Journal of Personality and Social Psychology
1983, Vol. 45, No. 4, 792–804

Bayesian Inference: Combining Base Rates With Opinions of Sources Who Vary in Credibility

Michael H. Birnbaum
University of Illinois at Urbana-Champaign

Barbara A. Mellers
University of California, Berkeley

Subjects made judgments of the probability of an event given base-rate information and the opinion of a source. Base rate and the source's hit and false-alarm rates were manipulated in a within-subjects design. Hit rate and false-alarm rate were manipulated to produce sources of varied expertise and bias. The base rate, the source's opinion, and the source's expertise and bias all had large systematic effects. Although there was no evidence of a "base-rate fallacy," neither Bayes' theorem nor a subjective Bayesian model that allows for "conservatism" due to misperception or response bias could account for the data. Responses were consistent with a scale-adjustment averaging model developed by Birnbaum & Stegner (1979). In this model, the source's report corresponds to a scale value that is adjusted according to the source's bias. This adjusted value is weighted as a function of the source's expertise and averaged with the subjective value of the base rate. These results are consistent with a coherent body of experiments in which the same model could account for a variety of tasks involving the combination of information from different sources.

The question, How *should* humans revise their beliefs? has been studied by philosophers and mathematicians, and the question, How *do* humans form opinions and revise them? has been investigated by psychologists. Early research that compared the two questions concluded that Bayes' theorem was a useful starting point for the description of human inference but that humans are "conservative," or revise their probability judgments in a manner less extreme than implied by Bayes' theorem (Edwards, 1968; Peterson & Beach, 1967; Slovic & Lichtenstein, 1971).

Edwards (1968) discussed three interpretations of conservatism: misperception, misaggregation, and response bias. Misperception includes the possibility that objective probabilities are transformed to subjective probabilities by a psychophysical function. Misaggregation refers to use of a non-Bayesian rule to combine evidence. Response bias refers to nonlinearity in the judgment function relating judged probabilities to subjective likelihoods. Early experimental work attempted to separate these interpretations, but the experiments could not yield definitive conclusions (Edwards, 1968).

Wallsten (1972) proposed a subjective version of Bayes' theorem to account for conservatism. In his formulation, subjective probabilities replaced objective probabilities, the response scale was assumed to be only ordinal (allowing for nonlinear response bias), but the aggregation rule was assumed to be Bayesian. However, Shanteau (1975) found that judges in the bookbags and poker chips paradigm revise their opinions even when they are given evidence that is nondiagnostic. Shanteau concluded that his judges were averaging prior probability with a subjective probability representing the implication of each new sample of evidence (see also Troutman & Shanteau, 1977).

Kahneman and Tversky (1973) argued that judges neglect diagnostic information such as base-rate information and the validity of sources of information. Recently there has been a debate over the claim of a "base-rate fallacy," the contention that subjects neglect base-rate information crucial to a Bayesian analysis (Ajzen, 1977; Bar-Hillel, 1980; Birnbaum, 1983; Carroll & Siegler, 1977; Fischhoff, Slovic, & Lichtenstein, 1979; Lyon & Slovic, 1976; Manis, Dovalina, Avis, & Cardoze, 1980;

We thank Kevin Lanning, James Shanteau, and Amos Tversky for comments on an earlier draft.

Requests for reprints should be sent to Michael H. Birnbaum, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

Tversky & Kahneman, 1980, 1982). This work led many to the conclusion that human inference is best described in terms of a list of biases and heuristics, regarded as illusions or discrepancies between what people should do and what they actually do.

Unfortunately, much of the recent research has used the objective form of Bayes' theorem as the null hypothesis, without allowing for misperception or response bias. In fact, even the objective form of Bayes' theorem has not been provided in its full complexity (Pitz, 1975). As Schum (1981) has shown, the algebra can accommodate many subtle patterns of conditioning of evidence. Birnbaum (1983) noted that research on the "base-rate fallacy" used an incomplete Bayesian analysis. Birnbaum showed that behavior described as "neglect of base rate" may be consistent with rational Bayesian utilization of the base rate. Thus, it is not at all clear that Bayes' theorem deserves the bad press it has received in recent years as a framework for the study of human inference.

The purpose of this article is to test two models of how judges use base-rate information. A subjective version of Bayes' theorem and scale-adjustment averaging model (Birnbaum & Stegner, 1979) are examined as descriptions of probability judgments.

Inference Task

Let $P(L)$ be the probability of the event L that a used car will remain in working condition, or last, for 3 years. The base rate, or proportion of cars that last, is assumed to vary for different types of cars. To denote varying base rates, r, we write $P_r(L)$ as the prior probability that the car will last, given base rate r.

A source examines the car and reports whether the car seems to be in "good" shape ("G") or "bad" shape ("B"), mutually exclusive and exhaustive categories. The following characteristics of the source are known:

- $P(\text{"G"}|L)$ = probability that the source says the car is in good condition, given that the car lasts (L). In signal-detection theory, this would be termed the hit rate, $P(\text{HIT})$.
- $P(\text{"G"}|\bar{L})$ = probability that the source describes the car as good, given it actually fails to last ($\bar{L}$). $P(\text{"G"}|\bar{L})$ is termed that false-alarm rate, $P(\text{FA})$.

Suppose the source's hit and false-alarm rates are independent of the base rate. That is, suppose $P_r(\text{"G"}|L) = P(\text{"G"}|L)$ for all r; similarly, $P_r(\text{"G"}|\bar{L}) = P(\text{"G"}|\bar{L})$, for all r (see Footnote 1).

The judges in this experiment were asked to infer the probability that the car will last, given the base rate and the source's report. Their judgments were compared with Bayes' theorem and with a subjective version of Bayes' theorem that has the same algebraic structure but allows for conservatism due to misperception and response bias.

