Article:
McNair, S and Feeney, A (2014) When does information about causal structure improve statistical reasoning? Quarterly Journal of Experimental Psychology, 67 (4). pp. 625-645. ISSN 1747-0218

https://doi.org/10.1080/17470218.2013.821709

When does information about causal structure improve statistical reasoning?

Simon McNair[1,2] & Aidan Feeney[1]

[1] School of Psychology, Queen's University, Belfast BT7 1NN, United Kingdom.

[2] Centre for Decision Research, Leeds University Business School, Leeds LS2 9JT, United Kingdom.

Please address all correspondence to:

Simon McNair

s.j.mcnair@leeds.ac.uk

Tel: +44 (0)113 34 32685

Abstract

Base rate neglect on the mammography problem can be overcome by explicitly presenting a causal basis for the typically vague false positive statistic (Krynski and Tenenbaum, 2007). One account of this causal facilitation effect is that people make probabilistic judgements over intuitive causal models parameterised with the evidence in the problem. Poorly defined or difficult to map evidence interferes with this process, leading to errors in statistical reasoning. To assess whether the construction of parameterised causal representations is an intuitive or deliberative process, in Experiment 1 we combined a secondary load paradigm with manipulations of the presence or absence of an alternative cause in typical statistical reasoning problems. We found limited effects of a secondary load, no evidence that information about an alternative cause improves statistical reasoning, but some evidence that it reduces base rate neglect errors. In Experiments 2 and 3 where we did not impose a load, we observed causal facilitation effects. The amount of Bayesian responding in the causal conditions was impervious to the presence of a load (Experiment 1) and to the precise statistical information that was presented (Experiment 3). However, we found less Bayesian responding in the causal condition than did Krynski and Tenenbaum (2007). We conclude with a discussion of the implications of our findings and the suggestion that there may be population effects in the accuracy of statistical reasoning.

Introduction

One of the best-known conclusions from the Heuristics and Biases research programme (see Kahneman, Slovic & Tversky, 1982) was that when reasoning statistically about pairs of hypotheses, people fail to integrate the prior probability of the hypotheses with new evidence. Perhaps because this error, known as base rate neglect, was considered to be "...one of the most significant departures of intuition from normative theory" (Kahneman & Tversky, 1973, pg. 243), it has received an enormous amount of attention (for reviews see Koehler, 1996; Barbey & Sloman, 2007). Kahneman and Tversky (1973) proposed that base rate neglect occurs due to reasoners' inappropriate use of heuristics. A variety of more recent proposals suggest that base rate neglect is not evidence of poor statistical reasoning, but that it occurs due to the way that typical statistical word problems are formulated or worded. For instance, presenting statistical information as frequencies rather than percentages significantly decreases the tendency to neglect the base rate (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996). Other more recent work has shown that participants perform significantly better, irrespective of statistical format, when the nested set structure of the problem can be readily inferred; this can be achieved by simply rephrasing the problem (Macchi, 1995, 2000), or by providing a visual aid (Sloman, Over, Slovak, Stibel, 2003).

In this paper we will be concerned with Krynski and Tenenbaum's (2007) proposal that because people make statistical judgements in part on the basis of their understanding of the causal relationships that exist in the world, they appear to neglect information about base rates when the reasoning problem doesn't have a clear causal structure. In support of this proposal, Krynski and Tenenbaum presented evidence that when the causal relations in the problem are clarified, performance on a classic base rate neglect problem (Eddy, 1982) is significantly improved. This proposal is important because it derives from the claim that a Bayesian framework which also represents causal relationships between pieces of statistical evidence is more appropriate for evaluating and understanding people's statistical reasoning than is the traditional, purely statistical Bayesian framework. In this paper we describe our attempts to extend and generalise Krynski and Tenenbaum's evidence for their proposal and to examine whether the representations they posit are intuitively or deliberatively constructed during reasoning.

Base Rate Neglect

Consider the version of Eddy's (1982) widely-studied mammography problem that was used by Gigerenzer & Hoffrage (1995; for other uses of the problem see Cosmides & Tooby, 1996; Macchi, 1995, 2000; Lewis & Keren, 1999).

The probability of breast cancer is 1% for a woman at age forty who participates in a routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening.

What is the probability that she actually has breast cancer? ____ %

The correct response to the problem is 7.8%, yet Gigerenzer and Hoffrage (1995) reported that a majority of responses given by their participants were between 70% - 90%. To calculate the correct response one needs to use Bayes' rule (see Equation 1) which specifies a procedure for combining prior probabilities and new evidence in order to produce a statistical judgment.

$$P(H|E) = \frac{\mathrm{P(H)\ x\ P(E|H)}}{\mathrm{P(H)\ x\ P(E|H)} + \mathrm{P(\neg\ H)\ x\ P(E|\neg H)}} \qquad \text{Equation 1}$$

Reading Equation 1 from left to right, $\mathrm{P(H|E)}$ is the probability that the hypothesis H (i.e.

that the patient has breast cancer) is true in the light of the evidence E (a positive mammography). Crucial to this judgement is P(H), representing the base rate of breast cancer in the population (1%). Whilst P(E|H) – the probability of a positive mammogram given that the patient has breast cancer (i.e. true positive rate) – is high (80%), reasoners should realise that this figure is true only for 1% of the population that actually have the disease, but typically fail to do so. Positive mammograms also occur in those that don't have cancer, and so the correct judgement also requires reasoners to represent and combineP(¬H) –the base rate for the absence of cancer – and P(E|¬H) – the probability of a positive mammogram in the absence of cancer (i.e. false positive rate). Ultimately, the correct judgement computes the ratio of true occurrences of cancer and positive mammograms to the total instances (true + false) of positive mammograms. If one ignores the base rates, as did participants in Gigerenzer & Hoffrage's (1995) experiments, the estimate to the mammography problem is close to 90%.

Alternative Causes and the Causal Bayesian Framework

Krynski and Tenenbaum (2007) argue that, rather than simply extracting and computing the given statistical data in accordance with Bayes Theorem, people draw on causal knowledge when thinking about probabilistic evidence. According to their causal Bayesian framework of probabilistic judgement, errors such as base rate neglect will occur where the given statistical data cannot be clearly mapped onto a qualitative causal model representation of the evidence. The framework posits three sequential steps in statistical reasoning. First, people construct a qualitative mental representation which depicts the evidence as a series of causally connected nodes. Second, they parameterise these nodes with the given statistical information, and third, they compute the judgement in accordance with the Bayesian rule. Krynski and Tenenbaum argue that base rate neglect arises on the typical mammography problem as the given statistics do not clearly map with the causal representation that reasoners attempt to initially construct. Interestingly, they propose that this “mismatch” between nodes in the model and available statistical parameters occurs due to the false positive evidence typically not being given a causal status. In other words, people can connect cancer to positive mammographies in direct causal terms, but cannot represent false positive data about mammographies in equivalent terms. Without a means of including this evidence in one’s initial causal model, there is subsequently a “mismatch” between the nodes in the model and the statistical parameters, leaving people unsure of how to integrate the statistical data. Krynski and Tenenbaum assert that reasoners instead tend to focus on $P(E|H)$ and $P(E|\neg H)$, in many cases simply subtracting the latter from the former, leading to highly over-inflated estimates of the probability of cancer.

