COGNITION

Cognition 77 (2000) 197–213

www.elsevier.com/locate/cognit

Frequency versus probability formats in statistical word problems

Jonathan St. B.T. Evans[a,*], Simon J. Handley[a], Nick Perham[a], David E. Over[b], Valerie A. Thompson[c]

[a]Centre for Thinking and Language, Department of Psychology, University of Plymouth, Plymouth PL4 8AA, UK
[b]School of Social and International Studies, University of Sunderland, Sunderland, UK
[c]Department of Psychology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

Received 5 April 2000; accepted 16 June 2000

Abstract

Three experiments examined people's ability to incorporate base rate information when judging posterior probabilities. Specifically, we tested the (Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. Cognition, 58, 1–73) conclusion that people's reasoning appears to follow Bayesian principles when they are presented with information in a frequency format, but not when information is presented as one case probabilities. First, we found that frequency formats were not generally associated with better performance than probability formats unless they were presented in a manner which facilitated construction of a set inclusion mental model. Second, we demonstrated that the use of frequency information may promote biases in the weighting of information. When participants are asked to express their judgements in frequency rather than probability format, they were more likely to produce the base rate as their answer, ignoring diagnostic evidence. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Frequency; Probability; Statistical word problems

1. Introduction

There is a substantial body of evidence to show that human participants are prone to a range of errors and biases when given statistical word problems in the psycho-

[*] Corresponding author. Fax: +44-752-233-176.
E-mail address: j.evans@plym.ac.uk (J.St.B.T. Evans).

PII: S0010-0277(00)00098-6

logical laboratory. This has led to a major debate concerning the implications that this research has for human rationality (see, for example, Evans & Over, 1996; Funder, 1987; Gigerenzer, 1993; Lopes, 1991; Stanovich, 1999). Some authors – whom we will call pessimists – suggest that such results augur ill for people's ability to make competent decisions under uncertainty in the real word. Others – optimists – argue that the experimental methodology severely underestimates people's real world competence.

One powerful argument that has been advocated by the optimistic or pro-rationality camp is that people are much more competent in reasoning about frequencies than one case probabilities (see Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). Since most research on quantitative word problems has employed probabilistic formats, it is argued, a quite misleading picture of people's actual competence in statistical reasoning and risky decision making has emerged in this literature. Cosmides and Tooby (1996) have, in addition, linked the claimed superiority of reasoning about frequencies to their research programme on evolutionary psychology. They point out that although information about probabilities is freely available in the modern world, this would not have been the case when the human species was evolving. However, it would have been necessary for people to encode and apply frequency information available in their environment in order to make effective decisions. To support their arguments, Cosmides and Tooby (1996) present a set of experiments which appear to show that base rate information is utilized much more appropriately and effectively when problems are presented in terms of frequencies rather than probabilities.

We are led to question this account for two reasons. The first results from our differing theoretical perspective and the second concerns what we see as unsatisfactory controls in the experimental materials that Cosmides and Tooby (1996) presented to their participants. We consider the theoretical issues first. It is possible to argue from an evolutionary or rational analysis perspective that people – and indeed most animal species – can only operate effectively to achieve their goals in an uncertain world by learning and applying information about frequency distributions. We agree with this assertion and a detailed account of our perspective can be found in the discussions of Evans and Over (1996). Evans and Over's position can be regarded as a mixture of optimistic and pessimistic. It is optimistic in the sense that it is argued that reliability in achieving everyday goals is frequently achieved by the use of implicit cognitive processes. It is, however, relatively pessimistic in its assessment of people's ability to engage in abstract explicit reasoning in order to solve novel problems to which previous belief and experience cannot easily be applied.

The underlying difference here is between what has been called the massive modularity hypothesis of Cosmides and Tooby (Samuels, 1998) and dual process theories (Evans & Over, 1996; Fodor, 1983; Reber, 1993; Sloman, 1996; Stanovich, 1999; Stanovich & West, in press). Cosmides and Tooby (1994) argue that the mind contains no central, general-purpose processes following content-independent rules, but only consists of domain-specific modules that have arisen through evolution by natural selection to solve specific adaptive problems. Dual process theorists, on the other hand, distinguish between implicit cognitive systems which may be modular (domain dependent) but not necessarily innate in nature, and an explicit system allowing general-purpose reasoning. Evans and Over (1996), for example, have argued that probabilistic information may be represented by both the implicit and the explicit cognitive systems with differing psychological consequences.

