Do Subjects Discard Relevant Data? A Critical Test of Base Rate Neglect

Gernot D. Kleiter and Marianne Krebs
Department of Psychology, University of Salzburg
Hellbrunnerstr. 34, A-5020 Salzburg, Austria
Fax: +43-662-8044-5126, e-mail: gernot.kleiter@sbg.ac.at

Michael E. Doherty, Hugh Garavan, Randall Chadwick, Gregory Brake
Department of Psychology, Bowling Green State University, Ohio 43403, USA
Fax: +1-419-372-6013, e-mail: mdoher2@bgnet.bgsu.edu

Abstract

One of the most widely accepted findings of the heuristics-and-biases program is that people making probabilistic inferences are insufficiently sensitive to base rates. Recently, though, the proposition of base rate neglect has been questioned on empirical, methodological, and normative grounds. The present paper introduces a critical test of the hypothesis of base rate neglect. This method, which we will call the Partial Information Paradigm, has subjects select data relevant to, for example, diagnosis of a disease, D, based on a symptom, S. The question is whether subjects select those frequencies of cases for which information about the presence or absence of D is available, but for which information about the presence or absence of S is not. Such frequencies are relevant to the estimation of the base rate of D and to the probability of D given S. Four experiments ask subjects to select those frequencies relevant to diagnosis, one of which also had subjects select frequencies relevant to prediction of S from D. A fifth was concerned with inference of correlation. Very few subjects selected only the normatively correct information. Experiment 6 simplified the task, using a binary response mode in a situation in which normative considerations dictate that all subjects ought to select the same frequencies. In that study, the hypothesis of random behavior could not be rejected; subjects were no more likely to select the correct frequencies than the incorrect ones. These results, based on a frequency data format and qualitative dependent measures, strongly support the conclusion that subjects are insufficiently attentive to base rates.

A major thrust in the research on judgment under uncertainty is the heuristics-and-biases program (Kahneman, Slovic, & Tversky, 1982). According to this program, non-deductive reasoning is guided by heuristic principles that may lead to violations of the principles of rationality (Kahneman & Tversky, 1972, 1973, in press; Tversky & Kahneman, 1974; but see Hammond, in press). This program has been challenged by a second program that argues that such violations can be made to disappear if probabilistic information is expressed in terms of frequencies rather than probabilities, and maintains that reasoning is guided by rational principles that are fooled by tricky tasks (Gigerenzer, 1991; Gigerenzer, in press). There are three major empirical domains in which

the two approaches conflict: (a) the conjunction fallacy, which entails violation of the extensionality principle, (b) calibration research, in which people are shown to be overconfident, and (c) research on the base rate fallacy in probabilistic classification tasks. It is the last of these which is the concern of this paper. Classification of objects is a basic capability of our cognitive system. Often, classifications must be made in the presence of uncertainty, as when properties of the classes and relationships between classes and objects are known only imperfectly. To be optimal, inferences about class membership should be sensitive to all relevant information. Bayes' theorem is a normative standard for classification under uncertainty:

$$P(D|S) = \frac{P(D)\,P(S|D)}{P(D)P(S|D) + P(\neg D)P(S|\neg D)}\,. \tag{1}$$

Or, in words related to diagnostic inference, the posterior probability of the disease, D, given the symptom, S, is equal to the product of the prior probability and the likelihood divided by a normalizing sum in the denominator. In the research on judgment under uncertainty, subjects' judgments of posterior probabilities are typically compared with those calculated via Bayes' theorem. Much early research, for example, had subjects update probabilities sequentially in the light of new information. Conservatism was one of the effects found with the iterative application of Bayes' theorem; subjects' revisions were not as large as they should be, which prompted Edwards, Phillips, Hays, & Goodman (1968) to conclude that the subjects do not combine priors and likelihoods properly. They proposed (in those early days of artificial intelligence) that experts should estimate the priors and the likelihoods, but that integration thereof should be carried out not by human judgment, but by computers programmed to apply Bayes' theorem.

A scant four years later, Kahneman and Tversky (1972) proposed that the problem was not merely a quantitative bias that required a simple parameter adjustment; it was rather a deeper, qualitative difficulty that amounted to misspecification of the model. They first investigated the impact of base rates upon the judgment of classification probabilities in a classical series of experiments using tasks now known widely as the Lawyer-Engineer, Taxi Cab, and Tom W. problems (Kahneman, Slovic & Tversky, 1982). Based on these studies, Kahneman and Tversky concluded that base rates are largely ignored. The problem was not, they argued, that people have problems integrating general (base rate) and specific (individuating) information; people rely instead on the heuristic of representativeness. The neglect of base rates came to be one of the most widely accepted phenomena of the heuristics and biases approach, but, as noted above, it has now come under fire.

Gigerenzer (1991), in a critique of the heuristics and biases program, argued that uncertainty is cognitively encoded in terms of frequencies, not probabilities. Thus, if a task is presented in a probability format, subjects have difficulty because the presentation format is not compatible with the strategies people use to encode frequencies. He observed that presenting tasks in the frequency format generally leads to better performance. In a series of experiments, Cosmides and Tooby (1992) found the same result, and similar observations were made by Tversky and Kahneman (1983) with regard to the conjunction fallacy.

Koehler (in press), in a review of base rate research that included results obtained in replication studies on the classic base rate tasks, noted that the findings are not consistent; in some studies judgments were found to be sensitive to base rates, in others not. The thrust of Koehler's argument is that people do not completely neglect base rates, but that base rates often do not have the degree of impact on people's judged posterior probabilities demanded by the normative model. In a similar vein, Lynch and Ofir (1989) and Ofir (1988) proposed that base rates and likelihoods are utilized as cues similar to those in the non-Bayesian regression paradigm (Cooksey,1996; Slovic & Lichtenstein, 1971). For further reviews see Bar-Hillel (1980, 1982, 1990) and Fischhoff & Bar-Hillel (1984).

Some Notation and the Distinction Between Probabilities and Frequencies

In real life problems, probabilities must be estimated from frequencies, the accuracy of the estimates depending on the sizes of the samples from which the frequency counts were obtained. If the samples are small the resulting estimates are imprecise and accompanied by large confidence intervals; if the samples are large the estimates are precise and the confidence intervals small. We will express the information about an imprecise probability by a second order probability density function. In order to make the discussion concrete, the remainder of the paper is framed in terms of a simple medical diagnosis problem.

For a typical medical diagnosis task we introduce the following notation (Figure 1, right panel): Let τ denote the prior probability that a person is suffering from disease D (call it Tanner's syndrome), π_1 the conditional probability that a person suffering from D is showing symptom S (call it presence of the Beta protein), and π_2 be the conditional probability that a person not suffering from the disease (abbreviated by $\neg D$) is showing the symptom. These probabilities are treated as not directly observable, uncertain quantities, or random variables, and are estimated by frequency counts in a sample of cases, of which n_1 suffer from Tanner's syndrome and n_2 do not. Assume further that of the n_1 subjects suffering from Tanner's syndrome, a show the beta protein and c do not. Of those not suffering from Tanner's syndrome, b show the symptom and d do not.

Prior and Posterior Beta Distributions

Let us assume that we know very little about the uncertain quantities before we observe the actual frequencies. We express this vagueness by a relatively flat prior distributions. We could do this by using a uniform distribution, most conveniently by using a beta distribution, with the two shape parameters equal to one, i.e., $Be(1,1)$. For the ease of presentation, though, we use the improper beta distributions $Be(0,0)$, instead. This does not change any of the following arguments but keeps some expressions simpler.

After having observed the frequencies a, b, c, and d, (the notation for the cells is shown in

Figure 1: Structure of an elementary probabilistic classification; D=disease present, $\neg D$=disease absent, S = symptom present, $\neg S$ = symptom absent. On the left hand: numerical example in which more information is available about the base rate of the disease than about the likelihoods; in the middle: frequency notation; on the right hand: parameter notation.

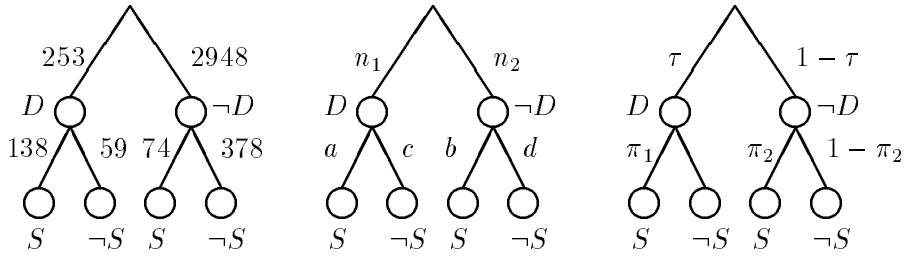

Table 1), thereby observing n_1 and n_2, we update the priors and obtain posterior distributions for each of the three probabilities τ, π_1 and π_2. We assume that the data generating processes for the frequencies are Bernoulli processes. From elementary (Bayesian) statistics (Bernardo, 1994; Kleiter, 1981) we know that the posterior distributions are beta distributions, given by:

$$\tau \sim Be(n_1, n_2)\,, \quad \pi_1 \sim Be(a, c)\,, \quad \text{and} \quad \pi_2 \sim Be(b, d)\,. \tag{2}$$

Table 1: Notation for the information presented to subjects. Call cells a, b, c, and d the core. For diagnostic inference, the normative cells are the core plus e and f, whereas cells g, h and z are irrelevant.