To test their account, Krynski and Tenenbaum compared performance on standard and causal versions of the mammography problem. In the causal version participants were told that false positive mammographies are caused by benign cysts. The argument now was that people could intuitively construct a causal model which connected both cancer, and cysts, to positive mammographies. The given statistical data now directly mapped to the parameters of the nodes in this model, in turn allowing participants to calculate the ratio of the probability of a positive test given that a woman has cancer to the overall probability of a positive result. Krynski and Tenenbaum carried out two experiments using these materials. In Experiment 1, only base rates for each cause were presented, and participants were told that most women who had either a benign cyst or breast cancer received a positive mammogram. In Experiment 2, participants were told that, of the 1% of women with cancer, 80% received a positive mammogram ($P(E|H) = .8$). In the causal condition participants were told that 30% of women had benign cysts and that 50% of those women tested positive, and in the standard condition that 15% of women without breast cancer tested positive. Thus, in both cases, the probability of a positive result in the absence of cancer ($P(\neg H)$ x $P(E|\neg H)$) was set at .15.

The results of these experiments are striking: in Experiment 1 where the problem was simpler, about 25% of participants who received the standard problem gave the Bayesian response whereas roughly 45% of participants who read the Causal version did. In their Experiment 2 where the problem was more complex, rates of Bayesian responding in the standard condition fell to roughly 15% whereas they remained at roughly 45% in the Causal condition. In both experiments Base rate neglect almost vanished in the Causal condition (< 10%) whereas about 30% of responses in the standard condition suggested that the base rates had been neglected.

Krynski and Tenenbaum's results support the adequacy of their normative account as a description of human statistical reasoning, and suggest novel ways in which good statistical reasoning might be facilitated, namely that the overall probability of the evidence should be calculated across all candidate causes of that evidence. Under their account, Equation 2 specifies how to arrive at a normatively correct estimate for problems such as the mammography problem. In this equation C refers to the focal cause (i.e. breast cancer) and alt C refers to the alternative cause.

$$P(C|E) = \frac{\mathrm{P(C) \ x \ P(E|C)}}{\mathrm{P(C) \ x \ P(E|C) + \ P(alt \ C) \ x \ P(E|alt \ C)}} \qquad \text{Equation 2}$$

Although Krynski and Tenenbaum (2007) designed their reasoning materials so that application of the statistical Bayesian norm and their alternative causal Bayesian norm gave the same answer, in many cases this will not be true. This is because the causal Bayesian account does not require a base rate for the complementary hypothesis (e.g. that a woman does not have breast cancer) but for an alternative cause (e.g. that a woman has a benign cyst). Because the base rate for the complementary hypothesis in their example is almost 1, application of the statistical Bayesian norm calls for the probability of a positive test and a benign cyst to be weighted by a probability close to one. However, if the base rate for cancer had been set at .2, then the probability of the complementary hypothesis would have been .8, and application of the statistical Bayesian norm would have required the probability of a positive test and a benign cyst to be weighted by .8. As the causal Bayesian framework does not require the $\mathrm{P(alt\ C)} \times \mathrm{P(E|alt\ C)}$term to be weighted by the base rate for the complementary hypothesis, the normative frameworks predict different answers in this case. So although Krynski and Tenenbaum (2007) cleverly designed their materials to that both normative frameworks would give the same answer, this is only possible when the base rate for the focal hypothesis is very low.

A consequence of the difference between the statistical and causal normative Bayesian accounts is that the statistics required to compute each answer may be different. In fact, the materials used in Krynski and Tenenbaum’s second experiment confound provision of information about an alternative cause with the particular statistical information that is described. So, participants in the causal conditions were provided with four statistics, including the base rate for the alternative cause, whereas participants in the standard condition were provided with three statistics and left to infer the base rate for the complementary hypothesis (P(¬H)). This observation is not meant as criticism; in order to provide participants with the percentage statistics required to calculate the causal Bayesian answer, Krynski and Tenenbaum had to provide them with information about the alternative cause rather than about the logically defined complementary hypothesis. One way to avoid a confound in the statistics that are presented would be to present participants in the causal condition with summary information about false positives. That is, rather than presenting participants with separate base rates and likelihoods for the alternative cause, they learn the percentage of participants who possess the alternative cause and a positive mammogram. As this statistic is identical to the false positive rate provided to participants in the standard problem, providing it to both sets of participants allows us to manipulate the provision of causal information whilst controlling for the particular statistics that are presented in each condition. We adopted this strategy in all three of the experiments to be described here.

Prior to Krynski & Tenenbaum's demonstrations, the role of causal information in base rate neglect had already been investigated. However, Krynski & Tenenbaum's claims differ from and go beyond that previous work. For example, Bar-Hillel (1980) showed that information about the likelihood of the evidence given the hypotheses appeared more causally relevant than the base rates and that simultaneously decreasing the causal relevance of the likelihoods and increasing the causal relevance of the base rates reduced base rate neglect (see also Ajzen, 1977). However, that earlier work was carried out in a strictly statistical Bayesian framework and treated causality as affecting the perceived relevance of statistical information. In line with other recent work carried out against the background of Pearl's (2000) ideas about causal models (for a review see Sloman, 2005), Krynski & Tenenbaum suggest that the success of participants' Bayesian judgments is contingent upon being able to intuitively represent the given evidence in causal terms; statistical problems which support this approach should yield higher levels of Bayesian responding.

Our initial rationale

To the best of our knowledge, Krynski and Tenenbaum's experiments are the only ones demonstrating causal facilitation effects (but see Hayes, Newell & Hawkins, 2013). Because of the potential theoretical and practical importance of those findings, our first aim was to extend and generalise them beyond the limited range of reasoning problems which were employed in the original experiments. Krynski and Tenenbaum (2007) propose that the causal Bayesian framework represents a more descriptive account of peoples' statistical reasoning than the traditional Bayesian norm; if this is the case then causal facilitation effects should be observed on novel problems which permit intuitive causal models to be easily parameterised with the appropriate data in Bayesian terms. Our second aim was to investigate, through the use of a secondary load manipulation, the extent to which the causal facilitation effect is intuitive in nature. Although Krynski and Tenenbaum do not claim that their effects have an intuitive basis, it is possible that providing participants with information about alternative causes improves statistical judgement because it permits an effortlessly constructed and parameterised causal model. Alternatively, the construction and parameterisation of a causal model may be effortful and require deliberative processing. In Experiment 1 we presented participants with sets of standard or causal problems whilst under cognitive load. If we assume that the imposition of a secondary load will arithmetic computation equally in causal and standard problems, then we should find an interaction between problem type and load size only if causal models are effortfully constructed and/or parameterised. However, other researchers (e.g. Ajzen, 1977; Tversky & Kahneman, 1980) have suggested that the effects of causal knowledge on statistical judgement are mediated via a causal heuristic, and this possibility is not disputed by Krynski and Tenenbaum (2007, see pg. 447). If the role of causal knowledge in statistical reasoning is intuitive or heuristic, then putting reasoners under load should not interfere with the size of the causal facilitation effect, that is, participants should be able to represent the evidence in clear causal terms, irrespective of load.

Experiment 1

Method

Participants: There were 64 participants (19 male), aged between 18 and 40 years. Participants received no payment and were recruited around Queen's University Belfast through word of mouth and online advertising.

Design: The experiment had a 2 (Problem Type: Causal vs. Standard) x 2 (Load: High load vs. Low load) mixed design. Problem Type was manipulated between subjects and Load within subjects.

Materials: Although each participant saw only eight reasoning problems, a total of 16 reasoning problems were used in this experiment (see Appendix 1). Two of these were taken from the literature, Gigerenzer and Hoffrage's (1995) adapted version of Eddy's (1982) mammography problem, and Tversky and Kahneman's (1982) cab problem. The remainder were new problems. We assigned each of the 16 problems at random into one of two sets of eight problems, and equal numbers of participants saw each set. There was a Causal and a Standard version of each problem. The only difference between these versions was whether an alternative cause was given for P(E|¬H). As problem type was manipulated between subjects, participants saw either eight Standard or eight Causal problems. Because of the likely difficulty of reasoning statistically under load, unlike Krynski & Tenenbaum (2007), we presented statistical information in natural frequency format rather than percentage format, although we asked participants to provide a percentage answer. Presenting data in natural frequency form has been shown to significantly improve Bayesian performance (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996), mainly because it permits a simpler Bayesian calculation on the basis of fewer statistical parameters (3 as opposed to 4). Our Causal mammography problem can be found below.

