Theory and Decision (2005) 59: 307–344
DOI 10.1007/s11238-005-0946-8

PHILIP T. DUNWOODY, ADAM S. GOODIE and
ROBERT P. MAHAN

THE USE OF BASE RATE INFORMATION AS A FUNCTION OF EXPERIENCED CONSISTENCY

ABSTRACT. Three experiments examine the effect of base rate consistency under direct experience. Base rate consistency was manipulated by blocking trials and setting base rate choice reinforcement to be either consistent or inconsistent across trial blocks. Experiment 1 shows that, contrary to the usual finding, participants use base rate information more than individuating information when it is consistent, but less when it is inconsistent. In Experiment 2, this effect was replicated, and transferred in verbal questions posed subsequently. Despite experience with consistent base rates increasing sensitivity to base rates in word problems, verbal responses were far from normative. In Experiment 3, participants' use of base rates was once again moderated by its consistency, but this effect was itself moderated by the diagnosticity of base rate information. Participants were highly accurate in estimating experienced base rates. These studies demonstrate that base rate usage is complex and a function of how base rates are presented (experienced versus summary statistics) and response format (choice proportions versus probability estimates). Knowledge of base rates was insufficient for proper usage in verbal word problems. Although choice proportions showed a sophisticated sensitivity to experienced base rate information, participants seemed unable to demonstrate a similar sophistication when given typical word problems indicating that base rate neglect is a function of information representation and not an inherent processing bias.

KEY WORDS: base rate neglect, decision making, learning, choice, direct experience, diagnosticity

Koehler (1996) argued that the conditions under which base rates are neglected have themselves been neglected, arguing that there is little understanding of "how the ambiguous, unreliable, and unstable base rates of the real world are and should be used" (p. 1). The present research explores the adaptive usage of base rates under conditions of direct experience with unstable and differentially reliable base rates in a simulated environment.

When people are given specific (or individuating) information about a case, plus information about the population distribution from which the case was drawn (or base rate information) they often underweight the base rate information when judging the likelihood of an event. The cab problem is one of the most common problems used to demonstrate base rate neglect (Kahneman and Tversky, 1972; Bar-Hillel, 1980; Tversky and Kahneman, 1980). In this example, the base rate is more extreme than the witness is credible and therefore the cab is more likely to be Green than Blue despite the witness indicating that the cab was blue. Despite this fact, the typical response corresponds with the individuating information (the witness testimony) and is unaffected by the base rates and of green and blue cabs (Tversky and Kahneman, 1982). There are many other examples in the literature of participants ignoring the base rates and instead relying predominantly on the individuating information (Kahneman et al., 1982). This finding was considered so robust that Bar-Hillel (1980) stated, "The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (p. 215).

However, this strong view about the robustness of base rate neglect has not lasted. Koehler (1996) argues, "We have been oversold on the base-rate fallacy in probabilistic judgment from an empirical, normative, and methodological standpoint" (p. 1). He argued that few studies demonstrate a total neglect of base rates. More common are studies where base rate usage is less than is prescribed by Bayes's Theorem. He also questioned the appropriateness of Bayes's Theorem as a universal judgment criterion, arguing that there are few examples in the real world where Bayes's Theorem can unambiguously be mapped to the problem space.

Hammond (1996) points out that there is an important difference between evaluating the coherence of a judgment (e.g., how well it matches with Bayes's Theorem) and the correspondence of a judgment (how well it predicts events in the world). Research on base rate usage that emphasizes the rationality, or coherence, of a judgment typically evaluates judgment performance against a normative criterion (i.e., Bayes's Theorem). Research that emphasizes the correspondence of a judgment evaluates how well decision makers adapt to different environments, and implies that there is often no single normative criterion but rather performance should be evaluated on the basis of its adaptiveness (Gigerenzer et al., 1989; Cooksey, 1996).

Most base rate research has been of the coherence variety, assessing the relationship between judgments and a mathematical model, Bayes's Theorem. While the coherence of a judgment is undeniably an important concern, so too is the correspondence of a judgment. Less consideration has been given to how base rate usage reflects and adaptively responds to events in the world.

Bjorkman (1984) and Hammond (1996) note that most people never encounter Bayes's Theorem. What they encounter are events. These events may be summed up in descriptive statistics or reflected in formulae but what people usually experience are events. A similar argument is made by Gigerenzer and Hoffrage (1995), who noted that "organisms did not acquire information in terms of probabilities and percentages until recently" (p. 686). They argue that the information representation (percentages and probabilities) may not align with people's internal representations. Gigerenzer and Hoffrage proposed that a more natural format would increase people's Bayesian responses and provided participants with a series of word problems using summary statistics in absolute frequency terms, "...assum[ing] that as humans evolved, the 'natural' format was frequencies as actually experienced in a series of events, rather than probabilities or percentages" (p. 686). Using the same world problems and changing only the numerical representation, they demonstrated that participant Bayesian responses increased significantly when a frequency, rather than a percentage, representation was used.

A literature has emerged on base rate neglect under direct experience, which extends the representation argument introduced by Gigerenzer and Hoffrage. Where frequency formats seek to summarize experience, direct experience provides it. The results of base rate studies using direct experience have mostly reflected underweighting of base rates in comparison with normative models. Edgell et al. (1996), Estes et al. (1989), Gluck and Bower (1988), Kruschke (1996), Nosofsky et al. (1992); and Shanks (1990) have all reported base rate neglect under direct experience, under at least some conditions. These studies all provided information by means of direct experience and then elicited verbal probability estimates. Goodie and Fantino (1995, 1996, 1999a) and Lovett and Schunn (1999) have also provided participants with direct experience, and then measured differential sensitivity to base rates and individuating information in predicting future events, and observed greater sensitivity to the individuating information than to base rates. This effect depends on a pre-existing relationship between the cue (the individuating information) and the outcome (Goodie and Fantino, 1996), but is robust to monetary incentives for correct predictions (Goodie and Fantino, 1995), super-extended training of 1600 trials (Goodie and Fantino, 1999b) and symbolic pre-existing relationship (Goodie and Fantino, 1996, Experiment 3).

Goodie and Fantino (1999b, p. 327) posed the question: "...when base rates change relatively often and cue accuracy relatively seldom isn't it just as well to underweight base rates, since they're liable to change at any moment? Todd and Goodie (2002) tested this claim in a Monte Carlo simulation, and found that an "irrational" strategy like base rate neglect could rival Bayesian integration in performance given the right environment. Todd and Goodie created a variety of different environments where base rates and cue accuracies were manipulated and created numerous simulated agents that used a variety of decision making strategies. In simulated environments where base rates changed more often than cue accuracies, simulated participants who neglected base rates correctly predicted the outcome 72% of the time while simulated participants who used Bayesian integration correctly predicted the outcome 74% of the time. Despite the lack of coherence in the strategy of base rate neglect, responses still provided empirical accuracy, or correspondence, within a certain environment. These views suggest it can be reasonable to neglect base rates in some environments.

Consider the possible rationale for neglecting base rates in the following example adapted from Nisbett and Ross (1980). Imagine a hypothetical choice between buying a Volvo or a Saab, assuming Volvos have a better long-term repair record. Now imagine you hear from a friend who drives a Volvo that their car just died on the highway-now which brand would you be likely to pick? Here, it is usually argued that a single new piece of information should not overwhelm the evidence of cumulative repair records (base rates) – but in similar sorts of experimental tasks it often does. There are many manipulations that mitigate this effect (Koehler, 1996), often to a remarkable degree (Gigerenzer, 1991), but it has proven difficult to eliminate entirely. Nisbett and Ross argued that most people generally put too much weight on the individual case of the neighbor's experience, and not enough on Volvo's superior base rate of reliability. However, it is not clear that people should always use the base rate as Nisbett and Ross's Volvo–Saab example suggests. It is possible that while Volvo may have a superior long-term reliability record, they may be having production problems. They may have opened a new plant with an unusually high defect rate or they may be under new management and stream-lining their production facilities. There are many potential reasons why their short-term reliability record may be substantially lower than their long-term base rate for reliability. The neighbor's bad experience with the Volvo can be viewed as a cue that the reliability of Volvo is not what it once was. In short, the base rate may be changing.

