Memory & Cognition
2002, 30 (2), 171-178

The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle

GAËLLE VILLEJOUBERT
Leeds University Business School, Leeds, England

and

DAVID R. MANDEL
University of Victoria, Victoria, British Columbia, Canada

In judging posterior probabilities, people often answer with the inverse conditional probability—a tendency named the *inverse fallacy*. Participants ($N = 45$) were given a series of probability problems that entailed estimating both $p(H|D)$ and $p(\sim H|D)$. The findings revealed that deviations of participants' estimates from Bayesian calculations and from the additivity principle could be predicted by the corresponding deviations of the inverse probabilities from these relevant normative benchmarks. Methodological and theoretical implications of the distinction between inverse fallacy and base-rate neglect and the generalization of the study of additivity to conditional probabilities are discussed.

In this article, we examine posterior probability judgment, which involves one's assessing the likelihood of an event by updating a prior probability in light of new evidence. A normative model for calculating posterior probabilities is Bayes's theorem. This theorem states that $p(H|D)$, the posterior probability that hypothesis H is true given datum D, can be calculated as follows:

$$p(D|H) = \frac{p(D|H) \cdot p(H)}{p(D|H) \cdot p(H) + p(D|\sim H) \cdot p(\sim H)}, \qquad (1)$$

where $p(D|H)$ and $p(D|\sim H)$ refer to the conditional probability of observing D, given that hypothesis H is true and given that the mutually exclusive, alternative hypothesis, $\sim H$, is true, respectively. In Bayesian terms, these probabilities are called *likelihoods,* whereas the probabilities $p(H)$ and $p(\sim H)$ are called *prior probabilities.* Posterior probability judgments are fundamental to belief revision and are involved in many consequential real-world situations such as medical diagnosis or juror decision making. Consider, for example, a physician who knows, *prior* to the examination of an individual patient, the probability that a person will have disease X. If the patient presents a diagnostic symptom, she will have to update the probability that he has the disease, given this new observation. Suppose the physician knows that (1) only 5% of the overall population suffers from disease X, (2) 85% of patients who have the disease show the symptom, and (3) 25% of healthy patients show the symptom. According to Bayes's theorem, the posterior probability that a patient who shows the symptom has the disease can be calculated as follows: $p(\text{disease} | \text{symptom}) = (.85 \times .05)/[(.85 \times .05) + (.25 \times .95)] = .15$. Thus, there is only a 15% chance that the patient examined has the disease even though he presents a highly diagnostic symptom.

The Inverse Fallacy

That both lay and expert judges often confuse a given conditional probability with its inverse probability has been noted in many studies. This tendency has been alternatively labeled the *conversion error* (Wolfe, 1995), the *confusion hypothesis* (Macchi, 1995), the *Fisherian algorithm* (Gigerenzer & Hoffrage, 1995), and the *inverse fallacy* (Koehler, 1996a). In the present article, we adopt Koehler's term to refer to the tendency for judges to confuse any of the following: $p(H|D)$ with $p(D|H)$, $p(\sim H|D)$ with $p(D|\sim H)$, $p(H|\sim D)$ with $p(\sim D|H)$, or $p(\sim H|\sim D)$ with $p(\sim D|\sim H)$. Although there are other algorithms that participants can use when they estimate posterior probabilities (see, e.g., Gigerenzer & Hoffrage, 1995), the inverse fallacy is often the most frequent error observed.

As early as 1955, Meehl and Rosen reported that clinicians considered that the probability of the presence of a symptom given the diagnosis of a disease was on its own a valid criterion for diagnosing the disease in the presence of the symptom. This result was later experimentally demonstrated by Hammerton (1973), who observed that me-

We thank Vittorio Girotto, Evan Heit, Jay Koehler, David Over, and Frédéric Vallée-Tourangeau for their feedback on this research. The data for this research were collected in 1999 at the Department of Psychology, University of Hertfordshire, Hatfield, United Kingdom while the first author was completing her doctoral thesis. Portions of this research were presented at the Fourth International Thinking Conference, Durham, United Kingdom. Correspondence concerning this article should be sent to G. Villejoubert, Leeds University Business School, Maurice Keyworth Building, University of Leeds, Leeds, West Yorkshire LS2 9JT, England (e-mail: gv@lubs.leeds.ac.uk).

—Accepted by previous editorial team

dian judgments of p(disease | symptom) were almost equal to the presented value of the inverse probability, p(symptom | disease). Liu (1975) replicated those results by varying the value of $p(D|H)$ in a between-subjects design. Similarly, Eddy (1982) investigated how physicians estimated the probability that a woman has breast cancer, given a positive result of a mammogram. Approximately 95% of clinicians surveyed gave a numerical answer close to the inverse probability.

In Kahneman and Tversky's (1972) *taxicab problem* (see also Bar-Hillel, 1980; Lyon & Slovic, 1976; Tversky & Kahneman, 1980), participants were asked to estimate the probability that a cab had been involved in an accident given that it was Blue rather than Green. When asked to estimate $p(H|D)$, most participants answered with a value that matched the inverse probability, $p(D|H)$. More recently, Dawes, Mirels, Gold, and Donahue (1993) demonstrated that this fallacy extended to individuals' beliefs inherent to their implicit personality theory.

