Word Count: 25,024

A Mixed-Methods Analysis of Bayesian Reasoning: Nested Sets versus Causal Framing

Stephen H. Dewitt [1], Anne S. Hsu [2], David A. Lagnado [3], Saoirse Connor Desai [4], Tom Kenny [5], Alice A. Alberici [6], Norman E. Fenton[7]

[1]Department of Psychology. King's College London.
[2]School of Electronic Engineering and Computer Science. Queen Mary University of London.
[3]Department of Experimental Psychology. University College London.
[4]Department of Psychology. City University London.
[5]Shared Assets. London.
[6]Medical School. University of East Anglia.
[7]School of Electronic Engineering and Computer Science. Queen Mary University of London.

This work was supported by the European Research Council Advanced grant number ERC-2013-AdG339182-BAYES-KNOWLEDGE and the Engineering and Physical Sciences Research Council grant number EP/L50483X/1. We thank Martin Neil and Magda Osman for feedback and guidance to the first author at a number of crucial stages.

Correspondence concerning this article should be addressed to Stephen H Dewitt. dewitt.s.h@gmail.com

Abstract

How do people solve Bayesian word problems, and how is this affected by different presentational formats? We empirically compare the efficacy of the natural frequency / nested sets and causal approaches, using think-aloud analyses to examine the underlying cognitive processes. Experiment one demonstrates an increase in accuracy with a nested sets / natural frequency framing but not with a causal framing. From the think aloud data, a single five-stage solution process (the 'nested sets' process) is observed modally among individuals providing the normative Bayesian answer across all conditions. In experiment two the nested sets approach is evaluated using a problem with greater ecological validity and the increased accuracy effect is preserved. Experiment two also finds that spontaneous conversion of the problem by participants to real numbers (natural frequencies) is highly associated with accuracy, but is not essential. Experiments one and two also provide a mixed-methods analysis of the most common erroneous processes individuals undertake and find the confusion hypothesis a more fitting explanation of results than base rate neglect. Experiment three confirms the null causal finding of experiment one in a modified design and also demonstrates that the mere presence of the think aloud protocol increases accuracy. Experiment four experimentally tests whether prompting problem conversion to real numbers, and prompting individuals to follow the nested sets process improve accuracy. No effect is found for conversion, but an effect is found for the nested sets process prompt.

Keywords: Natural Frequencies; Nested Sets; Causal Reasoning; Bayesian Reasoning; Base Rate Neglect; Confusion hypothesis

General Introduction

The presentation of Bayesian problems

The ability to make simple Bayesian inferences from given statistics is rapidly becoming a necessary skill in modern society (Meder and Gigerenzer, 2014). Key domains where Bayesian inference abilities are increasingly important include medicine and law (Barrett and McKenna, 2011; Fenton et al., 2014; Forrest, 2003; Gigerenzer and Edwards, 2003; Meder et al., 2009). For example, in medicine, both doctors and patients need probabilistic inference to make accurate assessments of the risk of the patient having a specific condition given statistical information about the prevalence of the condition in combination with diagnostic test results (Barrett and McKenna, 2011; Meder et al., 2009; Wegwarth et al., 2012). In law, there is increasing use of statistical forensic evidence; however, since forensic experts have been discouraged from explicitly using Bayes to present the conclusions of their analyses in the courtroom (Donnelly, 2005), it is left to both lawyers and juries to perform for themselves the necessary Bayesian calculations in order to understand the true impact of the evidence (Fenton et al., 2014). Ineffective presentations of such statistics greatly increase error rates in comprehension. The consequences include poor patient decisions (Gigerenzer and Edwards, 2003; Navarrete et al., 2014) and miscarriages of justice (e.g. Forrest, 2003; Mehlum, 2009).

Approaches used so far

There have been many attempts to improve Bayesian reasoning through altering problem framing. The three approaches with the greatest advocacy are the natural frequency approach (Brase, 2008, 2013; Gigerenzer and Hoffrage, 1995; Johnson and Tubau, 2013), the nested sets / partitive / subset approach (Evans et al., 2000; Fiedler et al., 2000; Girotto and Gonzalez, 2001; Johnson-Laird et al., 1999; Lewis and Keren, 1999; Macchi, 2000; Mellers and McGraw, 1999; Sloman et al., 2003; Tversky and Kahneman, 1983) and the causal approach (Hayes et al., 2013; Krynski and Tenenbaum, 2007; McNair and Feeney, 2014a,b). We illustrate the three approaches using the classic 'medical diagnosis problem' (e.g. Eddy, 1982). Here, the problem solver is asked to compute the probability of a woman having a disease given prior knowledge of its prevalence in combination with the results of a diagnostic test with less than 100% reliability. A version of the problem which employs none of the above three approaches and is taken from Gigerenzer and Hoffrage (1995) is presented below.