Consider how one ought to use base rates in the following problem introduced by Gigerenzer (1991), which has the same formal structure as the Volvo problem:

You live in a jungle. Today you must choose between two alternatives: to let your child swim in the river, or to let it climb trees instead. You use only one criterion for that choice, your child's life expectancy. You have information that in the last 100 years there was only one accident in the river, in which a child was eaten by a crocodile, whereas a dozen children have been killed by falling from trees. Just yesterday your neighbor told you that her child was eaten by a crocodile. Where do you send your child? (p. 106)

Gigerenzer argued that most people would put considerable weight on the new information from the neighbor, just as they put weight on the friend's report of a bad Volvo, and that they would arguably be right to do so. "...the parents may assume that the river world has changed, and that the important event (being eaten or not) can no longer be considered as an independent random drawing from the same reference class. Updating 'old' base rates may be fatal for the child" (p.107). The world has likely changed, and with it the relevant base rates, and so narrow Bayesian updating is not a convincingly normative way to address the problem. In making this case Gigerenzer followed in the tradition of Brunswik (1956, 1957), who argued that researchers should not assume participants treat stimuli as perfectly consistent information sources.

Viewing past base rates and individuating information as imperfect indicators of future events leads to the question of whether people are differentially sensitive to base rates and individuating information when the consistency of these imperfect indicators is manipulated. Based on the argument above, it is reasonable to assume that base rates change and that humans should be sensitive to such changes. We should expect people to use base rates and individuating information more when history has shown them to be consistent indicators than when history has shown them to be inconsistent indicators. The present experiments sought to test the role of base rate consistency more directly, testing whether participants use base rates more when experienced base rates have been consistent than when they have been inconsistent. No studies to date have examined base rate usage as a function of experienced base rate consistency. The present experiments mark a beginning of this effort.

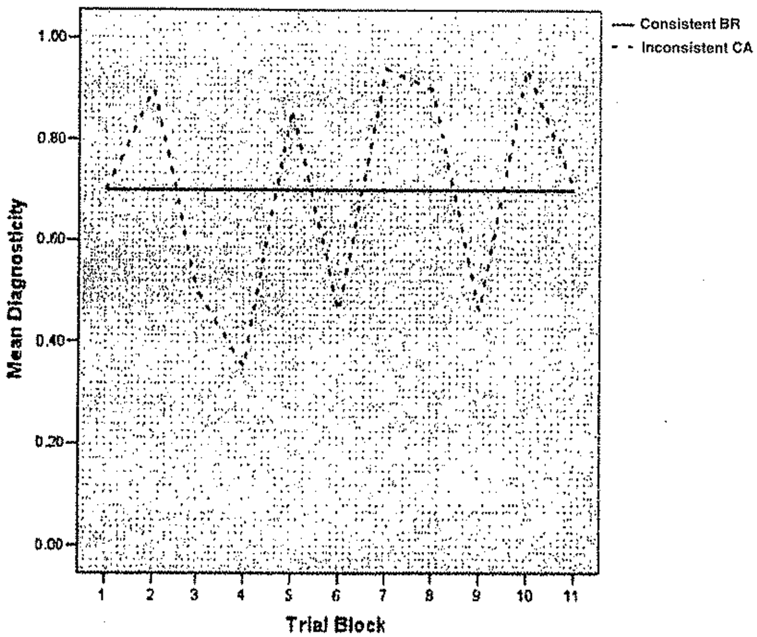
1. EXPERIMENT 1

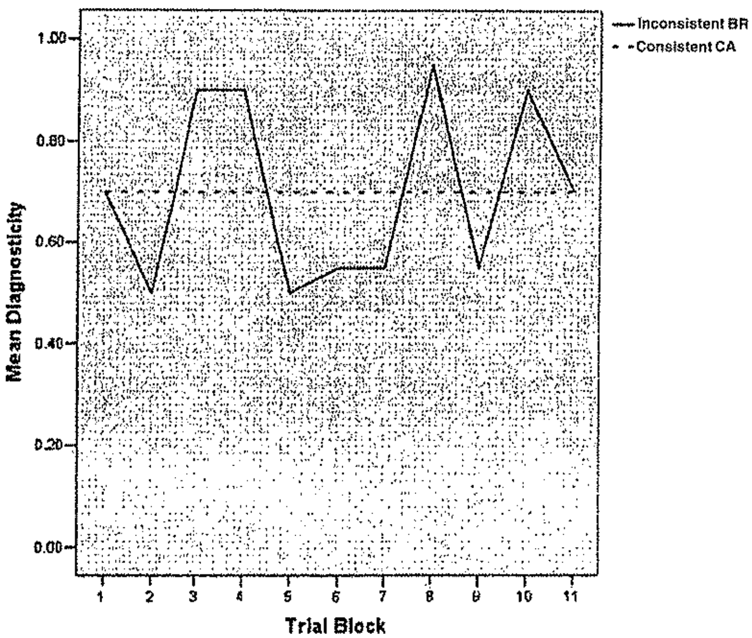
Experiment 1 addressed the issue of information use as a function of information consistency. Two environments were created and both environments contained a cue and base rate information in the form of how often a particular outcome was correct. In one environment the cue was a consistent predictor of the outcome but the base rate of the outcome was inconsistent. In the second environment the cue was an inconsistent predictor of the outcome but the base rate of the outcome was consistent. It was hypothesized that people would use consistent information more than inconsistent information. This usage pattern should result in greater use of the cue in the first environment described above and greater use of base rates in the second environment described above. In the present context, there was no long-term performance advantage for choosing one strategy over the other. While reliance on the inconsistent information source would provide for more short-term variations in performance, these variations in the inconsistent source average to make it equal to that of the consistent source over the course of the experiment. Consequently, even though the information sources varied in their consistency, they were equal in diagnosticity over the long run. This design allowed for a comparison of cue and base rate usage as a function of their relative consistency.

1.1. *Method*

1.1.1. *Participants*

Forty-four undergraduates participated for partial course credit. Upon arrival participants were randomly assigned to the consistent base rate/inconsistent cue accuracy or the inconsistent base rate/consistent cue accuracy condition. Cue accuracy served as the individuating information and is analogous to the witness testimony in the cab problem. Cue accuracy is operationally defined below. Twenty-two people participated in each condition. The mean age was 19.4 ($s = 1.44$), nine were male, 31 were female, and four participants did not indicate their gender.


Mean Diagnosticity
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8 9 10 11
Trial Block
Consistent BR
Inconsistent CA


Mean Diagnosticity
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8 9 10 11
Trial Block
Inconsistent BR
Consistent CA

1.1.2. *Materials*

The surface characteristics of the stimuli were identical for both groups while the base rates and cue accuracies varied. Base rate was defined as the proportional reinforcement for a particular response. That is, predicting green outcomes was reinforced 70% of the time. Cue accuracy was defined as the probability of a correct cue given the outcome, or $p(c|o)$. The first 20 trials had a base rate of 70% green and a cue accuracy of 70%. After the first 20 trials either the base rate or cue accuracy (varied between participants) changed every 20 trials. Figure 1a and b lists the base rates and cue accuracies in effect in each phase, and reflect that for one group (Figure la) the base rate was always 70% while cue accuracy fluctuated while maintaining an average value of 70%. For the other group (Figure lb), cue accuracy was always 70% and the base rate fluctuated but maintained an average value of 70%. Participants performed 200 learning trials with either the cue accuracy or the base rate fluctuating every 20 trials. They then performed 100 test trials in which the cue accuracy and base rate were again 70%. All analyses were conducted on performance data in the last 100 trials. Participants were unaware of which trials were learning trials and which were test trials.

The structure of the task is depicted in Figure 2, which reflects contingencies in effect when both the cue accuracy and the base rate of green outcomes are 70%. After an inter-trial interval the green cue was presented 58% of the time with the blue cue presented the other 42% of the time. When the green cue was presented, green predictions were reinforced 84% of the time with blue decisions being reinforced 16% of the time. However, when the blue cue was presented, green and blue decisions were reinforced equally (50%). The crucial comparison, and hence dependent measure, is choice proportions following a blue cue. Here, if participants' strategies emphasized base rates they would predominantly

Figure 1. (a) Base rates and cue accuracies experienced by the consistent base rate groups in Experiments 1 and 2. (b) Base rates and cue accuracies experienced by the inconsistent base rate groups in Experiments 1 and 2.

