BEHAVIORAL AND BRAIN SCIENCES (1996) **19,** 1–53
Printed in the United States of America

The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges

Jonathan J. Koehler
Department of Management Science and Information Systems, Graduate School of Business, University of Texas at Austin, Austin, TX 78712
Electronic mail: *koehler@mail.utexas.edu*

Abstract: We have been oversold on the base rate fallacy in probabilistic judgment from an empirical, normative, and methodological standpoint. At the empirical level, a thorough examination of the base rate literature (including the famous lawyer–engineer problem) does not support the conventional wisdom that people routinely ignore base rates. Quite the contrary, the literature shows that base rates are almost always used and that their degree of use depends on task structure and representation. Specifically, base rates play a relatively larger role in tasks where base rates are implicitly learned or can be represented in frequentist terms. Base rates are also used more when they are reliable and relatively more diagnostic than available individuating information. At the normative level, the base rate fallacy should be rejected because few tasks map unambiguously into the narrow framework that is held up as the standard of good decision making. Mechanical applications of Bayes's theorem to identify performance errors are inappropriate when (1) key assumptions of the model are either unchecked or grossly violated, and (2) no attempt is made to identify the decision maker's goals, values, and task assumptions. Methodologically, the current approach is criticized for its failure to consider how the ambiguous, unreliable, and unstable base rates of the real world are and should be used. Where decision makers' assumptions and goals vary, and where performance criteria are complex, the traditional Bayesian standard is insufficient. Even where predictive accuracy is the goal in commonly defined problems, there may be situations (e.g., informationally redundant environments) in which base rates can be ignored with impunity. A more ecologically valid research program is called for. This program should emphasize the development of prescriptive theory in rich, realistic decision environments.

Keywords: base rates; Bayes theorem; fallacy; judgment; natural ecology; probability

Introduction

Suppose you are the coach of an Olympic basketball team. The score is tied and your team has possession of the basketball with just 5 seconds left in the game. After calling a time-out, you must decide which of two players, Murphy or McGee, will take the final shot that could catapult your team to victory. You know that Murphy made 60% of his shots this year, whereas McGee made 40% of his shots, although neither player had much experience playing against the type of defense that you expect to encounter here. You also suspect that McGee is a slightly better shooter in pressure-packed situations such as this, although your supportive data are limited. Whom should you choose to take the last shot, Murphy or McGee? Are there principles that you, as a coach, can use to maximize the chances that the selected shooter will come through with a game-winning basket? Are there principles that bettors should follow to predict which team will win the game?

Psychologists, particularly decision theorists, routinely use Bayes' theorem[1] as a normative model for aggregating base rate[2] and other probabilistic information (Fischhoff & Beyth-Marom 1983; Slovic & Lichtenstein 1971; von Winterfeldt & Edwards 1986). Meehl and Rosen (1955) recommended this approach for clinical diagnoses, and researchers in other fields – including accounting (Joyce & Biddle 1981), forensic science (Evett 1986; Sullivan & Delaney 1982), and on occasion law (Fienberg & Schervish 1986; Finkelstein 1978; Kaye 1989) – likewise recommend greater use of base rates and Bayesian information aggregation methods.

Although an increased attention to base rates and Bayesian methods will sometimes improve predictive accuracy, this target article rejects the idea that there is a single, clear normative standard for using base rate information in most realistic decision situations. This is not because classical probability theory leads to unresolvable paradoxes or is logically flawed, as some have intimated (Brilmayer & Kornhauser 1978; Cohen 1981a; Jonakait 1983). It is because base rate information does not map unambiguously into the Bayesian framework in most real world problems. There is a big difference between identifying a theoretically sound normative rule for aggregating probabilistic information of a certain sort (i.e., prior odds and likelihood ratios) and fairly *applying* that rule to a broad class of base rate tasks. Applicability of the rule depends critically on how well the task or, more precisely, the decision-maker's representation of the task, meets the assumptions of the rule. Where key assumptions are violated or unchecked, the

 0140-525X/96 $9.00+.10

superiority of the normative rule reduces to an untested empirical claim.

Empirical research on base rate usage has been dominated by the perspective that people ignore base rates and that it is an error to do so. For many years, the so-called base rate fallacy, with its distinctive name and arsenal of catchy and counterintuitive illustrations (e.g., the lawyer-engineer and cab problems), had a celebrity status in the literature. Through the mid-1980s, psychologists routinely extolled the robustness of the empirical phenomenon, whereas issues related to the applicability of the Bayesian normative model received little consideration.[3] One influential review confidently concluded that, "The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact" (Bar-Hillel 1980, p. 215; see also Borgida & Brekke 1981).

Recently, some (but not all) commentators have softened their views. Conditions under which base rates are properly used or properly ignored have been considered, and some – including the architects of the base rate fallacy – concede that the normative issue may be more complicated than previously supposed (Bar-Hillel 1990; Tversky & Kahneman 1982; see also Payne et al. 1992). This perspective is pursued and extended here. It is argued that the case for a general base rate fallacy has been overstated at both the descriptive and normative levels. Not only is there little evidence that base rates are routinely ignored, but a critical review of the recent literature shows that base rates usually influence judgments and often do so in reasonable ways.

In addition, the present article suggests that the existing paradigm, within which base rate research has been conducted, is at least partly responsible for misperceptions of base rate phenomena. Our knowledge and understanding of how people should and actually do use base rates is unlikely to increase substantially from examining the results of additional laboratory experiments that search for performance errors relative to a narrow and abstract Bayesian norm. Instead, a more ecologically valid program of research must be pursued. Such a program would examine base rate usage in a variety of real world domains in light of the task representations and aims of the decision makers. The primary goals of this new research program would be (1) to identify the conditions under which decision makers make more and less use of base rates in the natural ecology, (2) determine whether and when people would make "better" decisions – as measured by some ecologically relevant performance criterion – by attaching more or less weight to base rates than they actually do, and (3) to develop guidelines that people can use to make better decisions in information aggregation problems. Despite its limitations, the as-yet-unsynthesized base rate literature provides a valuable starting point for addressing the first goal. Progress on the second and third goals awaits empirical study of base rate use and decision quality in the natural ecology.

An important but infrequently asked question in the judgment literature is: What is meant by decision quality and how should it be operationalized? Sometimes it is reasonable to equate decision quality with predictive accuracy. If investor A is able to make more accurate stock market forecasts than investor B, investor A can make more money than investor B. Here, we may be comfortable measuring decision quality in terms of a simple performance criterion: profitability. Other times, policy considerations apart from accuracy affect the quality or appropriateness of a decision. In U.S. courtrooms, for example, arguments based on diagnostic but "unfair" base rate evidence are often rejected (Koehler 1992; 1993a). This suggests that conclusions about how well people use base rates and prescriptions for improving decision making should depend on the extent to which different degrees of reliance on base rates would lead to better and worse outcomes in particular situations. Unlike the current approach, this situation-specific, outcome-based approach can lead to useful prescriptive guidelines.

Sections 1 and 2 below address the descriptive component of the base rate fallacy. Section 1 evaluates, and ultimately rejects, the strong claim that base rates are ignored in the paradigmatic lawyer-engineer problem and in social judgment tasks. Section 2 reviews the experimental base rate literature and links this diverse set of studies through a general theory of base rate usage. The theory, which emphasizes task features and cues that sensitize decision makers to the base rate and a willingness to use base rates in tasks that can be represented in a frequentist manner, provides a starting point for examining base rate usage in the natural ecology. Sections 3 and 4 address normative issues. Section 3 examines the base rate reference class problem, giving special attention to L. Jonathan Cohen's (1981a) provocative – but ultimately unsustainable – thesis that base rates derived from insufficiently relevant reference classes ought to be disregarded. Section 4 considers the assumed relation between base rate tasks and the Bayesian model and concludes that normative claims are restricted by the lack of an unambiguous mapping of tasks onto this model. Section 5 explores issues related to moving the base rate paradigm out of the laboratory and into the natural ecology. After arguing that the application of a narrowly construed Bayesian rule for base rate use in the natural ecology will be unproductive, it identifies some conditions under which failures to attend to base rates in the natural ecology will be more or less costly. Section 5 also calls for the development of performance measures that take account of features other than predictive accuracy where appropriate. Section 6 offers a summary and conclusion.

1. Are base rates ignored?

Hundreds of laboratory studies have been conducted on the use of base rates in probability judgment tasks. Although this research has not produced a simple picture of when, why, or how base rates are used, investigators frequently conclude that base rates are universally ignored: "Recent psychological research suggests that people in general, including trained statisticians, ignore base rate probabilities" (Christensen-Szalanski & Bushyhead 1981, p. 931); "information about base rates is generally observed to be ignored" (Evans & Bradshaw 1986, p. 16); "it has repeatedly been shown that people commit the base-rate fallacy, that is, that they ignore base-rate frequencies and, instead, base their judgments solely on the similarity between the individual's personality and the prototypes of the categories under consideration" (Ginossar & Trope 1987, p. 464); "base rate information concerning categories in a

Table 1. *Base rate usage in the lawyer-engineer problem*

	KT '73	SGLS '76	WH '78	GT '80	FB '84			H '84	GT '87		GHB '88				
Diagnostic individuating information[a]															
High base rate (70%) posteriors	.55	.83	.79	.70[b]	.04	1.0	1.0	.57[c]	.35	.80	.62	.61	.36	.81	.38
Low base rate (30%) posteriors	.50	.66	.71	.68[b]	.00	.96	.70	.43[c]	.22	.72	.46	.45	.34	.71	.25
Observed difference	.05	.17	.08	.02[b]	.04	.04	.30	.14[c]	.13	.08	.16	.16	.02	.10	.13
Deviation from individualized Bayesian analysis	—	—	.30	—	—	—	—	—	—		.17	.16	.20	.16	.11
Nondiagnostic individuating information[a]															
High base rate (70%) posteriors	.50[c]	.70	.54	.59[b]		.73		.70[c]	.65			.61			
Low base rate (30%) posteriors	.50[c]	.45	.36	.31[b]		.38		.30[c]	.45			.60			
Observed difference	0	.25	.18	.28[b]		.35		.40[c]	.20			.01			
Deviation from individualized Bayesian analysis	—	—	.40	—		—		—	—			.30			

Sources: Kahneman & Tversky (KT) 1973; Swieringa et al. (SGLS) 1976; Wells & Harvey (WH) 1978; Ginossar & Trope (GT) 1980; Fischhoff & Bar-Hillel (FB) 1984; Hamilton (H) 1984; Ginossar & Trope (GT) 1987; Gigerenzer et al. (GHB) 1988.
Note: Fischhoff and Bar-Hillel (1984) used 10 diagnostic and 10 nondiagnostic "profiles." Due to limited space, only the results of profiles identical to those used by Kahneman and Tversky (1973) are reproduced here. The results for profiles used in Ginossar and Trope (1987) and in Gigerenzer et al. (1988) are listed separately. Results for the remaining studies are averaged across profiles as reported in the original studies.
[a]All cell values are means unless otherwise noted.
[b]Cell values are estimated medians from Figure 1 and Table 1 in the original study.
[c]Cell values are medians.

population is ignored in estimating category membership of a sample of a population" (Nisbett & Borgida 1975, p. 935); "many (possibly most) subjects generally ignore base rates completely" (Pollard & Evans 1983, p. 124). These characterizations of the base rate literature – some of which are made by researchers whose own work shows attentiveness to base rates – are dreadfully misleading. Even if one overlooks the low ecological validity of much of this literature, *few studies have shown* that base rates are completely disregarded by most or even some people (Bar-Hillel 1990; Lynch & Ofir 1989; Manis et al. 1980; Wells & Harvey 1978).

