Do Subjects Understand Base Rates?

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

Investigations of the degree to which people neglect or use base rates typically require subjects to make a judgment based on presumptive integrations of base rates and likelihood ratios. The present paper deals with a logically prior issue, whether people understand what data are needed to constitute a proper base rate. The 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. Only the former frequencies are relevant to the estimation of the base rate of D, hence to the probability of D given S. Six experiments are reported. 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 by using a binary response mode, but subjects were no more likely to select the correct frequencies than the incorrect ones. The subjects perceive the relevance of the information as invariant in respect to the problem structure. They do not recognize that the relevance of information depends upon the type of inference to be drawn. These results, all based on a frequency formats and qualitative dependent measures, strongly support the conclusion that subjects do not understand the concept of base rates.

Two research programs dominate the field of reasoning under uncertainty: the *heuristics and biases* program (Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996) and the *ecological rationality program* (Gigerenzer, 1991, 1996; Cosmides & Tooby, 1996). According to the former program, reasoning under uncertainty is guided by a small number of cognitive heuristics such as availability, representativeness, and anchoring, which often yield reasonable judgments but sometimes lead to systematic errors and characteristic biases (Kahneman & Tversky, 1972, 1973, 1996; Tversky & Kahneman, 1974). The ecological rationality program argues that human reasoning is well adapted under ecologically relevant conditions and that such errors and biases can largely be made to disappear if the probabilistic information provided to the subjects is reformulated and presented in terms of *frequencies*. In fact, there is empirical evidence that the subjects show *better performance* in experiments run in the frequency program than in the heuristics and biases program. The question is why.

According to the ecological rationality program, the frequency format leads to better performance because people understand frequencies better than probabilities. We argue that for the base rate fallacy this attribution is not cogent. Usually, the frequency format is employed in a special form only, so called *natural sampling* (Kleiter, 1994, 1996; Pasini & Kleiter, 1995), which means mainly that all cases that have been sampled have complete information. In the natural sampling condition, the frequencies corresponding to the base rates are irrelevant. The frequency format is not confined to this specific simple structure, but more complex "non-natural" sampling forms have not been investigated. The term "non-natural" sampling means that some observations have only partial information. There is no implication that the sampling procedure is one that people do not naturally experience. If the frequency format is the actual cause of improved sensitivity to base rates, and if that improvement is actually due to subjects' understanding of the nature of base rates, then the subjects should select the relevant data needed to construct the appropriate base rates in the Partial Information Paradigm, described below.

Koehler (1996), 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 probabilities demanded by the normative model. Given such inconsistency of behavioral results, perhaps subjects simply do not understand base rates in a fundamental way, but rather are responding to variation simply because it is variation. This is an especially salien possibility in within subjects designs. 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). In the next section we discuss several characteristics that distinguish frequency and probability formats of base rate tasks. We

introduce and motivate our new version of base rate task, the Partial Information Paradigm.

Task Analysis

This section first shows that the term "frequency format"'is not a univocal one: A given problem may be constructed in a variety of frequency formats that differ greatly among themselves in psychological complexity. We then go on to explore some statistical distinctions between natural and non-natural sampling.

Task formats We first consider the frequency and the probability formats. Here is a typical example of a base rate task presented in the *probability format*, taken from Gigerenzer & Hoffrage (1995).

> 1 % of women at age forty who participate in routine screening have breast cancer. 80 % of women with breast cancer will get positive mammographies. 9.6 % of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

When this task is reformulated in the *frequency format* it reads as follows

> 10 out of every 1,000 women who undergo a mammography have breast cancer. 8 out of every 10 women with breast cancer who undergo a mammography will test positive. 99 out of every 990 woman who do not have breast cancer will test positive. Imagine a new representative sample of women who had a positive mammography. How many of these women do you expect actually to have breast cancer? (Sedlmeier & Gigerenzer, 1995)

In the second version, so it is claimed, more subjects are sensitive to the base rates than in the first one. Replacing probabilities by frequencies seems to change the cover story of the task, but not its underlying structure. The two tasks seem to be isomorphic. But are they really? [1]

Some reflection shows that the probability and frequency versions differ in complexity, specifically with respect to the number of data generating processes implicated in each. The natural sampling condition usually involves a single data generating process only, while in general the frequency format involves several processes.

[1] We note that the numbers in the frequency version are not, strictly speaking, frequencies. When the cover story says '10 out of 1,000 women have breast cancer' this is understood as a proportion standardized to 1,000 women. This is a common style in newpapers and other media.

Number of processes The frequency example may be modeled by random sampling from a *single* urn containing 1,000 balls, each one corresponding to one person. A certain number of balls is associated with cases suffering from breast cancer and a subset of these are associated with cases testing positive etc. The frequency format cover story quoted above describes an example corresponding to a *single urn* model. The task, though, can easily be described in the frequency format as a *four urn* example.

> In an epidemiological study investigating 236,465 women it was found that 2,365 had breast cancer. In a second study investigating the medical history of 212 women with breast cancer it was found that 170 tested positive and 42 tested negative. Finally, in a third study involving 7,267 healthy women it was found that 726 women tested positive and 6,541 tested negative. How many of the next 300,000 women to be screened in the forthcoming year and testing positive would you expect actually to have breast cancer?

Other versions with two or three urn models can easily be constructed. No doubt, a four urn example is more psychologically complex than a one urn example. The simplification from four urns to one urn can be done by statistical theory. If relative frequencies are used in place of probabilities – a reasonable procedure with the large sample sizes – the solution of the task is the same as in the single urn example. The point is that casting a probability task into a frequency task is not unique. Casting it into a natural sampling task introduces the most simple single urn structure.