Symptom	Disease		
	D	$\neg D$	$?D$
S	a	b	g
$\neg S$	c	d	h
$?S$	e	f	z

The conceptual distinction between probabilities and frequencies is, of course, important. Frequencies are used to update our knowledge about probabilities. Probabilities are in turn used to predict frequencies. Clearly, frequencies and probabilities are not the same. This distinction is not made explicit in some frequentistic models of uncertainty judgment (Gigerenzer, 1991; Cosmides & Tooby, 1992).

The problem in a medical diagnosis task is to determine the probability that a patient who shows some symptom is suffering from a disease. We denote this posterior probability by μ. If τ, π_1 and π_2 are known exactly, then μ is determined precisely by Bayes's theorem:

$$\mu = \frac{\tau\pi_1}{\tau\pi_1 + (1-\tau)\pi_2}\,. \tag{3}$$

If the point values of τ, π_1 and π_2 are known only up to a second order probability distribution, then μ is known only up to a second order distribution. The probability distribution of μ is a function of the three quantities, τ, π_1 and π_2. The problem is to obtain the distribution of μ from the distributions of τ, π_1 and π_2 (Kleiter, 1992).

Natural Sampling

It may be shown that under the assumptions made so far the probability distribution that a patient showing the symptom S is suffering from D is given by

$$\mu \sim Be(a, b)\,. \tag{4}$$

The proof of the result is given in (Kleiter, 1994). This result may be startling at first: all probabilistic information about μ is contained in the frequencies a and b. The base rate determined by n_1 and n_2 is normatively irrelevant. Let us return to the example. We assumed that we have one large sample of n observations, consisting of two subgroups with n_1 and n_2 cases. The n_1 cases split into a and c observations, and the n_2 cases into b and d observations. No cases were incomplete. Thus $n_1 = a+c$, $n_2 = b+d$, and $n = n_1 + n_2$. We call this condition 'natural sampling,' which means simply that all cases that have been sampled have information about both D and S. In the natural sampling condition, the frequencies corresponding to the base rates are irrelevant.

The result is interesting for the investigation of base rates. It shows that base-rate neglect can be rational. Thus, neglecting the base rates is not necessarily a non-optimal heuristic that sometimes leads to good judgments and sometimes to clear errors. Under specifiable circumstances, base-rate neglect is an optimal strategy. How often it may be optimal depends upon the prevalence of the natural sampling condition, which calls to mind Brunswik's (1956) generally unheeded call for ecological surveys. Of course, even in a natural sampling condition, the actual judgment is correct if and only if a and b are combined properly. The point estimate of the subjects should be close to the mean of the beta distribution $Be(a, b)$ which is $a/(a+b)$.

Natural Sampling Cannot Provide a Proper Test Bed For the Hypothesis of Base Rate Neglect

The mathematical result shows that the experimental investigation of base rates under natural sampling conditions is insufficient as a critical test for the optimality of the judgments. An experiment by Christensen-Szalanski and Bushyhead (1981, experiment 2) illustrates the problem. In

that paper, they use what we are calling in this paper natural sampling, and even show statistically in their equation 3 that all of the information necessary for the computation of the posterior probability is in cells a and b, yet they conclude that their data show 'that physician's do use base rate information.' Situations in which base rates are normatively irrelevant simply do not warrant good agreement between calculations from Bayes' theorem and subjects' judgments as support for the proposition of base rate sensitivity; the experimental task *must* violate natural sampling. The completely additive cardinality decomposition of the reference set is violated if some data is missing or additional data are available. The critical point is that the additivity between the cardinalities in the super - and subclasses are broken apart. A person who is properly sensitive to base rates uses *all* available base rate information, that is relevant to the inference to be drawn, which may include data for which no likelihood information is available. A subject who neglects such information or who explicitly declares it to be irrelevant is insensitive to base rates. In a nutshell, the critical condition for an understanding of the role of base rates consists in asking the subjects whether they consider the partial information providing information on the base rates only, but not about the likelihoods, as relevant or not.

Partial Information as the Critical Test Condition

Given that all of the information is from frequentistic data, the critical condition for testing base rate sensitivity is the non-natural sampling condition. Sampling is non-natural if the frequencies of the disease and the symptom are not additive. We may know more about the base rates than about the conditional symptom probabilities, as reflected in the left panel of Figure 1. For example, we may have more knowledge about the epidemiology of a disease than about the likelihoods of the symptoms. That case would occur when we have observed cases for which we know the presence or absence of the disease but not the presence or absence of the symptom. Such a situation could be framed as additional information about the base rates or missing information about the conditional symptom probabilities, which are mathematically (but probably not psychologically) equivalent. We will refer to such conditions under the umbrella term 'partial information.'

On one hand, it may be shown that cases for which the presence or absence of the disease is known but information concerning the presence or absence of the symptom is unknown are relevant for the diagnosis. We denote a case in which information that the disease is present or absent but the symptom information is missing by $D\ ?S$ and $\neg D\ ?S$, respectively. The frequencies of such cases are essential to proper estimation of the base rate, and therefore relevant for the distribution of μ. On the other hand, cases with missing disease information are irrelevant for diagnosis. That is the frequencies of $?D\ S$ and $?D\ \neg S$ cases are irrelevant for the distribution of μ (the proof is given in the appendix).

The Partial Information Paradigm

In the present paper we propose a new method investigating sensitivity to base rates. Subjects are asked to select frequentistic case information that they consider relevant to inference; the key question is whether or not they select normatively relevant frequencies and discard the irrelevant ones. For some of the cases data are incomplete. If the subjects are truly sensitive to base rates they should select one type of partial information as relevant, and discard another type as irrelevant. Consider the following hypothetical medical diagnosis problem, which is a prototype of the ones that our subjects were presented in the experiments described below.

> Imagine that you have been hired by a large, urban medical center for the summer, and you have been given the task of organizing certain of their records. The medical center has been doing blood workups on patients being examined for a particular heart disease for a number of years, but the records have been very poorly managed; they are in disarray and many are incomplete. Before giving you the job of organizing the records, your supervisors have scanned some of them, and they suspect that there may be an important relationship between a blood protein that had been thought irrelevant to the heart disease in question. We'll call it the Beta protein. They think that the presence or absence of the Beta protein might be useful for diagnosing Tanner's syndrome. Some of the records are incomplete because information had been taken from files but not returned. Some are incomplete because the full blood workup was not done, others because the patient was referred elsewhere before the diagnosis was completed, or because the patient simply did not come back before all tests were completed. Others are incomplete in that they contain neither the presence or absence of the Beta Protein nor the diagnosis. The diagnoses, when made, can be considered highly reliable and valid, and were made independently of the presence or absence of the Beta protein.
>
> A summary list of every category of case that is in the files is provided on the form called CASE SUMMARY. We would like you to indicate whether the various types of cases shown on the CASE SUMMARY form are relevant to the diagnosis of this heart disease, and to rate the importance of those you indicate as relevant ...

Compare this with the taxicab problem; both deal with inference in the face of uncertainty; both ask the subject to combine a base rate with a likelihood; both ask the subject to do so without the aid of tools, etc. But there are fundamental differences; in the taxicab problem and many other base rate problems, the base rate is given as a percentage obtained from another source, and the subject draws a probabilistic inference that requires implicit integration of the information. In the problem investigated in this paper, on the other hand, the base rate information is frequentistic, from the same source and essentially of the same type as the information that constitutes the likelihoods, and the subject selects the information that would be needed to draw the appropriate inference.

This is, to the best of our knowledge, a completely new method of studying sensitivity to base rate information, and we assert that it provides a critical test whether subjects have a useful degree of insight into the implications of the base rate.

Analysis of Subjects' Judgments in the Partial Information Paradigm

In Experiments 1 through 5, subjects are presented nine different Information Types. There are three task types; data selection for diagnosis, for prediction and for the inference of correlation. If the subjects are sensitive to the critical information and to the critical information only, they should select the core information (a, b, c, and d) for all tasks. They should also select e and f (but not g, h, or z) for diagnosing D from S, g and h (but not e, f, or z) for predicting S for D, and e, f, g, and h (but not z) for a correlation task. (The notation is as in Table 1.) If in a diagnosis task the subjects are discarding e and f as irrelevant, then they are showing base rate neglect. If they select e and f but also select g and h, then we would conclude that they are selecting information indiscriminately rather than than showing base rate sensitivity. We call the cells a, b, c and d the core, and the six cells a through f the normative cells for diagnostic inference.