Causal Mammography Problem

> Suppose the following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

> 10 in every 100 have breast cancer at the time of the screening. 8 in every 10 of those with breast cancer will receive a positive mammogram. However, a dense but benign cyst, which looks like a cancerous tumour on the X-ray, can cause positive mammographies in those without cancer; this occurs in 20 out of every 90 mammograms without cancer.
>
> Of those that receive positive mammographies, what % would you expect to have cancer?

The Standard version read identically save for one line: *"However, 20 in every 90 of the remaining cases without cancer will still receive a positive mammogram."* Our Causal problems presented reasoners with the number of those with cancer (base rate), the number of those with cancer and positive mammographies (true positive), and finally the number of those without cysts who receive positive mammographies (false positive). Note that the same statistical information is supplied in both versions of the example problem above, and that in both cases the correct answer is the same regardless of whether one applies a statistical or a causal Bayesian norm:

$$\frac{\text{Breast cancer and positive result}}{\text{Breast cancer and positive result} + \text{Cyst and positive results}}$$

Before each reasoning problem, participants were shown a 3 x 3 grid containing a pattern of dots (see also Franssens and De Neys, 2009; Verschueren, Schaeken & d'Ydewalle, 2004). In the complex condition (see Figure 1a) the patterns contained four dots in randomly determined grid positions. In the low load condition (see Figure 1b) these patterns were simple continuous lines of three dots. Each grid was displayed for 1000ms, giving participants enough time to perceive the pattern clearly. Their task was to keep the displayed pattern in memory whilst calculating an answer to the reasoning problem. After submitting an answer to the reasoning problem participants then had to recall the dot pattern, from memory, on a blank 3 x 3 grid, and received feedback as to their recall performance.

Problem order of presentation was randomised for each participant. However, Load was blocked so that half of the participants attempted four low load trials first and the other half received the high load trials first.

Figure 1. Example dot patterns used in (a) high load, and (b) low load conditions of Experiment 1.

Procedure: The experiment took place in a small computer lab in QUB, where participants were tested individually or in groups of up to six. Participants were told that there was no time limit for completion of the experiment, and that use of calculators or pens was prohibited. Most participants completed the experiment within 25 minutes. All parts of the experiment were presented in E-Prime.

Results

Data Coding: To allow for calculation errors, responses within 5% of the correct Bayesian answer were coded as Bayesian, whilst those within 5% of the statistical Bayesian account estimate produced using only the likelihoods were coded as base rate neglect. Responses within 5% of the estimate produced by using the focal base rate only were labelled as likelihood neglect. We coded a likelihood neglect category because previous work has shown that people sometimes overweight the base rates instead of neglecting them (see Evans, Handley, Over & Perham, 2002). All remaining responses were labelled as other. The problems we used in the experiment were constructed so that these response categories never overlapped. Note also that our system for coding responses is different from that used by Krynski and Tenenbaum (2007) who categorised only exactly correct answers as Bayesian, and answers greater than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H)}$ as base rate neglect. They coded all other answers as other.

Secondary Task Analysis: In order to test whether participants attended to the secondary task in this experiment, we carried out a 2 (Problem Type) x 2 (Load) mixed design ANOVA on the number of dot patterns that participants correctly remembered. Consistent with participants having attended to the patterns, we found a main effect of Load, F(1, 62) = 451.03, p < .001, $\text{eta}^2 = .45$, such that fewer complex patterns were remembered (mean = 2.64, S.D. = 1.10) than simple patterns (mean = 3.64, S.D. = .57). None of the other effects tested by the analysis approached significance. Examination of secondary task performance revealed large individual differences. For example, five participants correctly completed four or fewer response grids, 13 correctly completed five response grids and only 15 participants correctly recalled all eight. These individual differences suggest considerable variation in the degree to which participants attended to the secondary task. Accordingly, in all of the analyses of performance on the statistical reasoning tasks described below, we included overall performance on the secondary task as a covariate.

Primary task analysis: Examination of Appendix 2 suggests that performance across the 16 different reasoning problems used in this experiment was similar. The mean number of each type of response, broken down by Problem Type and Load, is to be seen in Figure 2. We carried out separate 2x2 mixed design ANCOVAs on Bayesian, base rate neglect and likelihood neglect responses and in each case overall performance on the secondary task was included as a covariate. The analysis of Bayesian responses revealed a marginally significant main effect of Load, F(1, 61) = 2.99, p < .09, eta^2 = .05, with reasoners producing fewer correct responses under high load (mean = .88, S.D. = 1.22) than under low load (mean = 1, S.D. = 1.08). In addition, the interaction between Load and secondary task performance was marginally significant, F(1, 61) = 3.92, p < .06, eta^2 = .06. None of the other effects tested by the ANCOVA approached significance.

Figure 2. Mean rates of each Response Type (out of 4), broken down by Load and Problem Type, for Experiment 1. Error bars represent the standard error of the mean. N.B. BRN = base rate neglect, LN = likelihood neglect.

To explore the marginally significant interaction we performed a median split on overall performance on the secondary task and found an effect of load on Bayesian responding in participants who performed well on the secondary task, t(28) = 2.77, p < .01, Cohens d = .51, but no effect of load in participants who performed less well on the secondary task, t(34) = .13, p > .85, Cohens d = .03. As recently recommended by Simmons, Nelson, & Simonsohn (2011), we reanalysed our data excluding the covariate. Unsurprisingly, given that the marginally significant results in the ANCOVA involved the covariate, there were no significant effects detected by this analysis (all Fs < 1.5, all ps > .2). The analysis of base rate neglect responses revealed a significant effect of Problem Type only, F(1, 61) = 7.69, p <

.01, $eta^2 = .11$ (all other ps > .1), with significantly more base rate neglect on Standard problems (mean = 1.85, S.D. = 1.41) compared to Causal problems (mean = .94, S.D. = 1.21). This effect remained even when the covariate was omitted from the model, F(1, 62) = 7.53, p < .01, $eta^2 = .11$ (all other Fs < 1, all ps > .4). None of the effects tested by the analysis of likelihood neglect responses were significant.

Data on individual differences in responding are to be found in Table 1, where it may be seen that almost all participants gave mixed responses. Only one participant gave the Bayesian response on every trial and only eight gave a majority of Bayesian responses.

(Table 1 here)

Discussion

Experiment 1 failed to produce any evidence that providing information about an alternative cause for false positives improves reasoning. At best, participants were significantly less likely to commit base rate neglect when information about an alternative cause had been provided. We found some evidence that imposition of a secondary load affected participants' responding. In particular, amongst participants who did well on the secondary task there were significantly fewer Bayesian responses under heavy load. This suggests that our failure to find an overall effect of load may have occurred because a substantial number of participants prioritised the primary over the secondary task.

Without finding a causal facilitation effect, it is difficult to comment on how exactly heavy load interfered with reasoning. Although there appear to be individual differences in the degree to which participants engaged with the secondary task, it is also possible that the relatively weak effects of our Problem Type manipulation may have been due to the imposition of a secondary load. For example, it is possible that even a light load resulted in substantial decrements in performance on the primary task. Because the results of this experiment were unexpected, we ran a second experiment in which we manipulated problem type and load entirely within participants. Forty participants completed 16 reasoning problems, four in each cell of the design. Load size was blocked and block order was counterbalanced. We found no effects of either variable in this experiment and when collapsed across load conditions, rates of correct responding to the Causal (25%) and Standard (22%) problems were very similar to those observed in Experiment 1. In the light of this second failure to find either a causal facilitation effect or a clear effect of load, we adopted new aims for the remaining experiments.