Since reasoning about quantitative word problems is an explicit process, we see no basis for predicting that quantitative word problems will generally be easier when couched in terms of frequencies rather than probabilities, in apparent contrast with the findings reported by Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995). We therefore seek to explain their findings with reference to our theory of how the explicit reasoning system operates, via the construction and manipulation of mental models representing possible states of affairs. In order to explain the detail of our argument we need now to consider the experiments reported by Cosmides and Tooby (1996). They started out with the medical diagnosis problem of Cascells, Schoenberger, and Graboys (1978). The original problem was worded by Cascells et al. (1978) as follows:

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? __%

The correct Bayesian answer here is (about) 2%. This is most easily explained by imagining that there are 1000 people of whom 50 will test positive and of whom one actually has the disease. Participants giving answers classified as correct generally say either one in 50 or one in 51 have the disease. Strictly speaking neither is exactly correct as the false positive rate applies to the 999 who do not have the disease, not to the entire 1000, although the effect of this on the calculation is minimal with such a low base rate. This is nevertheless a potential cause of bias which was overlooked by Cosmides and Tooby but which is investigated in our own Experiment 3, reported below. For the time being we will treat answers of approximately 2% as correct. Only a minority of participants actually gave this answer in the study of Cascells et al. (1978) and the replication by Cosmides and Tooby (Experiment 1). The predominant choice was 95% ($1 - \text{FPR}$) which Cosmides and Tooby refer to as the *base rate neglect* response where FPR is the false positive rate. Presumably this is because participants misinterpret the false positive rate (5%) as the overall error rate of the test and therefore assume that it is correct 95% of the time. We tabulate this response as $1 - \text{FPR}$.

In their first experiment, Cosmides and Tooby compared this problem with a frequency version which was deliberately altered in a number of respects. The problem started as follows:

> One out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive (i.e. the 'true positive rate' is 100%). But sometimes the test also comes out positive when it is given to a

> person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease (the 'false positive' rate is 5%).

Participants are then told that a sample of 1000 Americans were selected by lottery and given the test. They then had to answer the question:

> How many people who test positive for the disease will actually have the disease?__ out of __

On this version 56% gave the correct 2% response as compared with only 12% on the original version. Of course, this frequency version also improves the wording of the original Cascells et al. (1978) version in other respects. For example, it does not presuppose that people understand terms such as 'prevalence' and 'false positive rate'. Cosmides and Tooby go on to present probability versions with improved wording and clarification but none of these produce performance of such a high level as the frequency version shown above.

Our problem with these experiments is that the frequency versions used do not simply present frequency information. They also a cue a simple mental model of set inclusion by focusing every statistic on a group of 1000 people. None of the probability versions used in their paper appear to assist the participant to construct such a model. Hence, we are not convinced that it is frequency information per se which is responsible for the facilitation. Our own experiments involve carefully controlled comparisons of problems which differ only in the format with which specific parts of the evidence are presented. In Experiment 1, we compare a probability version of the false positive information with two frequency versions which we call easy and hard. The frequency easy (FE) version describes the false positive rate as 50/1000 to make it easy to combine with the base rate evidence (1/1000). The frequency hard (FH) version presents this information as 1/20 which makes it more difficult to combine with the base rate in a single mental model. If frequency information per se is what makes the problem easy, then the FH version should be easier than a probability (P) version. If it is cueing of a set inclusion model which causes the facilitation, then the FH version should be no easier than a P version. Finally, we varied the format of the question that was asked, such that it requested information as either a frequency or a probability.

2. Experiment 1

2.1. Method

2.1.1. Participants and design

Two hundred and fifty-five undergraduate students of the University of Plymouth participated as volunteers in Experiment 1 and were tested in large groups. They were run in six independent groups each receiving one of six versions of the medical diagnosis problem. The design was 2 × 3 independent groups with the factors question format (probability or frequency) and false positive format (frequency easy, frequency hard, or probability).

2.1.2. Materials and procedure

The problem format for the frequency easy problem with frequency questions was as follows:

> One out of every 1000 people has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.Imagine that we have selected a random sample of 1000 people. Given the information above:On average, how many people who test positive for the disease will actually have the disease: __ out of __

In the frequency hard version, the last sentence of the first paragraph read:

> Specifically, out of every 20 people who are perfectly healthy, 1 of them tests positive for the disease.