Bayesian Model

Because the source's hit and false-alarm rates are assumed to be independent of base rate, Bayes' theorem can be written:

$$P_r(L|\text{"G"}) = \frac{P_r(L) \cdot P(\text{"G"}|L)}{P_r(L) \cdot P(\text{"G"}|L) + [1 - P_r(L)] \cdot P(\text{"G"}|\bar{L})}, \quad (1)$$

where $P_r(L|\text{"G"})$ is the probability that the car lasts given the source says the car is in good shape, and the base rate for this type of car is r. The prior odds that the car lasts given only the base-rate information follows:

$$\Omega_r = \frac{P_r(L)}{1 - P_r(L)}. \quad (2)$$

The posterior odds, $\Omega = P_r(L|\text{"G"})/[1 - P_r(L|\text{"G"})]$, given the source's opinion can be expressed as follows:

$$\Omega = \Omega_r \frac{P(\text{HIT})}{P(\text{FA})}. \quad (3)$$

If the source had said the car was bad ("B"), the hit/false-alarm ratio would be replaced by the miss/correct-rejection ratio $[P(\text{"B"}|L)/P(\text{"B"}|\bar{L})]$.

Equation 3 implies that the effects of base rate should multiply the source information. Taking logarithms of both sides yields

$$\log \Omega = \log(\Omega_r) + \log[P(\text{HIT})] - \log[P(\text{FA})]. \quad (4)$$

Equation 4 shows that if the Bayesian model is descriptive of human inference, then it should be possible to monotonically transform the subjects' probability judgments to an additive decomposition of base rate, source's hit rate, and false-alarm rate for each level of the source's opinion. This implication follows because log odds (log Ω) is a monotonic function of posterior probability [$P_r(L|$"G"$)$].

If the Bayesian model fails due to misperception (i.e., subjects combine information consistent with Bayes' theorem but are "conservative" in their probability estimates), one could replace the objective probabilities in Equations 1 and 2 with subjective scale values. According to this hypothesis, the model still might provide an accurate representation of the process by which the subject combines or aggregates the information. Similarly, if subjects differ from Bayes' theorem due to a response nonlinearity (e.g., a reluctance to use extreme values on the response scale), then it should still be possible to find a monotonic function of the judge's responses, such that the additive alegbraic structure is satisfied. However, if the data cannot be transformed to fit the subjective Bayesian model, then one can reject the Bayesian models as theories of information aggregation.

Scale-Adjustment Averaging Model

Birnbaum and Stegner (1979) developed a model for the process by which judges combine information from sources of varying expertise and bias. With the assumption that expertise and bias depend on hit rate and false-alarm rate, the scale-adjustment averaging model can be adapted to the present task, as shown in the following equation:

$$P_{rxbo} = \frac{ws + w^{*}s_r + w_{xb}s_{bo}}{w + w^{*} + w_{xb}}, \qquad (5)$$

where P_{rxbo} is the judged probability the car lasts, given that a source of expertise (x) and bias (b) gave an opinion (o), and the base rate was r. The weights w, w^{*}, and w_{xb} are the weights of the initial impression, the base-rate information, and the source's opinion given by a source with expertise x and bias b, respectively. The scale value of the source's report, s_{bo}, is assumed to depend on the opinion ("G" or "B") and the source's bias. The other scale values, s and s_r, are for the initial impression and the base rate, respectively.

When a piece of information (either base rate or source's opinion) is not presented, its weight is set to zero. Thus, the scale-adjustment averaging model implies that the effect of a piece of information (e.g., base rate or source's report) is directly proportional to its weight and inversely related to the number and weight of other pieces of information. Further, the model implies that the effect of a source's bias is amplified by the source's expertise. On the other hand, Bayesian models imply that the effect of the base rate (Equation 4) should be independent of the characteristics of the source, P(HIT) and P(FA). Similarly, the effect of a source's opinion should be the same regardless of the value (or even absence) of base-rate information.

Method

To assure that the used car problem is comparable to those used in previous research on the base-rate fallacy, a preliminary study used the same procedure as in those studies; that is, each subject received only one problem. The main portion of the experiment used a within-subject design, in which each subject received many problems.

Single-Judgment Task

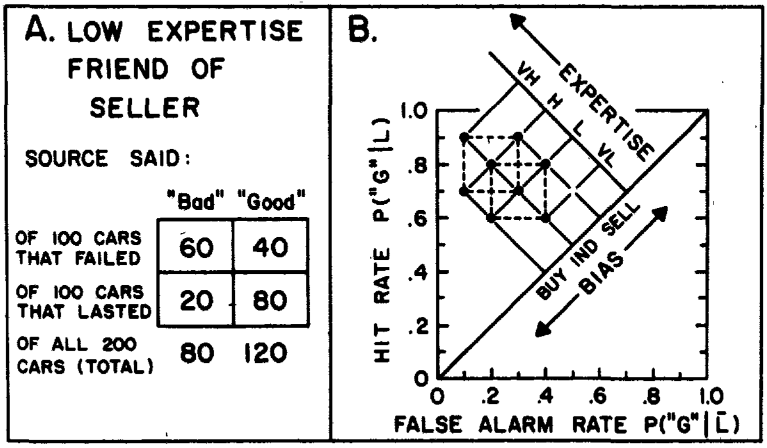
Sixty-five undergraduates received a single problem similar to that used in previous research on the base-rate fallacy. Instructions were modified from the cab problem (see Tversky & Kahneman, 1980) and read as follows:

> In this experiment, you will be asked to judge the probability that a used car will remain in working condition for 3 years. For cars of the same year, make, model, and mileage as the car you will be judging, 30% are expected to last for 3 years.
>
> A source, who examined the car, gave his opinion. He had previously been tested in his ability to distinguish between cars that actually lasted for 3 years and those that failed. Of 100 cars that actually lasted, he correctly identified 80 as being in "good shape." Of 100 cars that failed to last, he correctly judged 80 as being in "bad shape." The source stated that the car is in "good shape." What is the probability expressed as a percentage that the car will remain in working condition for 3 years?

Multiple-Judgment Task

Subjects were asked to make many judgments of the probability that cars would last given base-rate information and/or the reports of sources who examined the cars. Trials varied in base rate, sources' opinion, and source characteristics (hit and false-alarm rates).

Instructions. Base rates were represented by blue book values. Judges were told to assume that the probability a used car would last was directly proportional to its blue book value. They were told that of 100 cars with blue book values of $300, 30 are expected to last 3 years and the remaining 70 are expected to fail; with blue book values of $400, 40 of 100 are expected to last, and so on.