Some researchers have interpreted these findings in terms of a *base-rate fallacy*. The inverse fallacy is then understood to be the result of people's tendency to consistently undervalue, if not ignore, the base-rate information presented as a proxy for prior probabilities (e.g., Bar-Hillel, 1980; Dawes et al., 1993; Kahneman & Tversky, 1973; Pollard & Evans, 1983). Other researchers, however, have proposed that the base-rate effect was in fact originating from the inverse fallacy and not the reverse (e.g., Hamm, 1993; Koehler, 1996a; Wolfe, 1995). In support of this notion, Wolfe (1995, Experiment 3) found that participants who were trained to distinguish $p(D|H)$ from $p(H|D)$ were less likely to exhibit base-rate neglect compared with a control group. We agree that base-rate and inverse fallacies are different. The inverse fallacy entails not only the neglect of the base-rate information but also that of $p(D|\sim H)$. To illustrate this argument, consider the diagrams shown in Figure 1. Each diagram depicts two categories H and ~H. Their base rates are the proportion of space occupied by each category, respectively. The sample space delimited by the hatched areas represents the proportion of elements having the feature D. These diagrams indicate that the inverse fallacy relies on a different representation and integration of available information than does the base-rate fallacy. Moreover, as shown in Diagram 2, the integration of base rates into the final judgment is unnecessary when they are equal. Therefore, if judgment accuracy were only undermined by the neglect of base-rate information, judgments involving equal base rates should be normative.

If people commit the inverse fallacy in judging posterior probabilities, one would expect that posterior probability estimates would be systematically biased as a function of the deviation between the posterior probability and its inverse probability. Consider the physician who needs to estimate p(disease | symptom). If she commits the inverse fallacy, she will answer with the value of the inverse probability p(symptom | disease) = .85, rather than the Bayesian value of p(disease | symptom) = .15. In this case, the deviation between the estimate and the Bayesian value of p(disease | symptom) is .70. Although some preliminary research indicates that the inverse fallacy is a distinct contributor to deviations from Bayesian judgment, no study has yet examined whether the deviations between posterior probabilities and their inverse probabilities can be used to predict people's deviations from Bayesian judgment. This was a key objective of the present study.