> The probability of breast cancer is 1% for women at age forty who participate in 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? __%

In an early study, Casscells et al. (1978) gave a similar problem to 60 students and staff at Harvard Medical School. Only 18% of solvers gave the correct answer. Such low levels of accuracy on Bayesian word-problems became the accepted norm in the literature over the next decade (e.g. Bar-Hillel, 1980; Eddy, 1982). Each of the three approaches used these early studies, and the paradigms employed within them, as the benchmark for improvements.

Natural frequencies and nested sets. The natural frequencies and the nested sets / partitive / subset approaches possess a largely entwined history. The natural frequency approach, devised by Gigerenzer and Hoffrage (1995) attempts to improve reasoning on Bayesian problems by presenting the solver with the real number values they would observe if they sampled the given population in the problem one by one, making a record of the presence or absence of each feature of interest (e.g. cancer / no cancer, positive / negative test result). This process was termed ‘natural sampling’ and if carried out for the example given above, produces a set of categorical figures which can be represented visually as the tree diagram in Figure 1 below. Even in the absence of this assistive diagram, presenting the variables in this natural frequency format has been reliably shown to improve accuracy on Bayesian word problems when compared to probabilistic, percentage or normalized frequency formats (Brase, 2002; Chapman and Liu, 2009; Garcia-Retamero and Hoffrage, 2013; Hill and Brase, 2012).

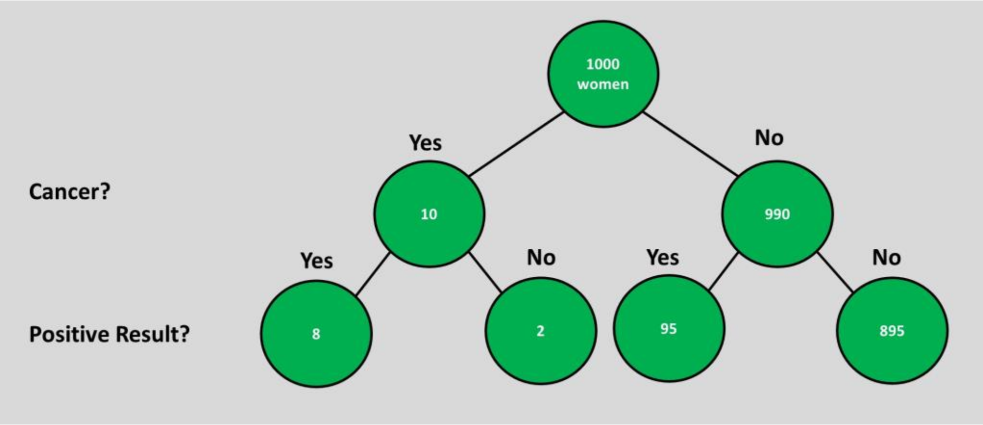


Figure 1. A depiction of a natural frequency representation of the mammogram problem, adapted from Gigerenzer and Hoffrage (1995). The 1000 women are sub-divided into those with and without cancer, and then further subdivided into those with and without a positive test result.