Some of the confusion may be attributable to the unfortunate use of the term "ignore" by some investigators to describe data suggesting only that subjects attach *relatively less weight* to base rate information than to descriptive, individuating information. In their classic and paradigmatic lawyer-engineer experiment, some of Kahneman and Tversky's (1973) subjects were told that a panel of psychologists had written personality descriptions of 30 engineers and 70 lawyers based on the results of personal interviews and personality tests. Subjects were then told that five descriptions had been chosen at random from this pool. After reading these descriptions, subjects assessed the probability that each of the persons described was an engineer. Other subjects were given the same descriptions and task but with 70 engineers and 30 lawyers. Although a small, but statistically significant, main effect for the base rate was found ($p < .01$), this study is widely cited as a convincing demonstration that baser rates are "ignored" (e.g., Fagley 1988; Nisbett & Borgida 1975). But even if one regards small base rate effects as evidence that base rates are ignored, additional work on the lawyer-engineer problem suggests that the conclusions others have drawn from this study[4] should be viewed with caution.

1.1. Inconsistent results in the lawyer-engineer problem

1.1.1. Empirical level. There have been numerous attempts to replicate the Kahneman and Tversky (1973) lawyer-engineer results. Table 1 presents the posterior probability estimates given by subjects in eight lawyer-engineer experiments that were conducted and reported in a comparable manner. These experiments were identical or nearly identical to Kahneman and Tversky's experiment and examined the impact of both diagnostic and nondiagnostic individuating information.[5] The top panel of Table 1 uniformly indicates that base rates influenced subjects' judgments in the presence of diagnostic individuating information. Differences between high and low base rate groups ranged from 2% to 30%, with an average near 11%.

The bottom panel of Table 1 does not present a consistent picture of whether or how base rates influenced judgments in the presence of nondiagnostic individuating information. In Kahneman and Tversky (1973) and in Gigerenzer et al. (1988), judgments for high and low base rate groups were virtually indistinguishable, suggesting that base rates had little impact on final judgments. However, the remaining four studies revealed strong base rate effects. Indeed, Hamilton (1984) concluded that subjects in the nondiagnostic individuating information conditions based their judgments exclusively on the base rates provided.

Gigerenzer et al. (1988) also noted the presence of large between-study differences in some of the lawyer-engineer experiments. They proposed that subjects who saw diag-

nostic individuating information in previous problems would continue to make extensive use of individuating information at the expense of base rates, because they would adopt a cognitive "set" for these problems. Subjects who had not seen diagnostic individuating information would not adopt such a set and be more likely to use base rates. Fischhoff and Bar-Hillel (1984) and Ginossar and Trope (1987) provided data consistent with the cognitive set explanation.

1.1.2. Normative level. The lawyer-engineer problem was designed, in part, to permit comparison of subjects' performance against a normative standard. However, the failure of most researchers to devise individualized normative criteria restricts the validity of these comparisons. As Wells and Harvey (1978) explained, normative performance criteria derived from summary statistics are not appropriate for problems such as the lawyer-engineer problem. Individual subjects may give perfectly Bayesian responses, yet their mean responses may look distinctly non-Bayesian. This can occur because the Bayesian model predicts a curvilinear relation between the two base rate groups, whereas means, for example, are derived from a linear process. As the variance in subjects' judgments about the diagnostic value of the individuating information increases, the plotting of means on a curvilinear Bayesian prediction line becomes increasingly misleading. Normative performance criteria based on medians are also problematic in that they make incomplete use of the full range of subjects' responses.

Wells and Harvey (1978) devised a normative analysis based on individual responses to avoid these pitfalls. Gigerenzer et al. (1988) also adopted this procedure. For each probability estimate (p) made by subjects in one base rate condition, a probability estimate (p^*) is computed to determine how subjects would have responded if they had been in the other base rate condition and responded in a Bayesian manner.[6] The mean of the Bayesian (p^*) scores in one base rate condition is then compared with the mean of the actual (p) scores provided by subjects in the other condition. A comparison of these individualized normative analyses with the empirical data obtained in Wells and Harvey and in Gigerenzer et al. appears in the last two rows of the top and bottom panels of Table 1. The data indicate that observed differences between the high and low base rate groups in Wells and Harvey and in at least one of the Gigerenzer et al. vignettes (No. 3) were smaller than would be predicted by the Bayesian model. However, the empirical data in at least three of the Gigerenzer et al. vignettes (Nos. 1, 2, 5, and possibly 4) were approximately in line with normative Bayesian analyses.

Regardless of how one accounts for outcome variation in the lawyer-engineer problem, the data do not provide strong support for the conclusion that base rates are ignored or even "largely ignored" (Kahneman & Tversky 1973, p. 242). Instead, the data indicate that, at least under some conditions, base rates influence probability judgments considerably, and sometimes to a degree that satisfies an abstract normative standard.

1.2. Social judgment studies

By the mid-1970s, social judgment studies suggested that base rates were ignored here as well, but further investigation weakened the claim.

In one frequently cited study, Nisbett and Borgida (1975) argued that subjects ignore consensus information when making causal attributions. In their study, some subjects were told the results of two experiments, one concerning shock taking (Nisbett & Schachter 1966) and the other concerning helping behavior (Darley & Latane 1968). The counterintuitive results produced by these studies – many subjects accepted high levels of shock in the former and failed to help a person in distress in the latter – were designed to serve as base rate information for Nisbett and Borgida's experiment. Subjects were then provided with a description of (or interview with) a target person and asked to rate the extent to which the target's behavior was caused by his personality or by the situation. Nisbett and Borgida found that these judgments were not affected by knowledge of the base rates as given in the shock-taking and helping experiments.

But is it reasonable to expect subjects to accept and use experimenter-supplied base rates when making locus of control attributions about a target case? Two arguments suggest otherwise. First, if subjects are provided with base rates that are counterintuitive, inconsistent with their experiences, or based on what are believed to be unrepresentative samples, the base rates probably will and should be discounted. In the Nisbett and Borgida (1975) study, many subjects may have had prior beliefs about the likelihood of taking strong shocks and helping persons in distress that were markedly different from the counterintuitive base rates. When Wells and Harvey (1977) repeated this study with subjects who received reassurances about the validity and applicability of the base rate, strong base rate effects on causal attributions were observed. A second reason for challenging Nisbett and Borgida's conclusion is that there is no logical or necessary relation between base rate frequency for a behavior and the causal locus (i.e., internal or external) of that behavior. As Wells and Windschitl (1994, sect. 2.3) wrote, "as the public learns that the base-rate for child abuse is much higher than previously thought, there is no rational requirement that attributions about abusers must shift toward external causes."

Locksley and her colleagues have argued that social stereotype base rates are disregarded when individuating information is made available (Locksley et al. 1980; 1982). Subjects in the Locksley et al. studies reportedly ignored sex stereotypic beliefs about assertiveness and other individually held stereotypic beliefs (e.g., traits of "day people" and "night people") when making personality predictions in the presence of minimally diagnostic individuating information.

Recent studies challenge the robustness of this conclusion, however. For example, sex stereotypic beliefs about height apparently exert a strong influence on height estimates made by children and adults for male and female targets, even when individuating target photo data are presented (Biernat 1993; Nelson et al. 1990). Krueger and Rothbart (1988) showed that predictions about a target person's aggressiveness were influenced both by stereotypes and by individuating information. Here, minimally diagnostic individuating information in the form of a single behavioral act was not sufficient to override the stereotype (base rate) effect. Krueger and Rothbart speculated that discrepancies between their results and those of Locksley et al. (1980; 1982) were due to differences in the diagnosticity levels of the information used in the two studies. That is, aggressiveness may be more stereotypic of gender

than assertiveness, and the behavioral acts described in the Locksley experiment may have conveyed more information than those used in the Krueger and Rothbart experiments. Hilton and Fein (1989), however, used a task similar to that of Locksley et al. (1980) and found robust effects for stereotypes on trait predictions. According to them, only individuating information that is "useful across many social judgment tasks" can reduce the impact of stereotype base rates (Hilton & Fein 1989, p. 201).

Can the Locksley et al. results be reconciled with these studies? Here again the problem appears to be a failure to use sufficiently individualized criteria for measuring the impact of different types of information on judgment. When Rasinski et al. (1985) repeated the Locksley et al. (1980, study 2) experiment, taking into account subjects' own stereotypes, the stereotypes were not disregarded. On the contrary, subjects "seemed to be overcautious in revising their stereotype-based judgments when they were presented with individuating diagnostic behavioral information" (Rasinski et al. 1985, p. 322).

In sum, there is little evidence either from the lawyer-engineer problem or from the stereotype literature to support the strong claim that base rates are routinely ignored when individuating information is made available. Indeed, when care is taken to develop an appropriately individualized criterion, even weaker forms of the base rate fallacy do not receive clear and convincing support. Add to this dozens of other studies that have identified extensive base rate use under a broad range of conditions (see sect. 2) and a fascinating puzzle emerges: How did the "base rates are ignored" misperception arise and sustain itself over the years?

1.3. Emergence of the myth

Two explanations for the emergence and persistence of the belief that base rates are widely ignored come to mind. The first invokes a Kuhnian account of scientific belief, and the second a heuristic account. Kuhn's (1962/1970) views concerning the nature of scientific progress and paradigm shift are well known. He stressed that a simple and powerful theory can withstand empirical challenge when the challenging data are not accompanied by a simple, general theory of their own.

Base rate research sprang from the heuristics and biases paradigm that dominated judgment and decision-making research in the 1970s and 1980s. The paradigm held that people's intuitive judgments about probabilistic events are made via simple error-prone heuristics. One heuristic – representativeness – suggests that people's judgments about the probability of category membership depend on how similar are the features of the target to the essential features of the category (Kahneman & Tversky 1972). Thus, judgments that Viki is an accountant depend upon how similar are Viki's interests, background, talents, and so on to those ordinarily associated with accountants.

As evidence for this heuristic mounted, base rate neglect became an easy sell. If people use the representativeness heuristic, and if base rates are less representative of a category's central features than individuating information, it follows that people will ignore base rates. Shortly after empirical support for this phenomenon appeared (Kahneman & Tversky 1973), the "base rate fallacy" (Bar-Hillel 1980) became a favorite instantiation of the heuristics and biases paradigm.

Subsequent research, however, failed to support a complete neglect of base rates. Some studies (e.g., the lawyer-engineer experiments) produced different results with nearly identical stimuli; some studies challenged the representativeness explanation on grounds that base rates receive little consideration even when they are no less representative than other sources of information; and some studies showed that people pay quite a bit of attention to base rates in certain tasks and contexts (see sect. 2). Simply put, the evidence did not support a simple and general base rate fallacy. But, without a compelling alternative explanatory theory, the underlying principle was too attractive to abandon on account of data. The result is a literature that has been simplified, misinterpreted (Lopes 1991), and selectively cited (Christensen-Szalanski & Beach 1984) by observers, researchers, and reviewers alike.