In the probability version this sentence read:

> Specifically, 5% of people test positive for the disease.

In asking for the answer in probability or frequency format, we did not alter the wording of the question in any way. The probability question version was simply as follows:

> On average, how many people who test positive for the disease will actually have the disease: __%

Table 1
Percentage of responses in Experiment 1[a]

	COR ($\cong$ 2%)	BR (0.1%)	1 − FPR (95%)	FPR (5%)	Other
Frequency question (*n*)					
FE (43)	35	45	2	2	16
FH (43)	40	28	0	7	25
P (43)	30	33	5	12	20
Average	35	35	2	7	20
Probability question (original) (*n*)					
FE (41)	44	12	7	12	25
FH (43)	9	5	23	16	47
P (42)	24	12	17	10	37
Average	26	10	16	13	36

[a] COR, correct; BR, base rate; FPR, false positive rate; FE, frequency easy; FH, frequency hard; P, probability.

2.2. Results

Participants provided a variety of answers which we classified into different categories (see Table 1). The main ones reported by Cosmides and Tooby were correct (2%), base rate neglect (95%) and base rate only (0.1%). In Table 1 we refer to the 95% as $1 - \text{FPR}$ (false positive rate), because we also found that some participants gave the FPR (5%) as an answer as well.

The closest version that we have to the frequency version of Cosmides and Tooby (Experiment 1) is frequency easy with frequency question. Here only 35% solved the problem as compared with their 56%. Cosmides and Tooby found a considerable number of participants who gave a base rate only response on this problem (28%). This response was much more common with our problem (45%). In order to provide evidence about the effect of our independent variables we carried out χ^2 analyses on more than one of the responses shown in Table 1, comparing in each case the frequency of people giving a particular choice (say base rate) with those who did not. We realize that these analyses are not strictly statistically independent of one another, but feel with this unusual data set this is the best way to elucidate the nature of the trends that are present.

First, we analyzed correct responses. We compared the effect of false positive format, collapsing across question format. This produced no overall significant effect. That is, there was no evidence to suggest that the probability versions of the problems were any more difficult than the frequency versions. However, when the analyses were repeated for each question type separately, a significant effect of FP format was found for the probability question (original) group only ($\chi^2 = 13.35$, d.f. $= 2$, $P < 0.005$). As can be seen from Table 1, most correct answers were associated with frequency easy and fewest with frequency hard with probability intermediate. These findings are only partially consistent with Cosmides and Tooby's findings: of the two frequency problems, only one was easier to solve than the probability versions, and then, only when the question itself was asked in probability format. Instead, these findings are consistent with our assertion that the format of the problem (frequency or probability) does not in itself determine difficulty which is instead related to the ease with which a set inclusive mental model may be formed.

We also carried out several analyses to investigate the effect of question formats (frequency versus probability) on responding, collapsing over false positive format. First, although there were more correct responses with frequency questions, the χ^2 (2.72, d.f. $= 1$) fell short of significance. It is apparent, however, from Table 1, that question format affects the kinds of erroneous responses which are given by participants. First, we see that the 'base rate only' responses occurred much more often when the question was asked in frequency format ($\chi^2 = 23.62$, d.f. $= 1, P < 0.005$). Second, we also found that the $1 - \text{FPR}$ response (or base rate neglect response, as described by Cosmides and Tooby) occurred significantly more often when the question was in probability format ($\chi^2 = 14.26$, d.f. $= 1$, $P < 0.005$).

We think that the key to understanding these trends is the processing of the false positive information. When the question asked is in frequency format, the frequency hard manipulations have little effect on response rates. As with frequency easy problems both correct responding and the base rate bias response are relatively high. We assume therefore that people focus initially on the set of 1000 and recode the one in 20 base rate as 50 in 1000. This either leads them to find the correct subset relationship or to fall into the base rate bias. When the probability question is asked, however, people are focused on percentages. Here we suspect they recode the one in 20 as 5%. If they think of this superficially as the *error rate* for the test, then when asked how often the test correctly classifies positive testers, they are prone to answer 95%, the 1 − FPR error. Having coded the FPR as 5% they also find it very difficult to integrate with the base rate information, as this coding obscures the subset relationships. Hence, with the probability question frequency hard is indeed very hard – just 9% of participants giving the correct answer. In the frequency easy version, with probability question, the compatibility of the base rate and false positive information reduces the influence of the question asked and makes an early integration of the two pieces of data more likely (44% correct solutions).