To express this kind of task complexity we may simply count the number of frequencies or probabilities that must be processed to solve a task. To estimate the diagnostic probability given an observed symptom in a binary natural sampling task we need only two frequencies (the two frequencies associated with the conditional symptom probabilities). To calculate the probability in a task in the probability format we need three values (the base rate and the two conditional symptom probabilities). In a non-natural three urn model task we need six frequencies.

Partially observed data The single urn frequency example contained *complete data*. For each of the 1,000 women the health state *and* the test data were reported. There were no women for whom only the mammography results were available, women for whom only the diagnoses were available, or women for whom both, health and test data, were unavailable. In real life we are often confronted with incomplete, missing, or additional data. Partially observed data can occur in the frequency format and such data are of primary concern in this paper. In statistics partially observed data can be difficult to process. Iterative procedures, e.g., expectation maximization (EM, Dempster, Laird, & Rubin, 1977) or iterative proportional fitting (IPFP, Bishop, Fienberg, & Holland, 1975) are used to obtain maximum likelihood estimates.

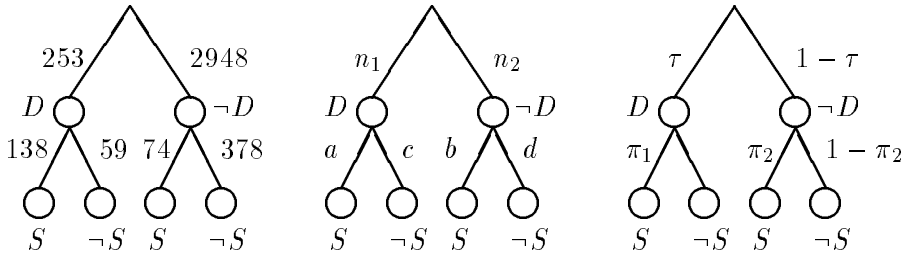
Precision In elementary textbooks on probability theory the probabilities contained in the examples are often assumed to be known exactly. The source of probabilities is not of primary interest in these books. The problems are typically such that the student simply has to find the appropriate rule or theorem to apply. This entails *deductive probability theory*, which involves deriving probabilities from a set of given probabilities. It is analogous to deductive logic where conclusions are derived from sets of premises. Psychology may investigate how competent humans are at doing deductive probability theory. If tasks are presented in a frequency format, though, this involves statistics. Statistics is concerned with, among others things, the question of how we arrive at reasonable probability estimates, how updating is done, etc. This is *inductive probability theory*. Psychology may also investigate how competent we are as intuitive statisticians. In real life problems, probabilities must be estimated from frequencies. The accuracy of the estimates depends on the sizes of the samples from which the frequency counts are 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.

Statistical Model of the Task For a medical diagnosis task we employ the following notation (Figure 1, right panel). Let τ denote the base rate 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 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, as in the middle panel, where 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.

Let us assume that we know very little about the uncertain quantities before we observe the actual frequencies. We express this vagueness by 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., $Be(1,1)$ (the reader unfamiliar with beta distributions is referred to Novick & Jackson, 1974 or to Bernardo & Smith, 1994). 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, (see 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 & Smith, 1994; Kleiter, 1981) we know that the

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.



posterior distributions are beta distributions, given by:

$$\tau \sim Be(n_1, n_2)\,, \quad \pi_1 \sim Be(a, c)\,, \quad \text{and} \quad \pi_2 \sim Be(b, d)\,. \tag{1}$$

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 distinction between probabilities and frequencies is important. Frequencies are used to update our knowledge about probabilities. Probabilities are in turn used to predict frequencies. 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:

$$\mu = \frac{\tau\pi_1}{\tau\pi_1 + (1-\tau)\pi_2}\,. \tag{2}$$

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).

We have discussed a number of task characteristics – the frequency and probability formats, the number data generating processes, complete and partially observed data, precision, and a statistical model – that warn us against directly comparing in a one-to-one fashion the probability and the frequency format of a task. We now turn to the criteria for saying a subject has understood base rates.

Understanding Conditionalization

Bayes' theorem is an elementary and non-controversial rule. The Bayesian approach employs the theorem in the process of *conditionalization*, which is at the very heart of probabilistic reasoning. What happens as we proceed from one conditional probability $P(A|X)$ to a second one $P(A|X,Y)$ by conditioning on a further piece of evidence? Probabilistic inference or the revision of beliefs in the light of evidence consists of two steps, a logical one and a numerical one (de Finetti 1974; Kleiter, 1991):

1. REMOVE step, the removal of non-instantiated possibilities: the truth of the conditioning event eliminates all possible worlds in which this event is false. If a positive test is observed in one or more patients all possible worlds in which this patient or these patients test negative are eliminated. Conditioning narrows the possibility space.

2. RE-STANDARDIZE step, the re-standardization of the probabilities: After eliminating alternatives the probabilities do not add up to 1.0 anymore. Thus, the probabilities need to be re-standardized to fit the new possibility space. This is done by Bayes' theorem.

The inductive process begins with a set of possibilities. Each incoming piece of information logically suppresses some of these possibilities, namely those which are incompatible with the new information and which should thus be excluded from further analysis. The second stage of induction deals with the remaining probabilities. Given a possibility space together with a probability distribution defined on it, the removal of possibilities leads also to a loss of probability mass. To fit the new situation the probabilities must be re-standardized to sum to 1.0. De Finetti (1974) showed that this re-standardization - if done in a coherent way - "automatically" follows Bayes' theorem.