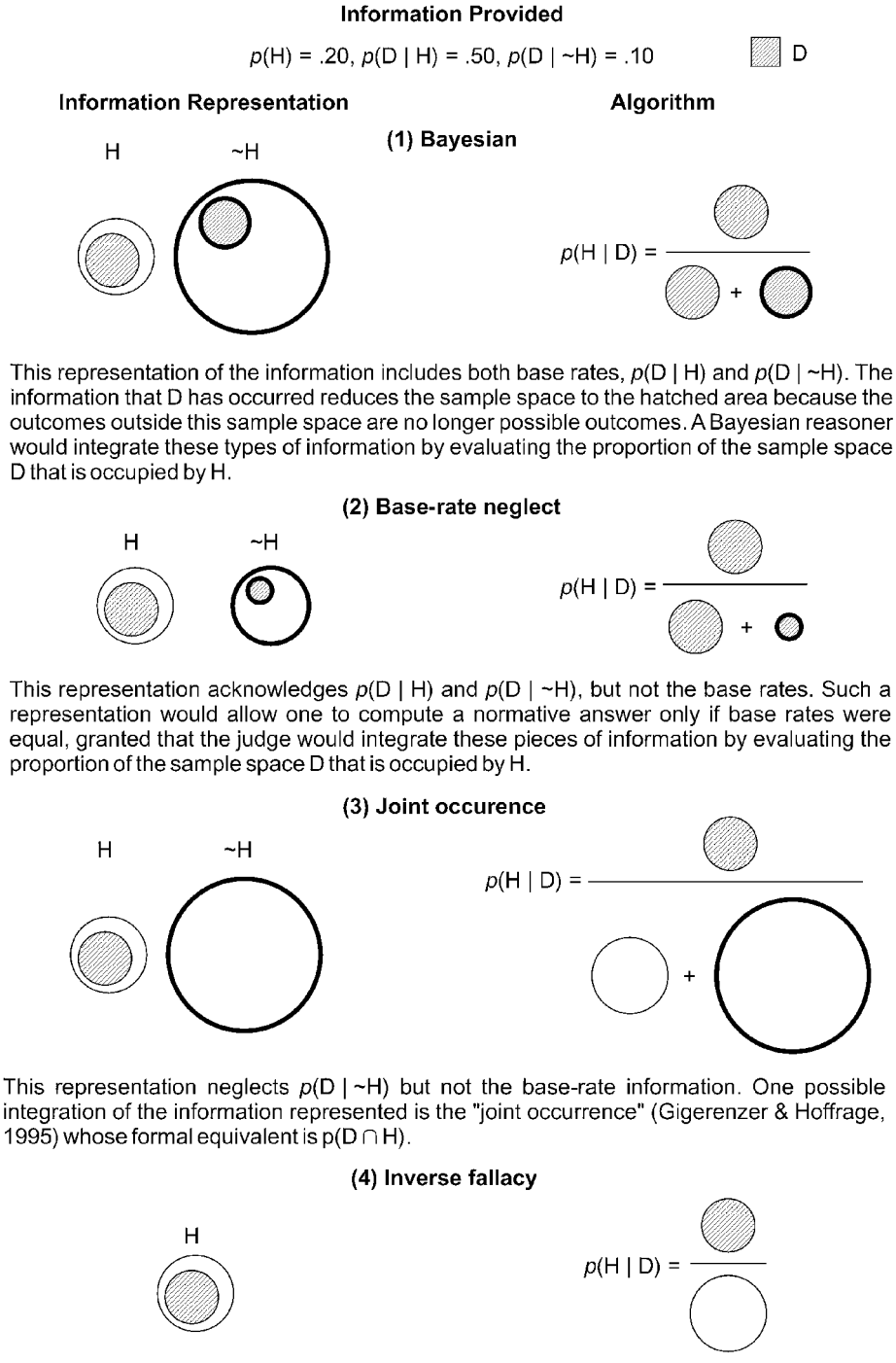
The Additivity Principle

The additivity principle states that the judged probabilities for complementary events should sum to unity. For instance, if one judges that p(disease | symptom) is .75, one also should judge that p(no_disease | symptom) is .25. From a descriptive standpoint, there is disagreement within the literature as to whether we should expect that judgments of two complementary probabilities will be additive. Some researchers (e.g., Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) have reported that the judged probability of a hypothesis and that of its complement are additive. Other researchers (e.g., Ayton, 1997) have reported that people's probability judgments tend to be *subadditive* (i.e., the sum of the individual estimates is greater than one). And evidence of *superadditivity* in the case of binary complementarity also has been found (see, e.g., Macchi, Osherson, & Krantz, 1999).

Although previous research has examined unconditional probability judgments, the additivity principle also applies to conditional probabilities because $p(H|D)$ and $p(\sim H|D)$ always add to one when H and ~H are exhaustive and mutually exclusive. In the domain of conditional probability judgments, Baratgin and Noveck (2000) suggested that the participants in Kahneman and Tversky's (1973) lawyer–engineer problem violated the additivity principle. The authors demonstrated that the participants integrated base rates more efficiently when they were induced to make complementary estimates. In the present study, we tested a stronger, empirically verifiable claim. Namely, that participants' complementary estimates will not be additive, mainly *because* people commit the inverse fallacy. Furthermore, the inverse fallacy offers a basis for a *systematic* prediction of the patterns of deviation from additive judgment. Our second objective, then, was to examine whether people's estimates of complementary posterior probabilities, and any deviations from the additivity principle, can be predicted on the basis of the inverse conditional probabilities.

EXPERIMENT 1

In this experiment, we examined whether evidence for the inverse fallacy would emerge even when base rate neglect could be ruled out. We used a problem adapted from Slowiaczek, Klayman, Sherman, and Skov (1992, Experiment 1A) in which participants were asked to estimate the posterior probability that an encountered alien creature



This representation of the information includes both base rates, $p(D \mid H)$ and $p(D \mid \sim H)$. The information that D has occurred reduces the sample space to the hatched area because the outcomes outside this sample space are no longer possible outcomes. A Bayesian reasoner would integrate these types of information by evaluating the proportion of the sample space D that is occupied by H.

This representation acknowledges $p(D \mid H)$ and $p(D \mid \sim H)$, but not the base rates. Such a representation would allow one to compute a normative answer only if base rates were equal, granted that the judge would integrate these pieces of information by evaluating the proportion of the sample space D that is occupied by H.

This representation neglects $p(D \mid \sim H)$ but not the base-rate information. One possible integration of the information represented is the "joint occurrence" (Gigerenzer & Hoffrage, 1995) whose formal equivalent is $p(D \cap H)$.

This representation implies the neglect of both $p(D \mid \sim H)$ and the neglect of the base-rate information. In this case, the judgment of $p(H \mid D)$ only relies on $p(D \mid H)$.

Figure 1. Sample-space representations of the information provided in textbook problems.

was one of two mutually exclusive types in light of the presence or absence of a diagnostic feature. Skov and Sherman (1986) noted that the use of natural groups imposes restrictions on the likelihood of a particular feature in these groups. In such cases, diagnosticity and likelihood are often confounded: A diagnostic trait (e.g., *likes parties*) would be frequent in the focal group (e.g., extroverts) and infrequent in the alternative group (introverts).