3. Experiment 2

In contrast with findings of Cosmides and Tooby (1996) we did not find much evidence in Experiment 1 that probability formats led to fewer correct responses than frequency versions. The only significant trend occurred when the question asked was in probability format and here only the frequency easy problem (similar to theirs and thought by us to cue a set inclusion model) was significantly easier. Participants actually gave fewer correct responses for one of the frequency problems (that is, the frequency hard group); in our view this occurred because the cues for a set inclusion mental model were removed. What is particularly striking in our findings, however, is that the kinds of errors people make are strongly influenced by the question format. In particular, the base rate neglect response (1 − FPR) was more common when people were asked to give their answers as a probability and the base rate only response much more common when people were asked to express their answers as a frequency. This confuses the issue for Cosmides and Tooby's theory, as a manipulation which should facilitate correct responding is actually introducing a new form of bias. Focusing on the base rate is just as much a failure to integrate information as focusing on the false positive rate. Although Cosmides and Tooby (1996) report a similar base rate only bias in the most frequent version of their Experiment 1, they give this little comment and no further investigation.

Why do our participants show a bias towards the base rate only response when the question asks for a frequency rather than probability response? Our discussion above of the processing of false positive rates gives an account of why the 1 − FPR answer is common with probability questions, but no real explanation for why frequency coding leads to the base rate bias. The simplest account of this trend that we considered was that it is due to the fact that the frequency question format matches that of the base rate information which in Experiment 1 – and in all of the experiments reported by Cosmides and Tooby (1996) – was presented in a frequency form: 1/1000. Hence, in Experiment 2 we presented this information in probability format – 0.1% – for half of the groups, to see if this would reduce the base rate bias. We also wondered whether the difference between probability and frequency versions was specifically connected with the idea of relative frequency versus one-off probabilities, as discussed in earlier papers on this topic. Hence, we devised two forms of question requiring a percentage, one of which asked for an individual probability and the other for a proportion. In this experiment we kept the format of the false positive information constant, and similar to the easy frequency format of Experiment 1.

3.1. Method

3.1.1. Participants and design

One hundred and forty-four students participated in Experiment 2, none of whom had taken part in Experiment 1, and were again tested in groups. There were again six independent groups, each given a different presentation format. However, this time we gave six replications of the problem to each participant which differed in the actual values of the base rates and false positive rates that were provided. Base rates were set at 0.1, 0.5 or 1% in different problems and false positive rates at 2 or 5% to produce the six repetitions. The presentation format was held constant for each participant.

3.1.2. Materials and procedure

The problem was similar in form to that given in Experiment 1. The opening sentence read ‘1/1000 (5/1000, 10/1000) people have disease X’ on frequency versions, or ‘0.1% (1%, 0.5%, 1%) of people have disease X’ on probability versions. The three forms of question asked were worded as follows:

> Frequency
> How many people who test positive for the disease will *actually* have the

Table 2
Percentage of responses given in Experiment 2[a]

	COR	BR	1 − FPR	FPR	Other
Frequency question (*n*)					
BR-F (24)	54	29	0	4	13
BR-P (24)	33	29	3	7	28
Average	44	29	2	5	21
Probability question (*n*)					
BR-F (24)	68	4	15	1	12
BR-P (24)	42	10	5	8	35
Average	55	7	10	5	24
Proportion question (*n*)					
BR-F (24)	62	4	0	11	23
BR-P (24)	29	26	8	10	27
Average	46	15	4	10	25

[a] COR, correct; BR, base rate; FPR, false positive rate; BR-F, as frequency; BR-P, as percentage.

disease? __ out of __

Individual probability
What is the chance that a person who tests positive for the disease will *actually* have the disease? __%

Proportion
What proportion of people who test positive for the disease will *actually* have the disease? __%

3.2. Results

In Experiment 2, we collected six responses for each participant. These were categorized as in Experiment 1, taking into account the actual statistics presented on each problem. For example an answer of 98% would be classified as $1 - \text{FPR}$ provided that the false positive rate on that particular question was set at 2%. The results are shown in Table 2. Each participant received a score from 0 to 6 in each of the response categories; the data were analyzed using a 2 (base rate format) × 3 (question format) between participants ANOVA. The analysis showed that the format of the base rate significantly influenced the rate of correct responses ($F(1, 140) = 15.49$, $\text{MSe} = 6.03$, d.f. $= 1$, $P < 0.001$): performance was substantially better when base rates were presented as frequencies rather than percentages. The question format, however, did not produce a significant main effect ($F(2, 140) = 1.01$, $\text{MSe} = 6.03$) and nor was there a significant interaction ($F < 1$).