The understanding of the REMOVE step is equivalent to understanding which possibilities should be eliminated and which not. Being able to distinguish the relevant from the irrelevant pieces of a problem is a prerequisite for its understanding. If in medical diagnosis a person tests positive, the possibility of a negative test can be eliminated; part of the data space can be "cut off." It would be against the laws of probability theory to cut off a piece of the hypothesis space though: To answer the question "what is the probability that a patient has disease X" we should not cut off the possibility of not-X. This is what subjects often do in the pseudodiagnosticity task (Doherty,

Mynatt, Tweney, & Schiavo, 1979). The REMOVE step is a precondition for proper inference, as is the proper combination of numbers according to some rules (Bayes' theorem, conjunction theorem, etc.). The distinction between the two steps is therefore psychologically relevant.

Closely connected with the two steps is a methodological question. Two response modes have been used in the literature to investigate probabilistic tasks:

1. Qualitative judgments: Asking subjects to select relevant information from a set of alternative options (Doherty, et al., 1979).

2. Numerical judgments: Asking subjects to give a probability or frequency rating. Data evaluation is done by comparing numerical judgments with "correct" numbers. There is little control for different strategies that may produce similar ratings.

It is natural to use qualitative judgments to investigate whether subjects understand the REMOVE step and its understanding. Similarily, it is natural to use the numerical judgments to investigate the RE-STANDARDIZE step. The experiments to be described below are concerned primarily with the qualitative question. Are subjects capable of selecting the relevant information? Smith, Langston, & Nisbett (1992, p. 3) discuss what it means to use a rule in reasoning. They list two requirements:

1. The reasoner should *recognize* that a piece of information is of a certain abstract kind so that it can be subsumed by the rule. A precondition to apply the modus ponens rule in deductive reasoning is to recognize in which way a given example corresponds to the components of the modus ponens. This involves the instantiation of an abstraction (Smith, Langston, & Nisbett, 1992).

2. The reasoner should *apply* the rule itself to the information.

The distinction can further be illustrated by the *recognize* - *act* cycle that governs the dynamics of production rules in production systems. A production rule has the form `if` $<$ *lhs* $>$ `then` $<$*rhs*$>$ where *lhs* represents a list of left hand side conditions that must be true to trigger *rhs*, the right hand side condition. The left hand side conditions must *match* contents contained in the working memory of a production system so that the action on the right hand side can fire. A person understands a rule if he or she recognizes which information matches the left hand side premises of the rule. The present study investigates the *recognize* step: Do subjects *recognize* what data constitutes a relevant base rate? If they do, they should select the relevant and discard the irrelevant data. That is, they should select information that matches the conditions in which the Bayes' theorem is properly applied.

Natural Sampling

Natural sampling is distinguished from *designed* sampling where observations are deliberately selected. Sampling in a 2×2 table, for example, may be done with random or with fixed marginals.

When the marginals are random they are informative about corresponding future frequencies. When the marginals are fixed and controlled by an experimenter they are noninformative. Moreover, natural sampling requires that the observations are *complete*. There are no missing data.

While practically all base rate experiments reported in the literature that present base rates in a frequency format use a natural sampling scheme, the Partial Information Paradigm used in our experiments uses a non-natural sampling scheme. We introduce missing (or additional) data. We note that the distinction between natural and non-natural sampling is not relevant to tasks presented in the probability format. The concept refers to properties of the data generating process but not to the underlying probabilities.

The distinction between natural and non-natural sampling is crucial for the interpretation of the results obtained from frequency format experiments. Natural sampling gives rise to dramatic simplifications when compared with non-natural sampling. This is especially true for the role of those frequencies that correspond to base rates: They simply drop out and can be ignored! It may be shown that under natural sampling the probability distribution that a patient showing the symptom S is suffering from D is given by

$$\mu \sim Be(a, b)\,. \tag{3}$$

The proof of the result is given in Kleiter (1994) and Kleiter & Kardinal (1995). This result may be startling to some: 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 $n = n_1 + n_2$. In the natural sampling condition, the frequencies corresponding to the base rates are irrelevant.

The result is essential for tasks in the frequency format. If the frequencies in the 2×2 table on which inferences are based were obtained under natural sampling, the marginals (base rates) do not contain any extra information that is not already contained in the cell counts in the core of the table. The left panel of Figure 2 shows the information available. In the right panel the irrelevant base rates of the disease and of the symptom are shaded out. The situation is different in the probabilistic format. The left panel of Figure 3 shows the information available. Note that conditional probabilities were put into the cells. There is no one-one correspondence between conditional probabilities and frequencies. The right panel shows the relevant information; the irrelevant information is shaded out again.

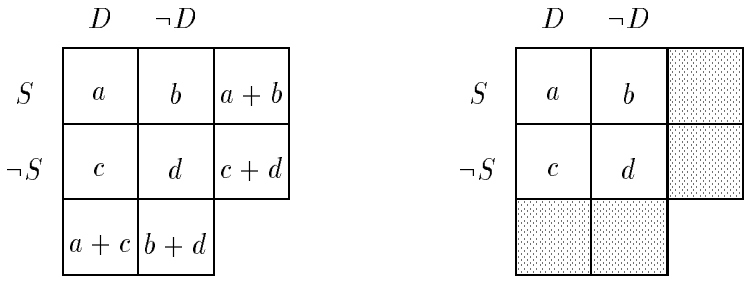


Figure 2. Left panel: 2×2 table containing the frequency of the presence and absence of a disease and the presence or absence of a symptom in a natural sampling condition. Right panel: relevant information to make inferences about the probability of the presence or absence of the disease given the presence or absence of the symptom; the information in the shaded boxes (base rates of the disease and the symptom) is not relevant in a natural sampling condition.