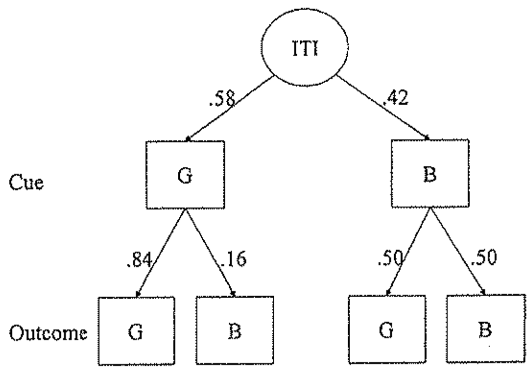


Figure 2. Incidence table representing contingencies when base rate and cue accuracy both equal 0.7 for Experiments 1 and 2. The "cue" is what was originally presented to the participant. They made their response and then the "outcome", or correct answer, was presented. ITI = inter trial interval. G = green. B = blue.

choose green since green responses are reinforced 70% of the time overall. If their strategies emphasized cue-matching they would predominantly choose blue since cue-matching is reinforced 70% of the time overall. If participant choices are conditional on the presence of the blue cue, it is possible that they will choose blue 50% of the time.

1.1.3. *Procedure*

Participants were told that they would be presented with either a green or blue square as a cue (the individuating information) and that they were to guess if the correct answer was a green or blue square. They were told that they would receive one point for each correct answer and it was their goal to amass as many points as possible.

On each trial, a green or blue square appeared at the top of the screen as a cue. Below were green and blue squares, and participants chose one of these as their response. Immediately following their choice they were given outcome feedback that said either; "Correct. You now have__points" or "Sorry, that is incorrect. You still have__points". Participants were asked if they understood the task and the feedback after the first 10 trials. If they did not, clarifications were made until they affirmed that the task and feedback were clear. A debriefing period followed the experiment.

1.2. *Results and Discussion*

Participants who experienced consistent base rates and inconsistent cue accuracies in the first 200 trials chose the base rate response (green) 56.0% (SEM = 5.7) of the time they were presented with a blue cue in last 100 trials. Participants who experienced consistent cue accuracies and inconsistent base rates chose the base rate response (green) only 36.7% (SEM = 4.7) of the time they were presented with a blue cue in the last 100 trials. This difference was statistically significant ($t(42) = 2.63$, $p < 0.05$). Previous research in this area (e.g., Goodie and Fantino, 1995) has shown a strong bias to match the cue, and this bias likely accounts for the lack of a symmetric difference favoring the consistent information between the two groups.

These results support the hypothesis that experienced information consistency influences the likelihood that participants will utilize that information source. In short, participants used base rate information more when it was consistent than when it was inconsistent. This finding highlights an ambiguity in the base rate literature. Base rates are typically defined as summary statistics of past events and treated as though subjects should use base rates as perfect indicators of future events. Typically only the base rate and individuating information means are provided. Experiment 1 shows that two environments with equal long-term base rate means can result in different choice patterns because decision makers are sensitive to the variability of base rates and cue accuracies. The variability of a cue is a relevant statistical characteristic of the environment and is clearly important if performance on a given task needs to be relatively stable. Fluctuations in the base rates may also indicate the beginning of a new base rate trend such as more crocodiles in the river this year or a drop in the reliability of Volvos. This finding highlights the importance of considering the base rate reference class when evaluating base rate usage under direct experience. Evaluating this problem with a simple Bayesian model that collapsed across past trials would fail to capture the environmental texture. Despite equal long-term base rates and cue accuracies, differential short-term values had a significant effect on participant choice behavior as predicted.

2. EXPERIMENT 2

Experiment 1 demonstrated that the consistency of base rates and cue accuracy, which were shown to be normatively relevant to behavioral decisions by Todd and Goodie (2002), do make a difference in how such decisions are made. However, the literature of base rate neglect has most often been concerned not with choice proportions, but with people's verbal answers to questions of summary statistics. It is not clear that experienced consistency of base rates would affect the verbal judgments that are most typical of the literature. Others have argued that "the abstract nature of the base-rate summary statistics may not be salient enough for the subjects to use" (Chrsitensen-Szalanski and Beach, 1982, p. 271; but see Nisbett and Borgida, 1975) and that direct experience with base rates is one way to make them more salient. For the above stated reasons, it is theoretically important to understand how direct experience with base rates and cues transfers to the types of word problems traditionally used in the base rate literature.

Christensen-Szalanski and Beach (1982) gave participants experience with base rates of pneumonia and the accuracy of test for pneumonia by having participants observe a series of slides. Each slide represented a single case and indicated that the person did or did not have pneumonia and the outcome of the test. The experienced base rate of pneumonia and cue accuracy was found to influence answers to the question "What is the probability that a person with a positive test result actually has pneumonia?" Participants' answers to this question were not significantly different from the

Bayesian criterion. However, the same participants had completed a typical base rate word problem prior to the trial-by-trial experience. The word problem and the experience had virtually identical probabilities but responses to the pre-experience word problem showed marked base rate neglect. Furthermore, even though participants responded with estimates consistent with a Bayesian criterion for the experienced events, they were unable to transfer this experience to novel base rate word problems. Christensen-Szalanski and Beach argued that it was possible for participants to respond in a Bayesian manner to the question about their direct experience because they were able to rely on experience. However, their lack of ability to transfer this experience to a novel word problem with different numerical values indicates they did not learn Bayes's rule.

The purpose of Experiment 2 was twofold. First, we sought to replicate the findings of Experiment 1 using a medical cover story. If we failed to replicate Experiment 1 because of a relatively superficial change in the framing of the task, then our conclusions from Experiment 1 would be undermined. On the other hand, successful transfer of the effect from a purely abstract task to one with a concrete context would represent an important step in extending it to the rest of the base rate literature. The second purpose of Experiment 2 was to help explicate the relationship between the experienced base rate environment used in Experiment 1 and the more traditional mode of base rate research, word problems. Hence, we tested whether experienced consistency of base rates and cue accuracies would influence responses to novel word problems traditionally used in base rate research. We imposed a medical cover story on stimuli that were otherwise identical to those in Experiment 1, measuring the same choice proportions as in Experiment 1, and then eliciting verbal answers to verbal questions on the basis of novel medical conditions.

As in Experiment 1, there were two groups that differed in which information source (base rates or cue accuracy) was consistent. Average values were 70% for both sources, for both groups, and the inconsistent parameter was manipulated according to the same schedule as was used in Experiment 1.

2.1. *Method*

2.1.1. *Participants*

Eighty-three new undergraduates participated for course credit. Upon arrival participants were randomly assigned to the conditions of variable base rate ($n=42$) and variable cue accuracy ($n=41$). None of the participants had participated in Experiment 1.

2.1.2. *Materials*

Although the stimuli in this experiment consisted primarily of cues, choice options and outcomes that were green and blue rectangles, identical to those used in Experiment 1, the following cover story was presented beforehand:

In this experiment, please imagine you're a doctor who specializes in a disease called Hyporia. There are two types of Hyporia, Blue Hyporia and Green Hyporia. Your task is to diagnose which form of Hyporia each patient has who comes to your office. There is a test that distinguishes Blue Hyporia from Green Hyporia. All of your patients have taken the test, and you'll know the results. However, like all medical tests, this one isn't perfect. So part of your task is to figure out how much credence to give to the test when making your diagnosis of each patient. You'll receive one point for each correct diagnosis.

The cue was presented with a label that said "Test". The choice options, rather than being labeled "Your choice", were labeled "Your diagnosis". The outcome was labeled "Correct" as it was in Experiment 1.

The learning phases were statistically identical to those used in Experiment 1, respectively, for the two groups (see Figure la, b). Afterwards, the following instructions were given:

On the basis of all your experience, you are now recognized as a leading authority on the diagnosis of Hyporia. Congratulations on your achievement. With recognition, however, comes additional responsibility. You are now in demand as a consultant to public health authorities in cities

where no doctor has your wealth of experience in this domain. You will be asked questions concerning Blue and Green Hyporia in four cities. Please answer them to the best of your ability, and then your participation in this experiment will be complete. Different cities may have different rates of the diseases, and not all cities have access to the same tests. So remember to pay close attention to the information that is provided about each city.
In City A, you are given the following information. Among people who have Hyporia, 80% have Blue and 20% have Green. The test that has been developed in City A correctly identifies both Blue and Green Hyporia 80% of the time, but makes errors 20% of the time.