Table 1
Percent of Gloms and Fizos Presenting Each of Twelve Features and Encountered Creature's Answer

Gloms	Fizos	Features	Response
98	58	plays the harmonica	Yes
2	42	exhales fire	No
90	50	wears hula hoops	Yes
10	50	gurgles a lot	No
80	40	have a flying license	Yes
20	60	gulp bluebottles down	No
42	2	smokes maple leaves	Yes
58	98	drinks petrol	No
50	10	has gills	Yes
50	90	eats iron ore	No
60	20	breeds scampi	Yes
40	80	climbs walls	No

Thus, the use of unnatural categories allowed us to control for the likelihood of the features.

Our first hypothesis was that the majority of participants would commit the inverse fallacy. Second, we hypothesized that the deviation between the participants' estimates and Bayesian answers could be predicted by the deviation between $p(D|H)$ and $p(H|D)$. Thus, participants were expected to overestimate $p(H \mid D)$ when $p(D|H) > p(H|D)$, and they were expected to underestimate the Bayesian posterior when $p(D|H) < p(H|D)$. Moreover, by experimentally manipulating the size of the deviation between posterior and inverse probabilities, we were also able to predict the magnitude of the participants' judgment inaccuracies. The third hypothesis tested was that when $p(D \mid H)$ and $p(D|{\sim}H)$ sum to less than one, the sums of participants' judgments of $p(H \mid D)$ and $p({\sim}H|D)$ would be superadditive. Conversely, the sums of their judgments of $p(H|D)$ and $p({\sim}H|D)$ were expected to be subadditive when $p(D|H)$ and $p(D|{\sim}H)$ exceeded one.

Our experimental design also allowed us to distinguish between the inverse fallacy and a simpler, matching heuristic. In the domain of logical reasoning, Evans (1998) defined a *matching bias* as a tendency to only consider as relevant the information whose lexical content matches that of the information presented in the propositional rule to be tested. By extension, the tendency to equate $p(H|D)$ with $p(D|H)$ could be defined as the tendency to estimate $p(H|D)$ on basis of the match between the experimental question and the information presented. Such a strategy, however, would lead people to answer with the displayed value of $p(D|H)$, when that value was explicitly provided, even when they were asked to estimate $p(H|{\sim}D)$. By contrast, the inverse fallacy account proposes that people will estimate $p(H|{\sim}D)$ with the value of $p({\sim}D|H)$. We therefore expected to be able to distinguish between the inverse fallacy and the matching heuristic in cases in which diagnostic response information indicated the absence of D (i.e., via *no* responses).

Method

Forty-five University of Hertfordshire undergraduates participated in the experiment for course credit. The participants were provided with a 13-page questionnaire. The first page presented their task as follows:

> Imagine you are visiting a planet called Vuma. There are two and only two types of *invisible* creatures that live on this planet. There are 1 million Gloms, and 1 million Fizos.
>
> You will randomly meet 12 creatures. Imagine you are particularly interested in guessing their identity. Each time you meet one of the invisible creatures, you want to know whether it is a Glom or a Fizo. You will walk with an interpreter who will ask each creature whether or not it possesses a certain feature. Each time, you will be provided with the percentages of Gloms and Fizos on Vuma possessing the target feature. The creatures cannot help but to tell the truth, so you can be sure you will get a truthful answer, which will provide you with some information about the creature's identity.

The participants were then provided with an example and were told: "On the basis of the creature's answer, you will be asked to estimate both the likelihood that it is a Glom and the likelihood that it is a Fizo. Turn the page for your first encounter." The participants "met" 12 creatures one by one on the subsequent pages. For each of these 12 encounters, the creature was asked about a different feature, and the participants were provided with a reminder of the number of Gloms and Fizos on the planet as well as the percentages of each type of creature that possessed the feature. In order to be consistent with Slowiaczek et al. (1992), posterior probability judgments were elicited by using a frequency question, and the participants were asked to estimate the "chances in 100" rather than the "probability" that H (or ~H) was true. A summary of the stimuli is presented in Table 1. Figure 2 shows the questionnaire layout corresponding to the first line of Table 1.

Stimuli and Design. Each of the 12 encounters represented a unique stimulus condition in a 2 (creature's response: no, yes) × 2 [expected direction of deviation: $p(D|H) < p(H|D)$, $p(D|H) > p(H|D)$] × 3 (expected magnitude of deviation: small, medium, large) fully crossed repeated-measures design with two dependent variables measured for each stimulus. Each of the 12 stimuli required two estimates, and each single estimate corresponded to distinct Bayesian and inverse values (see Table 2). The diagnostic probabilities presented

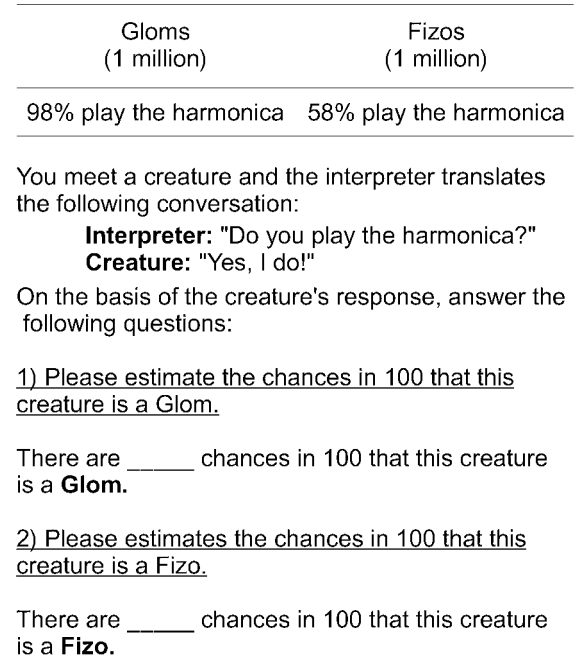


Figure 2. Representation of one of the stimuli used.