Ironically, psychologists' misperception of the base rate literature may also be due to heuristic thinking. In order to draw simple, punchy conclusions from a morass of complex and sometimes conflicting empirical studies, scientists may simplify the significance of the studies to the point where their conclusions misrepresent the data. Lopes (1991) has argued that whereas the heuristics and biases literature provides evidence that people think heuristically, it has been widely misinterpreted as a literature that demonstrates that people's judgments are generally poor. Todd and Morris (1992) observed that simple, general, but inaccurate statements about behaviorism literature have become more authoritative than either the existing data or claims made about the data by the original authors. The construction and acceptance of the Hawthorne effect similarly illustrates the point. Although studies in the late 1920s at the Hawthorne electrical plant are widely cited in authoritative texts and reviews as demonstrating that workers' productivity increased regardless of type of change made in their work environments, this is a serious distortion of the actual findings (Adair 1984; Gillespie 1991; Jones 1992). In a similar vein, the ubiquitous summary statements of base rate neglect distort the empirical literature. The following section considers this literature and attempts to frame it within the context of a general theory.

2. When are base rates used?

If base rates are not uniformly ignored, it is important to know when they are likely to be used to a greater and lesser extent. A great deal of laboratory research has been conducted on this question in recent years and some patterns have emerged. This section pulls together the diverse and as-yet-unsynthesized literature within a framework that focuses on task structure, task representation by the decision maker, and microfeatures of the experimental task. It is suggested that a base rate has its greatest impact in tasks that (1) are structured in ways that sensitize decision makers to the base rate, (2) are conceptualized by the decision maker in relative frequentist terms, (3) contain cues to base rate diagnosticity, and (4) invoke heuristics that focus attention on the base rate (see Fig. 1).

2.1. Task structure

In the typical base rate task, subjects are provided with a base rate summary statistic (e.g., 85% of the cabs in the city are Green) that they are expected to remember, trust, and

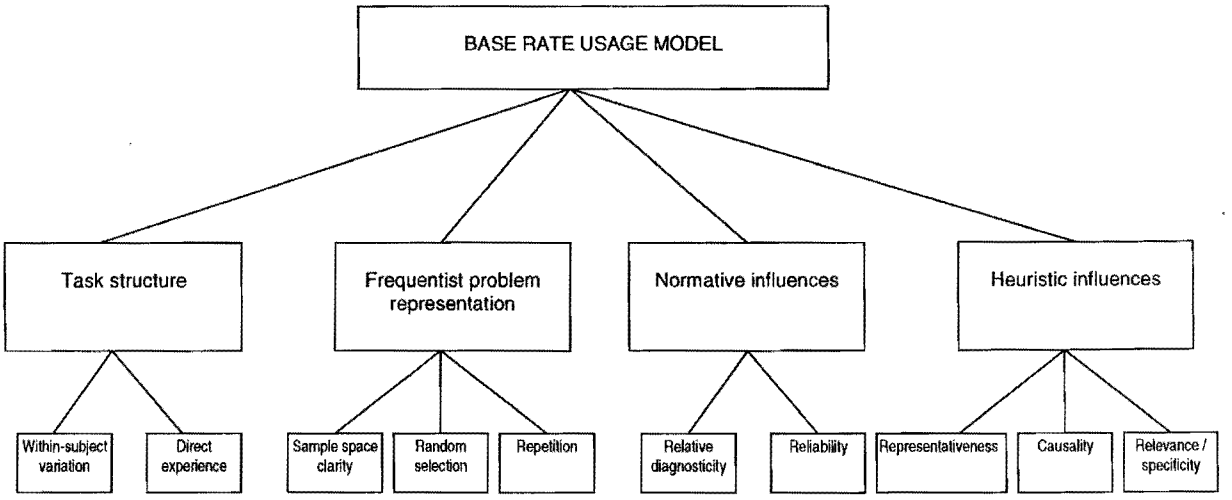

Figure 1. Base rate usage as a function of task structure, problem representation, normative influences, and heuristic influences.

use, even in the face of individuating information that supports an opposing hypothesis. It is not clear that the attention given to base rates in such tasks is representative of the attention they receive in tasks structured in other ways. Tasks that include within-subject base rate manipulations and opportunities for implicit base rate learning appear to sensitize people to base rates and result in greater use of base rates on final judgments.

2.1.1. Within-subject variation. Kahneman and Tversky's (1973) lawyer-engineer problem provided subjects with five descriptions of people under high or low base rate conditions. The individuating information exerted a large effect on subjects' judgments, whereas base rates exerted a smaller effect. But as Ajzen (1977, p. 312) explained, "The within-subjects manipulation of individuating information and the between-subjects manipulation of base rates [in Kahneman & Tversky's experiment] may have sensitized subjects to variations in the former and desensitized them to variations in the latter." That is, people may pay more attention to information that varies relative to information that does not.

Experimental support for the within-subject sensitization thesis has been amassed. Ajzen reported greater use of base rates on a lawyer-engineer type of task when a completely between-subjects design was used. Fischhoff et al. (1979) found greater base rate usage when it was manipulated within-subject than when it was manipulated between-subject. Birnbaum and Mellers (1983) found greater base rate usage in modified cab problems solved in the context of many problems. Schwarz et al. (1991, experiment 2) showed that subjects' reliance on both base rate and individuating information increased in a version of the lawyer-engineer problem when it was varied within-subject. Hinsz and Davidson (1993) obtained a similar result in a task concerned with predicting the success of job applicants.

One explanation for greater use of information that is varied within-subject is that people use variation as a cue to identify task focus and the experimenter's communicative intent (Schwarz et al. 1991). A second explanation is experimental demand (Dawes et al. 1993). Either way, exposure to a variety of base rates and problems can increase base rate use.

Some caution is required. First, not all researchers report a within-subject base rate sensitivity effect (e.g., Dawes et al. 1993; see also Gigerenzer et al. 1988, p. 519, n. 2). Second, this effect may compete with a cognitive "set" effect (discussed in sect. 1.1.1) that follows exposure to highly diagnostic individuating information. Of course, the cognitive set effect, if real, might also occur following exposure to highly diagnostic base rates, in which case the within-subject sensitization effect could be enhanced.

2.1.2. Direct experience. The way in which base rate information is learned may affect the way decision makers use this information. For example, when base rates are directly experienced through trial-by-trial outcome feedback, their impact on judgments increases (Lindeman et al. 1988; Manis et al. 1980; Medin & Edelson 1988). In Manis et al. subjects were provided with 50 yearbook pictures of male students and asked to predict their attitudes about each of two issues (marijuana legalization and mandatory seatbelt legislation). After each prediction, some subjects were told that 20% of the targets favored the proposed legislation. Other subjects were told that 80% of the targets favored the legislation. The influence of the feedback on base rate usage was quick and dramatic. After viewing 20 pictures, subjects who experienced the 80% base rate were more than twice as likely to predict that a target person favored the proposed legislation than subjects who experienced the 20% base rate.[7]

Studies in professional contexts indicate that the effects of directly experiencing base rates are not restricted to the laboratory. Butt (1988) showed that auditors learned and used the base rate for financial statement errors most easily by directly experiencing those errors. Christensen-Szalanski and Bushyhead (1981) showed that physicians who learned the low base rate for pneumonia (3%) from their clinical experience relied heavily on this base rate when making diagnoses. Christensen-Szalanski and Beach (1982) narrowed this result somewhat by showing that personally experienced base rates were used only by those who also experienced the relationship between the base rate and the diagnostic information.

Directly experienced base rates may be accorded more weight than indirectly experienced base rates because they

invoke an implicit rather than an explicit learning system. Information that is learned implicitly may be better remembered, more easily accessed, or otherwise more meaningfully instantiated than information learned explicitly (Holyoak & Spellman 1993; Medin & Bettger 1991; Medin & Edelson 1988; see also Weber et al. 1993). When the implicit learning experience comes in the form of trial-by-trial learning, the information at each trial may be encoded as a separate "trace" (Hintzman et al. 1982). In this way, multiple traces develop, and the information associated with these traces may be cognitively available. This contrasts with the explicit learning of a single summary statistic that does not produce multiple traces and that has been associated with less accurate judgments (Hintzman et al. 1982).

Medin and Edelson (1988) found that subjects in a categorization task who learned base rates experientially made extensive use of this information but did not make explicit reference to base rates. Instead, they typically reported that their responses were determined by the first category that came to mind. Spellman (1993) has argued that directly experienced base rates are particularly likely to be used in problems that require behavioral solutions (as opposed to verbal reports). To the extent that an implicit learning structure accounts for the direct-experience base rate effect, we might expect greater base rate use in the experientially rich natural ecology than in the laboratory, where base rates are usually provided as summary statistics.

A related explanation for the direct-experience base rate effect is that personally experienced information is more vivid or salient – hence more readily available – than data that are learned in other ways (cf. Borgida & Nisbett 1977; Brekke & Borgida 1988; Nisbett & Borgida 1975; Nisbett et al. 1982; Nisbett & Ross 1980). A third explanation is that people are more trusting of self-generated base rates, particularly those acquired through first-hand experience. If people are suspicious of the veridicality of experimenter-supplied base rates, a "bias" in favor of self-generated and directly experienced base rates should be expected. Although less cognitive than the implicit learning and vividness explanations, this third argument gains credence from studies showing greater base rate use when accuracy assurances are provided (Hansen & Donoghue 1977; Wells & Harvey 1977) and when base rates are self-generated (Ungar & Sev'er 1989).

2.2. Frequentist problem representation

Demonstrations of poor statistical reasoning in some contexts do not preclude the possibility of good statistical skills in others. Thus, although people sometimes behave as if they believe that evidence derived from small samples should be accorded the same value as evidence derived from large samples (Kahneman & Tversky 1972; Tversky & Kahneman 1971), others have noted that people demonstrate an intuitive appreciation of the law of large numbers principle in many everyday tasks. For example, Nisbett et al. (1983) pointed out that most people understand that a lower-quality football team might defeat a better team on any given day, but that they are unlikely to prevail over the longer series of games. More theoretically, these researchers argued that people often solve problems with the aid of intuitive, albeit imperfect, versions of sound statistical rules. According to the authors, these statistical heuristics are invoked when (1) the sample space and sampling process are clear, (2) the role of chance in the production of task outcomes is signaled clearly by the nature of the task, and (3) the culture prescribes that the solution to the task requires statistical reasoning. Nisbett and his colleagues have provided empirical support for this model (Fong et al. 1986; Jepson et al. 1983; Lehman et al. 1988).

The Nisbett et al. (1983) model provides guidance for identifying the conditions under which base rates will be used to greater and lesser degrees. Parts (a) and (b) of the model are characteristic of such prototypical frequency tasks as coin flipping, dice rolling, and card selection games. In general, frequency tasks have unambiguous sample spaces, employ random selection procedures, and are imperfectly predictable as a result of random variation. Frequency tasks are also characterized by their reference to outcomes over a series of independent trials. If statistical heuristics are more likely to be employed in tasks that have – or are perceived to have – frequency task features, it stands to reason that base rates will have more influence as well in problems that people represent as frequency tasks. Empirical support is provided by Gigerenzer and his colleagues, who argue that many cognitive illusions (including base rate neglect) "disappear" when subjects are able to represent the problems as relative frequencies rather than single-event probabilities (Gigerenzer 1991; in press; Gigerenzer & Hoffrage, in press; Gigerenzer et al. 1991).