Among people whose tests indicate they have Blue Hyporia:
How many actually have Blue Hyporia?
How many have Green Hyporia?

The fictitious disease Hyporia was adopted from McKenzie (1998). For each question, participants chose among options that included all multiples of 5% between 5% and 95%, plus 1% and 99%. The same four questions were repeated for Cities B, C and D. In City B, the base rate was 80% Blue and the test accuracy 60%. In City C, base rate was 60% Blue and test accuracy 80%. And in City D, the base rate and test accuracy were both 60%. In each city and for each test result, probabilities were recorded in log-odds form, partly to handle possible non-additivity.

2.2. *Results and Discussion*

Participants who experienced consistent base rates in the first 200 trials chose the base rate response (green) 47% of the last 100 trials (SEM = 4.6), while participants who experienced consistent cue accuracies (and inconsistent base rates) in the first 200 trials chose the base rate response (green) only 27% of the last 100 trials (SEM = 3.9), following a blue cue. This difference was statistically significant in the predicted direction ($t(81) = 3.29$, $p < 0.05$). Thus the primary effect observed in Experiment 1 was replicated. Also, these values are relatively stable from the beginning to the end of the last 100 trials. The

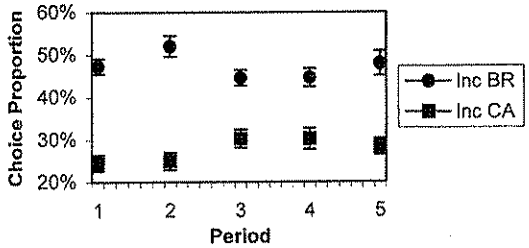


Figure 3. Proportion of choices that matched Blue cues in the test phase of Experiment 2, groups in 20-trial blocks. Inc BR = inconsistent base rates (with consistent cue accuracies). Inc CA = inconsistent cue accuracy (with consistent base rates).

proportions are displayed in Figure 3 for both groups in each 20-trial block of the last 100 trials. The difference between groups is evident and large in all blocks, and the effect is evident from the outset. The effect may be decreasing in magnitude across blocks, as the experience of the two groups, especially recent experience, becomes more similar. The interaction term is marginally statistically significant ($F(4, 320) = 2.05, p = 0.09$). Participants experiencing a variable cue accuracy made more total correct predictions (202.1 out of 300, on average, SEM = 1.99) than those experiencing variable base rates (198.6, SEM = 2.16). This difference was statistically significant ($t(81) = 2.15$; $p < 0.05$), although the practical significance of a 1.2% point difference in performance may be considered small.

Base rate responses averaged 27% when base rates were inconsistent in Experiment 2 but averaged 37% when base rates were inconsistent in Experiment 1. Likewise, base rate responses averaged 47% when base rates were consistent in Experiment 2 but averaged 56% when base rates were consistent in Experiment 2. Why are base rate responses about 10% lower across conditions in Experiment 2 than Experiment 1? We believe that the change in the decision frame likely accounts for this difference. In Experiment 1, participants were simply told that they would receive a cue and they had to guess the correct answer. In Experiment 2, participants were told that they would receive the results of a test that discriminated between two diseases and they had to make a diagnosis. Since the dependent measure is the proportion of green responses after a blue cue, it is reasonable that participants were more likely to diagnose someone as having "Blue Hyporia" when the test results say blue than they are to choose a blue response because a cue says blue. Although the difference in framing is subtle, it is likely that the cover story in Experiment 2 made the connection between the cue, or test result in this case, and outcome more salient.

The main question of interest in this experiment is whether group differences exist in responses to the verbal questions about novel cities. These questions were asked in complementary pairs, but these pairs did not produce uniform additivity. Only 54.0% of pairs summed to 100%; 28.0% were subadditive, and 17.9% were superadditive.

2.2.1. *Defining sensitivity*

To assess sensitivity to base rates and cue accuracy, difference scores between log-odds transformations of probability assessments were taken. Greater difference scores thus reflect greater sensitivity to base rates. This was done separately for cities with high and low cue accuracies, and these two difference scores were added together to obtain a global score of sensitivity to base rates. A parallel procedure was used to assess sensitivity to cue accuracy. In this case, difference scores were taken between cities of different cue accuracy, separately for each of the two base rates, and then these difference scores were summed to yield a global score of sensitivity to cue accuracy. The normative outcome would be equal sensitivity to both sources of information, regardless of the consistency of those sources.

When questioned about the implications of a blue test result, participants who experienced consistent base rates were more sensitive to variations in base rates, and those who experienced consistent cue accuracy were more sensitive to variations in cue accuracy. Participants who experienced consistent

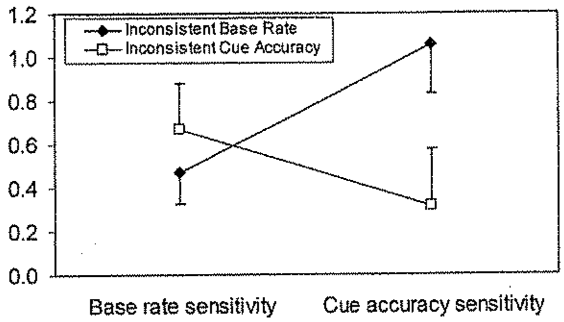


Figure 4. Sensitivity measure for word problem by experienced base rates and cue accuracies for Experiment 2. Larger values indicate greater sensitivity.

base rates and inconsistent cue accuracies had an average base rate sensitivity score that was 0.36 greater than their cue accuracy sensitivity score. Participants who experienced inconsistent base rates and consistent cue accuracies had an average base rate sensitivity score that was 0.58 less than their cue accuracy sensitivity score. This difference is visible in Figure 4 and is statistically significant ($t(81) = 2.40$; $p < 0.05$). The inferential test of difference scores was equivalent to testing the significance of the crossover effect that is evident in Figure 4.

Although experienced consistency of base rates and cue accuracies did influence responses to word problems with summary statistics, the responses in all conditions still demonstrated a lack of correspondence to the Bayesian criteria, and hence a lack of coherence. Table I lists the cell means for all combinations of cities and questions by group (consistent base rates–inconsistent cue accuracies versus consistent cue accuracies–inconsistent base rates). The top third of Table I lists the base rates and cue accuracies for each city. The middle and bottom third lists the Bayesian answer, mean responses along with the standard error of the mean for the two conditions and for each question. We added up the absolute difference between each participant's answer to each of these eight questions and the "correct" Bayesian answer. The smallest summed error was 57 percentage points, an average

TABLE I

Response to word problems by condition in Experiment 2

	City:	A	B	C	D
	Base rate blue:	80%	80%	60%	60%
	Cue accuracy:	80%	60%	80%	60%
Among people whose test indicate they have Blue Hyporia: How may actually have Blue Hyporia?	Bayesian Response	0.94	0.92	0.86	0.69
	IBR response	0.65 (0.02)	0.54 (0.03)	0.60 (0.03)	0.50 (0.02)
	ICA response	0.66 (0.03)	0.59 (0.03)	0.58 (0.03)	0.54 (0.03)
Among people whose test indicate they have Blue Hyporia: How may actually have Green, Hyporia?	Bayesian Response	0.06	0.08	0.14	0.31
	IBR response	0.33 (0.03)	0.40 (0.03)	0.35 (0.03)	0.44 (0.03)
	ICA response	0.29 (0.03)	0.31 (0.03)	0.33 (0.03)	0.38 (0.03)

IBR Response = average response in inconsistent base rate/consistent cue accuracy condition.
ICA Response = average response in inconsistent cue accuracy/consistent base rate condition.
Cells indicate mean response with standard error of the mean in parentheses.

of 7.1 percentage points error per question. Only two participants out of 83 averaged less than 10 percentage points of error per question, having summed error less than 80. The median summed error was 174 percentage points (21.8 points error per question). Thirty-three out of 83 total participants (39.8%) gave answers to these eight questions whose errors summed to more than 200 percentage points. There was no significant difference between the consistent and inconsistent base rate groups on their accuracy, $t(81)=1.19$, $p=0.24$.