Experiment 2

Although the initial rationale for our experiments was to extend evidence for a causal facilitation effect in statistical reasoning, and to examine whether the construction and parameterisation of causal models is effortful or effortless, the results of Experiment 1 and the follow-up experiment we have informally described, caused us to focus in our subsequent experiments on establishing whether, and under what circumstances, a causal facilitation effect could be produced with our materials. Our materials were subtly different to those used by Krynski & Tenenbaum, and it is possible that this subtle difference is obscuring the effects of providing information about an alternative cause. Alternatively, it may be that the load manipulation has interfered with our ability to detect the effect. In this experiment we abandoned the load manipulation, but continued to use our versions of the problems where participants learned either $P(E \cap \neg H)$ in the control condition, or $P(E \cap \text{alt } C)$ in the causal condition.

In addition to abandoning the load manipulation, in Experiment 2 we manipulated the nature of the response required from participants. In order to reduce the processing load required to give a correct response, we asked half of the participants in Experiment 2 to choose the correct response from an array of four. We expected this change to response requirements to produce an overall improvement in reasoning. Of greater interest was whether it would produce greater improvement with Causal problems than with Standard problems.

Experiment 2 also afforded us the opportunity to examine the effects on our results of coding participant responses in exactly the same way as Krynski & Tenenbaum. Whereas in Experiment 1 we used a four category coding system (Bayesian, Base Rate Neglect, Likelihood Neglect and Other), Krynski & Tenenbaum used three categories (Bayesian, Base Rate Neglect, and Neither). We also used different criteria for assigning responses to categories than did Krynski & Tenenbaum. We used the same decision rule for each category (to allow for calculation errors, when assigning responses to categories we accepted those within 5% of a particular answer), whereas they used different decision rules. In Krynski & Tenenbaum's coding scheme, only exactly correct answers were placed in the Bayesian category, whereas answers greater than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H)}$ were assigned to the base rate neglect category (see also Macchi, 1995, 2000). Because, for some of the problems in Experiment 1, the Bayesian probability estimate was higher than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H\,)}$, we were unable to compare results obtained using the two coding systems. In the problems selected for Experiment 2, the Bayesian estimate was always lower than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H\,)}$, so we will be able to consider the effects of coding strategy on our results.

Method

Design: The experiment had a 2 (Problem Type: Causal vs. Standard) x 2 (Response Format: Free vs. Multiple Choice) between subjects design.

Participants: One hundred participants (52 male), who ranged in age from 18 to 45 years, were recruited at a number of locations at Queen's University Belfast.

Materials: Participants attempted a subset of four of the problems that had been used in Experiment 1(see Appendix 2). Half of the participants saw Standard versions of these problems and half saw Causal versions. Reasoning problems were now presented on paper with participants giving written responses. At the end of each problem participants in the Free Response condition were asked to state the percentage chance of a particular outcome (e.g. Of those that receive positive mammographies, what % would you expect to have cancer?). Participants in the Multiple Choice condition were given a selection of four potential answers presented as percentages. Each of the four answers reflected one of the four categories of response for which we coded in Experiment 1. "Other" responses in this case were random numbers that were separated from each of the other answers by at least 10%.

Procedure: Participants were approached and asked whether they would participate. They were then given a page which contained the four reasoning problems. Participants were asked not to use calculators, and were advised not to take longer than ten minutes to complete the problems.

Results

So as to be able to compare the effects of coding scheme on our results, we coded responses using both the four-category system and Krynski & Tenenbaum's three-category scheme.

Four-category coding: Responses were initially coded in the same way as for Experiment 1. Examination of Appendix 2 reveals that participants' responses were similar across reasoning problems. Figure 3 shows the mean level of each response broken down by Problem Type for each Response Format. A 2 (Problem Type) x 2 (Response Format) ANOVA on Bayesian responses revealed a significant main effect of Problem Type, F(1, 96) = 23.24, p < .001, eta^2 = .2, with significantly more Bayesian responses on Causal problems (mean = 1.56, S.D. = 1.09) compared to Standard problems (mean = .82, S.D. = .80). There was a significant main effect of Response Format on Bayesian responses, F(1, 96) = 39.01, p <.001, eta^2 = .29, such that more Bayesian responses were given for Multiple Choice problems (mean = 1.7, S.D. = .94) than for Free Response problems (mean = .69, S.D. = 86).

A 2 x 2 ANOVA on rates of base rate neglect revealed a significant main effect of Problem Type, F(1, 96) = 17.13, p <.001, eta^2 = .15, with participants neglecting the base rate more often on Standard problems (mean = 1.28, S.D. = 1.05) compared to Causal problems (mean = .62, S.D. = .66). There was also a significant main effect of Response Format, F(1, 96) = 13.49, p = <.001, eta^2 = .12, such that participants neglected the base rate more often when they had to produce a response (mean = 1.25, S.D. = .98) than when they had to choose one (mean = .65, S.D. = .8).

Figure 3. Mean rates of each Response Type (out of 4), broken down by Response Format and Problem Type, for Experiment 2. Error bars represent the standard error of the mean. N.B. BRN = base rate neglect, LN = likelihood neglect.

None of the other effects tested by these analyses were significant nor were any of the effects tested by analyses of rates of likelihood neglect and other responses. Analyses of individual differences in consistency of responding (see Table 1) showed that all but two participants gave mixed responses across the four problems in this experiment.

Three-category coding: Free responses were recoded using Krynski and Tenenbaum's (2007) three-category coding scheme. Only the exact answer was coded as a Bayesian response, any response greater than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H)}$ was coded as a base rate neglect response, whilst any remaining responses were coded as other. Failing to make allowances for calculation errors when assigning responses to the Bayesian category resulted in near-negligible rates of Bayesian responding on Standard problems (mean = .08, S.D. = .28), with little improvement on Causal problems (mean = .15, S.D. = .37). A series of 2 x 2 ANOVAs revealed no effect of Problem Type on Bayesian responses ($p < .4$) but an effect of Problem Type on the tendency to neglect the base rate, $F(1, 48) = 6.46$, $p < .02$, $\text{eta}^2 = .16$ (Causal mean = 1.65, S.D. = 1.06. Standard mean = 2.63, S.D. = 1.21). Problem Type also significantly affected the rate of other responses, $F(1, 48) = 8.835$, $p < .01$, $\text{eta}^2 = .16$ (Causal mean = 2.19, S.D. = 1.02. Standard mean = 1.29, S.D. = 1.12).

Discussion

The results of Experiment 2 show a causal facilitation effect: participants who received problems describing an alternative cause were more likely to approximate the Bayesian response and less likely to neglect the base rate than were participants who did not receive information about an alternative cause. Although we found an effect of whether participants produced or chose a response, the size of the causal facilitation effect was the same in both response formats. These results show that even when participants are presented just with information about an alternative cause for the false positive rate rather than with information about the base rate and likelihood for the alternative cause (as was the case in Krynski and Tenenbaum’s study), a causal facilitation effect may be observed.

It is important to note that, for participants asked to produce responses, the rate of Bayesian responding in the Causal condition (24%) was very similar to that observed in Experiment 1. On the other hand, the rate of Bayesian responding in the Standard condition (11%) was considerably lower than we observed in Experiment 1. We will return to this finding in the General Discussion.