A similar analysis was run on the base rate only response. As can be seen from Table 2, this bias was strongly present when the question was asked in frequency form, regardless of the format of the base rate information. In the ANOVA both base rate format ($F(1, 140) = 3.95$, $\text{MSe} = 3.1$, d.f. $= 1$, $P < 0.05$) and question format ($F(2, 140) = 6.83$, $\text{MSe} = 3.1$, d.f. $= 2$, $P < 0.01$) were significant. Overall, the base rate response was given more often when the base rate statistic was presented as a probability rather than a frequency. Of course, there is more scope for bias here as correct responding is more common with the frequency format. The effect of the question format, as Table 2 shows, was that the base rate response was most common with the frequency question (29%), least common with the probability question (7%) and intermediate with the proportion question (15%). A Scheffé test revealed that only the difference between frequency and probability versions was significant on the paired comparisons. The $1 - \text{FPR}$, 'base rate neglect' response, was remarkably uncommon in this experiment, although most often observed, as in Experiment 1, when a probability question was asked.

3.3. Discussion

Experiment 1 confirmed our hypothesis that probability problems are not necessarily more difficult than frequency problems, and showed that in some cases that may actually be easier. The crucial factor seems to be the presence of cues to facilitate a set inclusion model: when present, problems are relatively easy and when absent, they are relatively harder, regardless of the presentation format. The findings of Experiment 2 also provide little encouragement for the hypothesis advanced by Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995) that frequency formats per se are easier than probability formats. Asking participants to express answers as a frequency did not lead to more correct responding compared with answers expressed as probabilities or proportions. It is true that in this experiment more correct answers were given when the base rate was expressed as a frequency rather than a probability, as Cosmides and Tooby would predict. However, since the format used was equivalent to the frequency easy types used in Experiment 1, we would also predict this trend on the basis of our set inclusion hypothesis. The relatively high correct responding even when the base rate was in percentage form suggests that many participants succeeded in recoding 0.1% as one in 1000, thus producing a frequency easy representation.

The most striking finding of Experiment 1 was that the question format produced a substantially different kind of bias when in frequency form than in probability form. Specifically, participants tended to give the base rate only when expressing their answers as a frequency. We had thought that this might be due to the congruence with base rate information which was also expressed as a frequency (one in 1000) in these experiments. The findings of Experiment 2 eliminate this explanation. The base rate only response was again most prevalent in the frequency question condition, but equally marked here with either type of base rate format. We return the reasons for this effect in Section 5, following presentation of our final experiment.

4. Experiment 3

The medical diagnosis problem used by Cosmides and Tooby (1996) and in our experiments to date involves a low base rate of one in 1000 throughout. As mentioned in Section 1, this could disguise a possible source of bias and lead us to overestimate our participants' competence in solving this kind of problem. We said earlier that the correct answer to the standard medical diagnosis problem was *about* 2%. Let us now look more precisely at the normative calculation involved. The question asked requires a posterior probability judgement: what is the likelihood of some hypothesis after inspecting some relevant data, in this case the result of a diagnostic test. Three parameters are relevant to this calculation:

HIT RATE (HR): The probability of the test coming up with a positive result when the hypothesis is true. This is set to 1 in the medical diagnosis problem.

FALSE POSITIVE RATE (FPR): The probability of the test coming up with a positive result when the hypothesis is false. This is set to 0.05 on the standard problem.

BASE RATE (BR): The prior probability that the hypothesis is true. This is set to 0.001 in the standard problem.

Positive test results can occur in one of two ways. First the individual may have the disease and will then always test positive. The chance of this, call it P1, is BR × HR ($0.001 \times 1 = 0.001$). Second, the individual may be free of the disease but nevertheless test positive. The chance of this, call it P2, is $(1 - \text{BR}) \times \text{FPR}$ ($0.999 \times 0.05 = 0.04995$). The posterior probability required is the chance that someone has the disease, given a positive rest result. This is given by P1 divided by P1 + P2 and this computes to be 0.0196. Hence the correct answer is 1.96% and not 2% as stated by Cosmides and Tooby.