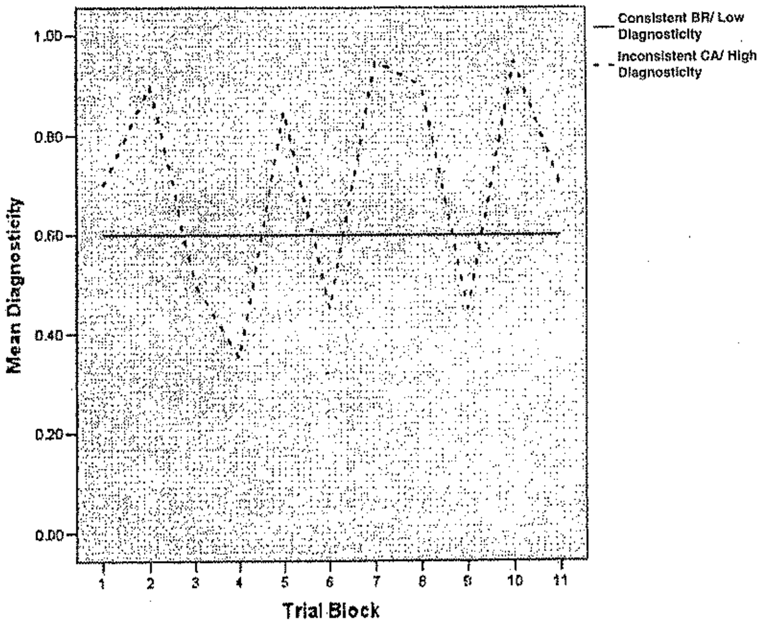
In Experiment 2, the effect of base rate consistency on choice proportions, observed in Experiment 1, was replicated. In addition, participants who had experienced consistent base rates were subsequently more sensitive to base rate information on verbal questions about novel environments. Despite this increased sensitivity, responses were still far from normative when compared with the Bayesian criteria. It has been argued that base rate neglect may be due to a lack of saliency of base rate information (Nisbett and Borgida,1975). The increased sensitivity to base rates presented in word problems demonstrates that experienced information consistency affects saliency in subsequent word problems. Despite the increase in information saliency, participants still failed to use the information in a normative way.

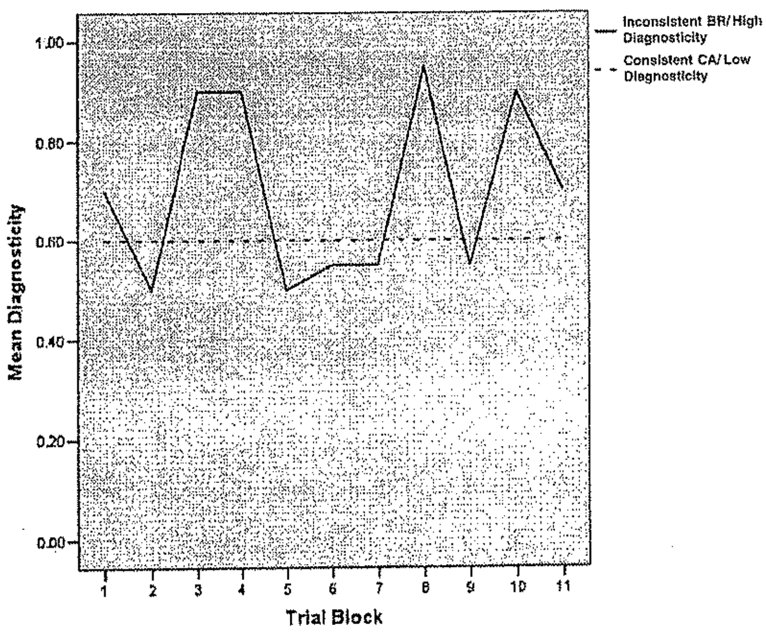
The results of Experiment 2 are consistent with those of Christensen-Szalanski and Beach (1982). Even though experience with base rates and cue accuracies may increase the saliency of such information, participants still fail to use this information in a normative manner when presented with novel base rate word problems. Christensen-Szalanski and Beach's comment that "subjects seldom have any real grasp of what the numbers mean" (p. 271) has considerable appeal. Our results, like those of Christensen-Szalanski and Beach, seem to indicate that base rate word problems "may be testing subjects' knowledge of a statistical rule or their ability to understand what are, after all, very complexly worded problems, rather than their ability to logically use probabilistic information". (Christensen-Szalanski and Beach, 1982, p. 277). This conclusion is consistent with Gigerenzer and Hoffrage's (1995) argument that people may fail to act in a Bayesian manner simply because they do not understand formal probabilities. Even though participants in Experiment 2 failed to give normative Bayesian responses to word problems, their responses on the word problems were influenced more by the previously experienced consistent information source than the inconsistent information source. Also, their choice proportions were clearly influenced by the consistency of the information, thereby replicated the finding of Experiment 1.

3. EXPERIMENT 3

We have thus far argued that base rate neglect may be adaptive in some environments, and Experiments 1 and 2 demonstrate that under conditions of direct experience participants neglect base rates or use base rates when it is arguably adaptive to do so. However, recall that the structure of Experiments 1 and 2 was such that, while an inconsistent information source provided for erratic short-term performance, there was no long-term performance gain by using the consistent or inconsistent information source. Figure la and b shows that the inconsistent and consistent information sources resulted in 70% reinforcement over the course of the entire experiment. What if participants were placed in an environment where the two information sources were not equally reinforced over the length of the entire experiment? Specifically, when a consistent information source is reinforced less over the length of the entire experiment than an inconsistent information source (depicted in Figure 5a and b), which will people utilize? Utilizing the consistent indicator would offer stable reinforcement during the entire experiment but use of the inconsistent indicator would offer inconsistent but higher reinforcement throughout the length of the experiment. Experiment 3 was designed with the main goal of testing whether the effect of information consistency on choice proportions demonstrated in Experiments 1 and 2 was robust to simultaneous manipulation of long-term reinforcement (diagnosticity).

As in Experiments 1 and 2, information consistency was manipulated, and the two consistency levels were also crossed with two diagnosticity levels to create four conditions. This 2-by-2 design is represented in Figure 6. Consistency was defined as it was in Experiments 1 and 2, with each parameter


Consistent BR/ Low Diagnosticity
Inconsistent CA/ High Diagnosticity
Mean Diagnosticity
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8 9 10 11
Trial Block


Inconsistent BR/ High Diagnosticity
Consistent CA/ Low Diagnosticity
Mean Diagnosticity
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8 9 10 11
Trial Block

Base Rate Diagnosticity	Base-Rate Consistency: Consistent	Base-Rate Consistency: Inconsistent
Hi (70%)	(a) Adapt to base rate 0.58 (0.05)	(b) Either: 1) Adapt to higher diagnosticity (BR) 2) Adapt to consistent information (CA) 0.69 (0.05)
Low (60%)	(c) 0.26 (0.04) Either: 1) Adapt to higher diagnosticity (CA) 2) Adapt to consistent information (BR)	(d) 0.31 (0.04) Adapt to cue accuracy

Figure 6. Two (base rate consistency)-by-two (base rate diagnosticity) design for Experiment 3. Conditions are specified by base rate manipulations. Cue accuracies are always orthogonally manipulated within each condition (e.g., high diagnostic base rate conditions have low diagnostic cue accuracies and consistent base rate conditions have inconsistent cue accuracies). Mean proportion of base rate responses after a blue cue are listed for each cell with SEM in parentheses.

remaining consistent or varying every 20 trials. High diagnosticity, for both base rates and cue accuracy, was defined as a 70% reinforcement rate over the course of the entire experiment (300 trials), and low diagnosticity was defined as a 60% reinforcement rate over the course of the entire experiment. As in Experiments 1 and 2, consistent and inconsistent indicators were always inversely related within each condition such that when base rates were consistent, cue accuracies were inconsistent. Likewise, base rate and cue diagnosticity were always inversely related within each condition such that when base rates were of high diagnosticity (70%), cue accuracies were of low diagnosticity (60%).