A. LOW EXPERTISE FRIEND OF SELLER

SOURCE SAID:

	"Bad"	"Good"
OF 100 CARS THAT FAILED	60	40
OF 100 CARS THAT LASTED	20	80
OF ALL 200 CARS (TOTAL)	80	120

Figure 1. Source characteristics. (Panel A shows the format used to present conditional probabilities in within-subjects design. Panel B shows hit rate and false-alarm rates of the eight sources used. Expertise is hit rate minus false-alarm rate. Bias is hit rate plus false-alarm rate. VL = very low; L = low; H = high; and VH = very high.)

They were told that the blue book value represented a class of cars rather than any individual car. Judges were told to assume that for the entire population of cars they would be judging, there was an even chance that a car would last 3 years.

Specific information for each car was represented by the opinion of a source of known expertise and bias who examined the car. Instructions stated that each source was given a test of ability to distinguish between cars that actually lasted or failed. The test results were said to be independent of base rate (blue book value); that is, the conditional probabilities that the source reports the car is in good shape, given it lasted or failed, are the same for all levels of blue book value.[1]

Each source was described in terms of hit rate, $P(\text{"G"}|L)$, and false-alarm rate, $P(\text{"G"}|\bar{L})$, and also described in terms of expertise and bias. There were four levels of expertise (very low, low, high, very high) and three levels of bias (friend of the buyer, independent, or friend of the seller). Instructions similar to those of Birnbaum and Stegner (1979) discussed expertise and bias.

For example, Figure 1 (panel A) shows how the test results for a friend of the seller with low expertise were presented. Note that of 100 cars that lasted, this source correctly identified 80 cars as being in good shape; that is, $P(\text{"G"}|L) = .8$. Of 100 cars that failed, the source wrongly identified 40 to be in good shape; that is, $P(\text{"G"}|\bar{L}) = .4$.

The source's bias is reflected in the column totals (hit rate plus false-alarm rate). In the example, the friend of the seller said 80 cars were in bad shape and 120 cars were in good shape. The source's expertise is reflected in the column differences (hit rate minus false-alarm rate). When the low-expertise source said "good shape," there were 40 more correct than incorrect judgments (80–40).

Each point in Figure 1 (panel B) represents one of the eight sources in terms of false-alarm and hit rates. For example, the low-expertise friend of the seller is illustrated in panel B as a point with coordinates (.4, .8) on the false-alarm and hit rate axes, respectively. The solid diagonal lines in Figure 1 show that each source can be represented in terms of bias (hit + false alarm) and expertise (hit − false alarm). The low-expertise friend of the seller is at the intersection of "*sell*" on the bias curve and "*L*" on the expertise curve. The two squares of dashed lines show that the eight sources form two 2 × 2 factorial designs of Hit Rate × False-Alarm Rate. The solid lines show that six of the sources form a 2 × 3 (Expertise × Bias) factorial design, in which expertise is either high or low and bias is either seller's friend, independent, or buyer's friend. The other two sources are independent, providing four levels of expertise for unbiased sources.[2] Judges were instructed

[1] Birnbaum (1983) noted that previous demonstrations of the base-rate fallacy did not give information crucial to a proper Bayesian analysis. Unless the ratio of the source's hit rate to false-alarm rate is known to be independent of the base rate, one cannot simply multiply prior odds by a fixed likelihood ratio. Given reasonable assumptions concerning the signal-detection behavior of the source, it can be shown that Bayes' theorem implies a solution that has previously been described as "neglect of base rate."

[2] The concepts of expertise and bias, as defined by Birnbaum and Stegner (1979) are analogous to the concepts of discriminability and bias used in signal-detection theory. When it assumed that the stimuli produce normal distributions of sensation with equal variance, the index of discriminability, d', is defined as $z[P(\text{HIT})] - z[P(\text{FA})]$, where $P(\text{HIT})$ and $P(\text{FA})$ are the hit rate and false-alarm rates, respectively, and z is the inverse standard normal distribution function. The labels in Figure 1 (panel B) treat expertise as hit rate minus false-alarm rate, which would be d' if the distributions were rectangular with equal variance. For the purposes of fitting the model, a separate weight was estimated for each source.

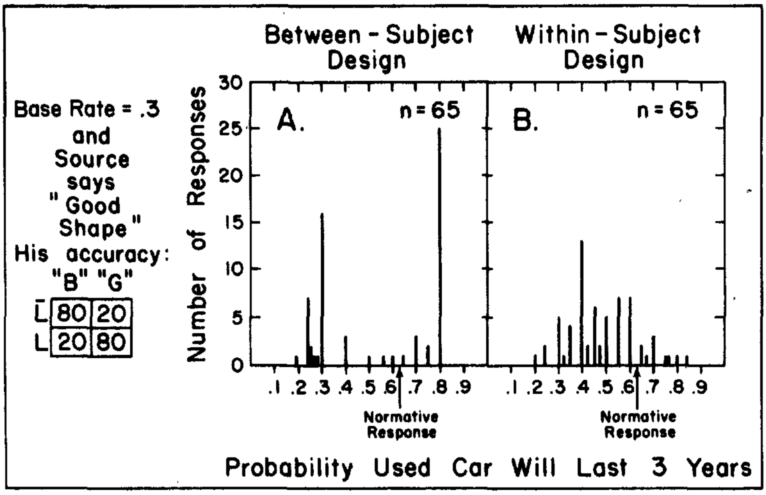


Figure 2. Comparison of between- and within-subjects response distributions for one problem. ("B" = bad shape; "G" = good shape; $\bar{L}$ = car fails to last; L = car lasts.)

to use a page with matrices as in Figure 1 (panel A) for each of the eight sources as an aid in making their judgments.

Design and procedure. Eighty test trials were constructed from an 8 × 2 × 5 (Source × Source's Opinion × Base Rate) factorial design, in which the eight sources of Figure 1 (panel B) gave opinions of either good shape or bad shape, and the five levels of base rate were .1, .3, .5, .7, or .9. In addition, the opinions of sources were also presented alone (without explicit base-rate information) to produce 16 trials—from an 8 × 2 (Source × Source's Opinion) factorial design. On five additional trials the subjects made their judgments given base rate alone.