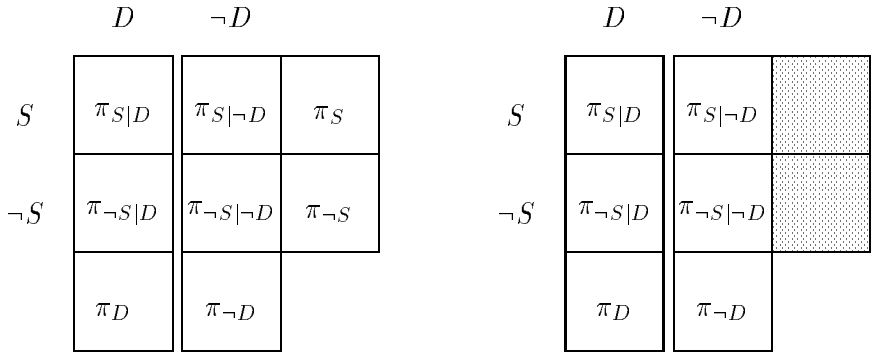


Figure 3. Left panel: conditional symptom probabilities and base rates of the presence (D) or absence ($\neg D$) of a disease and the presence (S) or absence ($\neg S$) of a symptom S. Right panel: relevant information to make inferences about the probability of the presence or absence of the disease given the presence or absence of the symptom; the information in the shaded boxes (base rate of the symptom) is not relevant.

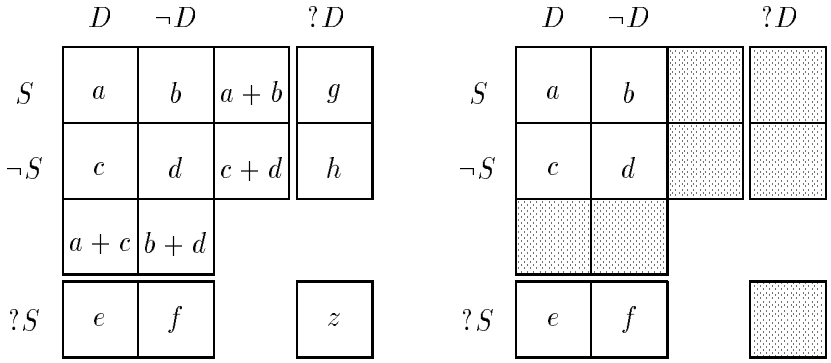


Figure 4. Left panel: frequencies of the presence and absence of a disease and the presence or absence of a symptom; additional frequency information about the base rate of the disease (e and f) and about the base rate of the symptom (g and h) in a non-natural sampling condition. Right panel: relevant information to make inferences about the probability of the presence or absence of the disease given the presence or absence of the symptom; the information in the shaded boxes (base rate of the the symptom) is not relevant in a natural sampling condition.

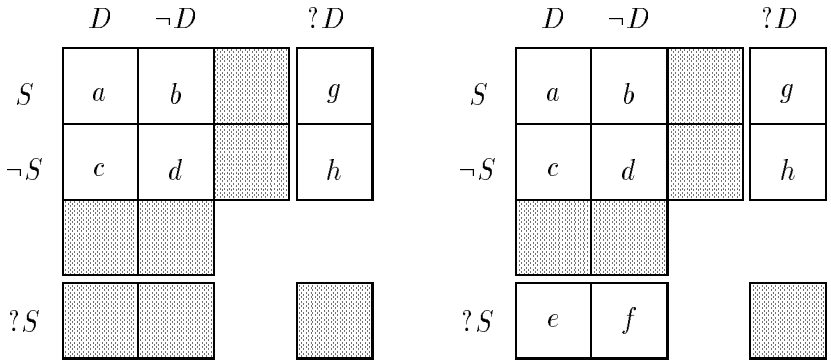


Figure 5. Left panel: relevant information to make predictive inferences about the probability of the presence or absence of the symptom given the presence or absence of the disease; the information in the shaded boxes (base rate of the the disease) is not relevant. Right panel: relevant information to make inferences about the correlation between the disease and the symptom; the information in the shaded boxes is not relevant.

Base rate frequencies are only relevant in non-natural partial sampling conditions. The left panel of Figure 4 contains the additional frequencies e, f, g, and h. In the panel on the right hand side irrelevant information is shaded out. The frequencies e and f provide relevant additional information about the disease so that the base rates are estimated by the sums $a + c + e$ and $b + d + f$. Compare the structure of the shaded boxes with that in the left panel in Figure 5. Rows and columns exchanged roles. The direction of conditioning is reversed now. This corresponds to a situation in which we make a *prediction* about the presence or absence of a symptom given the presence or absence of a disease. Now the additional symptom base rate frequencies g and h are relevant. Finally, the right hand panel in Figure 5 shows the relevant information for inferences about the *correlation* between the disease and the symptom. Now all four additional marginals are relevant.

A subject who understands base rates recognizes the difference in the direction of conditioning, that is, the difference between diagnoses and prediction. A subject who understands base rates does not confuse the base rate of the disease and the base rate of the symptom. The Partial Information Paradigm provides a framework to test this understanding. With minimal changes in the cover stories a task can be presented one time as a diagnosis, a second time as a prediction, and a third time as a correlation problem.

The Partial Information Paradigm may further be motivated by an analogy with the investigation of children's concept of the cardinality of a set. Consider a Piagetian task on children's understanding of the cardinality of sets. Marbles are filled into several glasses. A pair of glasses is selected by the experimenter. The children are asked to tell which one of the two glass contains more marbles. To test the understanding of the cardinality concept it is essential to vary the diameter of the glasses. If they have the same diameter the height to which the glasses are filled is a perfect criterion to judge the cardinality. No understanding of cardinality is needed to find "right" answers with glasses having the same diameter. Likewise, no understanding of base rates is needed in a natural sampling task to find the right answer. To test the understanding of which frequencies constitute a proper base rate for a given inference task non-natural sampling conditions are essential.