Table 2
Summary of Stimuli, Design, and Results: Bayesian and Inverse Probabilities, Expected and Observed Deviations as a Function of Magnitude and Direction of Deviation for Glom and Fizo Measures

Expected Magnitude of Deviation	Probabilities				Deviations					
					Bayes's Theorem					
	Bayesian		Inverse		Glom		Fizo		Additivity	
	Glom	Fizo	Glom	Fizo	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.
Expected Overestimation of Bayesian Values										
Large	.63	.37	.98	.58	.35	.14	.21	.09	1.56	1.23
Medium	.64	.36	.90	.50	.26	.10	.14	.06	1.40	1.16
Small	.67	.33	.80	.40	.13	.00	.07	.05	1.20	1.05
Expected Underestimation of Bayesian Values										
Large	.96	.04	.42	.02	−.54	−.53	−.02	.05	.44	.56
Medium	.83	.17	.50	.10	−.33	−.33	−.07	.00	.60	.72
Small	.75	.25	.60	.20	−.15	−.22	−.05	.10	.80	.78

Note—Bayesian probabilities refer to p(type | response) and inverse probabilities refer to p(response | type). Exp., Expected; Obs., Observed. The expected deviations from Bayes's theorem are given by the deviations between Bayesian and inverse values. The expected deviations from additivity are given by the deviations between the sum of the inverse values for Gloms and Fizos, and unity.

were chosen so that the creature's response would have no effect on the participants' estimates. Thus, the set of *yes* stimuli were perfectly matched to the set of *no* stimuli in terms of the Bayesian and inverse values. The order of stimulus presentation was randomized, and question order was counterbalanced across participants.

Results

Classification of estimates. In order to test whether the participants committed the inverse fallacy, each estimate was compared with the value of its corresponding inverse and Bayesian probability, respectively. An estimate was classified as Bayesian or as inverse if it was equal to its corresponding Bayesian or inverse value, respectively, within a margin of error of ±.02. Figure 3 shows the distribution of participants according to the number of their estimates classified as Bayesian or as inverse. The vast majority (80%) of participants provided no more than 4 Bayesian estimates out of a total of 24. By contrast, and in support of our first hypothesis, as many as 51% of the participants had 20 or more estimates that were equal to the corresponding inverse value. This is a powerful result that demonstrates the prevalence of committing the inverse fallacy in judging posterior probabilities. Finally, as expected, the participants did not rely on a matching heuristic, since most estimates of $p(H \mid \sim D)$ and $p(\sim H \mid \sim D)$ relied on the values of $p(\sim D \mid H)$ and $p(\sim D \mid \sim H)$, respectively. Within the set of *no* stimuli, 55% of the participants provided at least 7 (out of 12) estimates equal to the inverse value, whereas only 1 participant provided 7 estimates consistent with a matching heuristic.

Deviations from Bayesian judgment. The participants' estimates were converted to deviation scores, d, by subtracting the corresponding Bayesian value. Thus, $d = 0$ for accurate Bayesian estimates, $d > 0$ for overestimations of Bayesian probabilities, and $d < 0$ for underestimations of Bayesian probabilities. Analyses including order of stimulus presentation and order of question presentation as between-subjects variables revealed no significant effect of this manipulation, and the data were collapsed across order. The deviation scores were subjected to a 2 (creature's response) × 2 (expected direction of deviation) × 3 (expected magnitude of deviation) doubly multivariate repeated-measures analysis of variance (ANOVA). Hypothesis 2 stated that the participants' deviations from Bayesian judgment would be predicted by the deviation between the inverse and the Bayesian values. Accordingly, a significant two-way (expected direction of deviation × expected magnitude of deviation) interaction was observed [multivariate $F(4,41) = 47.29$, $p < .0005$]. Univariate F tests revealed that this interaction was statistically significant only for the Glom measure [$F(2,88) = 106.69$, $MS_e = 0.02$, $p < .01$, $\eta^2 = .71$]. The 2 (expected direction of deviation) × 3 (expected magnitude of deviation) interaction is shown in Table 2 (in the "Bayes's theorem-