Each test booklet contained 11 pages of instructions, 12 representative warm-up trials, and 101 randomly ordered test trials. Page orders were shuffled to provide different trial orders for different subjects.

Subjects. The judges were 65 undergraduates at the University of Illinois, who received extra credit in introductory psychology and were tested in 2-hour sessions.

Results

Single-Judgment Task

Figure 2 compares frequency histograms of responses to a single problem presented alone (on the left) and to the same problem embedded among many other problems in the within-subject design (on the right). The base rate in this case was .3, the source's characteristics are given by $P(\text{“G”}|L) = .8$ and $P(\text{“G”}|\bar{L}) = .2$; the source said "good shape." Bayes' theorem (Equation 1) applied to the objective probabilities implies $P(L|\text{“G”}) = .63$.

The modal response for subjects who judged a single problem is .80, which is the hit rate of the source. The second most frequent response is .30, the base rate, and the third most frequent response is .24, the product of .30 and .80. Fewer than 10% of the subjects responded between .4 and .6. The results are similar to those of Kahneman and Tversky (1973), Bar-Hillel (1980), and others. Therefore, this inference problem appears similar to those that led to the notion of a base-rate fallacy.

In sharp contrast, subjects who received the same problem embedded among many others give a very different distribution of responses. Most of the responses fall between .4 and .6, and only one person responded .8. The mode is .4, which is actually closer to the base rate than to either the normative response or the source's hit rate. Thus, in the one-judgment task, subjects appear to respond with one of the values given, whereas when given many problems, they appear to integrate the information. Fischhoff et al. (1979) also concluded that the response distribution differs for between- and within-subject designs.[3]

[3] There are many possible interpretations of the different results for the within- versus between-subjects experiments in Figure 2. Although the same subject population and general procedures were used, the amount of practice, the variation of variables, and the instructions were different for the two conditions (see also Fischhoff et al., 1979).

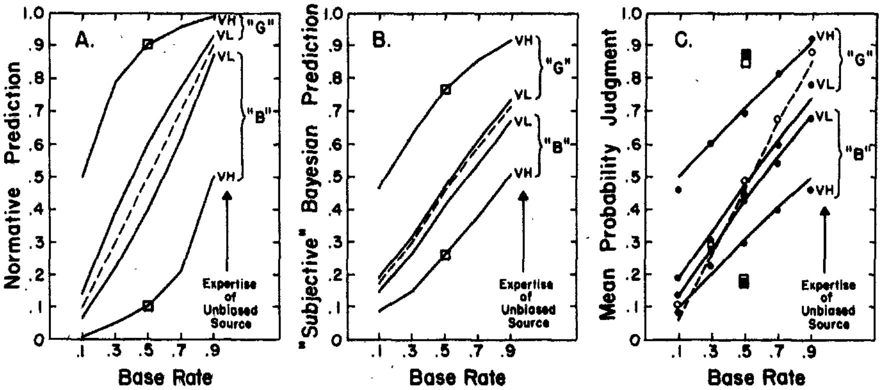


Figure 3. Predictions of three theories as a function of base rate. (Separate curves represent predictions for very high [VH] or very low [VL] expertise independents who gave reports of "good" ["G"] or "bad" ["B"], as in Figure 1. Dashed curves in all panels are for base-rate information alone. Panel A shows predictions of Bayes' theorem applied to the numerical values. Panel B shows predictions of best-fit subjective Bayesian model. Curves in Panel C show predictions of scale-adjustment averaging model. Open squares show prediction when VH expertise independent said either "good" [top square] or "bad" and no base-rate information was given. Solid squares, solid circles, and open circles in Panel C show mean judgments.)

Within-Subject Results

Figure 3 shows predictions of three models and data (panel C) for a portion of the within-subject design (independent sources of very high and very low expertise). Predictions of Bayes' theorem with objective probabilities are shown in panel A; predictions of a least squares fit of a subjective Bayesian model (with subjective scale values instead of objective probabilities) are shown in panel B. The lines in panel C show best-fit predictions of the scale-adjustment averaging model (Equation 5). The model fitting (predictions in panels B and C) is discussed in later sections.[4]

In each panel of Figure 3, predictions are plotted against the base rate with a separate curve for independent sources with very low or very high expertise who gave opinions of good or bad shape. Solid lines show the predicted effects when both the base rate and the source's report were provided. The dashed line shows the predicted effect of base rate alone. The open squares show the predictions for a very high expertise independent source, $P(\text{"G"}|L) = .9$, $P(\text{"G"}|\bar{L}) = .1$, who said either "good" or "bad," in the absence of specific information concerning the base rate (although the overall base rate was specified as .5).

The solid circles in panel C of Figure 3 show mean judgments based on source's report and the base rate. The open circles show mean judgments given only base rate; the solid squares show means for two judgments based only on the source's opinion (the very high expertise source said either "good" or "bad").

The slope of the solid points in Figure 3 (panel C) shows the effect of the base rate. If there were no effect of base rate, the curves would be horizontal. The vertical spread between the curves shows the effect of the source's report. Note that the vertical spread between the two curves labeled VH (very high expertise source) is much greater than the vertical spread between the two curves labeled VL (very low expertise source). The data in

[4] A computer program was written to estimate weights and scale values, to calculate predictions, and to compute fit. The sum of squared deviations between predicted and obtained judgments can be expressed as a function of the parameter estimates as follows:

$$F(w, w^*, w_{xb}, s, s_r, s_{bo}) = \sum (P_i - O_i)^2,$$

where F is a function of the estimated parameters, P_i is the predicted mean response, O_i is the observed mean judgment, and the summation is over all 101 cells in the experimental design. The best-fit parameters are those that minimize F. The value of F at the minimum is an index of the badness of fit of the model. Predictions for each observed datum can be calculated by substituting best-fit parameter estimates into the model (e.g., Equation 5). The computer program reads in the O_i, defines P_i, defines F, and initializes the parameter estimates; this program drives the subroutine STEPIT (Chandler, 1969), which varies the parameters in order to minimize F.