Consider first a subject who considers the core counts only, but who applies the normative integration algorithm to the frequency data of Table 2. Such a subject would obtain the following three first order probabilities for the disease, given the symptom is present or absent, respectively:

$$\begin{aligned} E(\mu|S) &= (.30 \times .70)/(.30 \times .70 + .70 \times .16) = 138/(138+74) = .651 \\ E(\mu|\neg S) &= (.30 \times .30)/(.30 \times .30 + .70 \times .84) = 59/(59+378) = .135 \end{aligned}$$

If the symptom is unknown the probability estimate for the disease to be present is obtained by the base rate estimate $E(\mu) = 197/(197+452) = .304$.

Table 2: Frequency information presented to subjects in experiments 2 and 5. The cells are the frequencies of the various Information Types.

Symptom	Disease		
	D	$\neg D$	$?D$
S	138	74	314
$\neg S$	59	378	1672
$?S$	56	2496	2812

The second order distributions are $p(\mu|S) \sim Be(138, 74)$, $p(\mu|\neg S) \sim Be(59, 378)$, and $p(\mu) \sim Be(197, 452)$, respectively. Note that the parameters of the three beta distributions can be read

directly from the core frequencies. Consider next a perfectly normative subject, who considers the frequencies contained in all of the normative cells of Table 2 and integrates them properly. Such a subject would obtain the following three first order probabilities for the disease given the symptom is present or absent, respectively:

$$\begin{aligned} E(\mu|S) &= (.08 \times .70)/(.08X.70 + .92 \times .16) = .27 \\ E(\mu|\neg S) &= (.08 \times .30)/(.08 \times .30 + .92 \times .84) = .03 \end{aligned}$$

The base rate estimate is $E(\mu) = 197/(253 + 2948) = .08$.

The second order distributions are $p(\mu|S) \sim Be(76, 207)$, $p(\mu|\neg S) \sim Be(62, 2027)$, and $p(\mu) \sim Be(253, 2948)$, respectively. Note that the parameters of the first two beta distributions were calculated by an approximation developed in Kleiter (1992). Only the third (marginal) distribution can be read directly from the sums of the appropriate cell counts, since it is only for the marginals that we have complete data.

Implications of Subjects' Selection Patterns for the Hypothesis of Base Rate Neglect

Note that the cells with $?D$ information, namely the g, h and z cells, are normatively irrelevant. These cases give additional information about the base rate of the symptom. However, this information is worthless for diagnostic inference. The neglect of the symptom base rate is normative. We give two explanations, an intuitive one in the following paragraph, and a formal one in the appendix which proves that the additional symptom information leaves the first and the beta distributed second order probabilities unchanged.

Here is the intuitive argument. Sure knowledge cannot be improved by probabilistic knowledge. The situation may be compared with the toss of a coin: if I already know the outcome of the toss, I cannot improve my knowledge by getting information about the probability of that event. We are considering a diagnosis *given* the presence of S. Given that it is present, information about its prior probability can therefore not improve our knowledge about S.

In the following series of experiments, we ask subjects to select the frequencies relevant to either diagnosis of D from S, prediction of S from D, or the correlation between D and S. Data selection is used as the primary measure of sensitivity to the importance of base rates because, unlike judged posterior probability, data selection requires only a qualitative understanding. Frequencies are used rather than probabilities in light of Gigerenzer's criticism that uncertainty is coded in terms of the frequencies, and not probabilities. Furthermore, frequencies are patently easier to understand, since no question arises of what is being conditioned on what. A variety of subjects, both in terms of educational level and nationality, are used to support generalization across subject populations, and a variety of tasks are used to support modest generalization across tasks. An important feature of the task designs, in addition to the critical inclusion of partial but relevant information described

above, is that some partial but irrelevant information is also included. The inclusion of irrelevant information provides a kind of internal control, or a baseline against which the selection of relevant information can be assessed.

General Method - Experiments 1 through 5

Each subject received a booklet that constituted the experimental task. The booklet was printed in English for the American students and in German for those in Austria. First there was a brief introduction and a section asking for demographic information. Then followed a description of the study as involving diagnostic reasoning, and an indication that the subject's task was to organize the records of a large number of case records that were in disarray, many of which were incomplete, with the goal of using the data to construct a diagnostic system.

The experimental task was similar to our introductory example. The cover story indicated that there was a suspected relationship between 'Tanner's syndrome' and 'Beta protein,' explained why many records were incomplete, and noted that the diagnosis of Tanner's syndrome, where the diagnosis was available, was reliable and valid and had been made independently of the presence or absence of the Beta protein. Next, the subjects were told that the various types of records were shown on the next page, called CASE SUMMARY, that they were to mark which types were relevant to the diagnosis of Tanner's syndrome and to rate the importance of each type of case record, which we will refer to as Information Types. Finally they were to describe how such case data ought to be used to make diagnoses. Most of the subjects were shown frequencies of the various Information Types on the CASE SUMMARY page. Those subjects were also asked to make diagnoses of three new cases, as described below.

The CASE SUMMARY page listed the nine possible outcomes of the 3×3 table formed by crossing the categories of disease present / absent / unknown with the categories of symptom present / absent / unknown (see Table 3). Some of the subjects had a text-only version of the CASE SUMMARY page; there was no column headed Number of cases. For those subjects who had the numeric version, the next page had three test cases, one with the Beta protein present, one with it absent and one with it unknown. They were asked to check whether they thought the patient had Tanner's syndrome, and to state the probability thereof. All subjects had a page asking for a suggested diagnostic strategy, and a page with NOTES at the top and an expression of thanks for their cooperation at the bottom, but that was otherwise blank.

Table 3: A sample CASE SUMMARY page

Categorization of the 1,400 cases found in the records showed that there were 9 types of records. Please note with a check mark in the appropriate circle whether each type of case is relevant to the diagnosis. If you check that a case type is relevant, please place a rating from 1 to 7 on the rating line next to the circle. Let 1 mean 'Slightly Relevant' and 7 mean 'Extremely Relevant.' If you check that a case type is not relevant, just leave the rating line blank.

UNKNOWN means that the data are missing.

Number of cases	Tanner's syndrome present?	Beta protein present?	Relevant (checkone) YES	NO	Rating
120	YES	YES	◯	◯	___
80	YES	NO	◯	◯	___
50	YES	UNKNOWN	◯	◯	___
100	NO	YES	◯	◯	___
350	NO	NO	◯	◯	___
300	NO	UNKNOWN	◯	◯	___
300	UNKNOWN	YES	◯	◯	___
50	UNKNOWN	NO	◯	◯	___
50	UNKNOWN	UNKNOWN	◯	◯	___

Experiment 1

Method

Subjects

The faculty and graduate students in the Psychology Department at Bowling Green State University were asked by intra-departmental mail to serve as participants.

Materials

Six different forms of the numerical booklet were used. In one, the order of the Information Types on the CASE SUMMARY page was as in Table 3, and the frequencies were those shown in Table 4. In the other five, the order of the Information Types was randomized, but the numerical values associated with a given type was held constant. For example, no matter where in the column of Information Types D S occurred, the frequency associated with it was 120. For each of the orders just described, there was a corresponding non-numeric, or text only, version. The two forms were identical except that in the non-numeric version there was no column labeled Number of Cases, and no posterior probability judgments were called for. One sixth as many non-numeric forms as numerical forms were distributed for two reasons: (1) a major concern of the investigation was what subjects would do with the numerical values, and (2) nonoptimal behavior in the non-numeric version would admit of alternative explanations in terms of what assumptions subjects might be making about the distributions of missing information.

Table 4: Frequency information presented to subjects in experiments 1, 3, and 4. The cells are the frequencies of the various Information Types.

Symptom	Disease		
	D	$\neg D$	$?D$
S	120	100	300
$\neg S$	80	350	50
$?S$	50	300	50

Task structure

The frequencies were selected so that the missing information was crucial to the judgment to be made. Note that the frequencies in Table 4 yield $E(\tau) = .25$, $E(\pi_1) = .60$, and $E(\pi_2) = .22$. By Bayes' theorem we get $E(\mu|S) = .47$, which means that the presence of a symptom would

slightly contraindicate the presence of the disease. For a subject who assumes that missing data are irrelevant, however, $E(\tau) = .31$, and by application of Bayes' theorem to the inappropriate data set we get $E(\mu|S) = .55$. Hence the presence of the symptom would lead to the incorrect judgment that the disease is probably present. Thus, the dependent variable calling for the diagnosis of the three new cases provides another indication of whether subjects are attending to the base rates of the disease, with a YES response to the question of whether the disease is present, given the symptom, providing an additional, though weaker, source of evidence concerning attention to the base rate. Of course, in making an actual diagnostic decision in which there are considerations of costs and payoffs, it is the actual value of the posterior probability that is involved, not whether it is greater or less than .50.

Procedure

All faculty and graduate students, except those who were involved in the research and those who had served as pilot subjects, had booklets placed in their departmental mailboxes with a request to return it to an appropriately labeled box in the mail room. Reminders were distributed. Of 30 faculty who were provided forms, 4 completed them. Of about 100 graduate students, 28 completed them, of whom two were dropped for giving evidence of misunderstanding the task. Needless to say, we were disappointed in the response rate. There are a number of possible reasons for such a low rate. One is that the booklets were distributed very early in the semester, which is a hectic time. Another is that the potential respondents know that several of the authors study judgmental biases, knowledge that would exacerbate the concern that people have about having their professional expertise evaluated, even though all responses were anonymous.