In short, a frequentist problem representation thesis is proposed, whereby the presence of such prototypical frequency task features, as random and repeated selection of targets from reference classes whose size and composition are stable and known, facilitates a frequentist representation. This, in turn, promotes rule-based statistical thinking that includes reliance on available base rates.

2.2.1. Unambiguous sample spaces. The frequentist representation thesis predicts that base rates will be used to a greater extent when they are derived from well-defined sample spaces or reference classes. In one study, subjects were asked to guess which of two cages random samples of lettered balls were drawn from, in light of base rate information about the distribution of lettered balls in each cage (Grether 1980). Subjects' judgments were influenced by both the sample information and the base rates, although the latter were accorded relatively less weight.

In Grether's (1980) experiment, base rates were clearly derived from well-defined reference classes. However, most base rates encountered in the natural environment, and some provided in laboratory studies, are derived from more ambiguous reference classes. Consider, for example, the well-known cab problem:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: (i) 85% of the cabs in the city are Green and 15% are Blue. (ii) A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases. Question: What is the probability that the cab involved in the accident was Blue rather than Green? (Tversky & Kahneman 1980, p. 62.)

Here there may be some confusion as to whether an appropriate reference class is "cabs in the city," "cabs in accidents," "cabs in accidents at night," and so on. When

people are not provided with reference class information they regard to be appropriate, they may have trouble representing the problem at all, let alone representing it in frequentist terms. When Ginossar and Trope (1987, experiment 6, p. 472) reframed the cab problem and the lawyer-engineer problem as games of chance with unambiguous sample spaces, they observed "robust base rate effects." Whereas the differences in mean probability estimates for the high and low base rate groups in the original form of these problems were small (2% in the lawyer-engineer problem, 4% in the cab problem), the differences between high and low base rate groups in the reframed versions were substantial (24% in the lawyer-engineer problem, 25% in the cab problem). Apparently, clarifying the reference class facilitated a frequentist representation of the task, which led to greater use of the available base rates.

2.2.2. Random selection. A recognition that chance plays an important role in producing outcomes also promotes a frequentist problem representation. However, a body of research suggests that people do not fully appreciate the role of randomness in their environment. People see patterns where none exist (Gilovich et al. 1985; Tversky & Gilovich 1989) and respond to single-shot chance events as if they could be controlled (Langer 1975). It is not surprising, then, that the use of base rate information depends, in part, on the clarity and credibility of the random component.

In the lawyer-engineer problem, subjects are typically told that the descriptions provided were selected at random. But the descriptions are *not* randomly selected and subjects may suspect as much once exposed to their stereotypical content. When Gigerenzer et al. (1988, experiment 1) ran a version of the lawyer-engineer problem in which subjects performed and observed the random sampling for themselves, the influence of base rates was much stronger than when random sampling was only verbally asserted. Similarly, reassurances about the representativeness of base rate sample data in causal attribution studies promoted greater use of these data (Hansen & Donoghue 1977; Wells & Harvey 1977).

2.2.3. Repetition. Repetitive sampling from well-defined sample spaces also facilitates frequentist task representation. In one well-known study physicians were supplied with an extremely low base rate for a hypothetical disease along with individuating test data that strongly suggested the disease was present (Casscells et al. 1978). The physicians' probability judgments associated with disease presence in a single patient revealed little influence from the base rate.

But Cosmides and Tooby (in press) showed that people attach greater weight to the base rate when asked to make the judgment in a frequentist manner. After replicating the Casscells et al. (1978) result, Cosmides and Tooby (experiment 4) showed that when subjects were asked to estimate the number of patients out of 100 who really had the disease, a large majority (76%–92%) gave answers that were consistent with a Bayesian analysis (in which the disease base rate was presumed to equal each subject's prior probability).

Repetitive or multiple sampling also appeared to promote a frequentist task representation and sound statistical reasoning in the illusion of control paradigm. Koehler et al. (1994) provided subjects with an opportunity to bet on the outcomes of dice tosses in high-control (e.g., tossed die for oneself) or low-control situations (e.g., experimenter tossed die). Consistent with Langer (1975), subjects bet more on a certain set of outcomes under high-control conditions than under low-control conditions in a single roll of the die. However, this phenomenon disappeared in the context of multiple die rolls. Introduction of the multi-shot context may have enabled subjects to represent the task in relative frequentist terms. This representation may have cued them to the underlying random component of the task (i.e., equal base rates for all possible outcomes) and eliminated the illusion of control bias.

In sum, base rate use and sound statistical thinking depend, in part, on whether problems are or can be represented in frequentist terms. Tasks that are perceived to involve random and repeated sampling from well-defined sample spaces promote frequentist representations. Such representations, in turn, increase base rate usage.

2.3. Normative influences

The normative status of certain sensitizing features discussed above is questionable. Why, for example, should the impact of base rates depend on whether they are presented between- or within-subject? On the other hand, many studies indicate that people respond to various features of base rate problems in ways that are appropriate. Most notably, people show greater attentiveness to base rates that are (1) high in relative diagnosticity, and (2) obtained from reliable sources. These tendencies indicate that people can reason critically about the diagnostic implications of variations in base rate evidence.

2.3.1. Relative diagnosticity. Highly diagnostic information should have a greater impact on predictions, beliefs, and attributions than less diagnostic information. Because this normative principle relates to the content of information rather than to its form (e.g., base rate vs. individuating), the influence of base rates should be greater when they are paired with relatively less-diagnostic individuating data than when they are paired with relatively more-diagnostic individuating data.

Ginossar and Trope (1980) gave subjects versions of the lawyer-engineer problem with one of three base rates (30%, 50%, 70%) and one of several individuating descriptions. The descriptions contained either (a) stereotypical features of a lawyer or an engineer, (b) features of both lawyer and engineer stereotypes, or (c) features that were not characteristic of either lawyers or engineers. A fourth group (d) received the descriptions in (a) but was asked to discriminate between two similar professions (electrical engineers and aeronautical engineers). The results showed that whereas subjects paid little attention to the base rates when combined with highly stereotypical and diagnostic individuating information (group a), they paid close attention to the base rates when the descriptions were less diagnostic (groups b, c, and d). The authors reported that the predictions made by groups b, c, and d "were quite close to those prescribed by the Bayesian model" (Ginossar & Trope 1980, p. 236).

These results are consistent with other base rate diagnosticity studies (e.g., Davidson & Hirtle 1990; Fischhoff &

Bar-Hillel 1984; Ofir 1988) as well as studies on the effects of stereotyping. It has been found that strong stereotypic base rates (e.g., sex-based beliefs about aggressiveness or height differences) exert more influence on trait predictions than do weaker stereotypic base rates (e.g., sex-based beliefs about assertiveness. See Hilton & Fein 1989; Krueger & Rothbart 1988; Locksley et al. 1980; 1982; and Nelson et al. 1990).

Base rate extremity. Other things being equal, base rates that are extreme (i.e., close to zero or one) are more diagnostic than those that are not. Some of the early base rate studies concluded that people are not sensitive to base rates of any sort, including extreme ones. For example, Lyon and Slovic (1976) used a problem that was structurally similar to the cab problem and reported that a reduction in the base rate from .15 to .01 had no effect on subjects' judgments. In both instances, subjects apparently paid little attention to the base rate, possibly because they regarded it as irrelevant. But when subjects believe the base rate is at least somewhat relevant, they are sensitive to its extremity. Ofir (1988) and Powell and Heckman (1990) gave subjects versions of the cab problem in which the cab color base rates were varied: 10% and 90% in Ofir (1988), 55% and 90% in Powell and Heckman (1990). Both studies reported strong main effects for base rates. Hamilton (1984) also reported greater use of extreme base rates in versions of the lawyer-engineer problem.

Similar, but less pronounced, effects have been reported in the social judgment literature. When assurances of base rate representativeness were provided, Wells and Harvey (1977) found that subjects' predictions of the percentage of others who would accept a maximum shock were higher for high and moderate base rates (26/34 and 16/34, respectively) than for low base rates (1/34). Differences between high and moderate base rates groups were directionally appropriate but did not reach statistical significance.

In sum, it appears that when base rates are made more extreme or when individuating information is made less diagnostic, the impact of base rates on judgments increases. However, these diagnosticity variations may go unheeded if the base rate is initially perceived to be nondiagnostic.

2.3.2. Reliability. Reliable evidence should have a greater impact on judgments than less-reliable evidence. Because the quantity and source of evidence affect its reliability, evidence derived from large samples or obtained from reliable sources and processes should have a greater impact on judgments than evidence that lacks these features.

As noted earlier, people are sometimes insensitive to sample size concerns. However, when information derived from small and large samples are made available in a single prediction task, people are more sensitive to this important variable. Kassin (1979) presented subjects with a description of a helping experiment and two conflicting sets of results, one derived from a small sample ($n = 10$), the other from a large sample ($n = 50$). Helping was reportedly high in one base rate condition and low in the other. The results revealed a main effect for sample size on subjects' predictions of helping behavior, indicating that the large sample base rate had an appropriately greater impact on judgments than the small sample base rate.

The credibility of the information source is also an important indicator of reliability and diagnosticity. Credibility is defined here in terms of accuracy: the predictions of credible sources are more accurate than those of less-credible sources. Ginossar and Trope (1987, experiment 5) manipulated the credibility of the source of individuating information and showed that base rates exert a greater impact on predictions when individuating personality descriptions are obtained from a low-credibility source (e.g., a beginning student or a palm reader) than when they are obtained from a high-credibility source (e.g., a group of trained psychologists). Similar results were obtained by Birnbaum and Mellers (1983). The impact of base rate source credibility has received less attention.

Whether people understand the difference between the credibility of information and its diagnosticity is another matter. Lynch and Ofir (1989) asked subjects to estimate the probability that a 5 year old Peugeot would remain free of mechanical problems for at least one year after purchase. Subjects were provided with a base rate as well as individuating information from a mechanic who examined the car. The mechanic's accuracy was given as 15%, 58%, or 90%. The results showed a strong interaction between values of the base rate and individuating information. As the numerical value of one type of information was lowered, the main effect for the other type of information increased. Similarly, Hinsz et al. (1988) used a variation of the cab problem to show that as the accuracy of the source of individuating information is incrementally reduced from 80% to 20%, the individuating information has less impact on probability judgments and is perceived as increasingly irrelevant for making those judgments.

In dichotomous judgment tasks it is appropriate to attach less weight to information as its credibility is reduced toward 50%. But as credibility drops below 50%, the diagnosticity of the information begins to increase. The opinions of a mechanic who is right 1% of the time are less credible, but more diagnostic, than those of a mechanic who is right 51% of the time. Decision makers who hear the advice of the high-credibility mechanic can be 51% accurate; decision makers who hear the advice of the low-credibility mechanic can be 99% accurate by not following the advice. This point apparently was lost on subjects in the Lynch and Ofir studies (1989) and in the Hinsz et al. (1988) studies, who attached weight to individuating information as a function of its credibility rather than its diagnosticity.