There has been much debate surrounding the psychological mechanisms underlying this increase in accuracy. Some authors have presented an evolutionary argument, claiming that the natural frequency format is that which humans would have been exposed to over their evolutionary history, and therefore would likely have adapted mental mechanisms specifically to process (Brase and Hill, 2015; Cosmides and Tooby, 1996; Hoffrage et al., 2002). An additional possible reason, noted by Gigerenzer and Hoffrage (1995) in their original paper, was that this format is computationally simpler than classic probabilistic or percentage formats, requiring fewer computational steps. While other work has suggested this is unlikely to be the sole, or even main, cause of improvement (Brase, 2002), it remains a potential contributory factor. Finally, other authors have suggested, initially through a misunderstanding of the format (see Hoffrage et al., 2002, for a review), that it is the fact that the natural frequency format reveals the ‘nested sets’ (Sloman et al., 2003) ‘subset’ (Evans et al., 2000; Johnson-Laird et al., 1999), ‘partitive’ (Macchi, 1995; Macchi and Mosconi, 1998; Macchi, 2000) or ‘outside’ (Fiedler et al., 2000) nature of the Bayesian problem that is the cause of its assistive effect. Each of these terms reflects the same underlying concept, and in line with recent reviews (e.g. Johnson and Tubau, 2015; McNair, 2015) the term ‘nested sets’ shall be used henceforth. It has been proposed by these authors that any format which reveals the nested sets structure of a Bayesian problem will show equal improvements in accuracy to natural frequencies, minus any possible benefit gained from the computational simplification. This nested sets structure can be represented as the overall conceptual structure of Figure 1 (i.e. the representation of the problem as sub-divided groups of units) but is thought not to require any specific unit in each node (e.g. frequencies / real numbers).

Each of these ‘nested sets’ approaches has also had success in improving accuracy when compared to ‘inside-perspective percentage’ or probability formats typical of the early studies in the field (e.g. Bar-Hillel, 1980; Casscells et al., 1978). Macchi (2000) presented participants with a version of a Bayesian word-problem which used ‘outside-perspective percentages’ to encourage the sub-dividing representation, without relying upon natural frequencies. With the ‘inside percentage’ format presented above, typical of early work and frequently employed in real settings (Meehl and Rosen, 1955; Bar-Hillel, 1980; Casscells et al., 1978; Eddy, 1982) the statistic for the probability of obtaining a false positive on the mammogram test, for example, would be presented from the point of view of a single woman e.g. ‘If a woman does not have breast cancer, there is still a 10% chance (or ‘probability’) that she will get a positive mammography.’ However, with Macchi’s ‘outside-percentage’ approach this statistic would be presented from the point of view of a group of women e.g. ‘Out of all of those women who do not have breast cancer, 10% will also get a positive mammography’. The latter, focused on divisions of groups, was thought by Macchi to encourage the mental construction of the sub-divided nested sets model (see Figure 1) unlike the former, focused as it is on an individual and the abstract concept of ‘chance’, or ‘probability’. Importantly, unlike with a natural frequencies approach, both versions were presented with the same numerical format (percentages) and so did not suffer from any computational confound, or other potentially confounding issues surrounding preference / familiarity with different number formats. With this approach Macchi found a large significant difference in accuracy, with 6% in the non-nested sets (inside / probability) framing and 33% in the nested sets (outside / group) framing.

Furthermore, this nested sets condition was compared to a natural frequency version and was found not to be significantly different. This firstly suggests, in line with Brase (2002) that the computational simplicity of natural frequencies may not actually be a large factor in its success. It also provides tentative evidence that the assistive effect of natural frequencies may in fact be due to the revelation of the nested sets representation.

However, this interpretation of Macchi's (2000) results has been criticized (Hoffrage et al., 2002). These authors claimed that Macchi's nested sets format merely encouraged individuals to construct a natural frequency version of the problem for themselves, which then ultimately caused the increased accuracy. This alternative interpretation cannot be ruled out on the present evidence in the literature as it contains a paucity of studies focused on participants' solution procedures. However, Macchi's approach retains experimental value as it removes any concern of a computational, or other, confound as the statistics presented are precisely equal to the 'inside percentage' format in all ways other than the 'outside-perspective' re-framing. For this reason, and given the fact that both nested sets and natural frequencies advocates have claimed Macchi's findings as providing evidence for their position (the difference is a matter of interpretation, not method), this paper will test the nested sets / natural frequency approaches using this single presentation format. Further, by examining participants' problem solving processes, we will address the question of whether this format works by encouraging participants to construct a natural frequency format or whether success is achieved without such a conversion.

Causal approach. The ‘causal model’ approach to assisting Bayesian reasoning is highly distinct from either the natural frequencies or nested sets approaches. It was devised by Krynski and Tenenbaum (2007), building upon developments in computer science that emphasised the role of causal models in probabilistic representations (Pearl, 2000; Sloman and Lagnado, 2005). Krynski and Tenenbaum argued that purely statistical models, such as those employed by both the nested sets and natural frequency approaches, were ineffective in most everyday reasoning situations due to the complexity and low levels of information-certainty those situations present. They therefore claimed that these models were unlikely to be good descriptive models of human reasoning, which is likely to be adapted to these complex, low-certainty environments that humans inhabit. Krynski and Tenenbaum theorised that people in fact normally approach Bayesian reasoning problems by firstly constructing a causal model of the scenario. This causal model is then populated with the appropriate statistics and the answer is computed via Bayesian inference.