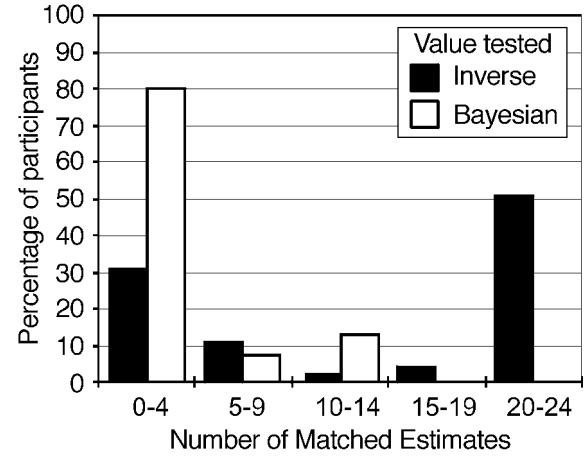


Figure 3. Distribution of participants as a function of the number of their estimates equating either the inverse probability or the Bayesian probability.

Glom-Observed" column). Deviation scores were positive and negative when overestimation and underestimation was expected, respectively. Moreover, the magnitude of these deviations varied as predicted. The multivariate analysis also revealed a significant main effect of response on deviation scores [multivariate $F(2,43) = 5.45$, $p < .01$]. It has already been suggested and observed that participants tend to be more influenced by positive answers than by negative answers (see Sherman & Corty, 1984; Slowiaczek et al., 1992), leading to a higher level of confirmation of the hypothesis tested.

Deviations from the additivity principle. Our third hypothesis was that the observed patterns of additivity would be a function of the sum of inverse probabilities. The sum of each participant's estimates of the complementary probabilities p(Glom | response) and p(Fizo | response) was computed. These sums were subjected to a 2 (creature's response) × 2 (expected direction of deviation) × 3 (expected magnitude of deviation) repeated-measures ANOVA. As expected, response did not significantly affect the scores [$F(1,44) = 0.12$, $MS_e = 0.05$, $p > .05$]. The significant main effects of expected direction and expected magnitude of deviation from unity are better explained by the predicted two-way (expected direction of norm deviation × expected magnitude of norm deviation) interaction [$F(2,88) = 36.09$, $MS_e = 0.05$, $p < .001$, $\eta^2 = .45$]. This interaction is illustrated in the last column of Table 2. As predicted, the participants' estimates were subadditive when complementary inverse probabilities summed to more than 1, and their estimates were superadditive when these probabilities summed to less than 1. The magnitude of these deviations also varied as predicted by the sum of inverse probabilities. Finally, a significant two-way (expected direction of norm deviation × creature's response) interaction was obtained [$F(1,44) = 5.96$, $MS_e = 0.14$, $p < .02$]. Further analyses revealed that this effect was due to a significant effect of response when superadditivity was expected [$M_{yes} = .65$, $SD = .22$ vs. $M_{no} = .73$, $SD = .24$, Bonferronni $t(88) = 14.44$, $MS_{pooled\ error} = .11$, $p < .001$, with a corrected α level of .025].

DISCUSSION

Consistent with previous research (e.g., Bar-Hillel, 1980; Eddy, 1982; Hamm, 1993; Wolfe, 1995), the results obtained in the present study reveal that roughly half of the sample equated the posterior and inverse probabilities on over 80% of the judgment trials. This is a powerful finding because, in past studies demonstrating the inverse fallacy, typically only one judgment per participant was solicited. By contrast, we have demonstrated that a sizeable proportion of judges consistently used an inverse fallacy algorithm over a set of judgment tasks that varied in terms of both the question asked and the available probability information. We were able to demonstrate that the direction and magnitude of deviations of participants' estimates from both Bayesian judgment and the additivity principle were successfully predicted by the deviation between the inverse and posterior probabilities, thus supporting our second and third hypotheses, respectively. Furthermore, deviations from normative benchmarks could not be accounted for by the base-rate fallacy because base rates for the two relevant categories were always equal.

Methodological Implications

Girotto and Gonzalez (2001) showed that some of Cosmides and Tooby's (1996) observations were erroneously classified as accurate judgments by these authors on the basis of parity between participants' posterior probability estimates and the numerical values computed by Bayes's theorem. In the present study, the collection of multiple estimates for each participant militated against such unwarranted conclusions resulting from a confusion between the outcome of a test and the computational process leading to this outcome. Such methodological precautions prevented us from mistakenly concluding that judgments of p(Fizo | response) demonstrated normative judgment (as they showed little deviation from the Bayesian norm, see Table 2). A more likely explanation, supported by the results observed for the Glom measure, is that those judgments were based on the inverse fallacy, whose outputs happened to be similar to those arising from Bayes's theorem.

The fact that the inverse fallacy is associated with nonadditive posterior probability judgments also has significant methodological implications. For instance, Slowiaczek et al. (1992, Experiment 1A) assumed that their participants' judgments were additive and combined estimates of $p(H|D)$ with estimates of $p(\sim H|D)$ subtracted from 1. This recoding procedure may have induced a bias in their results. For instance, when $p(D|H) = .5$, $p(D|\sim H) = .10$ and D is present (a stimulus that was also used by Slowiaczek et al., 1992), $p(H|D) = .83$ and $p(\sim H|D) = .17$. A judge committing the inverse fallacy, would estimate $p(H|D)$ and $p(\sim H|D)$ to be .50 and .10, respectively. Recoding $p(\sim H|D)$ as a .90 (1 − .10) estimate of $p(H|D)$ results in a near-normative answer, whereas the judge committing the inverse fallacy would have been more likely to estimate $p(H|D)$ to be .50 [the value of $p(D|H)$], thus underestimating the normative value by more than a 30% difference.