The reader may think it is fussy to quibble over 0.04%, but our point is that with such a low base rate one cannot tell whether participants scored as correct are actually computing a correct Bayesian solution or not. The issue is whether they are applying the false positive rate to the entire 1000 people as opposed to the 999 who are free of the disease. We explore this issue in Experiment 3 by introducing problems with considerably higher base rates, where the difference between making this error or not becomes measurable. We also introduce a manipulation designed to explicate the subset to which the FPR applies which should help participants to formulate the correct set inclusion model of the problem. For example, if the base rate was 20% we made it clear from the wording that the FPR applied to the 80% who did not have the condition. In order to make higher base rates plausible, we introduced new problem content based on a search for genetic markers. We also presented different groups with frequency and probability versions, to permit further investigation of the effects of this variable.

In contrast with the earlier experiments, we used an open-ended response format in Experiment 3, so that participants could choose to express their answers in probability format (probabilities, proportions or percentages) or in frequency format (fractions or ratios using integers). We predicted that people who get the answer right would be more likely to express their answers in frequency format. This is because anyone reasoning about the correct subsets will come up with frequencies which are more easily combined into a fraction than expressed as a percentage which requires further computation. Conversely, those giving one of the common incorrect responses (e.g. base rate bias, or base rate neglect) do not need to combine information and hence have no computational difficulty in generating a percentage. For example, correct reasoning for the example problem presented in Section 4.1.2 is as follows:

> Out of 1000 people, 200 have the disease and will all test positive. 10% of the remaining 800 will also test positive, so there will be 280 positive testers in total. Hence, 200 out of 280 people testing positive will have the disease.

Such correct reasoning would normally lead to someone expressing their answer as 200 out of 280 or 20 out of 28. Why risk an arithmetic error in converting this to a percentage? On the other hand, someone who gives the base rate 200/1000 = 20% or the 1 − FPR response (90%) will have little difficulty in generating a percentage.

4.1. Method

4.1.1. Participants and design

One hundred and three student participants took part in Experiment 3, none of whom had been tested in our previous experiments, and were tested in groups. They were run in each of four separate groups in a 2 × 2 (between) × 2 × 4 (within) design. The two between participant variables were frequency/probability format and explication/no explication of the subset to which the FPR applied. Within participants we varied the base rate used on four levels (10, 20, 30 and 40%) and the FPR on two levels (10 and 50%). Thus, each participant solved eight problems, presented in an independently randomized order.

4.1.2. Materials and procedure

An example of the problems used in frequency format with a base rate of 20% and a FPR of 10% showing easy/hard variants is as follows:

> A test has been designed to detect the presence of genetic marker D. It is known that out of every 1000 people in the population 200 possess this marker. The test will produce a positive result in all such cases. However, the test also produces false positive results. Specifically, out of the 800 people who do not carry the marker, 80 will produce a positive test result (Easy). Specifically, out of 100 people who do not carry the marker, ten will produce a positive test result (Hard).

The probability version of the same problem was as follows:

> A test has been designed to detect the presence of genetic marker D. It is known that 20% of people in the population possess this marker. The test will produce a positive result in all such cases. However, the test also produces false positive results. Specifically, out of the 80% of people who do not carry the marker, 10% will produce a positive test result (Easy). Specifically, 10% of people who do not carry the marker will produce a positive test result (Hard).

In both cases the question was asked in an open-ended way, not cueing frequency or probability format:

> What is the chance that someone who tests positive will actually carry genetic marker D: __

The eight variants were presented on separate pages of a booklet in randomized order. They differed only with regard to the letter assigned to the genetic marker and the values of the base rate and false positive rate used. As in previous experiments, the hit rate was set to 100% throughout.

4.2. Results and discussion

In Table 3, we show the correct posterior probability calculations for our eight problems together with the erroneous ones which result from applying the FPR to

Table 3
Correct and incorrect posterior probability calculations for the problems used in Experiment 3

Base rate	False positive rate	Posterior probability (correct)	Posterior probability (incorrect[a])
10	10	0.526	0.500
20	10	0.714	0.667
30	10	0.811	0.750
40	10	0.870	0.800
10	50	0.182	0.167
20	50	0.333	0.286
30	50	0.462	0.375
40	50	0.571	0.444

[a] Computed with FPR applied to the whole set rather than to 1 − BR.

the whole set rather than to 1 − BR. It can be seen that due to the much higher base rates used in this study, there is a substantial divergence between the answers produced by the two sets of figures. Hence, it will be possible to discover whether or not participants make this error, even allowing a small margin in the answers given for approximation.