How concerned should we be about the credibility/diagnosticity confusion? From the standpoint of real world decision making, the answer is probably, not very. Situations in which information from low-credibility sources are more diagnostic than those received from high-credibility sources are probably rare in the natural ecology. Information provided by a low-credibility source for dichotomous judgments generally approaches random performance (i.e., a hit rate of 50%) rather than systematically poor performance (i.e., a hit rate approaching 0%). Consequently, people may not only be unaccustomed to extracting more information from less-credible sources than from more-credible ones, but there may be little reason to worry about the consequences of this limitation for real world decision making. (This general theme is discussed in section 5.)

Taken as a whole, the studies reviewed in section 2.3 give reason to be optimistic about people's willingness and ability to employ principles of statistical reasoning. Sensitivity to variations in base rate diagnosticity has been demonstrated repeatedly. The studies reviewed in sections

2.1 and 2.2 suggest that this sensitivity is heightened by certain types of information presentations and internal representations. Sometimes, however, subjects in laboratory studies may become confused by the wording of probability judgment problems and their responses appear nonnormative. To the extent that these responses are the result of *semantic* confusion, we should be skeptical of the cognitive explanations that permeate this literature.

2.3.3. Semantic confusion and the inverse fallacy. People often confuse the hit rate conditional probability P(D|H) with its inverse, the posterior probability P(H|D) (Braine et al. 1990; Chapman & Chapman 1959; Connell 1985; Eddy 1982; Hamm 1993; Kaye & Koehler 1991; Wagenaar 1988; for discussion, see Margolis 1987). This confusion or "inverse fallacy" produces responses that are descriptively consistent with the assignment of no weight to base rates (and any information aside from the hit rate, for that matter). Recent studies, however, suggest that those who provide P(D|H) in tasks that seek P(H|D) may often be more properly classified as victims of semantic confusion rather than as perpetrators of a base rate fallacy.

Wolfe (1992, experiment 2) found that fewer than one in seven subjects had "an appropriate understanding of the hit rate concept" (p. 22), and that 74 of 96 subjects (77%) confused the hit rate P(D|H) with the posterior probability P(H|D) in verbal reports. Confusions of the contrapositive P(D|−H) with P(−H|D) were also common. Macchi (1995, experiment 3) reformulated some classic base rate problems to minimize the chance of semantic confusion between hit rates and posterior probabilities. Under these conditions, subjects made extensive use of available base rates and "base rate neglect" was reduced on word problems from 70%–76% to 10%–27%.

In criminal trials that include presentation of "matches" between a person and traces of genetic material recovered from violent crime scenes (e.g., blood, semen, or hair), forensic scientists may testify about the probability that a person who was not the source of the trace material would nonetheless "match." This is represented by the conditional probability P("Match"|Not Source). When asked to elaborate on the meaning of this probability, forensic scientists (and others) commonly (but mistakenly) equate it with its inverse, P(Not Source|"Match").[8] Koehler (1993c) found that under some conditions commission of this error in a hypothetical murder case dramatically increased mock jurors' willingness to return guilty verdicts.

The inverse fallacy carries with it potentially serious consequences in many areas. Equating conditional probabilities with their inverses effectively denies base rates a role in final judgments. But in some cases, much of the "bias" against base rates may be removed by stating the features of probabilistic judgment problems in ways that are clear to people who may not have a Ph.D. in statistics or in decision theory. Indeed, the fact that information can be formulated in ways that lead people to confuse P(D|H) with P(H|D) says more about the importance of effective communication of unfamiliar probabilistic notions than about the existence of an attentional or cognitive flaw (see Macchi 1994).

2.4. Heuristic theories

Early theories of base rate neglect invoked one or more heuristic explanations (e.g., representativeness, causality, specificity; see Bar-Hillel 1983; Tversky & Kahneman 1982). Although these heuristics provide insight into how people make a broad range of judgments, there is reason to doubt their scope and validity for the base rate phenomena of interest here.

As discussed in section 1.3, representativeness was the first explanation offered for base rate neglect (Kahneman & Tversky 1973). According to this theory, base rates will be ignored in category membership prediction tasks, because statistical base rates are less representative of a category's central features than nonstatistical individuating information. But when subsequent studies reported minimal base rate usage even when controlling for relative representativeness, other explanations were offered. For example, Ajzen (1977) argued that base rates were not heavily weighted, because they are perceived to have little causal relevance to the focal judgment. Thus, increasing the causal relationship between a base rate and the focal case should produce greater base rate use. Although intuitively compelling, empirical support for the causality heuristic has been mixed. Tversky and Kahneman (1980) showed that modifying the base rate reference class in the cab problem from "cabs in the city" to the more causal "cab accidents in the city" yielded greater use of base rates. However, Connell (1985, experiment 2) failed to replicate these data in a series of studies that used both within- and between-subjects designs. Macchi (1995, experiments 1 and 2) varied both the causal strength of the base rates and the wording of the probability question. She reported that base rate usage was affected by the rewording but not by causal strength manipulations. However, Wolfe (1995) reported that subjects were more likely to seek out base rates that had high-causal relevance than those that had low-causal relevance. In the face of such mixed and limited evidence, it is unclear whether and when causal relevance influences base rate usage.

2.4.1. Relevance and specificity. Bar-Hillel (1980) proposed a more general version of the causal relevance thesis. She argued that people order pieces of information by their perceived degree of relevance, allowing high-relevance information to dominate low-relevance information. When the perceived relevance of individuating information is lowered, or when the perceived relevance of base rates is raised, people will attach greater weight to base rates.

Bar-Hillel (1980) argued further that "specificity" is a determinant of perceived base rate relevance. However, she did not test the hypothesis that base rates derived from increasingly specific reference classes will be used more than those derived from less specific classes. To test this thesis, I conducted an experiment in which 234 introductory psychology students at Stanford University were randomly assigned to one of three base rate reference-class conditions in a version of the cab problem: (1) citywide, (2) accident propensity, or (3) highly specific. The wording of the citywide and accident propensity base rate problems was identical to that used in Tversky and Kahneman (1980). Subjects in the highly specific base rate reference-class condition were told that "85% of the cab accidents in the city that have all of the relevant characteristics of the present accident involve Green cabs, and 15% involve Blue cabs." The reference-class specificity hypothesis would receive support from a finding that judgments made by subjects in the highly specific base rate condition showed

the greatest base rate influence (i.e., lower estimates), whereas judgments made by subjects in the citywide base rate condition showed the least base rate influence (i.e., higher estimates). The results did not support this hypothesis. Significant between-group differences were not found ($F(2,231) < 2$; citywide $M = 53.8$ ($n = 84$); accident propensity $M = 58.6$ ($n = 71$); highly specific $M = 52.7$ ($n = 79$)).

Although these data do not support the specificity theory, they do not necessarily contradict the perceived-relevance thesis. Indeed, the broad form of the relevance thesis is consistent with the general framework and empirical literature described throughout section 2. It has been shown that base rate usage varies as a function of task structure, task representation, and various microfeatures of the experimental task. It may be that, in those environments where base rates have their greatest impact on final judgments, they are also perceived to be most relevant for the particular judgment under consideration.

3. The problem of reference-class specificity

The role played by base rate reference-class specificity presents important normative issues as well. How, if at all, should decision makers take reference-class specificity into account in probabilistic judgment tasks?

3.1. Cohen's base rate relevance argument

In a *BBS* target article, philosopher L. J. Cohen argued that base rate data ought to be ignored, except where it is known that the focal case "share[s] all the relevant characteristics" of the reference classes from which the base rate was derived (Cohen 1981a, p. 329). For Cohen, a relevant characteristic is one that is "causally connected" to the case in question (Cohen 1979, p. 397). Thus, age, sex, and income level might be among the characteristics relevant to estimating the chance that a U.S. citizen will make a campaign contribution to the Republican party.

In general, base rates that are derived from specific reference classes (e.g., wealthy middle-aged men) have stronger causal connections to events than base rates derived from more general reference classes (e.g., U.S. citizens). This is because specific reference classes (1) incorporate additional information, some of which may be causally relevant to the focal case, and (2) eliminate extraneous information that is present in the more general class. Certainly, then, a very informative base rate should have a greater impact on one's probability estimate than a less-informative one when both are available.

But Cohen took the argument to an extreme, and in doing so became vulnerable to severe criticism. He argued that base rates that fail to meet an extreme standard of relevance are diagnostically worthless. For example, he claimed that if a person suffers from either disease A or disease B, and disease A happens to be 19 times as common as disease B, the probability that the person has disease A is exactly equal to the probability that the person has disease B "unless told that the probability of A is greater (or less) than that of B among people who share *all [of the] relevant characteristics*, such as age, medical history, blood group and so on" (Cohen 1981a, p. 329, emphasis mine). Similarly, Cohen (1981a; 1981b) argued that the base rate for "cabs in the city" ought to be disregarded in Tversky and Kahneman's (1980) cab problem:

> "Why on earth should it be supposed that subjects asked to estimate the unconditional probability[9] that the cab *involved in the accident* was Blue ought to take into account a prior distribution of colours that would at best be relevant only if the issue at stake was just about the colour of a cab that was said to have been seen somewhere, *not necessarily in an accident*, and was taken to be Blue?" (Cohen 1981b, p. 365).

Cohen concluded that the "85%–15% distribution in cab colours . . . is a very weak foundation for an estimate of the *relevant* base rate" (p. 365).

But what is meant by "the relevant base rate?" By Cohen's standards, the base rate for "cabs in accidents" is more relevant than the base rate for "cabs in the city," because the former takes into account a causal feature of seemingly great significance, namely, the propensity to get into accidents. But it might also be argued that the base rate for "cabs in accidents at night" would be even more relevant. And if one believes that geographical location is important, then the base rate for "cabs in accidents at night in the vicinity of this accident" would be more relevant still. In principle, such base rate refinements may be offered until the reference class reduces to a set of one (the focal case alone), or at least until it becomes so small that it does not allow for a reliable base rate estimate (Lanning 1987).

In short, there is no such thing as *the* relevant base rate. Some base rates may be preferable to others because they are derived from reference classes that are more reliable, causal, or specific. The preference for these more specific base rates derives from their incorporation of information that might otherwise be ignored or unreliably assimilated. However, there is no accepted standard for ascertaining what constitutes a sufficiently relevant base rate, and there certainly are no standards for identifying a single, correct base rate that should be used to the exclusion of all others. Even base rates that are derived from relatively unrefined reference classes contain at least some information that gives them diagnostic value. Ultimately, it is hard to see how any data – base rate or otherwise – could prove useful to decision makers who embrace Cohen's relevance argument (Krantz 1981).

3.2. Specificity and accuracy

Meehl (1954) was also concerned about the tradeoff between base rate specificity and reliability. He advises that when several base rate reference classes are available, each of which is a refinement of another, the "best class is always defined . . . [as] the smallest class . . . for which the N is large enough to generate stable relative frequencies" (p. 22). This is a good rule of thumb for selecting base rates in the special case where data from several, increasingly refined reference classes are available. From a statistical perspective, the more information one has about a problem, the more variance may be explained by these known factors. As the remaining error variability is reduced, the distributions about the best parameter estimate (i.e., the mean) become tighter, and one's confidence that the estimate is at or near the mean increases.