A causal facilitation effect was not observed when responses were coded using Krynski & Tenenbaum’s three-category scheme. This is because a near-negligible number of responses were assigned to the Bayesian category when only the exact answer was coded as Bayesian (4% on Causal problems and 2% on Standard problems). Whilst this may be a result of requiring people to make more difficult calculations than Krynski and Tenenbaum, a clear disadvantage of the three-category scheme here is that it does not make allowances for calculation errors. However, results obtained using the four-category scheme, which allows for approximation, strongly suggest that Causal problems produce more approximately Bayesian responses. It is striking that even when allowances are made for approximation, rates of Bayesian responding on causal versions of our problems are much lower than were observed by Krynski and Tenenbaum (2007). Experiment 3 was designed investigate one possible cause of this difference between our results and theirs.

Experiment 3

The experiments described thus far have used materials in which the statistical information presented in the Standard and Causal problems has been identical: the base rate for breast cancer, the frequency of positive tests given breast cancer and the frequency of positive tests given a benign cyst/the absence of breast cancer. Krynski and Tenenbaum, on the other hand, presented different statistics in the two conditions. In the Standard condition, participants learned the base rate for breast cancer, and the likelihoods for a positive mammography given the focal and complementary hypotheses. In the causal condition they learned the base rate for cancer, the base rate for the alternative cause, and the likelihood of a positive test given each of these causes. Although the problems were designed so that they produced the same answer, Krynski and Tenenbaum argued that the differences between the two forms of the reasoning problem would lead to different causal models parameterised in different ways. It is possible that the relatively low rate of Bayesian responding we have observed may be due to the difference between the statistics presented to participants in our Causal problems and those presented by Krynski and Tenenbaum. Accordingly, in Experiment 3 we attempted to directly replicate Krynski and Tenenbaum's Experiment 2. In addition, we included a second causal condition, where participants were presented with causal materials similar to those we have used in Experiments 1 and 2. In these materials participants learned $P(E \cap \text{alt } C)$. Although all statistics in this experiment were presented in the form of percentages, the problems in this last condition are structurally identical to the Causal problems in our earlier experiments. Comparison of this three parameter causal condition to Krynski and Tenenbaum's four parameter causal condition will allow us to draw conclusions about whether the precise way that the underlying causal model is parameterised is important to the rate of Bayesian reasoning that is observed.

To allow for a complete replication of Krynski and Tenenbaum's Experiment 2, participants in this experiment initially attempted one of the mammography problems from that experiment. Responses on these problems were analysed separately. Next participants in each of the three conditions solved a further three problems. The predictions for this experiment are straightforward: if the precise statistics are important to the causal facilitation effect, then we should observe more Bayesian responding in the four parameter causal condition than in the three parameter causal condition, and a bigger facilitation effect in the former case. On the other hand, if merely providing a causal basis for false positives underlies the effect, then we should find no difference between the two causal conditions and an approximately equal facilitation effect caused by both.

Method

Participants: 126 participants (57 male, 69 female) aged between 18 and 33 years old were recruited. Participants were approached at a number of locations on the university campus.

Design: The experiment had one between subjects independent variable, Problem Type, manipulated at three levels: four parameter Causal vs. three parameter Causal vs. Standard. As in Experiment 2, problems were presented in a booklet, and participants gave written percentage probability estimates.

Materials: Experiment 3 used the same four problem scenarios as in Experiment 2. There were three versions of each problem. We used Krynski and Tenenbaum’s own causal (i.e. four parameter, see below) and Standard version of the mammography problem, with our own three parameter causal version expanding upon the Standard version by explicitly explaining a causal basis for false positives. As in Krynski and Tenenbaum’s materials, all data were now presented as percentages.

Four parameter causal mammography problem

> Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer. From past statistics, the following is known:
>
> 1% of the women have breast cancer at the time of screening. Of those with breast cancer, 80% received a positive result on the mammogram.

> 30% of the women had a benign cyst at the time of the screening. Of those with a benign cyst, 50% receive a positive result on the mammogram.
>
> All others received a negative result.
>
> Suppose a woman gets a positive result during a routine mammogram screening. Without knowing any other symptoms, what are the chances she has breast cancer?

In the three parameter Causal problem, false positive information was presented as *"*Of those without cancer, 15% received a positive result on the mammogram due to having a benign *cyst."* In the Standard version, the information was presented as *"Of those without cancer, 15% received a positive result on the mammogram."* As the Standard and Causal problems used in Experiments 1 and 2 expressed data as frequencies, both permitted a simplified calculation of the Bayesian answer, however, calculation of the statistical Bayesian answer in the current experiment required reasoners to represent the given false positive rate as the conditional probability that anyone without cancer may receive a positive test, and to recognise that they must also utilise the compliment of the focal base rate (e.g. 99% of people don't have cancer). In the four parameter Causal problem, calculation of the causal Bayesian answer required integration of the base rate and likelihood for the alternative cause, and in the three parameter Causal problem, it required integration of information about P(C ∩ alt cause).

Because Krynski and Tenenbaum set the focal base in the mammography problem at 1%, the statistical and causal Bayesian answers to that problem were the same. In the remaining problems, we varied the base rate which meant that the statistical and causal Bayesian answers were different. However, the difference between the answers was never more than 6%.

Results

Data Coding: For all of the analyses reported here, we used the causal Bayesian equation (Equation 2) to determine the correct answer on three parameter and four parameter causal problems. We used the standard Bayesian equation (Equation 1) to determine the correct answer on standard problems. For all problem types, estimates greater than $\mathrm{P(E|H)} - \mathrm{P(E|\neg H)}$ were coded as base rate neglect, and all other responses were coded as Other. To allow for an exact replication of Krynski and Tenenbaum's Experiment 2, we analysed responses on the mammography problem separately using Krynski and Tenenbaum's coding method. In the analysis of the remaining problems, because of the very low rates of exactly correct Bayesian responses in Experiment 2, we allowed for calculation errors by coding answers within 5% of the Bayesian response as correct. Note that this analysis was complicated somewhat by the fact that Bayesian responses on the Standard problems were now different to Causal problems, because correct answers on the Standard problems were defined with respect to the statistical Bayesian norm. This did not affect the mammography problem where, as noted, Krynski and Tenenbaum circumvented the issue by setting a focal base rate of 1%.

Mammography problem: Results indicated that very few reasoners gave the exact Bayesian answer on the mammography problem. Although they performed poorest on the Standard mammography problem, and best on the four parameter causal problem (see Figure 4), chi square analysis did not yield a significant association between Problem Type and type of response (p >.3). In further analyses, the rates of each response were compared more directly across each combination of Problem Types. In each instance, analysis failed to indicate a significant association between Problem Type and type of response (all ps > .1). Thus, we have failed to replicate the causal facilitation effect using Krynski and Tenenbaum's materials and coding procedure.

Figure 2. Proportions of each Response Type on the mammography problem, broken down by Problem Type, for Experiment 3. N.B. BRN = base rate neglect.

