Chosen				Group Chosen			
Traits	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx
I.PC	99	1	0	0	99	0	0	0	98	1	1	1
I.PR	18	72	5	5	2	95	2	2	19	69	6	6
I	83	11	3	3	84	10	3	3	78	13	5	5
PC.PR	35	50	8	8	6	88	3	3	33	48	9	9
I.PC.PR	62	33	2	3	11	85	2	2	57	35	4	4
X.A	1	2	84	13	2	2	64	31	2	3	79	16
X.B	12	13	50	25	13	14	42	30	14	14	46	26
A	7	7	64	22	6	6	54	33	8	9	60	23
B	26	24	25	24	25	25	25	25	26	24	25	24
X	7	7	64	22	6	6	57	31	8	9	60	23
A.B	12	13	50	25	13	14	41	32	14	14	46	26
X.A.B	4	5	71	19	6	7	54	33	6	7	65	22

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., “PC.I” means that trait PC and trait I were presented together.

Table A8
Best fits of ELMO to data from Experiment 2 (Table 4). Corresponding parameter values are shown in Table A4.

	Fit Mainly To Inverse Base Rate Probes				Fit Mainly To Illusory Correlation Probes				Fit Equally To All Data			
	Group Chosen				Group Chosen				Group Chosen			
Traits	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx	Ci	Ri	Cx	Rx
I.PC	88	12	0	0	86	8	2	4	88	9	1	2
I.PR	12	75	6	7	8	65	10	18	9	69	8	15
I	71	29	0	0	68	29	1	2	70	28	0	1
PC.PR	29	58	6	7	25	44	11	20	27	49	8	16
I.PC.PR	71	29	0	0	68	29	1	2	70	28	0	1
X.A	0	2	50	48	2	7	61	30	1	5	62	32
X.B	0	11	44	45	5	20	27	48	2	18	29	51
A	0	2	50	48	2	7	61	30	1	5	62	32
B	6	35	28	30	7	26	23	44	5	27	23	45
X	0	2	50	48	2	7	61	30	1	5	62	32
A.B	0	11	44	45	5	20	27	48	2	18	29	51
X.A.B	0	2	50	48	2	7	61	30	1	5	62	32

Note: Traits separated by a dot (i.e., period) were presented together in the display; e.g., “PC.I” means that trait PC and trait I were presented together.