Nevertheless, the decision maker should be warned that the reduction of variability associated with more specific base rates does not guarantee more accurate final judgments – even over a large series of predictions – within any particular set of parameters. Imagine, for example, a ver-

sion of the cab problem in which one is estimating the likelihood that a Green cab caused an accident at night on the south side of town. A series of strikingly different base rates may exist for each of several, increasingly refined reference classes. Although 900 out of 1,000 (90%) cabs in a city may be Green, the Green drivers may be more careful (or luckier) than the rest and account for only 200 out of 1,000 (20%) accidents. However, the Green company may account for almost all of the nighttime taxi driving in this city, hence be responsible for 150 out of the 200 (75%) nighttime accidents. But the Green company may operate primarily on the north side of town and therefore not be responsible for any of the 35 nighttime accidents that occurred on the south side. Consequently, the probability that a Green cab was responsible for the accident under investigation is zero or very close to zero. In this admittedly rigged example, the probability given by the less-refined "cabs in accidents" base rate would be closer to the true probability over a long series of repeated cab color predictions for nighttime accidents on the south side than the probability given by the more-refined "cabs in accidents at night" base rate.

The point is not that decision makers should disregard base rates derived from highly specific reference classes, or that base rates are more likely to confuse and mislead than to inform. The point is rather than base rates derived from increasingly narrowed reference classes do not necessarily converge on a single "relevant" base rate that, if used, will increase predictive accuracy. My intuition suggests that those who use base rates that incorporate more information will generally be better off than those who do not. But this is not an issue that can be resolved through recourse either to intuitive arguments or mathematical principles. Instead, the conditions under which the different base rate strategies are more and less likely to yield better judgments and decisions in real world tasks is an *empirical* matter.

Several points emerge. First, Cohen's advice to ignore insufficiently relevant base rates is unacceptable, because meaningful standards for assessing "relevance" are not available. But even if such standards did exist, Cohen's rejection of all base rates except those that contain "all" of the relevant characteristics of the focal case is too restrictive and would produce calamitous consequences if broadly implemented. Second, increasing the specificity of a reference class reduces error variance, although it does not guarantee a base rate with greater predictive accuracy. Furthermore, reference classes that are *too* specific may be statistically unreliable, because of their small sample space. In these cases, and perhaps in others whose parameters have yet to be identified, a decision maker might be better off using base rates derived from more general reference classes. Indeed, there may be times when decision makers will be better off ignoring base rates altogether (see sect. 5).

4. Difficulties mapping the normative model onto judgment tasks

The issues associated with reference-class specificity indicate that determining how base rates should be used is not as simple as widely assumed. But even where reference-class specificity is not a concern, application of normative rules for base rate use is tricky. Specifically, care must be taken to ensure that the assumptions of the rule are met by the decision maker's construal of the task (Gigerenzer 1991; Hastie 1983; Navon 1978; Rasinski et al. 1985). Where this is not the case, or where critical assumptions remain unchecked, demonstrations of performance errors become suspect.

4.1. Base rates do not equal prior probabilities

A general problem in base rate studies is that subjects may not represent the task or process available information in ways that experimenters assume. For example, in order to compare subjects' judgments to a Bayesian normative standard, experimenters typically assume that all subjects hold prior beliefs (i.e., prior probability estimates) that equal the available base rate. This assumption may not be reasonable either in the laboratory or in the real world. Because they refer to subjective states of belief, prior probabilities may be influenced by base rates and by any other information available to the decision maker prior to the presentation of additional evidence. Thus, prior probabilities may be informed by base rates, but they need not be the same. Indeed, it would be unusual if a decision maker had no information other than a single base rate on which to form a prior probability estimate in realistic tasks. Prior to hearing the daily weather forecast, your estimate that it will snow tomorrow is likely to be based on a great deal of information other than the relative frequency for snow on this day (or similar days) in previous years. Your estimate would take into account such factors as yesterday's temperature, cloud formation, and so on. In such cases it may be quite inappropriate for an available base rate to serve as one's prior belief.

When reasons exist for adopting a prior probability that differs from an available base rate, it is difficult to determine how much the belief should deviate from the base rate.[10] The difficulty arises, in part, because many of the influences on prior beliefs are subtle and difficult to quantify. In addition, the optimal combination of multiple items of information is likely to be computationally complex (see Schum & Martin 1982), and people may not adopt beliefs that are appropriate representations of the available information.

Even when there are no obvious sources of information available other than a base rate, a decision maker's prior probability may differ from the base rate. Bar-Hillel (1990) gives the example of a patient who must choose a surgeon for a dangerous operation. After being told that Surgeon A's patient-mortality base rate is twice as high as that of surgeon B, one may still reasonably believe that the probability of death is greater with surgeon B if one infers that higher mortality rates reflect greater experience with dangerous cases. Here the base rate itself provides the decision maker with information that may lead to the adoption of a different prior probability value. In other cases, the absence of information other than base rates may lead to the adoption of prior probabilities that differ from the base rates. In the courtroom, a party's failure to provide data in support of or in addition to a disputed base rate when such data are available and relevant to a claim, may itself be regarded as evidence against the claim.

Finally, some people may adopt prior probabilities that differ from the base rate even when good reasons for doing so apparently do not exist. Consider, once again, Tversky and Kahneman's (1980) taxi cab problem. According to the

experimenters, Bayes' theorem provides the "correct answer" to this problem, that is, 41%.[11] However, this solution presumes that all subjects adopt the base rates as their prior probabilities. Although this may seem appropriate, whether and when subjects adopt base rates as their priors is an *empirical* question. When 114 Stanford University undergraduates were asked to solve the cab problem after being presented with the base rate information alone, 30% did not answer in accord with the base rates.[12] For subjects whose priors were not equal to the base rate, the Bayesian solution to the cab problem is not 41%.

In short, base rates do not necessarily translate into prior probabilities for obvious reasons, for subtle reasons, and sometimes for no good reason at all. This translation problem is crucial to the normative component of the base rate fallacy contention, because Bayes' theorem combines prior probabilities – not base rates – with likelihoods to obtain normative solutions.

4.2. Task context matters

A related concern is that subjects may extract information from the problems and the experimental context in ways that are not considered by the experimenters or the normative models. Gabrenya and Arkin (1979) serendipitously observed that subjects sometimes regard information to have diagnostic value, that experimenters assume to be worthless. Subjects were presented with a variation of the lawyer-engineer problem that included the following description: "John, a man of 42, . . . is married and has one child. He enjoys his work and likes to work in his yard" (Gabrenya & Arkin 1979, p. 4). Although the investigators believed the description was nondiagnostic, their subjects disagreed and found it to be significantly more diagnostic of an engineer than of a lawyer.

Of greater theoretical interest, unwritten rules of discourse and conversation may focus subjects' attention on one type of information rather than on another. Such context-dependent behavior is normatively defensible if one accepts that conversational rules provide cues to diagnosticity. Schwarz et al. (1991) reported that subjects relied more on individuating information in the lawyer-engineer problem when the instructions defined the task as a psychological rather than a statistical problem. Likewise, Zukier and Pepitone (1984) found that subjects paid less attention to base rates when they were told to approach the task as clinicians rather than as scientists.

Apparently, subjects in these studies used the psychological or clinical context of the tasks as a cue that the forthcoming individuating information would be especially relevant and diagnostic. It is interesting to note that the assignment of extra weight to individuating information in such contexts is consistent with Bayesian tenets: subjects used *all* available information, including information derived from previous experience with similar contexts.

4.3. Information dependencies

Finally, strict application of the Bayesian model for combining prior probabilities and likelihoods requires independence (Fischhoff & Beyth-Marom 1983). If prior odds estimates influence likelihood ratios, then multiplying them to obtain posterior odds estimates will double count the priors.[13]

In a paper that received surprisingly little attention, Birnbaum (1983) pointed out that it is usually unrealistic to assume that likelihood ratios are independent either of base rates or of prior probabilities. Birnbaum reviewed empirical investigations in the signal detection literature, which showed that for human witnesses the ratio of the hit rate to false alarm rate (i.e., the likelihood ratio) depends on signal probabilities (i.e., base rates). That is, the accuracy of likelihood information derived from an observer's reports changes as the observer's knowledge of the base rates changes. Therefore, a witness who is aware that there are many more Green cabs than Blue cabs is probably predisposed to see Green cabs in ambiguous situations. In light of this dependence, the actual probability that a cab in an accident was Blue, given that a witness says so, may be closer to the median and modal responses given by untrained subjects (80%) than to the solution presented by base rate investigators (41%). Indeed, Birnbaum showed that one appropriate response for witnesses who are aware of the 85% base rate for Green cabs and who wish to minimize errors, is 82%.

In short, normative claims about data obtained from laboratory studies cannot be made without understanding how individual subjects represent the tasks and what informational assumptions they make. With this in mind, the Bayesian model serves as a normative benchmark in base rate and related judgment studies only to the extent that experimental tasks map onto it unambiguously.

5. Toward an ecologically valid research program

Even if such tasks are created that avoid the pitfalls described above, significant differences remain between the types of base rate problems used in the laboratory and those that arise in the natural ecology (Einhorn & Hogarth 1981). In laboratory experiments, subjects are typically provided with a single base rate that they are expected to treat as perfectly reliable. Failure to do so is regarded as an error. But base rates received from the natural ecology are not always perfectly reliable, and those who appreciate this may sometimes make better decisions than those who do not. Consider base rates in the medical and psychiatric domains.

As new diseases and cures appear, base rates fluctuate. Where these fluctuations are large, the predictive value of any single historical base rate diminishes. The utility of base rate is also compromised by disagreements about what constitutes the phenomenon of interest. For example, base rates for child abuse or personality disorders (e.g., schizoaffective disorder) will have limited diagnostic and treatment value so long as large disagreements exist within the expert communities about what comprises the condition or disorder (American Psychiatric Association 1987, p. 208; Caldwell et al. 1988). Still other times, a combination of factors compromises the usefulness of base rates. Snow (1985; see also Duncan & Snow 1987) reported that serviceable base rates for brain damage in neuropsychological settings are hard to come by, both because the disorder is hard to diagnose and because the incidence of cerebral dysfunctions has not remained stable over time. Absent empirical study, it is impossible to know whether brain damage base rates have any value at all.

Obviously, laboratory conditions do not correspond perfectly with conditions in the natural ecology. But the differ-

ences that do exist between laboratory and real world base rates, in combination with failures to consider the decision maker's assumptions, goals, and task representations, challenge the prevailing assumption that "errors" observed in the laboratory predict consequential errors in real world decision making (see also Ebbesen & Konecni 1980; Edwards 1990; Funder 1987; 1993). To the extent that important differences exist, training decision makers not to make the kinds of "errors" observed in some experimental tasks may not improve their judgments in real world problems. Indeed, Funder (1994) reports that subjects who are taught to make fewer errors in certain tasks sometimes become less accurate in broader contexts.

If we are to overcome these problems and offer a base rate research program that promises an understanding of real world decision behavior, a methodological shift away from the laboratory is needed. Patterns of base rate usage must be examined in more realistic contexts to determine when, if ever, people make consequential errors. As a first step, speculations about when such errors are most likely to occur are offered below.