Remaining problems: The proportions of each type of response on the three remaining reasoning problems, coded as described above, are presented in Figure 5. A one-way ANOVA on rates of Bayesian responses indicated a marginally significant main effect of Problem Type: F(2, 123) = 2.94, p <. 06, eta^2 = .05. Post-hoc t-tests on the means involved in the marginally significant main effect revealed significantly more Bayesian responses (mean = .67, S.D. = .82) in the four parameter Causal condition than in the Standard condition (mean = .29, S.D. = .71): t(82) = 2.28, p = <.03, Cohen's d = .49. The difference between the three parameter Causal condition (mean = .62, S.D. = .82) and the Standard condition was marginally significant (p < .06). As the means suggest, the difference between rates of Bayesian responding in the Causal conditions was non-significant. A one-way ANOVA on base rate neglect responses revealed a significant main effect of Problem Type: F(2, 123) = 5.72, p = <.01, eta^2 = .1. Post-hoc t-tests revealed a significant difference between the Standard (mean = 1.55, S.D. = 1) and three parameter Causal condition (mean = 1.1, S.D. = 1): t(82) = 2.02, p < .05, Cohen's d = .44. A significant difference also existed between the Standard and four parameter Causal condition (mean = .83, S.D. = .8): t(82) = 3.26, p <.01, Cohen’s d = .76. Analysis of the other response types revealed no significant differences between means. Figure 5 shows the mean number of each type of response according to Problem Type for the three non-mammography problems.

Figure 2. Mean rates of each Response Type (out of 3), broken down by Problem Type, on the non-mammography problems for Experiment 3. Error bars represent the standard error of the mean. N.B. BRN = base rate neglect.

Table 2 shows how participants varied in the numbers of each type of response given across each Problem Type.

(Table 2 here)

Discussion

Although we failed to replicate Krynski and Tenenbaum's effect using their materials, nonetheless Experiment 3 has provided more evidence that information about the causal basis of the false positive statistic facilitates Bayesian reasoning. In addition, although the rates of Bayesian responding observed in this experiment are quite similar to those observed in our earlier experiments, there was no evidence that the precise statistics provided to participants about the alternative cause increased the number of Bayesian responses. Participants were almost as likely to give a Bayesian response when told $P(E \cap \text{alt } C)$ as when told $P(\text{alt } C)$ and $P(E|\text{alt } C)$. It appears then, that the causal facilitation effect is due to provision of additional causal information rather than to the particular statistics that accompany that causal information.

General Discussion

The experiments in this paper were designed to further explore whether errors on typical statistical reasoning problems, such as the mammography problem, can be reduced by explicitly detailing a cause for false positive outcomes. We found that provision of additional causal information improved Bayesian responding in Experiments 2 and 3 and reduced the rate of base rate neglect in Experiments 1 and 2. These results suggest that the causal facilitation effect is real, but because the rate of Bayesian responding in our causal conditions never exceeded 25%, they also suggest that the effect of providing additional causal information may not be as universally powerful as was first assumed. The results of Experiment 3 show that the reduction in the rate of Bayesian responding in our causal conditions cannot be attributed to differences between our experiments and Krynski and Tenenbaum's in the nature of the statistical information that was provided. It appears that merely providing people with the cause for false positives allows some of them to integrate the statistics in the problem. One similarity between our results and Krynski and Tenenbaum's is the remarkable consistency in rates of Bayesian responding on causal problems found in each study. Despite changes to experimental design, procedures, and materials, roughly one in four of all responses to causal Bayesian problems in our experiments were approximately Bayesian. In both of their mammography experiments, 45% of all responses were exactly Bayesian.

One interpretation of the consistency within each set of experiments is that providing information about an alternative cause facilitates the intuitive construction and parameterisation of a causal model. That is, regardless of the number of statistics, the response format, or the number of problems attempted, when some participants read additional causal information they spontaneously include it in a qualitative causal model of the problem. Experiment 1 attempted to investigate this more directly through use of a dual task procedure found by Franssens and De Neys (2009) to increase base rate neglect errors as a result of intuitive thinking. It is noteworthy that even amongst those participants who were affected by a secondary load in Experiment 1, size of load did not interact with whether the problem was Causal or Standard in determining response type. This finding, together with the consistency in correct responding between load and non-load experiments, could be taken to suggest that the facilitating effect of additional causal information operates at a level that is not subject to factors such as working memory limitations, and is intuitive in nature.

Despite both our study and Krynski and Tenenbaum's indicating consistency in correct responding on causal statistical problems, important contrasts exist between our findings and theirs. In particular, we found less correct responding on causal problems (around 20% – 25%) than they did (around 45%). This at least partly accounts for the failure to find a causal facilitation effect in Experiment 1, as rates of correct responding on Standard problems were also over 20%, but fell sufficiently in latter experiments such that a significant difference existed between problem types. One possible explanation for the diminished rates of Bayesian responding that we observed on causal problems is that we sampled participants from a population with lower underlying ability than did Krynski and Tenenbaum (2007). Brase, Fiddick, and Harries (2006) have shown that undergraduate students from top-tier universities give more Bayesian responses than students from mid-tier universities, whilst still other work has indicated that differences in the ability of successive samples drawn from the same student population may also underlie differences in reasoning performance (Newstead, Handley, Harley, Wright, & Farrelly, 2004). More recently, it has been shown that the facilitating effects of presenting statistical information as natural frequencies are either limited to, or much more prevalent in participants with better numerical skills (e.g. Chapman & Liu, 2009; Galesic, Gigerenzer & Straubinger, 2009; Sirota & Jaunchich, 2011). All of this evidence, when considered alongside the differences between our results and Krynski and Tenenbaum’s, suggests that a study of individual or population differences in susceptibility to causal facilitation effects would be worthwhile.

Krynski and Tenenbaum’s original experiments were motivated by the view that people compute Bayesian statistics over causal models. This is a different normative model to the classical Bayesian model and calls for reasoners to consider the probability of alternative causes rather than the probability of a complementary hypothesis. The problems that they used confounded the provision of information about an alternative cause with provision of the statistics required to parameterise a causal model representing that alternative cause. That is, the problems that they used led to different models parameterised in different ways. One advantage of the materials that we have used here is that in both conditions, the correct solution was arrived at in the same way. In three experiments we found that alerting people to the causal basis for false positives either reduced particular types of error or significantly increased Bayesian responding. Experiment 3 showed no difference in the size of the causal facilitation effect based on whether participants attempted Krynski and Tenenbaum's materials or ours. This may be interpreted as evidence for the claim that the causal facilitation effect is due to the presence of additional causal information cueing some participants to integrate information about the focal and alternative causes in the scenario. The effect does not seem to be dependent on the particular statistics that are presented in the causal scenario.

In conclusion, the current research has supported Krynski and Tenenbaum's contention that expecting reasoners to conduct Bayesian inference purely on the basis of statistical data is at odds with how people intuitively think about statistical reasoning problems. Instead, Bayesian judgements appear to be facilitated whenever the given evidence provides a causal basis for false positive information, suggesting that being able to represent all of the evidence in a coherent causal representation is a key initial step in how people think about integrating probabilistic information. Compared to typical statistical word problems with poorly defined false positive data, causal problem materials produced significantly more accurate Bayesian judgements. The exact nature of how this information is expressed – be it as a single figure with a defined cause, or as a base rate and conditional probability – appears not to affect judgement accuracy. What is clear, then, is that participants' understanding of the false positive information has an importance influence on how they think about integrating the probabilistic data. That performance on causal problems in the current work was immune to variations in design and procedure perhaps supports the idea that reasoners intuitively represent causal structure when thinking about statistical judgements. Such a conclusion is consistent with other results which suggest that causal information affects statistical reasoning at an intuitive level (see Crisp & Feeney, 2009). However, the improvements in reasoning which come with clarifying the causal basis of the evidence may not be as large in every population as originally reported by Krynski and Tenenbaum (2007). Overall, the current work adds to the picture that is emerging in which causal relations play a central role in cognition (for reviews see Sloman, 2005; Gopnik, 2012; Waldmann & Hagmayer, in press). Future work might explore why additional causal information facilitates intuitively constructed representations in some people but not in others.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. Journal of Personality and Social Psychology, 35, 303-314. doi : 10.1037//0022-3514.35.5.303

Barbey, A. K., & Sloman, S.A. (2007). Base-rate respect: From ecological rationality to dual processes. Behavioral and Brain sciences, 30, 241-54. doi : 10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The Base Rate Fallacy in Probability Judgements. Acta Psychologica, 44, 211 - 233. doi : 10.1016/0001-6981(80)90046-3

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. Journal of Personality and Social Psychology, 45, 792-804. doi:10.1037//0022-3514.45.4.792

Brase, G. L., Fiddick, L., & Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. Quarterly Journal of Experimental Psychology, 59, 965-76. doi : 10.1080/02724980543000132

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. doi : 10.1016/0010-0277(95)00664-8

Crisp, A.K., & Feeney, A. (2009). Causal conjunction fallacies: The roles of causal strength and mental resources. Quarterly Journal of Experimental Psychology, 62, 2320 – 2337. doi : 10.1080/17470210902783638

Eddy, DM. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgement under Uncertainty: Heuristics and Biases (pp. 249 - 267). Cambridge, England: Cambridge University Press.