Results

Because of the small number of respondents, the data of all subjects and the various forms will be considered together. There might be sequence effects due to the order of the Information Types, but order was varied precisely to enable generalization of results beyond a single order. The overall frequency and relevance data are presented in Tables 5 and 6. Recall that normative considerations require that we know the likelihood ratio and the base rate of the disease. Hence the normative response is to select the first six rows of Table 3 and only those first six rows. (Note the importance of multiple orders.) Only three subjects chose the appropriate pattern of data.

A total of 24 subjects received the page requesting them to make diagnoses of three new cases. Five had incoherent probability values, e.g., saying the person probably did not have D but having $P(D|S) > .50$, and two did not respond to it. Of the remaining 17 subjects, 13 responded that the person *did* have D, with a mean $P(D|S)$ value of .56 to the case with the symptom present. As would be expected given virtually any construal of how people process this task, a majority of people responded in the negative to the S case ($N = 14$) and to the $?S$ case ($N = 16$).

Table 5: Frequencies of data selections for experiments 1 through 5. The cell designations are as defined in Table 1, the bottom row is the total number of subjects in the condition designated.

	Experiment or experimental condition[a]									
	1	2T	2N	3T	3N	4Dx9	4Pr9	4Dx5	4Pr5	5
Cell										
a	25	56	58	29	29	43	36	-	-	113
b	24	49	40	20	22	36	35	-	-	86
c	26	54	54	23	26	38	33	-	-	73
d	19	37	35	15	22	27	25	-	-	77
e	12	29	41	20	15	21	18	28	38	62
f	10	11	19	10	9	10	8	13	15	32
g	9	26	37	20	13	21	14	27	34	64
h	8	16	22	7	7	10	8	15	17	31
z	3	5	12	8	1	7	2	3	5	24
N	26	59	59	29	29	44	37	42	51	117

[a]N denotes numeric version, T = text-only version, Dx = diagnosis, Pr = prediction and the numbers after Dx and Pr refer to the number of Information Types from which subjects selected.

Table 5 contains in the first column the cells *a* to *z* and in the second column the number of Ss in Experiment 1 who marked them as relevant. Table 6 contains in the second column the according mean relevance ratings. Note in both tables there is a tendency for the subjects in this investigation to favor the *D* present cells, that is cells *a* and *c*. This 'positivity bias' shows up in the data on diagnostic inference in the studies described below, and is typical in this and related literatures.

Discussion

These data provide preliminary evidence that people are not sufficiently sensitive to the implications of the base rate of the disease. Only three subjects made the optimal selections of data for the diagnostic system, and the diagnoses of the new cases appear to be based on the likelihood ratio alone, with the mean probability value corresponding closely to the optimal value, assuming natural sampling. This is not to say that the behavior was completely inappropriate, or irrational. These subjects are all highly educated, and had had one or more courses in statistics. Of the 28 subjects, five explicitly recognized the problem as a Bayesian one. Fully 13 drew contingency tables, 12 of whom included disease absent data. Two calculated 2×2 tests of independence. Others raised issues

Table 6: Mean relevance ratings for experiments 1 through 5. The cell designations are as defined in Table 1, the bottom row is the total number of subjects in the condition designated.

	Experiment or experimentalcondition[a]									
	1	2T	2N	3T	3N	4Dx9	4Pr9	4Dx5	4Pr5	5
Cell										
a	5.9	5.4	5.6	6.5	6.2	6.8	6.6	-	-	5.7
b	5.4	4.3	3.5	3.9	4.2	4.8	5.2	-	-	4.0
c	6.3	4.8	4.8	4.1	4.9	5.2	4.8	-	-	3.2
d	4.2	2.9	3.3	3.3	4.4	3.8	4.0	-	-	3.4
e	2.3	2.2	3.4	3.4	2.5	2.5	2.2	3.4	3.7	1.9
f	1.7	.6	1.6	1.2	1.5	.9	.5	1.9	1.6	.9
g	1.6	1.7	2.5	2.8	2.5	2.7	1.9	3.4	3.5	2.3
h	1.4	.8	1.6	1.4	1.2	1.5	1.0	1.8	1.8	.8
z	.5	.3	1.0	1.1	.3	.9	.3	.8	.7	.7
N	26	59	59	29	29	44	37	42	51	117

[a]N denotes numeric version, T = text only version, Dx = diagnosis, Pr = prediction and the numbers after Dx and Pr refer to the number of Information Types from which subjects selected.

of sampling error, the reliability of the tests or whether there was a causal link between the Beta protein and Tanner's syndrome. These latter responses are irrelevant, given the narrow constraints of the task, but they are reasonable. The point of this research is not so much the fallibility of people as the extraordinary difficulty of applying formal models that have come on the scene only in the relatively recent past. But these models are, in fact, the appropriate models for some situations, and reasonable people, especially professionals, have to make the recognition of the impact of base rates part of their thinking. Although the data selections and diagnoses may be reasonable products of intelligent deliberation, they are wrong.

It might be argued that the nonoptimal behavior demonstrated in Experiment 1 is the result of extensive training, a sort of trained incapacity that results from so much attention being paid to the kind of thinking that underlies the very useful tool embodied in contingency tables. That is, one might argue that people who are relatively naive with respect to the task might perform better. We turn our attention to a replication with students in the early phase of their education, students in Introductory Psychology.

Experiment 2

Method

Subjects

128 students of an introductory psychology lecture at the University of Salzburg participated in the study. Ten subjects were not included in the data analysis because of incomplete answers, an obvious lack of understanding, etc. Of the remaining 118 subjects 84 were female and 34 male. The mean age was 23 years.

Materials and procedure

There were eight versions of the booklet, with two levels each of three variables: (a) numeric vs. text-only, (b) the left/right order of the D and S column, and (c) order of Information Types. Of the usable booklets, 59 subjects had the text-only version and 59 subjects the numeric version; 57 had the D on the left and 61 the S on the left. For 56 subjects the order was as in the example given above, that is, a, c, e, b, d, f, g, h, and z, for 62 the order was reversed. The study was run during a class session and used as an example for a lecture on methodology.

Results

The column 2T of Table 5 contains the frequencies with which the Information Types a to z were marked as relevant in the text-only version. The column 2N contains the according frequencies for the numeric version. The mean relevance rating are given in Table 6. We observe a positivity bias favoring the selection and the relevance of cell a. Only few Ss select the double question mark cell z which, of course, contains irrelevant information. Cell f, though, which contains relevant information is selected by only thirty of the 118 Ss. The frequencies and mean relevance ratings for the individual cells are of limited value. It is important to investigate different selection *patterns*. There are $2^9 = 512$ different possible selection patterns which are categorized in Table 8. The categorization is performed in terms of the number of posterior probabilities that can be computed with the help of the selected information: Both $P(D|S)$ and $P(D|\neg S)$, one only, that is $P(D|S)$ or $P(D|\neg S)$, or neither. Both posterior probabilities can be computed if (a) the core (a, b, c, and d) is selected, (b) the core plus irrelevant information (g, h, or z), (c) the core plus irrelevant plus relevant information, and, finally, (d) if all normatively relevant information is selected (the core plus e and f). Only two Ss selected the normatively relevant pattern. Twentyfour Ss selected information from which neither posterior probability can be calculated. Normatively cells e and f should and cells g and h should not be selected. The data in Table 5 and 6 do not show such preferences.

Note that the frequencies and mean ratings (Tables 5 and 6) are very similar for the numerical and text-only versions. While the observations in the cells are neither independent nor normally distributed, we can still use correlation coefficients as a device to describe the similarities among

Table 7: Correlations among the dependent variables across experiments and conditions. The upper half matrix presents the correlations among the relevance ratings; the lower half matrix the correlations among the selection frequencies. Only conditions in which 9 observations were required are represented. Decimals omitted.

	Experiment or experimental condition[a]							
	1	2T	2N	3T	3N	4Dx9	4Pr9	5
1		97	91	84	95	94	96	88
2T	97		96	94	96	99	98	94
2N	86	94		97	96	96	93	90
3T	77	88	96		94	96	94	96
3N	96	98	93	87		97	97	96
4Dx9	96	99	94	91	97		99	97
4Pr9	98	99	90	85	97	99		97
5	86	93	90	92	94	95	93	

[a]N denotes numeric version, T = text only version, Dx = diagnosis, Pr = prediction and the numbers after Dx and Pr refer to the number of Information Types from which subjects selected.

conditions. Table 7 shows that the correlation between numbers of numeric and text-only subjects selecting Information Types across the nine cells is .94, and the correlation between their mean relevance ratings is .96. This suggests that subjects were not strongly influenced by the specific frequency values in this experiment, and the data were pooled in order to have a substantial number of subjects for the next breakdown. The response patterns were categorized to allow some insight into the degree to which individual subjects were selecting data that would allow them to draw appropriate diagnostic inferences. The questions of most interest here are the number of subjects who selected only the optimal Information Types and the number who selected those Information Types that would be sufficient for them to infer the posterior probabilities. Table 8 shows that while only 2 of 118 subjects made optimal selections of Information Types, 64 selected the data necessary and sufficient to calculate $P(D|S)$ and $P(D|\neg S)$. Ten of the 59 subjects in the text-only version *explicitly* excluded cases with unknown values in their written comments!