Theoretical Implications

Additivity. The present experiment extended the study of additivity to conditional probability judgments. Specifically, we demonstrated that the pattern of subadditivity and superadditivity observed for the participants' judgments could be predicted from the sum of the inverse probabilities. Rottenstreich and Tversky (1997) specified that the binary complementarity predicted by support theory (Tversky & Koehler, 1994) applies to cases in which the alternative hypothesis is *explicitly described* as such. This precision implies that, in the present study, support theory would only predict additivity for measures of p(Glom | response) and p(∼Glom | response). Still, in the present

context, it was made clear to the participants that the creature could only be a Glom or a Fizo (and obviously, not both). Rottenstreich and Tversky suggested that "additivity is likely to hold" (p. 407) in such a case. Our results, however, strongly indicate that the latter proposition does not hold, at least in the context of conditional probability judgment. Future research might examine the effect of implicitly versus explicitly negating the alternative hypothesis on the additivity of posterior probability judgments in cases of binary complementarity.

Origin and extent of the inverse fallacy. The question of the underlying bases and the scope of the inverse fallacy still require investigation. The *sample-space framework* (Gavanski & Hui, 1992; Hanita, Gavanski, & Fazio, 1997; Sherman, McMullen, & Gavanski, 1992) proposes that the inverse fallacy is the result of a memory-based process. People access sets of information (sample spaces) from memory when judging probabilities. When participants are asked to estimate $p(H|D)$, they are required to base their judgments on the sample space defined by feature D. Gavanski and Hui argued that this sample space is unnatural because knowledge is partitioned by categories rather than by features. So, people may replace the unnatural sample space with a more readily accessible one—namely, the sample space of category H. This process would result in the inverse fallacy. The sample-space explanation, however, cannot fully account for the present results because the categories used were hypothetical, and judgments could not be based on sample spaces stored in memory.

Alternatively, Macchi (1995, 2000) proposed that the formulation of diagnostic information plays a key role in the interpretation of the data. Consider the following formulations: (1) The percentage of elements presenting the feature D is three times higher among H elements than among ~H elements. (2) In the group of elements presenting the feature D, the percentage of H elements is three times higher than the percentage of ~H elements. (3) The feature D is present in x% of H elements, the feature D is present in y% of ~H elements, and x is three times higher than y. Macchi (1995) proposed that a formulation such as (1) is interpreted as (2) by participants, as opposed to what it logically implies—namely formulation (3). It is this misinterpretation, Macchi argued, that leads to the inverse fallacy. Consequently, the tendency to estimate $p(H|D)$ with $p(D|H)$ results from the lack of clarity of the independence of base rate $p(H)$ and $p(D|H)$. Macchi recommended the use of formulations such as (3) to avoid ambiguity and demonstrated that it could reduce the proportion of inverse fallacies. Yet, this explanation for the origin of the inverse fallacy has its shortcomings. It is not clear why it is the ambiguous independence of $p(D|H)$ and $p(H)$ that would lead people to mistake $p(D|H)$ for $p(H|D)$, given that the combination of the base rate and the diagnostic information $[p(D|H).p(H)]$ results in $p(D \cap H)$, which is not the formal equivalent of $p(H|D)$, as is demonstrated by the diagrams presented in Figure 1. Furthermore, this explanation conflicts with the sample-space account because Macchi's suggestion that participants' interpretations [i.e., formulation (2)] rely on the sample space defined by feature D. Yet, according to the sample-space explanation, this is an unnatural and unlikely basis for probability judgments. Finally, even though our formulation of the diagnostic information was in line with Macchi's (1995) recommendations for reducing the inverse fallacy, we still found that 51% of participants made almost all their judgments in accordance with the inverse fallacy.

An alternative account relies on the frequency hypothesis. Gigerenzer and Hoffrage (1995; see also, e.g., Cosmides & Tooby, 1996) demonstrated that Bayesian answers can be elicited with the use of a frequency format for both the information and the question asked. Moreover, Thompson and Schumann (1987) showed that frequency formats reduced the number of inverse fallacies committed. A probability presentation format might increase the use of heuristics, but that still does not explain why most people tend to commit the inverse fallacy rather than use another heuristic (e.g., the joint-occurrence algorithm described in Figure 1). Finally, it is plausible that people simply confuse $p(H|D)$ with $p(D|H)$ because the latter *sounds* a lot like the former (J. J. Koehler, personal communication, April 2001). However, at present, no empirical research has tested this account.