Evans, J.St.B.T., Handley, S.J., Over, D.E. & Perham, N. (2002). Background beliefs in Bayesian inference. Memory & Cognition, 30, 179-190. doi : 10.3758/BF03195279

Fischhoff, B., Slovic, P., Lichtenstein, S. (1979). Subjective sensitivity analysis. Organisational Behavior and Human Performance, 23, 339-359. doi : 10.1016/0030-5073(79)90002-3

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. Thinking & Reasoning, 15, 105-128. doi : 10.1080/13546780802711185

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704. doi : 10.1037//0033-295X.102.4.684

Gopnik, A. (2012). Causality. In P. Zelazo (Ed.), The Oxford Handbook of Developmental Psychology. New York : Oxford University Press.

Hayes, B.K., Newell, B.R., & Hawkins, G.E. (2013). Causal model and sampling approaches to reducing base rate neglect. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Anuual Conference of the Cognitive Science Society. Austin, Texas: Cognitive Science Society.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237 - 251. doi : 10.1037/h0034747

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgement under Uncertainty: Heuristics and Biases. Cambridge, England: Cambridge University Press.

Koehler, D. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. Behavioral and Brain Sciences, 19, 1–16. doi : 10.1017/S0140525X00041157

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. Journal of Experimental Psychology: General, 136, 430-50. doi : 10.1037/0096-3445.136.3.430

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning. Psychological Review, 106, 411 – 416. doi : 10.1037/0033-295X.106.2.411

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. Quarterly Journal of Experimental Psychology, 48A, 188 - 207. doi:10.1080/14640749508401384

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. Organizational Behavior and Human Decision Processes, 82, 217-236. doi:10.1006/obhd.2000.2895

Newstead, S.E., Handley, S. J., Harley, C., Wright, H. & Farrelly, D. (2004). Individual differences in deductive reasoning. Quarterly Journal of Experimental Psychology, 57A, 33-60. doi : 10.1080/02724980343000116

Pearl, J. (2000). Causality: Models, Reasoning and Inference. New York: Cambridge University Press.

Sloman, S.A. (2005). Causal Models: How People Think About the World and Its Alternatives. New York: Oxford University Press.

Sloman, S.A., Over, D. E., Slovak, L., & Stibel, J. (2003). Frequency illusions and other fallacies. Organizational Behavior and Human Decision Processes, 91, 296-309. doi:10.1016/S0749-5978(03)00021-9

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359-1366. doi: 10.1177/0956797611417632

Tversky, A., & Kahenman, D. (1980). Causal schemas in judgement under uncertainty. In M. Fishbein (Ed.), Progress in Social Psychology (pp. 49 - 72). Hillsdale, New Jersey: Lawrence Erlbaum.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In Kahneman, D., Slovic, P., & Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 153-160). Cambridge: Cambridge University Press.

Verschueren, N., Schaeken, W., & D'Ydewalle, G. (2004). Everyday conditional reasoning with working memory preload. Proceedings of 26th Annual Meeting of the Cognitive Science Society (pp. 1399–1404). Mahwah, New Jersey: Lawrence Erlbaum.

Waldmann, M.R., & Hagmayer, Y. (in press). Causal reasoning. D. Reisberg (Ed.), Oxford Handbook of Cognitive Psychology. New York: Oxford University Press.

Appendices

Appendix 1: Reasoning Problems

The italicised text in each problem gives the wording for the presentation of false positive information in the Causal version.

Mammography Problem

Suppose the following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:
10 in every 100 have breast cancer at the time of the screening. 8 in every 10 of those with breast cancer will receive a positive mammogram. However, 20 in every 90 of the remaining cases without cancer will still receive a positive mammogram. However, a dense but benign cyst, which looks like a cancerous tumor on the X-ray, can cause positive mammograms in those without cancer; this occurs in 20 in every 90 mammograms of those without cancer.
Of those that receive positive mammographies, what % would you expect to have cancer?

Weather Forecast Problem

Suppose the following statistics are known about "Tornado Season" which occurs along the Gulf Coast of America from late February into late April.
Over the past 20 years, tornadoes have occurred on 20 out of every 100 days during “Tornado Season”. 15 out of those 20 tornado days have been correctly predicted 3 days in advance of their occurrence. However, tornadoes have been mistakenly predicted on 25 out of every 80 days. However, brief but severe fluctuations in storm activity can lead to a mistaken tornado prediction, and this has occurred on 25 out of every 80 days.
What % of predicted tornadoes would you expect to occur?

Counterfeit Money Problem

Suppose the following statistics are known about the banknotes currently in circulation.

15 in every 100 notes are counterfeit. Many retailers use special marker pens to test whether a note is counterfeit. These markers are successful in identifying 12 out of every 15 fake notes. However, 25 out of every 85 legal notes are incorrectly identified as counterfeit. However, the ink in the marker sometimes reacts with chemicals often found on legal notes, resulting in 25 out of every 85 legal notes being incorrectly identified as counterfeit.

Of those notes identified by marker pens as being counterfeit, what % would you expect to be counterfeit?

Suspect Problem

Suppose the following statistics are known about airport security:

It is estimated that 4 in every 100 people who pass through Heathrow airport are wanted by the police. Using CCTV images, facial recognition software will correctly identify 3 out of every 4 police targets. However, 12 out of every 96 ordinary travellers are identified as suspects. However, glare from the sun means that the software incorrectly identifies 12 out of every 96 ordinary travellers as suspects.

Of those identified as suspects by the software, what % would you expect to be wanted by the police?

Speed Camera Problem

Suppose the following statistics are true statistics about speed limit enforcement:

30 in every 100 drivers break the speed limit along a certain stretch of motorway. Speed cameras detect 20 out of every 30 offenders. 7 out of every 70 cars driving below the speed limit are mistakenly detected as speeding, however. However, rain on the camera lens can

distort the image a camera sees, so 7 out of every 70 cars driving below the speed limit are mistakenly detected as speeding.

What % of motorists identified as breaking the limit would you expect to have been speeding?

Sniffer Dog Problem

Suppose that the Department of Customs and Excise know the following statistics to be true: 10 out of every 100 pieces of luggage at Belfast City Airport contain drugs. Sniffer dogs correctly detect drugs in 9 out of every 10 cases where drugs are present. However, the presence of drugs is incorrectly detected in 2 out of every 90 cases. However, certain foodstuffs smell like drugs, and for this reason sniffer dogs will incorrectly detect the presence of drugs in 2 out of every 90 cases.

What % of luggage identified by sniffer dogs as containing drugs will actually contain drugs?