The experimental investigation of base rates under natural sampling conditions is insufficient as a test for the understanding of base rates. An experiment by Christensen-Szalanski and Bushyhead (1981, experiment 2) illustrates the problem. In the paper, the authors use 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 physicians 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. A person who is properly sensitive to base rates uses *all* avail-

able base rate information that is relevant to the inference to be drawn, which may include data for which no likelihood information is available. In a nutshell, a critical condition for discovering whether subjects understand 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.

The Partial Information Paradigm

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 of D, and therefore relevant for the distribution of μ. On the other hand, cases with missing disease information are irrelevant for diagnosis; the frequencies of $?DS$ and $?D\neg S$ cases are irrelevant for the distribution of μ (the proof is given in the appendix).

The Task Subjects are asked to select frequentistic case information that they consider relevant to inference. For some cases the 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 presented to our subjects 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 ...

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 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 (Figure 4, right panel), g and h (but not e, f, or z) for predicting S from D (Figure 5, left panel), and e, f, g, and h (but not z) for a correlation task (Figure 5, right panel). If in a diagnosis task the subjects are discarding e and f as irrelevant, then they are showing a lack of insight into the base rate. If they select e and f but also select g and h, then we conclude that they are selecting information indiscriminately rather than than showing base rate comprehension. 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 obtains the following three probability estimates (expected values or means of the second order distributions) 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 obtains the following three probability estimates for the disease given the symptom is present or absent, respectively:

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

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

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. 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.

Selection Patterns As we have noted, the cells with $?D$ information, namely the g, h and z cells, are normatively irrelevant for diagnosis. These cases give additional information about the base rate of the symptom. However, this information is worthless for diagnostic inference. We give two explanations, an intuitive one in the following paragraph, and a formal one in the appendix.

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 - EXPERIMENT 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. Then 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 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. Since the fundamental concern of this study was the investigation of the REMOVE step, these poserior probability judgments will not be treated further. 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 (check one) 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

Participants

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 DS 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 because 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

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. 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.

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 corresponding mean relevance ratings. Note that 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 do not sufficiently understand the implications of the base rate of the disease. Only three subjects made the optimal selections of data for the diagnostic system. This is not to say that the behavior was completely inappropriate, or irrational.

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.

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 $\neg D$ data. Two calculated 2×2 tests of independence. Others raised issues 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 it is 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

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.

better. We turn our attention to a replication with students in the early phase of their education, students in Introductory Psychology.

EXPERIMENT 2

Method

Participants

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 Table 3, 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

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. Column 2N contains the corresponding 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 a few Ss selected cell *z* which contains irrelevant information. Cell *f* which contains relevant information is selected by only 30 of the 118 Ss.

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.

Note that the frequencies and mean ratings (Tables 5 and 6) are 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 to describe the similarities among 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 presence of frequency values, and the data were pooled in order to have a substantial number of subjects for the next breakdown.

Of great interest 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. There are $2^9 = 512$ different possible selection patterns. These are categorized in Table 8. According to the number of posterior probabilities (including incorrect ones) that can be computed with 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 only normatively relevant information is selected (the core plus e and f). Twenty-four Ss selected information from which neither posterior probability can be calculated. 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 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 in experiment 2 of similar levels of age and education.

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 exp. 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.'

Participants

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 only the normatively correct selections of Information Types. The results of Experiment 3 are consonant with those of the first two experiments; subjects show little insight into the value of the partial information as contributing to a more precise 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.

Method

Participants

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 previously 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. Call the condition calling for subjects to select 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 sections were randomly assigned to the four versions of the task, with the restriction that there be two sections per task. 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 completing 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 only normatively relevant 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. Two of the 51 subjects in the Pr5 condition did so. 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 a $P(D|S)$ could be calculated 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 mean relevance ratings are both .99.

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

Participants

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 asymmetry in the explanation of the task, two diseases (Tanner's 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 versions were used, varying the order of the diseases (Brunner-Tanner vs. Tanner-Brunner), the order of the information types (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 as not together, and on the right: Brunner's disease and Tanner's syndrome always occur together. The task was conducted in several classroom settings and discussion groups.

Results

The frequencies of the various response patterns are shown in the last column of Table 8. Only five Ss selected the relevant information 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, compared with 64 % for cells c and d. The relevance ratings in Table 6 demonstrate the same tendency. The judgment of correlation shows a strong positivity bias. Relevant partial information is neither chosen systematically nor 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

We observe 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.

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 whatever 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 no information about the disease but 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 no information about the disease but did not have the 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 be more likely to select features that are present than to those features that are not present, or not known to be present. A related possibility comes from the fact that in 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 may be seen as contaminating the the observed part of the partial information. Hence, Experiment 6 used two semantically more balanced categories, and required a forced choice between them. Another possible criticism of the above studies is that the relevance ratings may have been reactive, and may have led subjects to believe that there was differential importance among the several cells. While this should not have influenced subjects who did have a firm understanding of base rates, the design of Experiment 6 did not call for such ratings.

Method

Participants

There were four groups of subjects, all run at Bowling Green: (a) 122 students in an introductory psychology course taught by the third author, (b) 14 introductory statistics students who had completed a unit on probability, (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 fourth 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) had 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 84 students who selected the normatively correct data and 84 who selected the incorrect data. For such data, of course, $\chi^2(1) = 0$.

Experiment 6 was replicated with an additional 80 undergraduates. In this replication no mention was made of missing information, and the problem was reframed in terms of additional information. Of the 80 Ss, 41 chose the appropriate data, 39 the inappropriate data. Thus, framing the partial information as missing or as additional observations led to practically the same results.