Figure 3 (panel C) show that judges are sensitive to the base rate, the source's opinion, and the expertise of the source. All of the main effects and interactions in panel C are statistically significant by analysis of variance. Of the 65 subjects, 64 gave higher probability inferences for larger base rates. Averaged over sources and opinions, the effect of base rate is also large and statistically significant, $F(4, 256) = 625.28$. (All main effects and interactions in Figures 4 and 5 are statistically significant, $p < .05$. Individual data were examined, and it was found that the means in Figures 3, 4, and 5 are representative of individual data.)

Fit of Subjective Bayesian Model

The subjective Bayesian model (Equation 1) was fit to the 101 mean judgments by means of a specially written computer program that utilized Chandler's (1969) STEPIT subroutine to minimize the sum of squared data-model discrepancies (see Footnote 4). The program estimated subjective probabilities for the five levels of base rate. For each of the eight sources, subjective hit rates and false-alarm rates were also estimated. It was assumed that probabilities of complementary events sum to one. An overall prior was also estimated, yielding 22 parameters to be fit to 101 judgments.

The subjective Bayesian model (panel B of Figure 3) gives a better approximation to the data than the Bayesian model with objective probabilities (panel A). Although it does a reasonably good job of describing the effect of base rate and source's opinions (compare panel B of Figure 3 with points in panel C), it makes some systematic errors.

Both versions of the Bayesian model imply that it should be possible to rescale monotonically the solid and dashed lines in panels A and B (Figure 3) to parallelism (Equation 4). However, the judgments in panel C show that the effect of base-rate information is smaller when the source's report is provided. The slopes of the curves connecting the solid points are less steep than the slope of dashed line connecting the open points. This crossover interaction is an ordinal violation of the Bayesian models and implies that deviations from the Bayesian models cannot be explained by misperception or response bias.

In addition, notice that if the base rate is .9 (without a source's opinion), the mean probability judgment is .87. However, when the very low expertise source gives a "good" report, the judged probability of the car lasting *drops* to .74. According to the Bayesian model, inclusion of that source's report should *raise* the probability judgment.

The Bayesian model also implies that the effect of the source's report alone should be the same as the effect of the source information when accompanied by base-rate information. However, the effect of source information is smaller when presented with base rate than without. To illustrate, note that if a very high expertise source says the car is in good shape, the mean judgment is about .9 (upper solid square in Figure 3, panel C). However, when the judges are *also* told that the base rate for this type of car is .5, the mean judgment is reduced to .7. According to the Bayesian model, these two probability judgments should be identical. (Note that open squares fall directly on the solid lines in panels A and B.) These aspects of the data can be explained by the scale-adjustment averaging model (open squares in panel C).

Fit of the Scale-Adjustment Averaging Model

The scale-adjustment averaging model (Equation 5) was also fit to the data by means of a program that iteratively minimized the sum of squared data-model discrepancies (see Footnote 4). This model also requires estimation of 22 parameters from the data. The sum of squared discrepancies was less than half that for the Bayesian model. The root-mean squared deviation was .022; the largest absolute discrepancy was .06.

The 22 parameters of Equation 5 are as follows: five values of s_r (base-rate scale values), eight values of w_{xb} (weights of the eight sources), six values of s_{bo} (scale values of the two opinions by three levels of bias), two values of w^* (weight of the base rate alone or with source's report), and the prior scale value, s. The weight of the initial impression (prior) was set to 1.00 with no loss of generality. Weight of information when it is not presented was set to 0.

The estimated weights and scale values for the scale-adjustment averaging model are given

Table 1
Estimated Parameters for Scale-Adjustment Averaging Model

	Source's bias		
Source	Seller	Independent	Buyer
	Weights of sources		
Expertise			
Very low	—	.082	—
Low	.111	.153	.129
High	.361	.582	.408
Very high	—	.935	—
	Scale values of opinions		
Opinion			
Good shape	1.309	1.359	1.577
Bad shape	−.166	−.003	.235

Note. Weight of initial impression (w) = 1.00; scale value of initial impression (s) = .368; weight of base-rate information (w^*) = 1.195; weight of base-rate information alone (w^* [alone]) = 2.903; scale values of base-rates (s_r) are −.049, .218, .504, .755, and 1.013, for objective base rates of .1, .3, .5, .7, and .9, respectively (see Footnotes 4 and 5).

in Table 1. Note that estimated weights of sources depend mostly on the source's expertise, but biased sources have slightly less weight than unbiased sources with the same value of hit rate minus false-alarm rate. The fit of the model was improved by allowing the weight of the base rate to increase (from 1.19 to 2.90) when only base-rate information was presented.

The scale values for the sources' opinions show higher values for good reports than bad reports, and scale values are higher for the buyer (who is presumably biased to say cars are in bad shape) than for the seller (who has the opposite bias). Likewise, the scale value of bad shape for the seller (who tends to say cars are in good shape), is lower than the scale values for the other two sources. The patterns of weights and scale values are similar to those of Birnbaum and Stegner (1979).[5]

Figures 4 and 5 show six two-way interactions among the four variables with data and predictions of the scale-adjustment averaging model in the same panels to allow examination of the model's success at describing other aspects of the data. Figure 4 shows three effects of source's expertise. The points represent mean judgments (averaged over other factors), and the dashed lines are the predictions of the scale-adjustment averaging model. In panel A, judgments based on source's opinion alone are shown with open points. Solid points are averaged over trials in which both source's opinion and base rate were given. The effect of source's opinion is predicted to be greater when opinion is alone than with base rate, because it is then proportional to $w_{xb}/(w + w_{xb})$, which is greater than $w_{xb}/(w + w_{xb} + w^*)$.

In panel B of Figure 4, the divergent Expertise × Bias interaction is predicted by the scale-adjustment averaging model (see Birnbaum & Stegner, 1979), because expertise multiplies the effect of bias, which is incorporated as an adjustment in the scale value. The greater the expertise of a source, the greater the effect of that source's bias.

In panel C of Figure 4, judgments are plotted against expertise of independent sources, with a separate curve for each level of base rate. The vertical spread between the curves is predicted to diminish from left to right because it is proportional to $w^*/(w + w^* + w_{xb})$, which decreases as the expertise of the source, w_{xb}, increases. This finding is consistent with previous results (Birnbaum, 1976; Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976).