We compared participants' base rate choice proportions after a blue cue for cells A and D of Figure 6, to see if the observed

Figure 5. (a) Base rates and cue accuracies experienced by participants in cell B of Figure 6 in Experiment 3. (b) Base rates and cue accuracies experienced by participants in cell C of Figure 6 in Experiment 3.

difference would be greater than that found in Experiments 1 and 2. Because this comparison was presumably augmented by the coupling of consistency with high diagnosticity, it was expected that the observed difference in base rate choice proportions between conditions A and D of Figure 6 would be greater than that found in Experiments1 and 2.

The more informative planned contrast compared the proportion of participants' base rate choices (green) after a blue cue for cells C and B of Figure 6. This comparison was designed to evaluate how choice judgments are affected when information consistency and diagnosticity are pitted against each other. Figure 5b represents cell C of Figure 6 where the base rate is of lower diagnosticity than the cue accuracy but the base rate is consistent. Participants may favor either the consistent information (the base rate) as they did in Experiment 1 or the inconsistent information which is higher in diagnosticity (the cue accuracy). Figure 5a represents cell B of Figure 6 where the base rate is of higher diagnosticity than the cue accuracy but the base rate fluctuates. Again, participants may favor either the consistent information (in this case the cue accuracy) as they did in Experiment 1 or the information higher in diagnosticity (in this case the base rate).

There are five possible outcomes that may be observed when participants' proportions of base rate choices are compared between cells C and B of Figure 6. First, participants in cells C and B could adapt to the consistent information despite the diagnosticity (Outcome 1). Participants may be satisfied with their level of performance using the consistent information or may simply fail to notice a diagnosticity difference of 10 percentage points. This would lead to greater base rate choice proportions in cell B than C. Second, participants in cells C and B could adapt to the information highest in diagnosticity despite the consistency (Outcome 2). Such a finding would mean that participants were sensitive to a diagnosticity difference of only 10% (60% versus 70%) and that this 10% difference is sufficient to override the already demonstrated effect of consistency. This would lead to greater base rate choice proportions in cell C than B. Third, participants could neglect the base rates and simply match the cue (Outcome 3), leading cells B and C to have equal choice proportions, both greater than 0.5. Fourth, participants could neglect the cue and use the base rates (Outcome 4), leading cells B and C to have equal choice proportions, both less than 0.5. The fifth possible outcome is the null hypothesis, leading cells B and C both to have choice proportions equal to 0.5. In order to test among these possibilities, a contrast was planned between base rate choice proportions in cells B and C. In the event they were not significantly different from each other, their pooled means were planned to be compared to 0.5.

3.1. *Method*

3.1.1. *Participants*

One-hundred undergraduates participated for course credit. Upon arrival participants were randomly assigned to the four conditions resulting in 25 participants per cell. The mean age was 19.9 ($s = 3.33$), 25 were male, 73 were female, and two participants failed to indicate their gender on the questionnaire. None of the participants had participated in Experiments 1 or 2.

3.1.2. *Materials*

The surface characteristics of the stimuli were identical to those in Experiment 1. The depth characteristics of the stimuli include the consistency manipulation in Experiment 1 with the addition of a diagnosticity manipulation. Participants in high base rate diagnosticity conditions experienced 70% base rate reinforcement over the entire experiment with a cue accuracy reinforcement of 60%. Participants in low base rate diagnosticity conditions experienced 60% reinforcement over the entire experiment with a cue accuracy reinforcement of 70%. The four conditions, two levels of consistency and diagnosticity, are represented in Figure 6. Part way through Experiment 3 a post-experimental questionnaire was added to assess the accuracy of participants' self-report strategies and awareness of base rates and cue accuracies Participants were asked the following questions;

Question 1 (*forced choice*):
Would you describe your decision making strategy as primarily...

(a) choosing the same color as the cue color
(b) choosing the opposite color of the cue color
(c) primarily choosing one color regardless of the cue color

Question 2 (*forced choice*):
If you chose answer 'c' above, what color did you primarily choose?
Questions 3 and 4 (*fill in the blank*):
Choosing a color that matched the cue was correct —% of the time.
Regardless of the cue color, green was the correct answer —% of the time.

3.2. *Results and Discussion*

3.2.1. *Choice proportions*

Participants who experienced base rates that were high in diagnosticity and consistent (cell A of Figure 6) in the first 200 trials chose the base rate response (green) 58.0% (SEM = 4.7) of the time they were presented with a blue cue in the last 100 trials. Conversely, participants who experienced base rates that were low in diagnosticity and inconsistent (cell D in Figure 6) in the first 200 trials chose the base rate response (green) 30.8% (SEM = 3.8) of the time they were presented with a blue cue in the last 100 trials. This difference was statistically significant ($F(1, 96) = 21.4$, $p < 0.05$). This planned contrast is consistent with the findings of Experiments 1 and 2. The mean difference for the consistency comparison in Experiments 1 and 2 is 19% and 20%, respectively, while the mean difference for the consistency/diagnosticity comparison in Experiment 3 is 27%. The larger mean difference in Experiment 3 suggests that the diagnosticity manipulation had an effect in addition to the consistency manipulation. The following contrast explicitly addresses the effect of diagnosticity compared to consistency. The greater reliance on the consistent and high diagnostic information would necessarily result in an increase in performance.

The results of the second planed contrast of Experiment 3 provide a clear indication whether more consistent information trumps more diagnostic information. Participants overwhelmingly utilized the information that was highest in diagnosticity whether or not that information was consistent Of the 50 participants that were in cells C and B, 43 (86%) had a choice proportion favoring the information highest in diagnosticity, 5 (10%) had a choice proportion favoring consistent information, and 2 (4%) had choice proportions that showed no strong preference (between 0.45 and 0.55). This is a large difference favoring information that is most useful, particularly when one considers that the diagnosticity difference was only 10 percentage points (60% versus 70%).

Participants who experienced base rates that were high in diagnosticity and inconsistent (Figure 5a and cell B of Figure 6) in the first 200 trials chose the base rate response (green) 69.3% (SEM = 4.5) of the time they were presented with a blue cue in the last 100 trials. Choices favoring the base rate in this condition were significantly different than a chance level of 50%, ($t(24) = 4.32,\ p < 0.05$). Conversely, participants who experienced base rates that were low in diagnosticity and consistent (Figure 5b and cell C in Figure 6) in the first 200 trials chose the base rate response (green) 25.6% (SEM = 3.7) of the last 100 trials. Choices favoring the base rate in this condition were significantly different than a chance level of 50%, ($t(24) = 6.58,\ p < 0.05$). The difference in base rate responses between groups (69.3% versus. 25.6%) was also statistically significant $F(1, 96) = 64.94,\ p < 0.05$).

While Experiments 1 and 2 demonstrated that base rate and cue usage is moderated by consistency, Experiment 3 demonstrates that use of the consistent indicator is itself moderated by the diagnosticity of the indicator. If diagnosticity is held constant, participants use the more consistent indicator which results in more stable performance throughout the duration of the task (seen in Experiments 1 and 2). If diagnosticity is manipulated, participants use the more diagnostic indicator (the one that gives them the best long term performance advantage) even when it is inconsistent between trial blocks and hence yields more erratic performance throughout the task (as seen in Experiment 3). This differential use of indicators based on their statistical characteristics argues for a fairly sophisticated sensitivity to the statistical characteristics of experienced base rate and cue accuracy information. Although participants do not necessarily use the information optimally, their strategies are clearly adaptive within the current simulated environments.

3.2.2. *Questionnaire*

Thirty-nine of the 100 participants in Experiment 3 completed the questionnaire. Each question is listed with a table indicating the group descriptives (Tables II–IV).

Question 1 (forced choice): Would you describe your decision making strategy as primarily: (a) choosing the same color

TABLE II
Answers to question 1 in experiment 3

	High diagnosticity and consistent base rates (cell A, Figure 6)	Low diagnosticity and consistent base rates (cell D)	High diaggnosticity and consistent base rates (cell B)	Low diagnosticity and consistent base rates (cell C)
(a) Choosing the same color as the cue color	36%($n=4$)	36%(4)	29%(2)	50%(5)
(b) Choosing the opposite color of the cue color	9%(1)	0%(0)	0%(0)	0%(0)
(c) Primarily choosing one color regardless of the cue color	55%(6)	64%(7)	71%(5)	50%(5)

Cells indicate percentage and number of participants in each group that chose each answer.