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 evidence that subjects do not fully appreciate the import of base rate information for diagnostic inference.

GENERAL DISCUSSION

The results of the above investigations are in agreement with one another. In each, the subjects failed to discriminate between relevant and irrelevant base rate information. In experiments 1 through 5 subjects were clearly not responding in some random fashion, which one might expect were the subjects not understanding the task. 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 diagnosis, prediction 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 diagnosis or prediction.

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, more strongly 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 probability with which a given piece of information is selected is a measure of its *perceived relevance*. Table 5 shows that the selection probability is low when a feature is unknown, medium when absent, and high when present. The frequencies in Table 5 suggest a simple linear dependence of the selection probabilities upon the feature categories. As a hypothesis, we assigned the "perceived relevance scores" 0, 1, and 2 to the categories "missing," "absent," and "present," respectively. The selection probabilities can now be expressed as a function of the perceived relevance score. Bayesian logistic regression analysis (using BUGS, Spiegelhalter, Thomas, Best, & Gilks, 1995) was used to investigate the relationship. Flat prior distributions were employed for the regression coefficients. As the estimates for the interaction coefficients were much smaller than those for the main effects, we report only results without interaction terms. Table 10 shows the point estimates of the regression coefficients for logits. The logits were re-transformed to probabilities. Figure 6 shows the predictions of the selection probabilities resulting from the regression models for experiments 1, 2, 3, 4Dx9, 4Pr9, and 5. The regression coefficients in Table 10 and the associated predictions in Figure 6 are very similar for the six experiments. Moreover, the regression coefficients for the disease and the symptom are very similar. A good fit results from the equation

$$\text{selection probability} = 0.1 + 0.2 \times x_1 + 0.2 \times x_2 \,,$$

where x_1 denotes the perceived relevance score (0, 1, or 2) of the disease and x_2 the perceived relevance score of the symptom. The intercept 0.1 may be interpreted as an unspecific selection tendency (a response bias to mark an item with a cross). The regression coefficients 0.2 and 0.2 for the disease and the symptom suggest that the relevance of the information in fact is perceived as a simple linear function of how many feature values are present. The result shows that in the present

series of experiments *subjects perceive the relevance of the information as invariant in respect to the problem structure.* They do not recognize that the relevance of information depends upon the type of inference to be drawn.

Table 10. Regression coefficients (logit units) of a two-dimensional linear logit analysis of the selections probabilities; α_0=intercept, α_1=slope of the disease, α_2 = slope of the symptom

	Experiment or experimental condition [a]					
Coefficients	1	2	3	4Dx9	4Pr9	5
α_0	-2.57	-2.04	-2.04	-2.42	-3.53	-1.79
α_1	1.89	1.34	1.33	1.36	2.02	.90
α_2	1.35	1.13	1.11	1.31	1.79	1.13

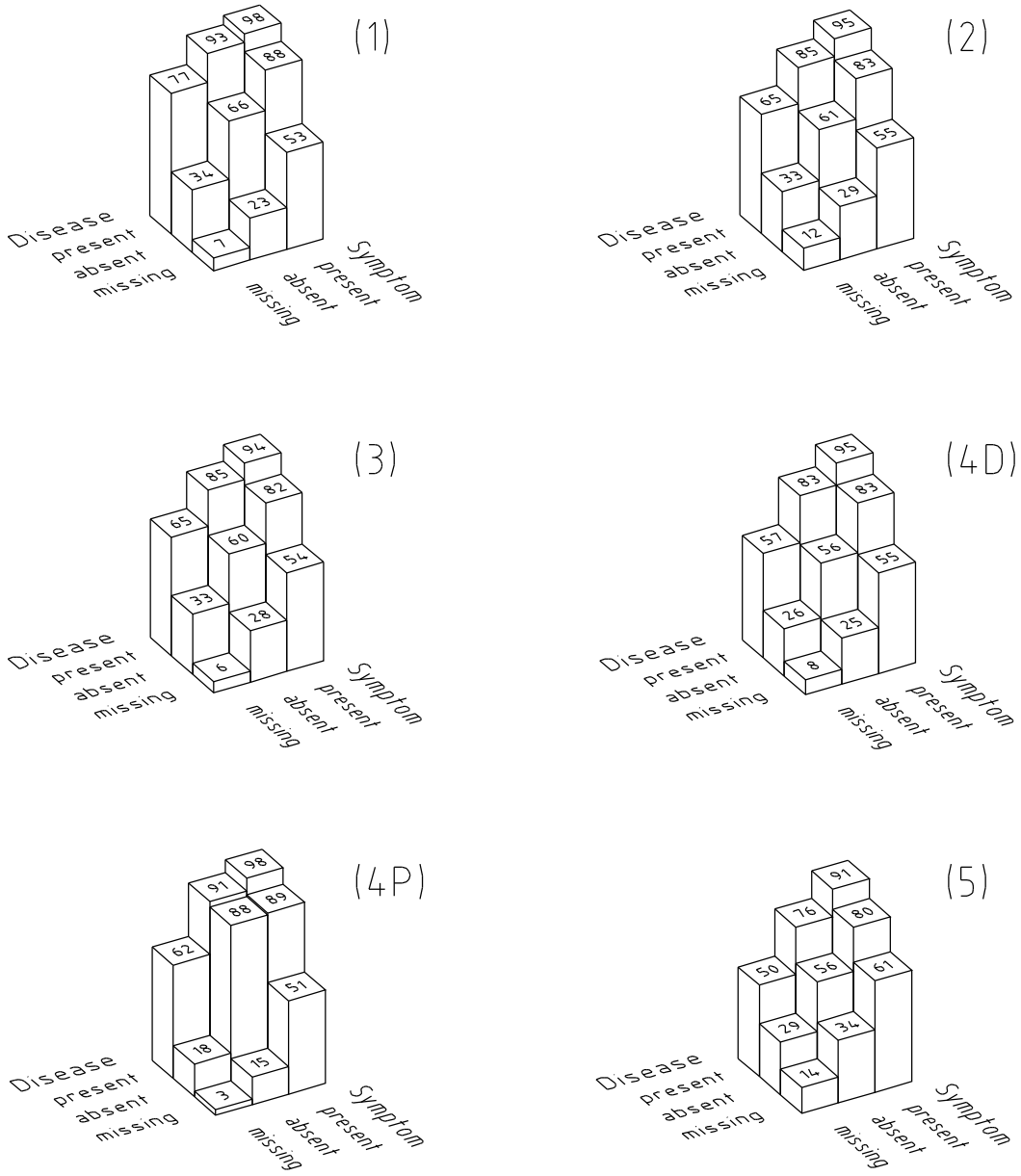
[a]Dx = diagnosis, Pr = prediction and the numbers after Dx and Pr refer to the number of Information Types from which subjects selected.