The scale-adjustment averaging model (Equation 5) implies no interaction between base rate and source's opinion. However, there is a slight divergent interaction in panel A of Figure 5, $F(4, 256) = 11.16$. The same interaction is more pronounced for the very high expertise source, shown in Figure 3 (panel C). This finding is analogous to previous results and may be accounted for by the configural-weight, scale-adjustment averaging model of Birnbaum and Stegner (1979), in which the lower valued source of information receives greater weight. It seems likely that the form of this interaction can be affected by changing

[5] In an averaging model the scale values often have greater range than the response scale. The most extreme response is assumed to be a weighted average of the initial impression, which is usually near the center of the scale, and the most extreme scale value, which must often be beyond the end point of the response scale. Scale values should not be compared with the response scale, as a judgment function intervenes (Mellers & Birnbaum, 1982).

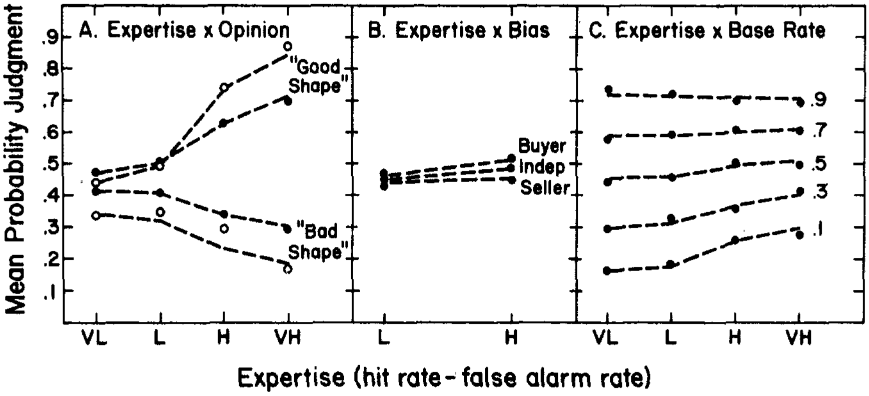


Figure 4. Effects of source's expertise. (Points are mean judgments; curves are predictions of scale-adjustment averaging model. Panel A shows probability judgments as a function of the expertise of independent sources with a separate curve for each level of opinion: Open circles show judgments based on source only; inner dashed curves are predictions averaged over levels of base rate. Panel B shows judgments as a function of source's expertise with a separate curve for levels of source's bias. Panel C shows judgments as a function of expertise of independent sources with a separate curve for each level of base rate. VL = very low; L = low; H = high; VH = very high.)

the judge's point of view (Birnbaum & Stegner, 1979).

Panels B and C of Figure 5 show the interaction between base rate and bias and between bias and opinion. In both cases, the model does a fairly good job reproducing the data. Because independents have higher weight than biased sources of the same expertise (Table 1), the effect of an independent's opinion (vertical separation between curves in Figure 5, panel C) is greater than that of either biased source. Furthermore, the weights explain why the effect of base rate (slope in Figure 5, panel B) is less when the source is independent than it is when sources are biased. In summary, despite some systematic deviations (panel A), the scale-adjustment averaging model generally gives a good fit to major features of the data.

Discussion

Base-Rate Fallacy

The term *base-rate fallacy* was coined to describe the supposed tendency to neglect base-rate information in favor of individuating information. Many of the studies that claimed to find evidence of a base-rate fallacy presented only one problem to a given subject. Our condition with one judgment yields similar data to those of Tversky and Kahneman (1980, 1982), Bar-Hillel (1980), and others. However, the present results show that when judges are presented with many problems, they utilize the base rate and are sensitive to expertise and bias of the source. Fischhoff et al. (1979) also found an effect of base-rate information in within-subjects designs.

It is difficult to interpret the effect or lack of effect of variables that have been manipulated between subjects. The problem is that when different subjects experience different stimulus contexts, responses cannot be compared without taking the different contexts into account. For example, by comparing judgments between subjects, it has been found that the number 450 can be judged greater than the number 550; however, one need not conclude that 450 actually seems greater than 550, because when judgments are compared within subjects, one finds that 550 is indeed judged greater than 450 (Birnbaum, 1974b, Figure 2). For a more complete discussion of between- and within-subjects designs, see Birnbaum (1982) and Mellers and Birnbaum (1982).

The case for a base-rate fallacy is also weakened by a theoretical analysis of the problems used in previous research. Birnbaum (1983) noted that several of the problems (cab problem, lawyer vs. engineer, light bulb) have a complex Bayesian solution in which behavior that has been described as "neglect" of base

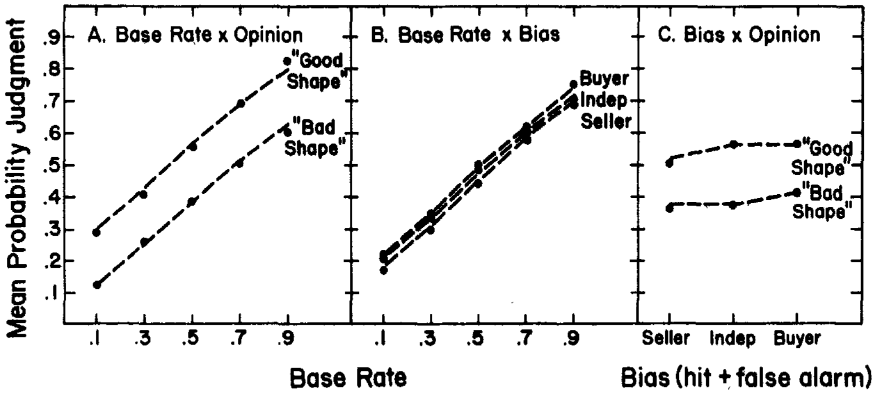


Figure 5. Effects of base rate, source's opinion, and source's bias. (Curves are predictions of scale-adjustment averaging model; points are data. Panel A shows mean judgments as a function of base rate with a separate curve for each level of source's opinion, averaged over the eight sources. Panel B shows mean judgments as a function of base rate with a separate curve for each level of bias of the sources, averaged over source's opinion and two levels of expertise. Panel C shows mean judgments as a function of source's bias, with a separate curve for each level of source's opinion, averaged over base rate and source's expertise.)

rate may actually be consistent with a rational Bayesian utilization of base rate.