In Table 4, we tabulate the answers given by our participants to these problems. The first three columns show the frequency of correct responses, base rate only responses and base rate neglect responses, for comparison with our earlier experiments. The column labelled ‘wrong subset’ shows the frequency of participants giving the erroneous response shown in Table 3 by applying the FPR to the whole set. As the response format was left open, we also show whether people chose to express their answers using frequencies rather than percentages. Any response showing fractions or ratios with the actual frequencies involved was classified as a frequency response.

We ran significance tests on three of these columns (the same cautionary

Table 4
Percentage of responses in Experiment 3

	Correct	Base rate	1 − FPR	Wrong subset	Positive tests (%)	Other	*Frequency response*
Frequency presentation (n)							
Subset explication (28)	29	32	6	1	7	25	*29*
No explication (24)	11	19	16	17	3	32	*29*
Average	21	26	11	9	5	28	*29*
Probability presentation (n)							
Subset explication (26)	24	9	14	3	3	47	*20*
No explication (25)	21	9	9	12	1	47	*15*
Average	23	9	11	8	2	47	*17*

comments about the non-independence of these analyses applies as with previous experiments). As predicted, solution rates were higher when the subset relationship was explicated (27%) than when it was not (16%) ($t(102) = 1.78$, $P < 0.05$, one-tailed). However, the use of frequency or probability formats had no effect on the rate of correct responding. This provides further evidence against the hypothesis that frequency formats are generally easier. In this experiment, again, the frequency versus probability comparison was properly controlled so that the problems differed in no other regard. An analysis of variance on correct responding revealed no other significant factors or interactions.

The base rate only bias was significantly associated with use of a frequency question format in our previous experiments. In this experiment, participants were given an open format with which to provide their responses. However, the ANOVA on the base rate bias showed that the bias was strongly associated with frequency presentation format in this experiment ($F(1, 102) = 7.57$, MSe $= 6.22$, $P < 0.01$). This is a different finding from the previous effect of the question format found in Experiments 1 and 3. An interpretation is offered below in Section 5. A new measure of interest in this experiment was the 'wrong subset' response, in which participants gave a response that would be correct, but for the misapplication of the false positive rate to the whole set (see Table 3). As expected, explication of the subset relationship in the problem presentation substantially reduced this error ($F(1, 102) = 11.62$, MSe $= 3.02$, $P < 0.01$). This error was made on 15% of answers where no explication was used, but on only 2% when explication was present. This confirms our hypothesis that use of a small base rate as in the study of Cosmides and Tooby (1996) and in our previous experiments leads to an overestimation of participants' competence in Bayesian reasoning.

The use of the open response format in this experiment provides a new measure of interest, shown in the final column of Table 4. We expected that correct answers would more often be expressed in frequency format, a prediction that was strongly confirmed. Of the 25 participants giving frequency format answers, 48% responded correctly, whereas of the 79 using percentage format, only 13% gave correct answers ($t(102) = 4.74$, $P < 0.001$). Table 4 also reveals a weak trend for participants in the frequency presentation group to give more frequency answers than those in the probability presentation group. This trend fell short of significance ($\chi^2 = 2.58$, d.f. $= 1$). There was also no significant association between production of the base rate only response and the format in which answers were expressed.

5. General discussion

We started out with the rather bold claim that despite the extensive evidence offered by Cosmides and Tooby (1996) and Gigerenzer and Hoffrage (1995) we did not believe that the use of frequency rather than probability formats in statistical word problems makes them easier to solve in itself. Our view was that none of the experiments in these papers compared frequency and probability versions that were precisely controlled in other respects. We also argued that the reason that participants were more successful on the frequency versions was that these problems cued a mental model of set inclusion which made the problems much easier to understand. We believe that we have accumulated a significant amount of evidence for our claims in the current paper, using carefully controlled comparisons with problems presented to just over 500 separate participants.