Discussion

There is little or no difference between the numeric and text versions with respect either to the Information Type selections or the relevance ratings. Even the high frequencies entered into the

Table 8: Patterns of individual responding in experiments 2, 3, 4 and 5, totaled over numeric and text-only conditions. Only conditions in which 9 observations were required are represented.

	Experiment or experimental condition [a]				
Pattern [b]	2	3	4Dx9	4Pr9	5
Both likelihoods					
core only	24	16	18	16	20
core + irrel	25	7	1	4	0
core + irrel + rel	13	5	5	2	23
core + rel only	2	0	2	0	5
$P(D\|S)$ only	24	21	10	10	23
$P(D\|\neg S)$ only	6	2	0	2	16
Neither likelihood	24	7	8	3	30

[a] Dx = diagnosis, Pr = prediction and the numbers after Dx and Pr refer to the number of Information Types from which subjects selected.

[b] There are 512 possible response patterns. The various possibilities are categorized in terms of the number of likelihoods that could be computed, broken down further within the category of 'Both likelihoods.'

critical e and f cells had no effect, as the correlation between the Information Type selections of Experiment 1 and the text-only condition of Experiment 2 is .97. Note that Experiments 1 and 2 differed in the ages, languages and educational levels of the subjects, the language of the booklets, the procedure in which they were run and the presence vs. absence of numerical information on the CASE SUMMARY page. Yet the pattern of Information Type selections was highly similar.

Experiment 3

Method

This experiment was run concurrently with Experiment 2, and provides a comparison with Experiment 1 within the same culture, but with different levels of age and education, as well as a comparison across cultures with subjects of similar levels of age and education.

Subjects

Students ($N = 58$) in a large class in Introductory Psychology at Bowling Green State University served as subjects for extra credit. About two thirds of the subjects were female.

Materials and procedure

The task was as in Experiment 1, but the instructions, originally written for faculty and graduate students, were simplified. The frequencies shown in Table 4 were used, as were the multiple forms of the booklet described in Experiment 1. The experiment was conducted in two groups outside of class times.

Results and Discussion

Of the 58 subjects, none made the normatively correct selections of Information Types. The results of Experiment 3 are consonant with those of the first two experiments; they show little insight into the value of the partial information as contributing to a more reliable estimate of the base rate. Tables 5, 6, 7, and 8 show that the American undergraduates behave much like American faculty and graduate students and Austrian undergraduates.

Experiment 4

One might argue that the task facing the subjects in the experiments thus far has been rather complex, even though no computation is entailed in the data selection dependent variable. Experiment 4 had two goals. One was to assess whether subjects would select data more in accord with the optimal model if the cognitive load was reduced. The reduction was accomplished by informing half the subjects that the core frequencies were relevant, checking those Information Types as relevant on the CASE SUMMARY page and assigning the four core frequencies maximum values of 7 on the relevance rating. This was intended to let subjects focus their attention completely on the partial information cells. The second goal was to provide a modest cross-task generalization. This was done by converting the task to a prediction task for half of the subjects, in that these subjects were given the task of selecting the data relevant to predicting the symptom from knowledge concerning the disease. The prediction of $P(S|D)$ is, of course, also a Bayesian problem, but for such predictions the relevant data are cells a, b, c, d, g and h; cells e, f and z are irrelevant. The intuitive explanation of this is the same as above; if you know the disease is present, then knowledge of how probable it is that the disease is present is manifestly irrelevant.

Method

Subjects

Students in an Introductory Psychology class at Bowling Green State University taught by the third author filled out the booklet during discussion sections, and received extra credit for doing so. The students had read text material prepared by the instructor and heard lectures on the logic of 2×2 tables on the import of base rates for judgment, and on Bayes' theorem. A total of 174 students provided usable data.

Materials

The materials were kept as similar as possible to those used in Experiments 1 and 3, save for the changes required by the modifications described above. The task described above was modified to create four versions. Two versions called for subjects to select the data relevant to diagnostic inference, as in the above experiments, and two called for subjects to select the data relevant to prediction of a symptom from knowledge of the disease. There were two forms of each type, one asking the subjects to select frequencies from all nine cells, as in the previous experiments, the other informing the subjects that the core was relevant and asking them to select which of the remaining five cells were also relevant. Relevance ratings were also asked for, as were three judgments of posterior probabilities on specific cases, as above. The three judgments were diagnoses or predictions, as appropriate. Let us call the condition calling for subjects to select the data relevant to diagnostic inference from all nine possible cells Dx9, and the others Dx5, Pr9 and Pr5.

Procedure

The subjects were run in eight discussion sections conducted by advanced graduate students. The eight class sections were randomly assigned to the four versions of the task, with the restriction that there be two sections per task. The booklets were handed out in the discussion sections, and any student wishing his or her data not be used in the analysis did the task and so noted on the form; one student did so. Students were given a number of points toward their grade for doing the task, plus bonus points for normatively correct selections, minus points for normatively incorrect selections. Multiple versions of each booklet type were used; all had numeric Information Types as in Table 4.

Results

The choice frequencies are presented in Table 5, and the relevance ratings in Table 6. Note that the frequencies and ratings are very similar for the diagnoses and predictions. The correlation between the frequencies for Dx9 and Pr9 across the nine cells is .95, while for Dx5 and Pr5 the correlation across the 5 cells is .99. If all subjects had chosen the normatively relevant cells and rejected the normatively irrelevant cells, these correlation coefficients would have been 0 and -.67, respectively. The correlations between the relevance ratings for Dx5 and Pr5 were .98 and .99, respectively.

Of the 42 subjects in the Dx5 condition none responded normatively correctly. Of the 51 subject in the Pr5 condition, two subjects responded normatively correctly. Table 8 shows that only two of of the Dx9 and none of the Pr9 subjects chose optimally. Again, however, a large majority of subjects did choose Information Types in such a way that $P(D|S)$ could be calculated correctly from the frequencies chosen, but they chose irrelevant information as well.

The frequencies of selections of the non-core cells tended to be much higher in the Dx5 and Pr5 conditions than their Dx9 and Pr9 counterparts. That is, if the core is designated as relevant,

more cases with only partial information are marked as relevant than when the core is not already so marked. However, selections of irrelevant cells increased as well as selections of relevant cells.

The correlations between the 4Dx9 and 4Pr9 selections and relevance ratings are both .99. In the diagnosis condition the subjects do not select the relevant disease (combined with missing symptom) more often then the irrelevant symptom (combined with the missing disease). Similarly, in the prediction condition the subject do not select the relevant symptom (combined with the missing disease) information more often then the irrelevant disease (combined with the missing symptom).

Discussion

As in the other experiments, we find no evidence for differential selection within the conditions. Providing the subjects with a partial solution of the task and telling them that the core is relevant was done to determine if focusing subjects' attention on the critical information, might enhance performance. It did not.

Experiment 5

This experiment is a departure from the others, in the sense that Experiments 1 through 4 tested subjects' sensitivity to base rates with respect to inferences of directed relations, either diagnoses or predictions. Experiment 5 deals with the selection of information relevant to the symmetric relation of correlation. For symmetric relations, both base rates are relevant. Hence, in the Partial Information Paradigm, partial information on both variables should be selected for optimal inference. (The proof is given in the appendix.)

Method

Subjects

A total of 119 students of the University of Salzburg took part in the experiment. The data of two Ss were excluded because they obviously had not understood the task properly. Of the remaining 117 Ss 50 were psychology students (13 male, 37 female) in the second year or later, and who had passed their basic statistics requirement. The other 67 students were from different fields, the majority from 'Publizistik' (mass communications), of whom 38 were male and 29 were female. The mean age of all subjects was 23.8 years.

Materials and procedure

To avoid any asymmetry in the explanation of the task two diseases (Tanner' syndrome and Brunner's disease) were used instead of one disease and one symptom. The subjects were asked to indicate which information they would select for making inferences about the correlation ('Zusammenhang') between the two diseases.

Eight different paper versions were used, varying the order of the diseases (Brunner-Tanner vs. Tanner-Brunner), the order of the check-list (forward vs. backward), and presence vs. absence of a question asking the subject to estimate the strength of the correlation on a graphical scale. This scale was like a ruler, but instead of numbers there were three anchors: on the left: Brunner's disease and Tanner's syndrome never occur together, in the middle: Brunner's disease and Tanner's syndrome occur equally often together and not together, and on the right: Brunner's disease and Tanner's syndrome always occur together). The task was run in several classroom settings and discussion groups.