In sum, despite the fact that the result of Kahneman and Tversky, Bar-Hillel, and others can be replicated in the single-judgment task, there are three reasons for a skeptic to doubt the claim that humans neglect base rate. The three reasons can be termed *empirical, methodological, and theoretical.* First, the empirical evidence concerning the base-rate fallacy is inconsistent. The major result does not appear in within-subjects designs or with certain versions of instructions (see Bar-Hillel, 1980; Birnbaum, 1983; Tversky & Kahneman, 1982). Second, the methodological problem is that comparison of numerical judgments between groups of subjects assumes that the mapping from subjective values to overt responses is the same for all subjects, an assumption that has been challenged by research on contextual effects (Birnbaum, 1974b; Mellers & Birnbaum, 1982). Third, there is a theoretical problem with much of the previous research. The problems posed to the subject have been vague enough to allow several rational solutions, depending on how the subject interprets the problem. Birnbaum (1983) has shown that the Bayesian solution to the cab problem used in previous studies requires the assumption that the hit to false-alarm ratio will be independent of base rate. However, research in signal detection shows that this ratio can increase as base rate decreases; therefore, Bayes' theorem does not necessarily imply an effect of base rate for problems such as the cab problem.

Bayes' Theorem and Conservatism

Bayes' theorem gives a better first approximation to the data than would the theory that subjects disregard base rate and source characteristics. For example, if subjects ignored base rate, the data curves in Figures 3 (panel C) and 5 (panels A and B) would be horizontal. If subjects ignored source characteristics, the curves marked VH and VL would coincide in Figure 3 (panel C), curves in Figure 4 would be horizontal, and the curves in Figures 4 (panel B) and 5 (panel B) would coincide. Bayes' theorem correctly predicts the direction of the major effects of base rate and source characteristics.

The objective form of Bayes' theorem (Figure 3, panel A) predicts judgments that are too extreme. The present data show a pattern of deviations between objective Bayesian predictions and judgments that might be described at first glance as "conservatism," as in Edwards (1968). For example, when the base rate is .9 and the very high expertise source says "good shape," the Bayesian prediction is .99, whereas the mean judgment is only .92 (Figure 3, panel C). The subjective form of

Bayes' theorem (Figure 3, panel B) gives a better second approximation to the data. By allowing the subjective impact of evidence to be less than the objective probabilities dictate, the subjective form of Bayes' theorem provides a better fit to the judgments.

However, even the subjective form of Bayes' theorem does not provide a satisfactory account of all of the data. The misperception and response bias interpretations of conservatism will not suffice to explain how increasing the evidence can decrease the response, and vice versa, (as in Figure 3, panel C). For example, when the very high expertise independent source says the car is in good shape, the Bayesian prediction is .9 and the mean response is .87. When subjects are also informed that the base rate is .7, Bayes' theorem implies an *increase* in the probability the car will last (to .96). However, the mean response was only .81, a *decrease* in probability. This example and others indicate that subjects are not simply being conservative: They can actually do the opposite of that implied by either form of Bayes' theorem.

The scale-adjustment averaging model can explain the above results as well as the overall conservatism. The model implies that additional faint praise can actually hurt the overall evaluation if the original evidence was already strong. Thus, humans appear to aggregate evidence by judging the average strength of evidence rather than by using Bayesian algebra in which a lot of weak evidence can imply a strong conclusion. The results in Figure 3 (panel C) are directly analogous to findings in intuitive regression (Birnbaum, 1976).

Conceptually, the scale adjustment averaging model can be interpreted as an attempt by the subject to strike a balance among the values of the evidence (Birnbaum & Stegner, 1979). The value of the evidence depends on the source's report and is adjusted according to the source's bias. This value is then multiplied by the source's expertise (or discriminability). These premises explain how expertise amplifies bias in Figure 4 (panel B). By interpreting the response as a balance between the source's report and the base rate, the model explains how weak but favorable evidence can decrease the overall response (Figure 3, panel C), and it explains how increasing source's expertise simultaneously amplifies the effect of source's opinion (Figure 4, panel A) and decreases the effect of base rate (Figure 4, panel C).

Research in Source Credibility

The present results fit in well with related research on how judges combine information from sources that vary in credibility (see Table 2). Birnbaum and Stegner (1979) found that the scale-adjustment averaging model gave a good description of estimates of the value of used cars based on estimates provided by sources. It is interesting that the same model describes both estimates of value and probability inferences. This finding appears consistent with Shanteau (1970) who concluded that judges use the same model for probability estimates and inferences based on samples drawn from an urn.

The scale-adjustment averaging model can be viewed as an extention of earlier averaging models of information integration (N. Anderson, 1971; Rosenbaum & Levin, 1968, 1969; Shanteau, 1975). It is important to note that there are a variety of different averaging models and there has been considerable theoretical discussion of how to represent deviations from early forms of the averaging model (T. Anderson & Birnbaum, 1976; Birnbaum, 1974a; Birnbaum & Stegner, 1979; Riskey & Birnbaum, 1974). At present, a configural-weighted, scale-adjustment averaging model appears to provide a consistent account of a wide variety of data (Birnbaum, 1982).

There now exists an array of consistent results linked by analogy, as in Table 2. The same model has given a good account of data in the following tasks: intuitive numerical predictions, based on independent numerical cues of known correlation with the criterion (Birnbaum, 1976); likableness of hypothetical persons described by adjectives provided by acquaintances of varied length of acquaintance with the target (Birnbaum et al., 1976); judgments of value of used cars based on blue book value and the opinions of sources who examined the cars (Birnbaum et al., 1976; Birnbaum & Stegner, 1979); predictions of IQs of adopted children based on IQs of biological and adoptive parents and environmental socioeconomic status (Birnbaum & Stegner,

Table 2
Analogies Among Variables in Source Credibility Research

Judgment	Messages (s)	Source expertise (w)
Numerical predictions[a]	Numerical cues	Cue-criterion correlations
Likableness of hypothetical people[b]	Adjectives	Length of acquaintance
Value of used cars[b,c]	Source's estimates; Blue book value	Mechanical knowledge
Predicted IQ[d]	IQs of biological and adoptive parents; SES of environment	Individual differences in importance of heredity vs. environment
Predicted exam performance[e]	IQ test; Study time	Reliability of IQ and study time measures
Probability inferences: $P(L\|\text{"G"})$[f]	Sources report, "G"; Base rates, $P(L)$	Hit and false-alarm rates of sources; $P(\text{"G"}\|L)$ and $P(\text{"G"}\|\bar{L})$

Note. Each row represents a different experimental situation.
[a] Birnbaum (1976). [b] Birnbaum, Wong, & Wong (1976). [c] Birnbaum & Stegner (1979). [d] Birnbaum & Stegner (1981). [e] Surber (1981). [f] Present experiment: For inferences, $P(L)$ = probability the car will last; "G" is the report of the source ("good shape").