Krynski and Tenenbaum (2007) noted that much previous work, especially on the medical diagnosis problem, had failed to consider the causal structure of the problem they were presenting to participants. In particular they had failed to provide the solver with a cause for the false positive rate: participants were typically told that positive test results could occur in the absence of cancer, but no reason or cause for this was given. This, Krynski and Tenenbaum argued, prevented solvers from constructing a causal model and thus from solving the problem in their preferred way, resulting in the low accuracies seen in early experiments in the field (e.g. Bar-Hillel, 1980; Casscells et al., 1978; Eddy, 1982). The authors found in two separate experiments that, by simply adding a single sentence providing this cause (‘harmless cysts look like cancerous tumours and can cause positive results on the mammography’), accuracy increased from around 20% to around 40%, a similar magnitude increase to that typically seen by the natural frequencies and nested sets approaches. Their explanation for this was that the addition of the second cause completed the causal mental model of the problem for participants (see Figure 2 below), allowing Bayesian inference.

Two further Bayesian-problem experiments drawing on the Taxi-Cab problem (Eddy, 1982) and one with a novel 2x2 design, both also showed a beneficial effect for providing a clear causal structure. Some replication success has followed this paper in subsequent years, with one study finding an effect for forced-choice but not open-ended answers (Hayes et al., 2013) one providing a null finding (McNair and Feeney, 2014a) and one providing support in two separate experiments, but only in a high-numerate sub-group (McNair and Feeney, 2014b).

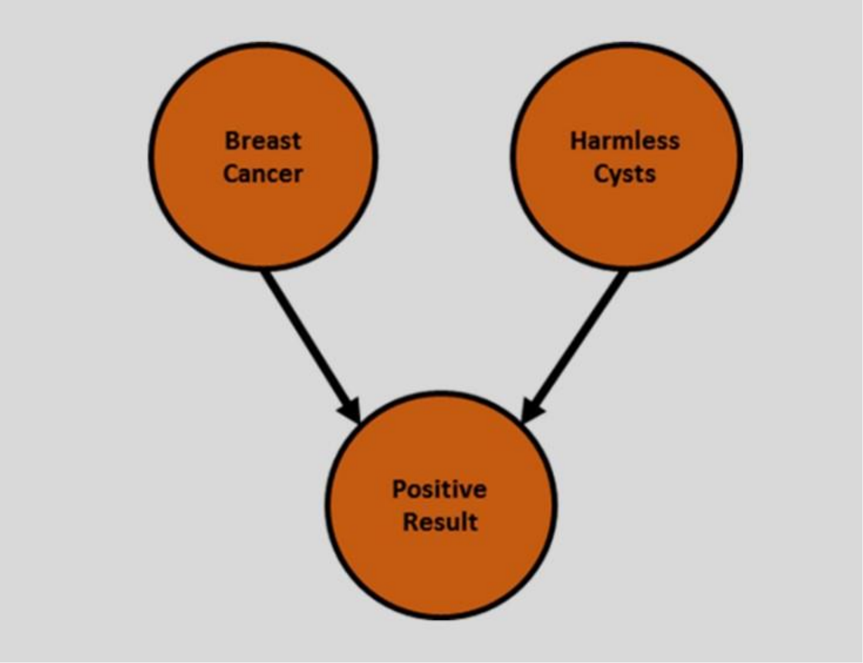


Figure 2. A simple representation of Krynski and Tenenbaum's (2007) causal model of the mammogram problem. The model contains the information that both breast cancer and harmless cysts are possible causes of positive mammogram test results.

Rationale

While the nested sets / natural frequencies and causal approaches to assisting Bayesian reasoning have the greatest advocacy in the field, no experiment has yet attempted to directly compare or combine them. Furthermore, despite over four decades of research examining the presentation of Bayesian problems, little consensus has emerged on the most effective method, or indeed on the means by which these methods achieve success. Even less agreement exists on the 'normal' way in which successful individuals solve Bayesian problems, or on why, and at what point, some individuals fail. In a recent appeal, McNair (2015) called for more research focused on understanding the processes that individuals undertake when faced with simple Bayesian problems. It was proposed this should provide vital information for how current presentation formats can be improved.