Results

The frequencies of the various response patterns are contained in the last column of Table 8. Only five Ss selected the relevant and the relevant information pattern only (cells a to h). Twenty (17 %) selected the core only. The last column of Table 5 shows that cell a is chosen by nearly all the Ss, while the cells c and d by about 64 % only. The relevance ratings in Table 6 demonstrate the same tendency. The judgment of correlation shows a strong positivity bias. Relevant partial information is not chosen systematically and not rated as relevant. The positivity bias can also be traced in the higher relevance attached to cells e and g (1.9 and 2.3) as compared to cells f and h (.9 and .8).

Discussion

A correlation coefficient is a concise one-number description of the relationship between two variables. According to the economy of encoding it may be easier to express information about the co-occurrences of two variables by correlation and not by conditional probabilities. If this were true, subjects should have superior understanding of correlation. We observe, though, a very similar selection pattern as in the diagnosis and the prediction tasks. There is no increased sensitivity for the additional base rate information in the judgment of correlation task.

Some anecdotal information is relevant. On several occasions we have informally observed explicit denial of the relevance of partially observed data for the judgment of correlation. Students with a statistical education argued that in the definition of the Pearson correlation the covariance in the numerator is the decisive quantity that determines the correlation coefficient. To calculate the sum of cross-products, they noted, one needs pairwise values on both variables. In their arguments the students did not consider means and variances. The first author has a friend who has published a book in the field of mathematical probability theory, and who has denied the relevance of partial information from a maximum entropy point of view. There is a very strong intuition that to make inferences about the co-occurence of two variables, only information about the frequency of the various co-occurences of these variables is relevant. The proof in the appendix shows that that intuition is incorrect.

Experiment 6

This experiment introduces a major methodological change by using a more traditional dependent variable. One might argue that the complexity of the versions of the task described above makes it rather insensitive, and precludes us from assessing fairly whatever level of insight subjects do have into the importance of the correct base rate for diagnosis. Furthermore, the free choice of how many cells to check may have introduced some unusual biases, as in Experiment 4 in which the proportion checked was influenced by the number to be checked. Hence, in Experiment 6, we simplified the wording as much as possible, made the task structure more transparent by presenting the core data in a contingency table format (see Table 9), and used a two-choice, forced choice response mode.

Table 9: An example of the contingency table shown to subjects on the second page of the booklet used in experiment 6.

	Tanner's Syndrome present	**Tanner's Syndrome absent**	**Tanner's Syndrome unknown**
Beta protein present	**120 patients** had the disease and the symptom	**100 patients** did not have the disease and did have the symptom	**A1** had noinformation about the diseasebut did have the symptom
Beta protein absent	**80 patients** had the disease and did not have the symptom	**350 patients** did not have the disease and did not have the symptom	**A2** had noinformation about the diseasebut symptom
Beta protein unknown	**B1** had the disease but had no information about the symptom	**B2** did not have the disease but had no information about the symptom	

Experiment 6 also addresses other possible criticisms. The experiments described above all used the *presence* and *absence* of a disease as two alternatives. It is well known that in human judgment there is a bias to accentuate the positive (Horn, 1989). Subjects may pay more attention to, hence be more likely to select, features that are marked as present than to those features that are not present, or not known to be present. A related possibility comes from the fact that the partially observed data the missing parts were explicitly described as missing. One might argue that subjects assume that the quality of such data is generally inferior. Somehow, the missing part contaminates

the actually observed part of the partial information. Hence, Experiment 6 used two semantically more equally balanced categories, and required a forced choice between them.

Method

Subjects

There were four groups of subjects, all run at Bowling Green: (a) 122 students in an introductory psychology course taught by one of the authors, (b) 14 introductory statistics students who had completed a unit on probability, and (c) a second group of 14 introductory statistics students in a course taught by the third author, who had completed a unit on probability with some emphasis on Bayes' theorem, and (d) 21 graduate students in psychology, enrolled in a course in methodology. This third group all had had at least one undergraduate course in statistics, and were currently enrolled in a graduate statistics course, and all had been exposed to Bayes' theorem in the previous lecture. All of the undergraduates had been exposed to Bayes' theorem, except the second group.

Materials

The booklet was reduced to two pages. The first page had a somewhat shortened version of the cover story, then a prominent 2×2 matrix with the core data (i. e., the four upper left cells in Table 9) preceded by an italicized statement that read 'It can be shown mathematically that all four of these categories of cases are relevant to the diagnosis of Tanner's Syndrome from the Beta protein.' On the second page the subjects were told that there were two categories of partial information, A and B, and that 'It can be shown mathematically that only one of these two categories of cases is important in determining the diagnosis of Tanner's Syndrome from the Beta protein.' The subjects were then asked to select one of the two categories. Table 9 shows one version of the page 2 matrix. The D information was category A in half the booklets and category B in the other half, and the 3×3 matrix (less the $?D$ $?S$ cell) was oriented with the disease on top for half the booklets and on the side for half the booklets. Four forms of booklets resulted.

Procedure

All subjects were run in classroom settings. The 122 introductory psychology students were run in small discussion sections, and received credit toward a course research participation requirement, unless they chose not to have their data used for research (one did so). One graduate student marked directly on the page 2 matrix, crossing out cell e and selecting cell c. Perhaps he or she did not accept our assertions about what could be shown mathematically. Another had prior knowledge of the experiment, leaving 19 graduate students with usable data. The data were collected in the two classes not taught by the third author as the first part of a guest lecture on the application of Bayes' theorem.

Results

Of the 121 Introductory Psychology students who provided usable data, 65 selected the normatively correct category of frequencies, 56 the incorrect category. Of the 14 students in the second group, 5 selected the normatively correct category of frequencies, 9 the incorrect category. For group 3 the frequencies were 6 and 8, respectively, and for the 19 graduate students they were 8 and 11. The possibility of between group differences was assessed by a 4×2 χ^2 test of independence, which yielded $\chi^2(3) = 2.48$, $p > .40$. The frequencies of the four groups were therefore combined, yielding totals of 84 students who selected the normatively correct data and 84 who selected the incorrect data. For such data, of course, $\chi^2(1) = 0$.

Discussion

These data are in accord with those of Experiments 1 through 5; subjects' selection behavior show little insight into the relevance of the frequencies of D $?S$ and $\neg D$ $?S$ for the assessment of the probability of D given S or $\neg S$. We take this as rather direct evidence that subjects do not fully appreciate the import of base rate information for diagnostic inference.

General Discussion

Substantive Conclusions

The results of the above investigations are in agreement with one another. In each, the subjects failed to show sensitivity to the implications of the base rate for inference. This is in spite of the fact that there was no cost to obtaining the base rate information, save the possible cognitive costs of having to think about the implications of base rate information for inference. The consistency of the findings is remarkable, given variation in the frequencies that constituted the table entries, in the number of cells from which subjects selected, in specifics of the instructions, in the educational levels, ages, and native languages of the subjects, and in the nature of the response mode.

More specifically, the above experiments found that:

1. Partial information is generally considered irrelevant, whether the information selections are in the service of diagnoses, predictions or the estimation of correlation.

2. Subjects' usage of conditional probability is not directionally sensitive, their selections of irrelevant information were the same whether those selections were in the service of inferences or predictions.

3. There is a strong positivity bias in respect to the relevance of the four cells of a two-by- two table; subjects tended to select information given the presence of a disease, for example, than given its absence.

4. For the estimation of covariance, only complete data in both variables is seen as relevant; marginal probabilities are judged as irrelevant for the estimation of correlation.

The conclusion that subjects failed in these investigations to show sensitivity to the implications of the base rate for inference is not to be taken as a claim that subjects are completely insensitive to the base rate. Clearly, as the literature reviewed by Koehler (in press) shows, people's use of base rates has been shown to be influenced in the appropriate direction, though typically not to the appropriate degree, by a wide variety of experimental manipulations. We do believe, however, that the Partial Information Paradigm provides a critical test of whether people have a sufficient understanding of the implications of the base rate so that they will use it relatively spontaneously. They do not.

How serious is the bias for real world inference? This is a difficult question to answer, absent information about the prevalence of natural sampling vs. partial information in the ecology in which important inferences are drawn. Two kinds of biases may be distinguished: adaptive and non-adaptive biases. Adaptive biases are violations of normative principles which are adaptive in specific environmental conditions. That means that they are not real biases if the environmental conditions are modeled properly by the experimenter (Anderson, 1990, 1991; Billman & Heit, 1988; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Kareev, 1995; Kornblith, 1993; Simon, 1967). Non-adaptive biases are the 'real' biases, in the sense that such biases would have negative consequences for one's life. We see there no compelling reason to know whether base rate neglect (as found in the research reported above) or insufficient sensitivity to the base rate, again absent ecological information, is a 'real' bias or not. Nor do we know how often the core frequencies would be adequate estimates of the marginals, that is, how often the missing information would bear the same ratios as the available information. It is our intuition that there are domains in which the differences are quite likely to be radical. One such domain might be psychodiagnosis, in the practice of which a clinician's base rates might be distorted by the selected sample, a problem which could well be exacerbated by the 'softness' of the data and the complex utility considerations involved in the diagnosis of pathology.

We feel strongly that one should not dismiss this bias as a laboratory artifact, using the reasoning that people get along so well in the world (Cohen, 1981). People have many strange beliefs, and many people do not get along in the world very well at all.