TABLE III
Answers to question 2 in Experiment 3

	High diagnosticity and consistent base rates (Group A)	Low diagnosticity and inconsistent base rates (Group D)	High diagnosticity and inconsistent base rates (Group B)	Low diagnosticity and consistent base rates (Group C)
Green	83%($n = 5$)	86%(6)	100%(5)	80%(4)
Blue	17%(1)	14%(1)	0%(0)	20%(1)

Cells indicate percentage and number of participants in each group that chose each answer.

as the cue color, (b) choosing the opposite color of the cue color, (c) primarily choosing one color regardless of the cue color. Thirty-eight percent of the participants chose (a), 3% chose (b), and 59% chose (c). See Table II for the cell means for each response by condition. More than half of the participants reported that their decision making strategy was primarily one of base rate usage (answer c).

Question 2 (forced choice): "If you chose answer 'c' above, what color did you primarily choose?" Of the participants who reported a primarily base rate decision strategy, the vast majority, 87%, reported choosing green. See Table III for the cell means for each response by condition.

Question 3 (fill in the blank): "Choosing a color that matched the cue was correct __% of the time." Participants who experienced a cue accuracy of 60% (low diagnosticity) estimated that choosing a color that matched the cue was correct 58.89% (SEM = 4.05) of the time while participants who experienced a cue accuracy of 70% (high diagnosticity) estimated that choosing a color that matched the cue was correct 64.52% (SEM = 3.98) of the time. Even though the estimates for cue matching differed in the direction consistent with the environment, the estimates were not significantly different from each other $t(37) = -0.99$, $p > 0.05$. Both the

TABLE IV

Answers to questions 3 and 4 in Experiment 3

	High diagnosticity and consistent base rates (Group A)	Low diagnosticity and inconsistent base rates (Group D)	High diagnosticity and inconsistent base rates (Group B)	Low diagnosticity and consistent base rates (Group C)
Choosing a color that matched the cue was correct __% of the time	$M=59.55$; $S=13.13$	$M=64.09$; $S=20.10$	$M=57.86$; $S=23.43$	$M=65.00$; $S=17.00$
Regardless of the cue color, green was the correct answer __% of the time.	$M=69.55$; $S=10.60$	$M=63.64$; $S=13.98$	$M=73.57$; $S=16.50$	$M=56.00$; $S=13.50$

Cells indicate mean response with standard deviation for each question.

mean estimated cue matching for the high diagnosticity condition ($M = 64.52$, SEM $= 3.98$) and the low diagnosticity condition ($M = 58.89$, SEM $= 4.05$) failed to differ significantly from the actual cue matching reinforcement in each condition, $t(20) = -1.38$, $p > 0.05$ and $t(17) = -0.27$, $p > 0.05$, respectively. See Table IV for the cell means for each response by condition.

Question 4 (fill in the blank): "Regardless of the cue color, green was the correct answer __% of the time." Participants who experienced base rates of 70% green gave a mean estimated response that green was the correct answer 71.11% (SEM $= 3.4$) of the time while participants who experienced base rates of 60% estimated that green was correct 60.00% (SEM $= 3.05$) of the time. These estimates were significantly different from each other, $t(37) = 2.565$, $p < 0.05$. The mean estimated base rates for the high diagnosticity condition ($M = 71.11$, SEM $= 3.04$) and the low diagnosticity condition ($M = 60.00$, SEM $= 3.05$) failed to differ significantly from the actual base rate in each condition, $t(17) = 0.37$, $p > 0.05$ and $t(20) = 0.00$, $p > 0.05$, respectively. See Table IV for the cell means for each response by condition.

Participant responses to the above questions were assessed against their actual choice behavior to determine self-report accuracy. Seventy-two percent of the participants had post-experimental questionnaire results that matched with their actual choice behavior. Participant estimates for cue matching reinforcement and base rates (obtained in question 3 and 4, respectively), were evaluated against the actual contingencies to see whether participants with accurate self-report decision strategies were more accurate at estimating base rates and cue matching reinforcement than participants with inaccurate self-report decision strategies. The absolute value of the response minus the actual value served as the dependent measure and was compared for accurate and inaccurate self-report participants for both base rates and cue matching. The mean difference score for cue matching was 12.5 (SEM $= 2.06$) for participants with accurate self-report decision strategies and 16.82 (SEM $= 3.83$) for participants with inaccurate self-report decision strategies. While the difference was larger for participants with inaccurate self-report decision strategies, the difference was not significant, $t(37) = -1.06$, $p > 0.05$. The mean difference score for base rates was 10.36 (SEM = 1.58) for participants with accurate self-report decision strategies and 10.91 (SEM = 2.22) for participants with inaccurate self-report decision strategies. This difference was not significant, $t(37) = -0.19$, $p > 0.05$.

Participants' estimates of cue-matching and base rate reinforcement were notably accurate and failed to significantly differ from the actual environmental contingencies. These data indicate that participants are indeed sensitive to both cue accuracies and base rates and that when these events are actually experienced by the participants they are highly accurate at estimating them retrospectively.

It is interesting to note that participants with accurate self-report decision making strategies were no more accurate at estimating cue matching reinforcement and base rates than participants with inaccurate self-report decision making strategies. In other words, the accuracy of self-report decision making strategies was not related to the accuracy with which environmental contingencies were encoded in, or recalled from, memory.

4. GENERAL DISCUSSION

Three experiments tested the effects of information consistency on the usage of base rate information. We hypothesized that the base rate neglect that has been widely observed in a vast literature may spring, at least in part, from a tendency to rely more heavily on consistent sources of information. In Experiment 1, participants underweighted base rates and relied more heavily on the cue when base rates were inconsistent and cue accuracies consistent, under conditions where both sources of information were equal in long-term diagnosticity. But, consistent with the hypothesis, participants for whom cue accuracies were inconsistent and base rates consistent weighted base rates more heavily than the cue. This shift in base rate and cue usage occurred even though the long-term diagnosticity of the indicators was identical between conditions. While a strategy that relied solely on the inconsistent indicator, whether it be the base rate or the cue, would provide the same long-term average performance as relying solely on the consistent indicator, such a strategy would provide for more erratic short-term performance. A more cautious approach would be to use base rate or cue information more when it is a consistent source of information. This strategy was observed in Experiment 1 and makes sense when one considers the error potential in relying on an inconsistent information source.

In Experiment 2, this effect was replicated, and participants were also found to be more sensitive to the experienced consistent information source when subsequently asked typical base rate word problems. Thus, when responding to the word problems participants who had experienced inconsistent base rates and consistent cue accuracies were more sensitive to cue accuracies than to base rates. Those who had experienced inconsistent cue accuracies and consistent base rates were more sensitive to base rate information than they were to cue accuracy information. This is consistent with our hypothesis, and contrary to a large body of research on base rate usage in verbal problems. The effect of base rate consistency is thus evident not only in choice proportions but also extends to hypothetical novel experiences that have been the more traditional vehicle of studies of base rate usage. Although experienced base rates did make the information more salient, as indicated by participant choice proportions and the sensitivity measure for the word problems, participant responses on the word problems were still far from Bayesian.

In Experiment 3, the effects of information consistency were not profound enough to override the effects of information diagnosticity. When consistency and diagnosticity are pitted against each other, 86% of participants used the information source highest in diagnosticity, even though this diagnosticity advantage was only 10 percentage points. This finding is consistent with that of Ginossar and Trope

(1987), who found that base rate usage varied as a function of the credibility of the individuating information. Ginossar and Trope manipulated the source of the individuating information (palm reader versus a psychologist) to manipulate credibility. The diagnosticity manipulation in the present study had the effect of making the base rate more or less credible than the cue accuracy. Participants experienced different levels of diagnosticity and used the base rate information as a function of its diagnosticity.

The highly accurate cue matching and base rate estimates in Experiment 3 add support to our conclusions about Experiment 2. Despite participants' clear sensitivity to base rates when actually experienced, participants do not act as if they understand the statistical and logical structure of typical base rate word problems. In Experiment 2, increased sensitivity to base rates, as demonstrated by their responses to base rate word problems and choice proportions, failed to lead to Bayesian responses when given typical base rate word problems. Given the accuracy of cue matching and base rate estimates in Experiment 3, it is most likely that participants in Experiment 2 would also have been able to accurately estimate the base rates and cue matching reinforcement. These findings are consistent with those of Christensen-Szalanski and Beach (1982) who found that accurate base rate estimates of experienced events did not translate into appropriate base rate usage on subsequent word problems. Our results from Experiments 2 and 3 provide a conceptual replication and extension of their findings.