First, we have provided a number of examples where frequency versions are no easier than the probability versions with which they are compared. In Experiment 1, frequency hard problems (lacking the set inclusion cues) were no easier (in fact, somewhat harder) than the probability versions and asking for responses in frequency rather than probability format was not significantly easier. In Experiment 2, asking for a response in frequency form led to no more correct answers than asking for a probability, whether the latter was expressed as a one case probability or a proportion. In Experiment 3 problems used either frequency or probability format throughout with an open-ended answer format: again there was no significant difference in accuracy between these conditions. We do not believe that these negative findings can be dismissed on statistical grounds in view of our large sample sizes: either the effect is not there, or it is very weak indeed.

We have also provided a second source of evidence in favour of our hypothesis that it is the cueing of a set inclusion mental model that facilitates performance on these problems. We introduced a new subset explication manipulation in Experiment 3. In this experiment we used base rates large enough to measure any error of application of the false positive rate. In the standard problem, the base rate is so low that people could incorrectly apply the false positive rate to the whole set rather than the subset excluded by the base rate, without measurable error. This means that Cosmides and Tooby (1996) could well have overestimated their participants' competence in Bayesian reasoning. We found in Experiment 3 that a number of participants gave answers consistent with this error. However, this erroneous answer was significantly reduced, and correct responses significantly increased when wording was used to explicate the subset to which the false positive rate applies.

In addition to the above results, we have another major finding in this study which goes against the hypothesis that people find it quite easy to reason with frequency versions of statistical word problems. Whilst the most often cited bias in the literature is 'base rate neglect', we have found substantial evidence of the opposite: a bias to give the base rate as the answer to the posterior probability question, neglecting the diagnostic evidence. This bias occurs repeatedly in our findings. In Experiments 1 and 2 it appears when people are asked to express answers in frequency format. In Experiment 3, where response mode was open-ended, the base rate bias was more marked on frequency presentations of the information. In all cases, the trends are large and significant.

We thought originally, given the results of Experiment 1, that the base rate bias might be due to the congruence of the question with the format of the base rate, but this explanation was eliminated by the findings of Experiment 2. Another explanation, consistent with all our findings, is the following. In order to express a probability as __ out of __ it is necessary to specify a denominator – the set of people out of whom those with the disease should be measured. This denominator is actually the total number of people with positive test results, a number which is not contained in the information given. (Girotto and Gonzales (2000), in a recent discussion of the medical diagnosis problem, also identify the need to focus on the correct subset of positive testers as the key to solving this problem.) If the participant takes 1000 as the denominator then they are doomed to failure. In Experiment 2 the number 1000 does appear in the text for all participants, including those who get the probability form of the base rate. This is because the frequency easy version was otherwise used throughout, so that the false positive information was described as 'out of every 1000 people who are perfectly healthy, X of them test positive for the disease'. In Experiment 3, the probability version was entirely probabilistic, with no reference at all to 1000 people, so the base rate only bias could only occur with frequency presentations.

In conclusion, our findings cast serious doubts upon the widely cited claim that frequency formats facilitate correct statistical reasoning in quantitative word problems. Instead we have shown that the manner of participants' responses to such tasks is strongly influenced by subtle variations in the presentation of task information which make it more or less easy to form a mental representation helpful to finding the normative solution.

Acknowledgements

This research was supported by a research grant (R000222085) from the Economic and Social Research Council of the United Kingdom.

References

Cascells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*, 999–1001.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: toward an evolutionary rigorous cognitive science. *Cognition*, *50*, 41–77.

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.

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

Fodor, J. (1983). *The modularity of mind*. Scranton, PA: Crowell.

Funder, D. C. (1987). Errors and mistakes: evaluating the accuracy of social judgements. *Psychological Bulletin*, *101*, 75–90.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow, & D. E. Over (Eds.). *Rationality*. London: Routledge.

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

Girotto, V., & Gonzales, M. (2000). Strategies and models in statistical reasoning. In W. Schaeken, G. De Vooght, & G. D'Ydewalle (Eds.). *Deductive reasoning and strategies*. Mahway, NJ: Erlbaum.

Lopes, L. L. (1991). The rhetoric of irrationality. *Theory and Psychology*, *1*, 65–82.

Reber, A. S. (1993). *Implicit learning and tacit knowledge*. Oxford: Oxford University Press.

Samuels, R. (1998). Evolutionary psychology and the mass modularity hypothesis. *British Journal for the Philosophy of Science*, *49*, 575–602.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stanovich, K. E., & West, R. F. (in press) Individual differences in reasoning: implications for the rationality debate. *Behavioral and Brain Sciences*.