Methodological Contribution

One of the main points in the introduction bears repeating. One cannot assess whether people are sensitive to base rates using an experimental design in which base rates are unnecessary. In these investigations we used a new method, in which the data were frequencies, the data were of the same sort as the data needed to calculate the likelihood ratios and could have been chosen as easily. They were not. But the essential element of the Partial Information Paradigm that makes this a methodological contribution to the literature is that it provides a critical test of the hypothesis of

base rate neglect; in a way that investigations using frequentistic data with natural sampling cannot.

Theoretical Issues

The conclusion that people are insensitive to base rates in the Partial Information Paradigm is a negative one; ideally we would have a positive explanation (Kahneman & Tversky, 1982), based perhaps on a protocol analysis (Ericsson & Simon, 1993) describing the process of selecting and discarding information. Alternatively, we might try to explain the results by a theory explaining effects in another information selection task. Relevance theory would be a good candidate. It has recently been used to model selections in Wason's four card problem (Sperber, Cara, and Girotto, 1995). Or we might remain within the heuristics and biases paradigm, and postulate that partially described cases are not considered representative of typical patients and therefore tend to be discarded.

However, if Evans' conception of heuristic, which refers to 'pre-attentive processes whose function is to select relevant information for analytic processing' (1984, p 452), is correct, then it will be very difficult to come up with a cogent explanation of why people do not attend to or select relevant information. We may simply have to limit our own attention to research aimed at discovering and elaborating the conditions under which people do attend to and use base rates, and those in which they do not. A similar conclusion and attribution to Evans was reached by Doherty, Chadwick, Garavan, Barr and Mynatt (in press), with respect to another controversial issue, people's sensitivity to the diagnostic implications of data. The psychological processes underlying inattention may be no more open to us than those underlying our difficulty with many other negative psychological events.

Conclusion

We have investigated information selection behavior in a probabilistic task with partial, frequentistic information. From a normative perspective information selection should be guided by the relevance axiom: a rational agent 'always takes into account all of the evidence it has to a question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains.' (Jaynes, under development, p. 114). The relevant evidence should lead one to update knowledge and beliefs according to the conditioning principle, and Bayes' theorem is an important rule through which the conditioning principle is implemented. We have investigated only a small section of a very large field. Which evidence do people select in a multivariate task containing partially overlapping data? What is the effect if the relationships are causal? Finally, we note that many psychologically stimulating questions arise from the rapidly developing field of Bayesian networks (Buntine, in press). It may be premature to engage in too many psychological speculations before more areas in the field of information selection are empirically investigated

References

Anderson, R. J. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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

Bar-Hillel, M. (1980). The base rate fallacy in probability judgments. *Acta Psychologica*, **44**, 211-213.

Bar-Hillel, M. (1982). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision Making under Uncertainty* (pp. 39-61). Amsterdam: Elsevier.

Bar-Hillel, M. (1990). Back to base rates. In R. M. Hogarth (Ed.), *Insights in Decision Making* (pp. 200-216). Chicago: University of Chicago Press.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory.* New York: Wiley.

Billman, D. O., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, **12**, 587-625.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley, CA: University of California Press.

Buntine, W. (in press). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering.*

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physician's use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 928-935.

Cohen, L. J., (1981). Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences*, **4**, 317-370.

Cooksey, R. (1996). *Judgment analysis: Theory, methods and applications.* San Diego: Academic Press.

Cosmides, L., Tooby, J. (1992). Are humans good intuitive statisticians after all? Department of Psychology, University of California, Santa Barbara, CA 93106.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society*, **B39**, 1-38.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (in press) On People's Understanding of the Diagnostic Implications of Probabilistic Data. *Memory and Cognition.*

Edwards, W., Phillips, L. D., Hays, W. L., & Goodman, B. C. (1968). Probabilistic information processing systems, design and evaluation. *IEEE Transactions on Systems Science and Cybernetics*, **4**, 248-265.

Ericsson, K. A., & Simon, H., A. (1993). *Protocol Analysis: Verbal Reports as Data.* (2nd ed.). Cambridge, Mass: MIT Press.

Evans, J. St B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, **75**, 451-468.

Fischhoff, B., Bar-Hillel, M. (1984). Diagnosticity and base-rate effect. *Memory and Cognition*, **12**, 402-410.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond 'heuristics and biases.' *European Review of Social Psychology*, **2**, 83-115.

Gigerenzer, G. (in press). On content-blind norms and vague heuristics: A rebuttal to Kahneman and Tversky. *Psychological Review.*

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models. A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

Hammond, K. R. (in press). *The competence of judgment: Coping with irreducible uncertainty, inevitable error, unavoidable injustice.* Oxford University Press.

Horn, L. R. (1989). *A Natural History of Negation.* University of Chicago Press, Chicago.

Jaynes, E. T. (under development). *Probability Theory: The Logic of Science.* Available: `ftp://bayes.wustl.edu/Jaynes.book`.

Kahneman, D., Slovic, P., Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-454.

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

Kahneman, D., & Tversky, A. (in press). On the reality of cognitive illusions: A reply to Gigerenzer's Critique. *Psychological Review.*

Kareev, Y. (1995). Positive bias in the perception of covariation. *Psychological Review*, **102**, 490-502.

Kleiter, G. D. (1981). *Bayes Statistik.* Berlin/New York: De Gruyter.

Kleiter, G. D. (1992). Bayesian diagnosis by expert systems. *Artificial Intelligence*, **54**, 1- 32.

Kleiter, G. D. (1994). Natural sampling: rationality without base rates. In G. H. Fischer, & D. Laming (Eds.) *Contributions to Mathematical Psychology, Psychometrics, and Methodology.* New York: Springer. 375-388.

Koehler, J. J. (in press). The base rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behavioral and Brain Sciences.*

Kornblith, H. (1993). *Inductive inference and its natural ground.* Cambridge, MA: MIT Press.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with Missing Data.* Wiley, NY.

Lynch, J. G., Jr., & Ofir, C. (1989). Effects of cue consistency and value on base-rate utilization. *Journal of Personality and Social Psychology*, **56**, 170-181.

Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes*, **42**, 343-363.

Simon, H. (1967). The logic of decision making. In N. Rescher (ed.), *The Logic of Decision and Action.* Pittsburgh: University of Pittsburgh Press, 1-20.

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

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, **57**, 31-95.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293-315.

Appendix

In this appendix we show which base rate frequencies are and which are not relevant for inferences in a 2×2 table containing partially observed data. For a more rigorous treatment of statistical analysis with missing data the reader is referred to Little & Rubin (1987) and to Dempster, Laird, & Rubin (1977).

The estimation of probabilities from partially observed data can be modeled by urn schemes. Imagine an urn containing balls on each of which a figure is painted. The *shape* of the figure is either a *diamond* or a *circle*, and its *color* is either *red* or *blue*. We denote the two binary variables by X and Y and encode the four possible (x, y) figures by $(1,1), (1,2), (2,1)$, and $(2,2)$, respectively. Balls are drawn randomly and with replacement. The frequencies of the four different outcomes are arranged in a two-way table. The outcomes contain partially observed data. The values of X or Y may be missing. Incomplete data is marked by a question mark in the appropriate position, $(x, ?)$ or $(?, y)$, respectively. The notation for the nine different frequencies is shown in Table 10. The total number of balls drawn is $N = a + b + c + d + e + f + g + h + z$, the number of complete data in the *core* is $n = a + b + c + d$.

Table 10: Frequency table containing the frequencies of fully (a, b, c, d) and partially (e, f, g, h, z) observed data.

		Y (Color)		
		1 (red)	2 (blue)	? (missing)
$X(Shape)$	1 (diamond)	a	b	g
	2 (circle)	c	d	h
	? (missing)	e	f	z

The urn scheme is used as a model of the diagnosis, the prediction, and the correlation problem. Each ball represents a patient. X corresponds to the Beta protein, Y to Tanner's disease, $x = 1$ denotes the absence of the symptom, $x = 2$ its presence, and so on.

Given partially observed data we want to estimate

1. the conditional probability of the disease Y given the symptom X

$$\pi_{y|x} = \frac{\pi_{xy}}{\pi_x} = \frac{\pi_x \pi_{y|x}}{\pi_{x=1}\pi_{y|x=1} + \pi_{x=2}\pi_{y|x=2}}, \tag{5}$$

2. the conditional probability of the symptom X given the disease Y

$$\pi_{x|y} = \frac{\pi_{xy}}{\pi_y} = \frac{\pi_y \pi_{x|y}}{\pi_{y=1}\pi_{x|y=1} + \pi_{y=2}\pi_{x|y=2}}, \tag{6}$$

3. and the 2×2 correlation between the symptom X and the disease Y,

$$\phi = \frac{\pi_{xy} - \pi_x \pi_y}{\sqrt{\pi_x \pi_y (1 - \pi_x)(1 - \pi_y)}} . \tag{7}$$

In the following sections we investigate which estimates are sensitive to which frequencies.