Pregnancy Test Problem

Suppose the following statistics about pregnancy true:

When a sexually active woman's period is over five days late, in 30 out of every 100 cases this is a sign that conception has occurred. Pregnancy tests will correctly indicate pregnancy in 26 out of every 30 cases. However, the tests incorrectly detect pregnancy in 20 out of every 70 cases. However, the dye in pregnancy tests near their expiration date can seep, resulting in positive results even if conception has not occurred. This happens in 20 out of every 70 cases.

What % of sexually active women whose periods are over 5 days late, and who have positive results, are pregnant?

Email Spam Problem

Assume the following statistics are true about a program for blocking spam e-mails:

40 of every 100 emails are spam e-mails. Spam-filters detect and block 20 out of every 40 spam messages. However, 18 out of every 60 non-spam e-mails are mistakenly blocked as spam. However, because the filters work by detecting keywords, some legitimate emails containing these keywords are mistakenly blocked as spam; this happens in 18 out of every 60 cases.

What % of blocked e-mails are spam e-mails?

Drug Test Problem

Suppose the following statistics about drug use are true:

A survey revealed that 15 of every 100 people employed by a particular company use drugs outside of work. Random urine tests were introduced which correctly indicate the presence of illegal drugs in 11 out of every 15 cases. However, there is a positive urine test in 31 out of every 85 innocent cases. However, chemicals in over-the-counter medications will trigger a positive urine test in 31 out of every 85 innocent cases.

What % of employees with a positive test has used illegal drugs?

Virus Scanner Problem

Suppose these statistics about an internet security package are true:

20 of every 100 files downloaded from the internet contain harmful viruses. 18 of every 20 infected files are detected by anti-virus scanners before they have a chance to do any damage. However, virus scanners incorrectly identify 8 out of every 80 harmless files as infected. As some programming code found in viruses is also found in harmless files, virus scanners incorrectly identify 8 out of every 80 harmless files as infected.

What % of files identified as harmful turn out to contain a virus?

HIV test problem

Suppose the following statistics about HIV are true:

15 out of every 100 people in southern Africa may have HIV without their knowledge. HIV tests that detect particular antibodies will correctly indicate a HIV infection in 13 of 15 cases. However, the tests come back positive in 30 out of every 85 cases where the HIV virus is absent. However, as other conditions can cause the presence of some antibodies associated with HIV, the tests come back positive in 30 out of every 85 cases where the HIV virus is absent.

What % of people with positive test results has the virus?

Swine Flu Problem

Suppose the following statistics about the spread of swine flu are true:

By December, 35 out of every 100 flu cases in the UK will involve Swine Flu. 24 out of every 35 of those infected will receive positive tests for H1N1. However, 18 out of every 65 people infected with normal flu will still receive positive tests for Swine Flu. As several of these genetic markers are also associated with normal flu, 18 out of every 65 people infected *with normal flu will receive positive H1N1 tests when they don't have Swine Flu.*

What % of patients with positive results has Swine Flu?

Water Contamination Problem

Suppose recent tests at a bottled water plant found:

12 in every 100 bottles produced are contaminated with harmful bacteria. A litmus test identifies 10 in every 12 of contaminated bottles. However, 18 in every 88 uncontaminated

bottles are identified as being contaminated. However, the litmus test is also sensitive to some harmless bacteria and incorrectly identifies 18 in every 88 uncontaminated bottles as being contaminated.

What % of bottles identified as contaminated are actually contaminated?

Cab Problem

Suppose two taxi cab firms operate in a city; the Blue company and the Green company.

25 in every 100 cabs in the city are Blue and 75 in every 100 are Green. A witness to an evening car accident identified a Blue taxi as being involved. The court tested the witness' ability to identify taxis under similar visibility conditions and found that they correctly identified 18 out of 25 Blue cabs. However, the witness identified 22 out of 75 Green cabs as Blue. However, due to faded paint, some blue and green taxis looked similar, and the witness misidentified 22 out of 75 green cabs as blue.

Given this information, what is the % chance the witness was correct and the taxi involved in the accident really was Blue?

Intrusion Detection System Problem

Suppose the following crime statistics are true:

30 out of every 100 museums in the UK are broken into each year. The Laser detection systems which all museums have nowadays will sound an alarm in 20 out of every 30 robbery attempts. However, every year alarms sound when there is no intruder in 6 museums out of every 70. However, rodents can trip the alarm and every year alarms sound when there is no intruder in 6 museums out of every 70.

Given a raised alarm; what is the % chance there has been a robbery attempt?

Fighter Jet Problem

Suppose that in preparation for the threat of attack from Vietnamese and Cambodian planes during the Vietnam War; the US Army trained lookout soldiers.

In every 100 training trials, 28 out of every 30 Vietnamese planes were correctly identified. However, 20 out of every 70 Cambodian planes were misidentified as Vietnamese. Because the training simulated night-time conditions, 20 out of every 70 Cambodian planes were misidentified as Vietnamese.

A soldier identified a Vietnamese plane as responsible for a recent attack; what is the % chance that the fighter jet was in fact of Vietnamese origin?

Appendix 2: Further results for each experiment, broken down by problem content

Experiment One: Response frequencies broken down by problem.

	Bayesian		Base Rate Neglect		Likelihood Neglect		Other	
	Caus	Stand	Caus	Stand	Caus	Stand	Caus	Stand
Mammography	4	3	0	5	1	0	11	8
Tornado	3	3	5	8	1	2	7	3
Money	4	1	2	4	4	4	6	7
Suspect	8	5	0	0	2	5	6	6
Speed Camera	3	3	2	1	1	2	10	10
Sniffer Dog	4	5	3	0	2	2	7	9
Pregnancy Test	2	2	3	6	4	0	7	8
Email	4	4	0	2	5	3	7	7
Drug	4	3	2	4	7	3	3	6
Virus	3	4	2	5	6	3	5	4
HIV	3	4	0	0	4	4	9	8
Swine Flu	3	3	3	3	4	2	6	8
Water	3	3	2	5	4	2	7	6
Cab	6	5	4	5	2	3	4	3
Intruder	4	3	1	5	2	2	9	6
Jet	6	5	1	6	1	3	8	2

Note: Row totals = 32 as half of sample received first eight problems and half of sample received second eight problems.

Experiment Two: Response frequencies broken down by Problem Type and Response Format.

Free Response

	Bayesian		Base Rate Neglect		Likelihood Neglect		Other	
	Causal	Standard	Causal	Standard	Causal	Standard	Causal	Standard
Mammography	7	2	5	9	3	6	11	7
Tornado	8	2	8	11	1	3	9	8
Pregnancy Test	6	2	5	11	4	6	11	5
Drug	6	2	4	12	5	1	11	9

Multiple Choice

	Bayesian		Base Rate Neglect		Likelihood Neglect		Other	
	Causal	Standard	Causal	Standard	Causal	Standard	Causal	Standard
Mammography	14	6	1	8	4	5	5	7
Tornado	14	5	2	5	3	5	5	11
Pregnancy Test	13	10	2	6	2	3	7	7
Drug	12	5	5	7	3	7	4	7

Note: Row totals = 50 as each half of the sample received all 4 problems in one of the two response format conditions.

Experiment Three**:** Response frequencies broken down by problem.

Standard

	Bayesian	Base Rate Neglect	Other
Mammography	5	20	17
Tornado	2	26	14
Pregnancy Test	4	23	15
Drug	6	17	19

Causal

	Bayesian		Base Rate Neglect		Other	
	3 p	4 p	3 p	4 p	3 p	4 p
Mammography	7	10	14	12	21	20
Tornado	12	10	19	18	11	14
Pregnancy Test	10	11	13	10	19	21
Drug	4	7	14	7	24	28

N.B. 3p = three parameter causal problem, 4 p = four parameter causal problem.