5.1. When is it more and less important to use base rates?

When base rates in the natural environment are ambiguous, unreliable, or unstable, simple normative rules for their use do not exist. In such cases, the diagnostic value of base rates may be substantially less than that associated with many laboratory experiments. But does this mean that real world decision makers can disregard base rates, or treat them idiosyncratically and expect to perform as well as those who do not? Does it mean that intuitive processing of real world base rates is as justifiable as any alternative approach? These are empirical questions, and what little evidence there is suggests that the answer is no.

The linear-models literature supports the superiority of simple linear representations over clinical or intuitive judgment across a broad spectrum of probabilistic environments (Dawes et al. 1989). Moreover, several medical studies have indicated that a "Bayesian approach" – in which base rates are treated as prior probabilities, and combined with likelihood ratios according to Bayes' theorem – can lead to more accurate diagnoses than unaided expert judgment (Balla et al. 1985; Duthie & Vincent 1986; Willis 1984). In each of these studies, experts attached too little weight to base rates relative to Bayesian prescriptions.

It would be premature, however, to conclude that this approach always produces superior performance and that decision makers should therefore use it. First, even when the Bayesian rule is well understood, and one's prior probability is identical to a single available base rate, people may have a difficult time estimating likelihood ratios. As noted previously, unless special care is taken when presenting probabilistic information, likelihood ratios are commonly confused with posterior odds ratios. This confusion could create havoc with the Bayesian approach.

Second, the intuitive averaging strategies that people seem to adopt when faced with probabilistic belief-updating tasks (Hogarth & Einhorn 1992; Lopes 1987) lead to responses that are nearly perfectly correlated with Bayesian responses across a range of conditions (McKenzie 1994). The conditions under which decision makers ought to abandon an intuitively appealing and reasonably accurate strategy, such as averaging, in favor of a potentially more accurate Bayesian approach have yet to be identified.

Finally, the natural ecology may present decision makers with situations in which a relative inattention to base rates will not impede judgmental accuracy. One such situation exists when information is presented to decision makers in what Gigerenzer and Hoffrage (in press) call a natural sampling format. When the frequency or probability of H&D and D are known,[14] decision makers may compute P(H|D) without explicitly attending to P(H) base rate information.

This follows from the fact that different formulae can be used to compute the same conditional probabilities. Thus, whereas Bayes' theorem explicitly uses P(H) to compute P(H|D),[15] P(H|D) may also be computed as the ratio P(H&D)/P(D). If P(H&D) and P(D) are available, P(H) may and should be ignored. To illustrate, suppose a stock analyst wants to estimate the chance that a particular stock doubled over a one-year period, given that its earnings were better than expected. The estimate can be obtained by dividing the number of stocks that doubled and those that turned in better than expected earnings by the number of stocks that turned in better than expected earnings. Separate consideration of the base rate probability that a stock doubled is unnecessary. To the extent that such natural sampling situations are common, we should be less concerned about a relative inattention to base rates in laboratory tasks where sample spaces are not easily accessed (see Gavanski & Hui 1992).

A second type of situation in which the natural ecology may forgive those who ignore base rates is the informationally rich environment. Real world decision environments are not typically as impoverished as those described in the pages of psychological stimulus materials. Instead, most real world situations are rich with information. Unlike most base rate problems that appear in the psychological literature, available base rate information is not ordinarily inconsistent with the bulk of other sources of information.

For instance, suppose you know that the summer-evening jazz festivals in the city are usually crowded. Perhaps the base rate for a large crowd is 80%. It is unlikely that individuating information relevant to an estimate of a crowd size will contradict this base rate. Traffic in the immediate vicinity of the festival will probably be heavy, not light. More police officers will be assigned to this area, rather than less. The lines at nearby restaurants will be longer, not shorter. In such cases, where much of the available information is consistent, the failure to incorporate base rates is probably *less* likely to hinder predictive accuracy.

Inattention to base rates is *more* likely to impede accuracy when the base rates conflict with other sources of information and are high in relative diagnosticity. In medical diagnoses, for example, inattention to reliable low base rates could lead to extensive overdiagnosis and excessive treatment. Consider that the general base rate for hypothyroidism is less than 1 in 1,000 among young adult males (De Keyser & Van Herle 1985). But the primary symptoms of this disease – dermatological problems, depression, and fatigue – are quite common. A doctor who disregards the base rate and relies solely on the individuating symptomatology and resultant likelihood ratios, will surely overdiagnose this disease.

Base rates need not be highly reliable or extreme to have

relative diagnostic value. When there is little or no additional information for making a judgment, decision makers should heed even moderately diagnostic base rates. For example, an inexperienced entrepreneur or investor may be well-advised to bear in mind the low base rate of success for new businesses prior to committing resources to an unknown venture.

In short, where information is naturally sampled, or where there is a good deal of redundant information, decision makers probably will not suffer from a relative inattention to, or underweighting of, base rates. But this strategy is riskier when reliable and extreme base rates are at odds with other data, or when little individuating information is available. It is interesting that the laboratory research reviewed in section 2 is generally consistent with this pattern. This suggests that whereas people may not use base rates optimally or even consistently, their strategies often promote accuracy.

5.2. Concerns other than accuracy

There are situations in which accuracy is not the sole criterion for evaluating judgmental quality. For instance, the U.S. legal system is concerned with a variety of fairness and process issues, some of which interfere with judgmental accuracy (Nesson 1985; Tribe 1971). Indeed, certain types of highly diagnostic evidence are routinely excluded at trial (e.g., illegally obtained confessions), because their admission undermines other judicial values. Likewise, some have argued that base rate evidence is inconsistent with the legal norm of individualized justice and should therefore be excluded as well.[16]

Even where base rate evidence clearly does not compromise fundamental values, cost of error considerations may persuade decision makers to make judgments that they believe are inaccurate. In U.S. criminal trials, guilt must be proved "beyond a reasonable doubt." This standard of proof reduces the chance of erroneous convictions relative to lesser standards (e.g., "preponderance of evidence"). A cost of this systemic value is that juries will often return not-guilty verdicts in criminal cases where they *believe* the defendant is guilty, but they are not convinced beyond a reasonable doubt. Similarly, physicians may wish to treat patients who probably do not have a serious disease as if they believed the disease were present, when failure to treat the disease could have serious consequences. Even when cost of error considerations are irrelevant and accuracy is the primary goal, decision makers might also take into account other decision costs, such as the time, mental effort, and money that may be required to improve accuracy (see Hogarth 1987).

From a prescriptive standpoint we must relax our notion of what constitutes an appropriate response in realistic base rate problems (cf. Bar-Hillel 1983). Many problems will have multiple solutions, depending on the value assigned to factors not related to accuracy per se. Indeed, whenever the assumptions, goals, and values of decision makers vary, people exposed to identical information may arrive at different solutions, none of which are necessarily erroneous. An empirical research program that replaces the existing, rigid performance criterion with one that is sensitive to various person- and situation-specific performance criteria will ultimately lead to a better understanding of base rate usage and promote the development of prescriptive models.

6. Summary and conclusions

We have been oversold on the base rate fallacy from an empirical, normative, and methodological standpoint. At the empirical level, a thorough examination of the base rate literature (including the famous lawyer-engineer problem) fails to support the conventional wisdom that people routinely ignore base rates. Quite the contrary, the literature shows that base rates are almost always used, and that their degree of use depends on task representation and structure. Tasks that can be represented in frequentist terms or that are structured in ways that sensitize people to base rates are more likely to be solved through reference to base rates than other types of tasks. Thus, although it is widely accepted that people attach little if any weight to base rates, this seems to be true only in certain tasks and contexts, many of which are quite unlike those that exist in the natural ecology.

This conclusion fits well with a growing body of work showing that people are capable of sound statistical reasoning when information is learned and presented in certain ways (Cosmides & Tooby, in press; Fong et al. 1986; Gigerenzer 1991; Gigerenzer & Hoffrage, in press; Gigerenzer et al. 1988; 1991; Nisbett et al. 1983). It also fits well with observations made from daily life. Base rates are commonly used in many arenas. Baseball managers routinely "play the percentages" by choosing left-handed batters to face right-handed pitchers and vice versa; police officers stop and detain suspected criminals, in part, on the basis of background characteristics; voters mistrust the political promises of even their most favored politicians. In each instance, base rate probabilities are considered and given substantial – if not determinative – weight.

At the normative level, the popular form of the base rate fallacy should be rejected, because few tasks map unambiguously into the narrow framework that is held up as the standard of acceptable decision making. Mechanical applications of Bayes' theorem to identify performance errors are inappropriate when key assumptions of the model (e.g., independence of prior probabilities and likelihoods) or the decision makers' representation of the task (e.g., equivalence of base rates and prior probabilities) are unchecked or grossly violated. Furthermore, the potential ambiguity, unreliability, and instability of base rates in the natural ecology can reduce their diagnosticity. Under these conditions, there is no single, theoretical normative standard for identifying appropriate base rate usage. Indeed, there may be situations (e.g., informationally redundant environments, natural sampling of information) in which base rates can be ignored with impunity, with little or no reduction in predictive accuracy. And even where certain formulae might increase predictive accuracy, prescriptive models for base rate usage should be sensitive to a variety of policy considerations (e.g., cost of error, fairness) where appropriate.

At the methodological level, we should discontinue the practice of searching for performance deviations from an inflexible standard in laboratory tasks. This approach is unlikely to yield additional benefits. Instead, we should pursue a more ecologically valid program of research that examines (1) patterns of base rate use in the natural ecology, and (2) when and how people would benefit from adjusting the intuitive weights they assign to base rates in real world tasks. In domains where research indicates that decision makers would benefit substantially from greater base rate

use, it may be helpful to encourage frequentist problem representations or to sensitize decision makers to base rates in one of the ways outlined in section 2 (e.g., provide for direct base rate experience).

A determination of whether and when people fail to make good use of base rates requires an examination of real world performance. It is one thing to report that, on a particular task, subjects do not give responses that conform with an arguably correct solution, because they attached less weight to one informational cue than to another. It is quite another thing to argue from such evidence that people "generally" reason fallaciously. Such a conclusion is linked not only to the number, breadth, and reliability of studies that support the phenomenon but also to the ecological validity of those studies.

If the stimuli, incentives, performance standards, and other contextual features in base rate studies are far removed from those encountered in the real world, then what are we to conclude? Shall we conclude that people would be richer, more successful, and happier if only they paid more attention to base rates? Shall we teach this in our schools? Shall we advise professional auditors – a population that already pays substantial attention to base rates (Koonce 1993) – that they too would benefit from attending more closely to base rates? And what shall we tell Olympic basketball coaches, jurors, weathermen, stockbrokers, and others who must sort through a morass of ambiguous, unstable, or conflicting base rates to estimate how likely an event did or will happen? These are the types of questions that need to be addressed in contexts that are richer, and with performance standards that are more comprehensive, than those that have been used to date.