The conclusion that subjects in the Partial Information Paradigm failed to comprehend 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 (1996) shows, people's use of base rate information is influenced in the appropriate direction, though typically not to the appropriate degree, by a 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 the base rate spontaneously and appropiately. They do not. It appears, on the other hand, that subjects do spontaneously use what might be considered a defective "natural sampling heuristic." That is, subjects appear to assume that whatever complete data they have are sufficient for diagnosis and prediction.

How serious is a base rate neglect in real world inference? The question raises two sub-questions: (1) What is the *sensitivity* of a heuristic natural sampling solution as compared to the EM-solution in non-natural sampling conditions, and (2) what is the *importance* of these conditions in everyday life? There is no general answer to the first question. It is obvious that the heuristic (natural sampling) point estimates are equivalent to the optimal estimates (EM iteration) if the base rate proportions in the complete and in the incomplete cases are the same (Pasini & Kleiter, 1995). The precision, though, is higher for the optimal solution as it is based on more observations.

It may be that less is known about the base rates than about the conditional symptom probabilities (Kleiter & Kardinal, 1995). We have not treated this case in this paper. At first sight, people often deny the logical possibility of this case. They argue that it cannot happen that the sample size of those cases in which the disease is observed is less than the sample size

Figure 6. Estimated selection probabilities obtained from logistic linear regression for experiments 1 to 5.



of those cases in which the conditional symptom frequencies are observed. Imagine, though, that fifty subjects inflicted with the disease X and fifty subjects not inflicted are investigated for testing positive or negative on a given symptom. If the sample sizes are completely controlled by the experimenter, then we learn nothing about the base rates, but only about the conditional symptom probabilities. As a matter of fact, if no other information about the base rates is available no diagnostic probabilities can be calculated. All values in the interval $[0, 1]$ are coherent.

Non-natural sampling seems to be quite common in the world of modern technology. Multicomponent technological systems require the integration of partial information, including subjective expertise, statistical data, and theoretical considerations. An example is probabilistic safety analysis (Cacciabue & Papazoglou, 1996). The total number of plane crashes is well known, the knowledge about their causes is partial and incomplete. The total numbers of crimes is better documented than many of their situational details, the motives of the criminals, or the characteristics of the victims. For new diseases (like BSE, Anderson et al., 1996), the epidemiology is much better understood than the diagnostic tests (at the time this paper is written no *in vivo* test of BSE is known).

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 no compelling reason to believe 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 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.

The conclusion that people do not understand 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; see also Evans & Over, 1996). 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.

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. A similar conclusion and attribution to Evans was reached by Doherty, Chadwick, Garavan, Barr and Mynatt (1996), 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.

The observed positivity bias, together with the way in which subjects perceive and weight the relevance of missing, absent or present disease or symptom values is in good agreement with predictions made by mental models theory. A fundamental representational assumption of this theory is that 'individuals normally represent explicitly only those situations that are true' (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1996, p. 5; Legrenzi, Girotto, & Johnson-Laird, 1993). We may assume that subjects in the Partial Information Paradigm do not represent the missing feature values, or only represent such absent feature values as a 'footnote' (Johnson-Laird et al., 1996). A similar suggestion was put forward in respect to the pseudodiagnosticity task (Doherty & Mynatt, 1990; Doherty, Schiavo, Tweney, & Mynatt, 1979). Closely related are the conclusions drawn by Horn (1989) from his study of negation in the history of philosophy and psychology.

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. 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. The essential element in our experiments is that we did not ask whether people *use* base rates, but whether people *understand* base rates. The main conclusion we draw from our experiments is, that they do not.

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.

Anderson, R. M., Donelly, C. A., Ferguson, N. M., Woolhouse, M. E. J., Watt, C. J., Udy, H. J., Mawhinney, S., Dunstan, S. P. Southwood, T. R. E., Wilesmith, J. W., Ryan, J. B. M., Hoinville, L. J., Hillerton, J. E. Austin, A. R., Wells, G. A. H. (1996). Transmission dynamics and epidemiology of BSE in British cattle. *Nature*, **382**, 779-788.

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.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Cacciabue, P. C. & Papazoglou, I. A. (eds.) (1996). *Probabilistic safety assessment and management, Vol. I-III.* London/Berlin: Springer.

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

De Finetti, B. (1974). *Theory of probability, Vol I.* London: Wiley.