Concluding Remarks

The present study demonstrated how the inverse fallacy can account for deviations from Bayes's theorem and the additivity principle. This fallacy might also explain other results on probabilistic reasoning observed within the literature. Koehler (1996b) showed that people confuse posterior odds ratios with likelihood ratios. Such confusion could be explained by the tendency to commit the inverse fallacy. Doherty, Mynatt, Tweney, and Schiavo (1979) demonstrated that people tend to choose diagnostically worthless information such as $p(D_1|H)$ and $p(D_2|H)$ to revise the probability that hypothesis H is true. This "pseudodiagnosticity" phenomenon may be explained by the fact that participants seeking $p(D_1|H)$ and $p(D_2|H)$ think that they are given $p(H|D_1)$ and $p(H|D_2)$. Finally, the present research spotlighted the need to distinguish judgment output from the judgment process and demonstrated that the examination of judgment over multiple trials was an effective method to do so.

REFERENCES

Ayton, P. (1997). How to be incoherent and seductive: Bookmakers' odds and support theory. *Organizational Behavior & Human Decision Processes,* **72**, 99–115.

Baratgin J., & Noveck, I. A. (2000). Not only base rates are neglected in the Engineer–Lawyer problem: An investigation of reasoners' underutilization of complementarity. *Memory & Cognition,* **28**, 79-91.

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

COSMIDES, L., & TOOBY, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition,* **58**, 1-73.

DAWES, R. M., MIRELS, H. L., GOLD, E., & DONAHUE, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science,* **4**, 396-400.

DOHERTY, M. E., MYNATT, C. R., TWENEY, R. D., & SCHIAVO, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica,* **43**, 11-21.

EDDY, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge: Cambridge University Press.

EVANS, J. ST. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning,* **4**, 45-82.

GAVANSKI, I., & HUI, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality & Social Psychology,* **63**, 766-780.

GIGERENZER, G., & HOFFRAGE, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review,* **102**, 684–704.

GIROTTO, V., & GONZALEZ, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition,* **78**, 247-276.

HAMM, R. M. (1993). Explanation for common responses to the blue/green cab probabilistic inference word problem. *Psychological Reports,* **72**, 219-242.

HAMMERTON, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology,* **101**, 252-254.

HANITA, M., GAVANSKI, I., & FAZIO, R. H. (1997). Influencing probability judgments by manipulating the accessibility of sample spaces. *Personality & Social Psychology Bulletin,* **23**, 801-813.

KAHNEMAN, D., & TVERSKY, A. (1972). On prediction and judgment. *ORI Research Monographs,* **12**.

KAHNEMAN, D., & TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review,* **80**, 237-251.

KOEHLER, J. J. (1996a). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral & Brain Sciences,* **19**, 1-53.

KOEHLER, J. J. (1996b). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios and error rates. *University of Colorado Law Review,* **67**, 859-886.

LIU, A. Y. (1975). Specific information effect in probability estimation. *Perceptual & Motor Skills,* **41**, 475-478.

LYON, D., & SLOVIC, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica,* **40**, 287-298.

MACCHI, L. (1995). Pragmatic aspects of the base rate fallacy. *Quarterly Journal of Experimental Psychology,* **48A**, 188-207.

MACCHI, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior & Human Decision Processes,* **82**, 217-236.

MACCHI, L., OSHERSON, D., & KRANTZ, D. H. (1999). A note on superadditive probability judgment. *Psychological Review,* **106**, 210-214.

MEEHL, P., & ROSEN, A. (1955). Antecedent probability and the efficiency of psychometric signs of patterns, or cutting scores. *Psychological Bulletin,* **52**, 194-215.

POLLARD, P., & EVANS, J. ST. B. T. (1983). The role of representativeness in statistical inference. In J. St. B. T. Evans (Ed.), *Thinking and reasoning* (pp. 309-330). London: Routledge & Kegan Paul.

ROTTENSTREICH, Y., & TVERSKY, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review,* **104**, 406-415.

SHERMAN, S. J., & CORTY, E. (1984). Cognitive heuristics. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 189-286). Hillsdale, NJ: Erlbaum.

SHERMAN, S. J., MCMULLEN, M. N., & GAVANSKI, I. (1992). Natural sample spaces and the inversion of conditional judgments. *Journal of Experimental Social Psychology,* **28**, 401-421.

SKOV, R. B., & SHERMAN, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology,* **22**, 93-121.

SLOWIACZEK, L. M., KLAYMAN, J., SHERMAN, S. J., & SKOV, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition,* **20**, 392-405.

THOMPSON, W. C., & SCHUMANN, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law & Human Behavior,* **11**, 167-187.

TVERSKY, A., & KAHNEMAN, D. (1980). Causal schemata in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49-72). Hillsdale, NJ: Erlbaum.

TVERSKY, A., & KOEHLER, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* **101**, 547-567.

WOLFE, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory account. *Journal of Behavioral Decision Making,* **8**, 85-108.

(Manuscript received January 23, 2001;
revision accepted for publication September 6, 2001.)