1981); and predictions of performance based on unreliable or reliable measures of IQ and study time (Surber, 1981).

In these studies, weight is used to represent cue-criterion correlation, source's length of acquaintance with the target person, source's mechanical expertise, perceived importance of heredity or environmental information, and reliability of information (length of test) of IQ test or study time sample, respectively. In the present case of probability inferences, weight is mostly a function of source's hit rate minus false-alarm rate. Thus, weight is based on a source's perceived ability to discriminate the true states of nature.

In this set of analogies, the scale value of information depends on the source's message: the value of the cue, the likableness of the adjective, the source's estimate, the value of parents' IQs, and the value of IQ test or study time, respectively. In the present study, scale value depends on the source's opinion (good or bad shape).

In the scale-adjustment averaging model, the source's bias also affects the scale value. In both the present study and that of Birnbaum and Stegner (1979), it was found that the effect of bias on the judgments was greater for sources of higher expertise. Thus, when a source is believed to be biased, judges appear to adjust the value of the source's message before combining it with other information.

Conclusions

In sum, subjects utilize base-rate information in the within-subject design but give data comparable to previous findings when they are given only a single problem. Although subjects utilize the base rate, neither the objective nor subjective versions of Bayes' theorem gives a satisfactory account of the judgments. Instead, a scale-adjustment averaging model gives a reasonably accurate account of the data. In this model, base-rate information is averaged with information from other sources. The scale value of the source's opinion depends on the report and the bias of the source, whereas the weight of the source depends primarily on the source's expertise.

References

Anderson, N. H. Integration theory and attitude change. *Psychological Review,* 1971, *78,* 171–206.

Anderson, T., & Birnbaum, M. H. Test of an additive model of social inference. *Journal of Personality and Social Psychology,* 1976, *33,* 655–662.

Ajzen, I. Intuitive theories of events and the effects of base rate information on prediction. *Journal of Personality and Social Psychology,* 1977, *35,* 303–314.

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

Birnbaum, M. H. The nonadditivity of personality impressions. *Journal of Experimental Psychology,* 1974, *102,* 543–561. (Monograph) (a)

Birnbaum, M. H. Using contextual effects to derive psychophysical scales. *Perception & Psychophysics,* 1974, *15,* 89–96. (b)

Birnbaum, M. H. Intuitive numerical prediction. *American Journal of Psychology,* 1976, *89,* 417–429.

Birnbaum, M. H. Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical judgment.* Hillsdale, N.J.: Erlbaum, 1982.

Birnbaum, M. H. Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology,* 1983, *96,* 85–94.

Birnbaum, M. H., & Stegner, S. E. Source credibility in social judgment: Bias, expertise and the judge's point of view. *Journal of Personality and Social Psychology,* 1979, *37,* 48–74.

Birnbaum, M. H., & Stegner, S. E. Measuring the importance of cues in judgment for individuals: Subjective theories of IQ as a function of heredity and environment. *Journal of Experimental Social Psychology,* 1981, *17,* 159–182.

Birnbaum, M. H., Wong, R., & Wong, L. Combining information from sources who vary in credibility. *Memory & Cognition,* 1976, *4,* 330–336.

Carroll, J. S., & Siegler, R. S. Strategies for the use of base rate information. *Organizational Behavior and Human Performance,* 1977, *19,* 392–402.

Chandler, J. P. STEPIT: Finds local minima of a smooth function of several parameters. *Behavioral Science,* 1969, *14,* 81–82.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Subjective sensitivity analysis. *Organizational Behavior and Human Performance,* 1979, *23,* 339–359.

Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review,* 1973, *80,* 237–251.

Lyon, D., & Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica,* 1976, *40,* 287–298.

Manis, M., Dovalina, I., Avis, N., & Cardoze, S. Base rates can affect individual predictions. *Journal of Personality and Social Psychology,* 1980, *38,* 231–248.

Mellers, B. A., & Birnbaum, M. H. Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance,* 1982, *8,* 582–601.

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. *Psychological Bulletin,* 1967, *68,* 29–46.

Pitz, G. F. Bayes' theorem: Can a theory of judgment and inference do without it? In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1975.

Riskey, D. R., & Birnbaum, M. H. Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology,* 1974, *103,* 171–173.

Rosenbaum, M. E., & Levin, I. P. Impression formation as a function of source credibility and order of presentation of contradictory information. *Journal of Personality and Social Psychology,* 1968, *10,* 167–174.

Rosenbaum, M. E., & Levin, I. P. Impression formation as a function of source credibility and the polarity of information. *Journal of Personality and Social Psychology,* 1969, *12,* 34–37.

Schum, D. A. Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance,* 1981, *27,* 153–196.

Shanteau, J. An additive model for sequential decision making. *Journal of Experimental Psychology,* 1970, *85,* 181–191.

Shanteau, J. Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica,* 1975, *39,* 83–89.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, *6,* 649–744.

Surber, C. F. Effects of information reliability in predicting task performance using ability and effort. *Journal of Personality and Social Psychology,* 1981, *40,* 977–989.

Troutman, C. M., & Shanteau, J. Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance,* 1977, *19,* 43–55.

Tversky, A., & Kahneman, D. Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1). Hillsdale, N.J.: Erlbaum, 1980.

Tversky, A., & Kahneman, D. Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press, 1982.

Wallsten, T. Conjoint-measurement framework for the study of probabilistic information processing. *Psychological Review,* 1972, *79,* 245–260.

Received February 10, 1982
Revision received February 24, 1983 ■