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. (1996) On People's Understanding of the Diagnostic Implications of Probabilistic Data. *Memory and Cognition*, **24**, 644-654.

Doherty, M. E. & Mynatt, C. R. (1990). Inattention to $P(H)$ and $P(D|-H)$: A converging operation. *Acta Psychologica*, **75**, 1-11.

Doherty, M. E., Schiavo, M. D., Tweney, R. D., & Mynatt, C. (1979). Pseudodiagnosticity. *Acta Psychologica*, **43**, 111-121.

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

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and Reasoning.* Hove, UK: Psychology Press.

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. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, **103**, 592-596.

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

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

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1996). Naive probabilistic reasoning: A mental model theory. Department of Psychology, Princeton University, NJ 08544.

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. and Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, **103**, 582-591.

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. (1991). Incomplete knowledge. Its representation in psychology and artificial intelligence. In F. Klix & E. Roth (eds.), *Kognitive Prozesse und geistige Leistung.* Deutscher Verlag der Wissenschaften: Berlin, 106-147.

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.

Kleiter, G. D. (1996). Critical and natural sensitivity to base rates. Commentary to a target paper by J. J. Koehler. *Behavioral and Brain Sciences*, **19**, 27-29.

Kleiter, G. D. & Kardinal, M. (1995). A Bayesian approach to imprecision in belief nets. In Mammitzsch, V. & Schneeweiß, H. (eds.), *Symposia Gaussiana, Conf. B.* Berlin: de Gruyter, 91-105.

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

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

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition*, **49**, 37-66.

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.

Novick, M. R. & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

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

Pasini, M. & Kleiter, G. (1995). La base rate fallacy nella categorizzazione di eventi sequenziali. *Giornale Italiano di Psicologia*, **22**, 641-661.

Sedlmeier, P. and Gigerenzer, G. (1995). Teaching Bayesian reasoning in less than two hours. University of Paderborn, Fachbereich 2 - Psychologie, 33095 Paderborn, Germany.

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.

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, **16**, 1-40.

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

Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (1995). BUGS. Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Inst. of Public Health, Cambridge, `ftp.mrc-bsu.cam.ac.uk`.

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

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 11. 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 11. Frequency table containing the frequencies of fully (a, b, c, d) and partially (e, f, g, h) 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

$$\pi_{y|x} = \frac{\pi_{xy}}{\pi_x} = \frac{\pi_x \pi_{y|x}}{\pi_{x=1}\pi_{y|x=1} + \pi_{x=2}\pi_{y|x=2}}, \tag{4}$$

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

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

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

$$\phi = \frac{\pi_{xy} - \pi_x \pi_y}{\sqrt{\pi_x \pi_y (1 - \pi_x)(1 - \pi_y)}} . \tag{6}$$

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 likelihood (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:

$$\begin{aligned} E(\hat{\pi}|a, b, z) &= P(i = 0|a, b, z)\,\hat{\pi}_1 + P(i = 1|a, b, z)\,\hat{\pi}_2 \\ &= \frac{a}{(a+b)}\frac{(a+1)}{(a+b+1)} + \frac{b}{(a+b)}\frac{a}{(a+b+1)} \\ &= \frac{a}{(a+b)}\frac{(a+1+b)}{(a+b+1)} = \frac{a}{a+b} = E(\hat{\pi}|a, b) . \end{aligned} \tag{7}$$

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 in a completely analogous fashin to the unconditional probability in the univariate case. Conditioning with respect to the X value $x = 1$ in Table 11 restricts the problem to the first line, a and b are equivalent to a and b in the univariate case, and g is analogous 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

$$\begin{aligned} E(\hat{\pi}_{y|x}|a,b,c,e) &= P(k=0|a,b,c,e)\,\hat{\pi}_{y|x}^{(k=0)} + P(k=1|a,b,c,e)\,\hat{\pi}_{y|x}^{(k=1)} \\ &= \frac{c}{(a+c+1)}\frac{a}{(a+b)} + \frac{a}{(a+c+1)}\frac{a+1}{(a+b+1)} \neq \frac{a}{a+b}\,. \end{aligned} \tag{8}$$

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 probability in cell $(1,1)$ we obtain the ML estimate

$$\hat{\pi}_{x=1,y=1} = \frac{a + e\frac{a}{a+c}}{a+b+c+d+e+f} \tag{9}$$

and an 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 Incomplete data in each of the two variables is relevant with 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. Each estimator now 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

$$E(\hat{\pi}_{x=1,y=1}|a,b,c,d,e,f) = \sum_{i=0}^{e+g} P(i|a,b,c,d,e,f)\frac{a+i}{a+b+c+d+e+f}\,, \tag{10}$$

and the probabilities are

$$P(i=0|a,b,c,d,e,f) = \frac{b}{(a+b)}\frac{c}{(a+c)}\,, \tag{11}$$

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

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

$$
\begin{aligned}
A^{(t+1)} &= a + g\frac{A^{(t)}}{A^{(t)}+B^{(t)}} + e\frac{A^{(t)}}{A^{(t)}+C^{(t)}}\,, \\
B^{(t+1)} &= b + g\frac{B^{(t)}}{A^{(t)}+B^{(t)}} + f\frac{B^{(t)}}{B^{(t)}+D^{(t)}}\,, \\
C^{(t+1)} &= c + h\frac{C^{(t)}}{C^{(t)}+D^{(t)}} + e\frac{C^{(t)}}{A^{(t)}+C^{(t)}}\,, \\
D^{(t+1)} &= d + h\frac{D^{(t)}}{C^{(t)}+D^{(t)}} + f\frac{D^{(t)}}{B^{(t)}+D^{(t)}}\,.
\end{aligned}
\tag{12}
$$

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 (6).

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.