We concur with Christensen-Szalanski and Beach's main conclusion that the typical base rate word problems test participants' "knowledge of a statistical rule or their ability to understand...complexly worded problems, rather than their ability to use probabilistic information." (p. 277). All three experiments demonstrate a clear and sophisticated use of probabilistic information.

Christensen-Szalanski and Beach concluded that such findings "raise a troubling doubt about the appropriateness of quantitative word problems in the study of human judgment"

(p. 277). We disagree with this conclusion. Given that information is routinely presented in the form of summary statistics, the use of word problems provides a convenient and informative format for evaluating people's use of information. Since experts ranging from policy makers to physicians rely on summary statistics to inform their decisions the emphasis in base rate research may benefit from a shift from demonstrations of error to reducing error through changes in the information format. Our results, coupled with those of Christensen-Szalanski and Beach, indicate that the format of the base rates (experience versus summary statistics) and the response format (choice proportions versus judgments from word problems) have a large impact on how base rate information is used. These findings are likewise consistent with those of Gigerenzer and Hoffrage's (1995) who found that information format greatly affects base rate usage.

The present experiments contribute to our understanding of base rate usage in several key ways. They demonstrate that base rate usage is a function of the relative degree of consistency of both base rates and cue accuracy. They demonstrate that base rate usage is complex and directly related to the statistical characteristics of the base rate information as well as the formats in which the base rates are provided and the responses are obtained. They demonstrate that knowledge of base rates is insufficient for appropriate use of base rates in the traditional format of summary statistics as used in word problems. They also demonstrate that the base rate event frame, long-term versus short-term, is an important characteristic of the ecology to which participants are sensitive.

Brunswik (1956, 1957) argued that experimenters should not assume that participants use cues as though the cues were perfect predictors because humans evolved to cope in uncertain circumstances where the consistency and diagnosticity of information sources varied. He argued instead that researchers ought to fluctuate the diagnosticity (or ecological validity in Brunswik's terms) of cues and evaluate whether ensuing decisions reflect a parallel shift in utilization. Base rates of previous events are a cue for predicting future events.

When past base rates are presented only as summary statistics, participants have only long-term base rate information available to them. The Volvo and crocodile examples discussed in the introduction may not be adequately described by the long-term base rates. Base rate as a category depends on the reference class. The base rate for rain this month may be very different from the average monthly base rate for rain this year. This fact makes it important that participants are sensitive to fluctuations in experienced base rate consistencies.

In sum, base rate usage was found to be a function of past experience, both of base rate diagnosticity and base rate consistency. The conservative decision maker, not knowing when base rates are high or low, may use them cautiously. Cautious usage may be an adaptive bias. Although this possibility was suggested by Goodie and Fantino (1999a) and tested in a Monte Carlo simulation by Todd and Goodie (2002), to our knowledge these are the first empirical studies to demonstrate that the use of base rates is based on base rate consistency as well as diagnosticity.

ACKNOWLEDGEMENT

We thank Erika Mitchell, Jamie Morrow and Katherine O'Brien for assistance in data collection.

Preparation of this manuscript was partially supported by Grant 1R01MH067827-01A2 from the National Institute for Mental Health to A.S.G.

REFERENCES

Bar-Hillel, M. (1980), The base-rate fallacy in probability judgments, *Acta Psychologica* 44, 211–233.

Bjorkman, M. (1984), From dualism to pluralism: The third world in judgment under uncertainty, in Legerspetz K.M.J. and Niemi P. (eds.), *Psychology in the 1990s*, Elsevier, Amsterdam.

Brunswik, E. (1956), *Perception and the Representative Design of Psychological Experiments*, University of California Press, Berkeley.

Brunswik, E. (1957), Scope and aspects of the cognitive problem, in Gruber, H. Hammond, K. R. and Jessor, R. (eds.), *Contemporary Approaches to Cognition*, Harvard University Press, Cambridge, Mass., pp. 5–31.

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

Cooksey, R.W. (1996), *Judgment Analysis: Theories, Methods, and Applications*, Academic Press, New York.

Edgell, S. E., Roe, R. M. and Dodd, C. H. (1996), Base rates, experience, and the big picture, *Behavioral and Brain Sciences* 19, 21.

Estes, W. K., Campbell, J. A., Hatsopoulos, N. and Hurwitz, J. B. (1989), Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, 556–571.

Gigerenzer, G. (1991), How to make cognitive illusions disappear: Beyond "heuristics and biases", in Stroche, W. and Hewstone, M. (eds.), *European Review of Social Psychology* Vol. 2, Wiley, New York, pp. 83–115.

Gigerenzer, G. and Hoffrage, U. (1995), How to improve Bayesian reasoning without instruction: frequency formats, *Psychological Review* 102, 684–704.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Kruger, L. (1989), *The Empire of Chance*, Cambridge University Press, New York.

Ginossar, Z. and Trope, Y. (1987), Problem solving in judgment under uncertainty, *Journal of Personality and Social Psychology* 52, 464–474.

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

Goodie, A. S. and Fantino, E. (1995), An experientially derived base-rate error in humans, *Psychological Science* 6, 101–106.

Goodie, A. S. and Fantino, E. (1996), Learning to commit or avoid the base-rate error, *Nature* 380, 247–249.

Goodie, A. S. and Fantino, E. (1999a), Base rates versus sample accuracy: competition for control in human matching to sample, *Journal of the Experimental Analysis of Behavior* 71, 155–169.

Goodie, A. S. and Fantino, E. (1999b), What does and does not alleviate base-rate neglect under direct experience, *Journal of Behavioral Decision Making* 12, 307–335.

Hammond, K. R. (1996), *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*, Oxford University Press, New York.

Kahneman, D., Slovic, P. and Tversky, A. (eds.). (1982), *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, Cambridge, Mass.

Kahneman, D. and Tversky, A. (1972), Subjective probability: a judgment of representativeness, *Cognitive Psychology* 3, 430–454.

Koehler, J. J. (1996), The base rate fallacy reconsidered: descriptive, normative, and methodological challenges, *Behavioral and Brain Sciences* 19, 1–53.

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

Lovett M. C. and Schunn C. D. (1999), Task representations, strategy variability and base-rate neglect, *Journal of Experimental Psychology: General* 128, 107–130.

McKenzie, C. R. M. (1998), Takin into account the strength of an alternative hypothesis, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24, 771–792.

Nisbett, R. E. and Borgida, E. (1975), Attribution and the psychology of prediction, *Journal of Personality and Social Psychology*, 32, 932–943.

Nisbett, R. E. and Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, Englewood Cliffs, NJ.

Nosofsky, R. M., Kruschke, J. K. and McKinley, S. C. (1992), Combining exemplar-based category representations and connectionist learning rules, *Journal of Experimental Psychology: Learning Memory and Cognition* 18, 211–233.

Shanks, D. R. (1990), Connectionism and the learning of probabilistic concepts, *Quarterly Journal of Experimental Psychology* 42A, 209–237.

Todd, P. M. and Goodie, A. S. (2002), Testing the ecological rationality of base rate neglect, in *From Animals to Animats 7: Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, MIT Press/Bradford Books, Cambridge, Mass.

Tversky, A. and Kahneman, D. (1980), Causal schemas in judgments under uncertainty, in Fishbein, M. (ed.), *Progress in Social Psychology*, Erlbaum, Hillsdale, NJ, pp. 49–72.

Tversky, A. and Kahneman, D. (1982), Evidential impact of base rates, in Kahneman, D. Slovic, P. and Tversky, A. (eds.), *Judgment under Uncertainty: Heuristics and Biases,* Cambridge University Press, Cambridge, pp. 153–162.

Address for correspondence: Philip Dunwoody, Department of Psychology, Juniata College, 1700 Moore Street, Huntigdon, PA 16652, U.S.A. Tel.: +1-814-641-5333; Fax: +1-814-641-3695; E-mail: drdunwoody@yahoo.com