ACKNOWLEDGMENTS

Helpful comments and suggestions were made on earlier versions of this article by Jon Baron, Ulf Bockenholt, James Dyer, Brian Gibbs, Roger Gould, Robert Josephs, Florence Keller, Joshua Klayman, Amy McCready, Jeffrey Rachlinski, David Rosenhan, David Schkade, Ron Westrum, Maria Wolverton and many reviewers. Skilled research assistance was provided by Audrey Chia and Sam Lindsey. A shorter version of this article appeared in *Psycoloquy* along with 11 commentaries and 2 rejoinders by the author.

NOTES

1. A "base rate" may be defined as the relative frequency with which an event occurs or an attribute is present in a population (Ginossar & Trope 1987; Hinsz et al. 1988; Lanning 1987). Hence the base rate for six figure annual incomes might be 95% among major league baseball players in the United States, but less than 1% among professional guitar players.

2. Bayes' theorem follows directly from the multiplicative rule of probability, which holds that the joint probability of two events H and E equals the product of the conditional probability of one of the events, given the second event and the probability of the second event. In mathematical notation:

$$P(H \cap E) = P(H \mid E)P(E)$$

$$P(H \cap E) = P(E \mid H)P(H)$$

$$\therefore \quad P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

where $P(E) = P(E \mid H)P(H) + P(E \mid -H)P(-H)$ for binary hypotheses

$$\text{Odds form:} \quad \frac{P(H \mid E)}{P(-H \mid E)} = \frac{\frac{P(E \mid H)P(H)}{P(E)}}{\frac{P(E \mid -H)P(-H)}{P(E)}} = \frac{P(E \mid H)P(H)}{P(E \mid -H)P(-H)}$$

$$= \frac{P(H)}{P(-H)} \times \frac{P(E \mid H)}{P(E \mid -H)}.$$

The letters *H* and *E* may be thought of as standing for Hypothesis and Evidence, respectively. P(H) and P(−H) refer to the probabilities of truth and falsity of a hypothesis H prior to the collection of additional evidence. In Bayesian terminology, P(H) and P(−H) are "prior probabilities," and their ratio is the "prior odds." P(E|H) and P(E|−H) represent the information value of the evidence if the hypothesis is true and false, respectively; their ratio is the "likelihood ratio." P(H|E) and P(−H|E) are the probabilities that the hypothesis is true and false in light of the evidence; their ratio is the "posterior odds," which represents the combination of the prior odds and likelihood ratio.

3. But some philosophers did not accept this framework (see Cohen 1981a; 1981b; Levi 1981; Niiniluoto 1981).

4. The conclusions *others* have drawn from the works of Kahneman and Tversky are often stronger than those offered by the two researchers in their seminal papers.

5. Other lawyer-engineer studies not described in Table 1 for reasons of noncomparability include Borgida and Nisbett (1977), Davidson and Hirtle (1990), Gabrenya and Arkin (1979), and Schwarz et al. (1991).

6. It must be noted that this technique, although better than normative analyses based on summary statistics, is also flawed, because it rests on the unverified assumption that subjects' prior probability estimates equal the base rate provided.

7. Bar-Hillel and Fischhoff (1981) reinterpreted the Manis et al. (1980) results and suggested that the base rates may have been influential only for cases in which the pictures were perceived to be nondiagnostic. In reply, Manis et al. (1981) argued that their data did not support this reinterpretation, although it could not be ruled out with certainty.

8. After testifying that a DNA match was found between blood from a murder victim and blood recovered from a blanket, an FBI scientist in a Florida case was questioned by a prosecuting attorney as follows:

Q: "And in your profession and in the scientific field when you say match what do you mean?"

A: "They are identical."

Q: "So the blood on the blanket can you say that it came from Sayeh Rivazfar [the victim]?"

A: "With great certainty I can say that those two DNA samples match and they are identical. And with population statistics we can derive a probability of it being anyone other than that victim."

Q: "What is that probability in this case?"

A: "In this case that probability is that it is one in 7 million chances that it could be anyone other than the victim." (Wike v. State, 596 So. 2d 1029 [Fla. S. Ct. 1992], transcript, p. 417–8).

For additional examples and discussion of related probabilistic errors and exaggerations at trial, see Koehler (1993b; Koehler et al. 1995).

9. Although Cohen refers to this probability as unconditional, it is a conditional probability because subjects are expected to make their estimates, given the evidence provided by the eyewitness' testimony.

10. In such cases, the construction of prior beliefs is itself a Bayesian problem, with base rate information serving as a kind of "pre-prior" probability, whereas all other information feeds into the likelihood ratio.

11. If B and G denote the hypotheses that the cab in the accident was Green and Blue, respectively, and "b" denotes the witness' reports seeing a Blue cab, then

$$\frac{P(B \mid \text{"}b\text{"})}{P(G \mid \text{"}b\text{"})} = \frac{P(B)}{P(G)} \times \frac{P(\text{"}b\text{"} \mid B)}{P(\text{"}b\text{"} \mid G)} = \frac{.15}{.85} \times \frac{.80}{.20} = \frac{12}{17}.$$

$$\text{Hence, } P(B \mid \text{"}b\text{"}) = \frac{12}{12 + 17} = .41$$

12. Similarly high proportions were reported by others with different problems. See Beyth-Marom and Fischhoff (1983, experiment 5), (36%), Ginossar and Trope (1980) (28%), and Schneider and Fishback (1994) (36%).

13. Dependencies between prior odds and likelihood ratios are not necessarily non-normative (Koehler 1993d).

14. Strictly speaking, P(D) itself need not be made explicit if both H&D and −H&D are known because P(D) = P(H&D) + P(−H&D).

15. For binary hypotheses, $P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D \mid H)P(H) + P(D \mid -H)P(-H)}$

16. Debate concerning the role base rates and other types of probabilistic evidence that should play at trial has raged for years (see Brilmayer & Kornhauser 1978; Cohen 1981c; Finkelstein & Fairley 1970; Kaye 1979; 1981; 1989; Koehler 1991; 1993a; Koehler et al. 1995; Koehler & Shaviro 1990; Nesson 1985; Saks & Kidd 1980–81; Shaviro 1989; Tribe 1971; and Wells & Windschitl 1994. Related papers can also be found in "Debate" 1991; "Decision" 1991; and "Probability" 1986).

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Cognitive algebra versus representativeness heuristic

Norman H. Anderson

Department of Psychology, University of California, La Jolla, CA 92093-0109. **nanderson@ucsd.edu**

Abstract: Cognitive algebra strongly disproved the representativeness heuristic almost before it was published; and therewith it also disproved the base rate fallacy. Cognitive algebra provides a theoretical foundation for judgment-decision theory through its joint solution to the two fundamental problems – true measurement of subjective values, and cognitive rules for integration of multiple determinants.

Koehler speaks rightly: "We have been oversold on the base rate fallacy." Indeed, the representativeness heuristic had hardly been published before it failed a critical test in the classic Bayesian two-urn task (Leon & Anderson 1974).

In this Bayesian task, subjects judged the probability that a sample of red and white beads had been drawn from one of two urns, each with known proportions of red and white beads. The representativeness heuristic asserts that urn proportions, which determine the base rate, have essentially no effect. But Figure 1 shows large effects of urn proportions – disproving the representative heuristic and therewith showing a fallacy in the base rate fallacy. Supportive results from social cognition are cited in Anderson (1986).

The Leon-Anderson (1974) experiment seems virtually forgotten, even though it has been well corroborated. An objection was raised to its use of within-subject design (see Slovic et al. 1977), but these same writers subsequently used a similar design and reached a softer but similar conclusion (Fischhoff et al. 1979). In doing so, however, they neglected to cite the earlier experiment.

Koehler thus seems right in reaffirming Christensen-Szalanski and Beach's (1984) suggestion that the representativeness heuristic has survived through selective citation of the literature. A similar conclusion holds for the availability and the anchoring and adjustment heuristics (Anderson 1986).

The Leon-Anderson (1974) experiment was constructive; the theoretical curves in Figure 1 fit the data well. This information integration model has been extensively supported (Anderson 1991; in press). Information integration theory not only corrects the error of the representativeness heuristic but goes beyond to provide a tested cognitive theory.

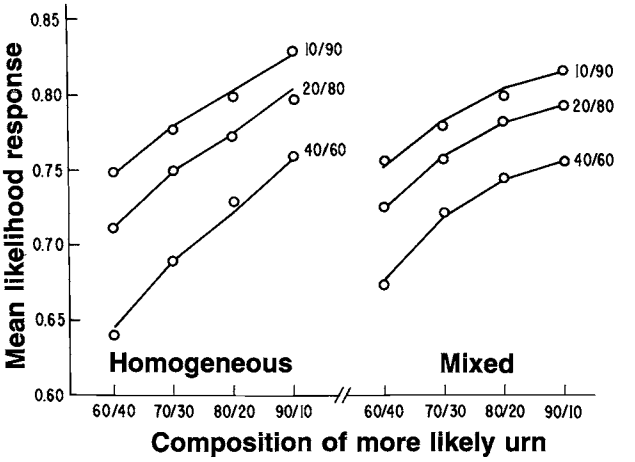

Figure 1 (Anderson). Large base rate effects disprove representativeness heuristic. Subject sees sample of beads drawn at random from one of two urns, each with specified numbers of red and white beads (listed on the horizontal axis and as a curve parameter, respectively.) Data points are mean judged probabilities that the sample comes from the more likely urn. Left panel averaged over four homogeneous samples (1, 2, 3, and 4 red beads); right panel averaged over four mixed samples (3:2, 3:1, 4:1, and 6:2 – red:white ratios). Base rate is determined by urn proportions. Both urn proportions have large effects, shown by upward slope of and vertical separation among the curves, contrary to representativeness heuristic. Base rate thus has visibly large effects. Curves are theoretical, from decision-averaging model of cognitive algebra. Theory-data fit appears good. (From: M. Leon & N. H. Anderson, "A ratio rule from integration theory applied to inference judgments," *Journal of Experimental Psychology*, 1974, 102:27–36. Copyright 1974 by American Psychological Association. Reprinted by permission.)

The judgment-decision field faces two fundamental problems: multiple determination and subjective value. *Multiple determination* is fundamental, because virtually all judgments and decisions are integrated resultants of multiple determinants. *Subjective value* is fundamental, because that is what is integrated. A's judgments and decisions cannot be understood in terms of B's values; A's integration rules can only be diagnosed in terms of A's personal values.

Substantial progress has been made on both problems with cognitive algebra and functional measurement. Algebraic integration rules for multiple determinants have been found by Lola Lopes, Dominic Massaro, Gregg Oden, Anne Schlottmann, James Shanteau, Ramadhar Singh, Friedrich Wilkening, and Yuval Wolf, among others cited in detail in Anderson (1986; 1991; in press) virtually in every field of psychology, from social development to psychophysics and language. This cognitive algebra has been possible because functional measurement allows for personal value.

Integration theory involves a different way of thinking. The traditional way of thinking appears in Koehler's attempt to bring order into the base rate issue by classifying empirical determinants, such as task structure and experimental procedure. This approach is useful, and Koehler does a fine job, although perhaps not accounting for the Leon-Anderson (1974) experiment, in which one condition did show no effect of base rate. More seriously, this traditional approach is inherently inadequate to handle the two fundamental problems: (1) measurement of subjective values, and (2) integration of multiple determinants. Only by facing these two problems can a general theory of judgment-decision be attained.