Sensitivity of probability estimates to partial information

The univariate case Consider an urn with an unknown composition π of red and blue balls. N balls are drawn independently and with replacement. We observe a red and b blue balls. The color of z balls is not registered. If $z = 0$, the maximum liklihood (ML) estimator of π is equal to the relative frequency, $\hat{\pi} = a/(a+b)$. If $z = 1$, there are two possibilities. The missing color may be either red or blue. Let i denote the number of red balls. The probability for the first possibility is $P(i = 1|a, b, z) = a/(a+b)$, and the ML estimator of the composition is $\hat{\pi}_1 = (a+1)/(a+b+1)$. The probability for the second possibility is $P(i = 0|a, b, z) = b/(a+b)$, and the ML estimator of the composition is $\hat{\pi}_2 = a/(a+b+1)$. The expected value of the two estimators is the weighted average:

$$\begin{aligned} E(\hat{\pi}|a, b, z) &= P(i = 0|a, b, z)\, \hat{\pi}_1 + P(i = 1|a, b, z)\, \hat{\pi}_2 \\ &= \frac{a}{(a+b)} \frac{(a+1)}{(a+b+1)} + \frac{b}{(a+b)} \frac{a}{(a+b+1)} \\ &= \frac{a}{(a+b)} \frac{(a+1+b)}{(a+b+1)} = \frac{a}{a+b} = E(\hat{\pi}|a, b) . \end{aligned} \tag{8}$$

The missing value has no influence on the estimator, and by induction, this holds for any $z \geq 1$. The probability weights $P(i|a, b, z)$ represent the so called 'predictive probabilities'. They assign a probability to each event in the sample space of missing data. They do this in the light of the sufficient statistics of the actually observed data a, b, and z. $\hat{\pi}_i$ is the ML estimator for the actually observed and the predicted data combined.

The bivariate case

Completely missing data Consider $e = f = g = h = 0$ and $z \geq 1$. The number of completely missing data z is irrelevant for the estimation of the joint, the marginal, and the conditional probabilities. Consider the joint probability π_{xy}. The probability of the remaining three cells is $1 - \pi_{xy}$. The case reduces to the univariate case treated in the previous section. The same holds for the marginal probabilities. The correlation ϕ is a function of the joint probabilities. As each of these probabilities is insensitive to z, the correlation is also insensitive to z.

The selective sensitivity of the conditional probabilities to base rates Consider the estimation of the conditional probability $\pi_{y=1|x=1}$. Without any partially observed data the ML estimator is $\hat{\pi}_{y=1|x=1} = a/(a+b)$. Assume that partially observed data about one more ball with a diamond figure but with an unknown color becomes available. We have $g = 1$ and $e = f = h = 0$. The conditional probability behaves completely analog to the unconditional probability in the univariate case. Conditioning with respect to the X value $x = 1$ in Table 10 restricts the problem to the first line, a and b are equivalent to a and b in the univariate case, and g is analog to z. Supplemental $(x, ?)$ observations are irrelevant in respect to $\hat{\pi}_{y|x}$.

Consider now the case where $e = 1$ and $g = h = f = 0$, that is the case of a supplemental $(?, y)$ observation. We have

$$\begin{aligned} E(\hat{\pi}_{y|x}|a,b,c,e) &= P(k=0|a,b,c,e)\,\hat{\pi}_{y|x}^{(k=0)} + P(k=1|a,b,c,e)\,\hat{\pi}_{y|x}^{(k=1)} \\ &= \frac{c}{(a+c+1)}\frac{a}{(a+b)} + \frac{a}{(a+c+1)}\frac{a+1}{(a+b+1)} \neq \frac{a}{a+b}\,. \end{aligned} \tag{9}$$

and thus $E(\hat{\pi}_{y|x}|a,b,c,e) \neq E(\hat{\pi}_{y|x}|a,b,c)$. The partially observed data $(?, y)$ is relevant for the estimation of the conditional probability of Y given X. The information improves our knowledge about the *base rate* of Y. We thus should be sensitive to the partially observed information. More generally, we have a two-way contingency table with one supplemental one-way margin (Little and Rubin, 1977, p. 173). For the joint probabiliy in cell $(1,1)$ we obtain the ML estimate

$$\hat{\pi}_{x=1,y=1} = \frac{a + e\frac{a}{a+c}}{a+b+c+d+e+f} \tag{10}$$

and analog for the other three cells. The estimate of the marginal probability is $\hat{\pi}_{x=1} = a/(a+b+c+d)$. The conditional probability of a red ball $(y = 1)$ given the ball has a diamond $(x = 1)$ finally is $\hat{\pi}_{y=1|x=1} = \hat{\pi}_{x=1,y=1}/\hat{\pi}_{x=1}$.

The sensitivity of the correlation to both kinds of partial information We show that incomplete data in each of the two variables is relevant in respect to the estimation of the ϕ correlation. It is sufficient to show that the estimators of the joint probabilities π_{ij} are sensitive to partially observed data. The case is more complex than the previous ones because each estimator depends on the frequency of all four incomplete data and on the *order* in which they can be obtained.

We investigate $\pi_{x=1,y=1}$. If $e = f = g = h = 0$, the ML estimator is $\hat{\pi}_{x=1,y=1} = a/(a+b+c+d)$. If $e = g = 1$ and $f = h = 0$, then none $(i = 0)$, one $(i = 1)$, or two balls $(i = 2)$ may belong to cell $(1,1)$. We have

$$E(\hat{\pi}_{x=1,y=1}|a,b,c,d,e,f) = \sum_{i=0}^{e+g} P(i|a,b,c,d,e,f)\frac{a+i}{a+b+c+d+e+f}\,, \tag{11}$$

and the probabilities are

$$P(i=0|a,b,c,d,e,f) = \frac{b}{(a+b)}\frac{c}{(a+c)}\,, \tag{12}$$

$$
\begin{aligned}
P(i=1|a,b,c,d,e,f) &= \frac{a}{(a+b)}\frac{c}{(a+c+1)} + \frac{c}{(a+c)}\frac{a}{(a+b+1)}\,, \\
P(i=2|a,b,c,d,e,f) &= \frac{a}{(a+b)}\frac{(a+1)}{(a+c+1)} + \frac{a}{(a+c)}\frac{(a+1)}{(a+b+1)}\,.
\end{aligned}
$$

There is only one way to obtain $i = 0$: the g-data goes to cell $(1, 2)$ and the e-data goes to cell $(2, 1)$. The probability for $i = 0$ is the product of $b/(a + b)$ and $c/(a + c)$. There are two ways how to obtain $i = 1$. (i) the g-data goes to cell $(1, 1)$ or (ii) the e-data goes to cell $(1, 1)$. The probability for $i = 1$ is therefore the sum of the two probabilities. There are two ways how to obtain $i = 2$. (i) the g-data goes to cell $(1, 1)$ first, so that the probability for the e-data depends on it, or (ii) the e-data goes to cell $(1, 1)$ first and the g- data depends on it. If $e, f, g, h \geq 1$ the expressions become quite complicated.

The iterative *expectation maximization* (EM) algorithm is used to find maximum likelihood estimators for the joint probabilities in the 2×2 table with supplemental data on both marginals (Little and Rubin, 1977, p. 182ff). Let $A^{(t)}, B^{(t)}, C^{(t)}$, and $D^{(t)}$ be the estimates of the frequencies in iteration t. We than have

$$
\begin{aligned}
A^{(t+1)} &= a + g\frac{A^{(t)}}{A^{(t)}+B^{(t)}} + e\frac{A^{(t)}}{A^{(t)}+C^{(t)}}\,, \\
B^{(t+1)} &= b + g\frac{B^{(t)}}{A^{(t)}+B^{(t)}} + f\frac{B^{(t)}}{B^{(t)}+D^{(t)}}\,, \\
C^{(t+1)} &= c + h\frac{C^{(t)}}{C^{(t)}+D^{(t)}} + e\frac{C^{(t)}}{A^{(t)}+C^{(t)}}\,, \\
D^{(t+1)} &= d + h\frac{D^{(t)}}{C^{(t)}+D^{(t)}} + f\frac{D^{(t)}}{B^{(t)}+D^{(t)}}\,.
\end{aligned}
\tag{13}
$$

In the first iteration we set $A^{(1)} = a$, $B^{(1)} = b$ etc. Usually, convergence is fast. During the iteration the frequencies of the complete cases a, b, c, d remain constant. The counts of the incomplete cases g, h and e, f are proportionally distributed to the cells. The proportion is determined by the current conditional probability factors $A/(A + B)$ and $A/(A + C)$ etc. The procedure generates expected frequencies for the 2×2 table. They can be used to estimate the ϕ coefficient by replacing parameters by estimates in Formula (7).

Author note

This research was conducted in part while the first author was a visiting professor at Bowling Green State University and the third author was a visiting professor at the University of Salzburg. It was supported by National Science Foundation grant SBR-9422253 to Bowling Green State University, Michael E. Doherty and Clifford R. Mynatt principal investigators. The authors would like to acknowledge the contributions to this paper made by Ryan Tweney.