This need for a focus on solution process has been echoed by Johnson and Tubau (2015) who also highlighted the need for greater understanding of the stages at which unsuccessful solvers fail, as well as the point at which they make each calculation in the solution process. McNair (2015) also encouraged researchers to stop relying on numerical responses as their sole source of data due to its inability to distinguish between two very different thought processes which coincidentally produce the same quantitative answer. Instead, both McNair and Johnson and Tubau championed a methodology advocated by Ericsson and Simon (1980) and later utilized by Gigerenzer and Hoffrage (1995), which they called ‘think aloud’, and in which participants verbalize or record their thought processes while solving the problem.

Finally, the vast majority of studies testing these approaches have been conducted on undergraduate university students and in relatively small samples. The former issue is particularly important because it is reasonable to suppose that young undergraduates are not necessarily representative of the wider population in their capacity to solve Bayesian problems. Salthouse (1996) showed that human ability to process information declines with age, peaking in the early 20’s. Moreover, even within this age group, Brase et al. (2006) showed that students from higher-ranking universities perform at a higher level on Bayesian problems. If this association between education and Bayesian reasoning ability extends outside of universities it is very unlikely that a wider age group with a greater variation in education level will perform at the same level as those who have principally been studied so far. In support of this conjecture, Micallef et al. (2012) who studied the general population, found only 6% accuracy on a natural frequency phrasing of the medical diagnosis problem, which is far below previously-found levels (e.g. Gigerenzer and Hoffrage, 1995; Johnson and Tubau, 2013).

Experiment one

The first aim of the present experiment was to directly compare the natural frequency / nested sets approach with the causal approach to improving Bayesian problem solving accuracy in large general population samples. Macchi's (2000) outside-framed percentage approach to revealing the nested sets structure was used to represent the former, while Krynski and Tenenbaum's (2007) approach was used to represent the latter. Within this aim it was hypothesised that a significant main effect of the nested sets framing would be found in the whole sample. It was further hypothesised, based on the 'high numerate only' finding of McNair and Feeney (2014b) that only individuals in the high numerate sub-group would improve with a causal framing. Thus, given the fact that the present sample was likely to have lower numeracy levels than previous work, we predicted no significant main effect of the causal framing in the sample as a whole. However, it was hypothesised that a high-numerate sub-group (using a median-split on the Berlin numeracy scale (Cokely et al., 2012)) would show a significant main effect of causal framing.

The second aim of the experiment was to combine the nested sets and causal approaches in a single condition to determine if the two effects are additive, or even super-additive. It was again hypothesised that no significant interaction between the two conditions would be seen in the sample as a whole, but a significant and positive interactive effect would be seen in the high-numerate sub-group split at the median.

The third aim of this experiment was to heed McNair's (2015) and Johnson and Tubau's (2015) appeals to examine problem-solving processes and individual differences by using a 'think aloud' methodology alongside a numeracy measure in order to gain greater insight into the processes that participants undertake when solving Bayesian problems. This analysis will be exploratory but will aim to uncover both the 'normal' or 'preferred' processes people undertake when approaching these problems (either successfully or unsuccessfully) as well as how the natural frequency / nested sets and causal framings affect these processes.

Method

Participants. The final sample size for experiment one was 113. From an original sample of 124, nine participants were removed due to a clear lack of engagement with the experiment as evident in their numerical and think aloud data. Demographic data for all four experiments can be found in Table 1. Participants for all three experiments were recruited through the Amazon MTurk service and were required to be in the United States and to have a greater than 95% HIT approval rating. Participants were paid an average of $6.40 per hour for taking part.

Ethical approval for all the studies described was provided by the Queen Mary Research Ethics Committee (REF: QMREC1328) and was deemed to be extremely low risk.

Table 1 Participant demographics for all four experiments.

	EXPERIMENT 1		**EXPERIMENT 2**		**EXPERIMENT 3**		**EXPERIMENT 4**	
	Numeric	Percent	Numeric	Percent	Numeric	Percent	Numeric	Percent
Total Sample	113	100%	521	100%	429	100%	364	100%
Gender								
Male	51	45.1%	232	44.5%	220	51.3%	212	58.2%
Female	61	54.0%	288	55.3%	208	48.5%	152	41.8%
Other	1	0.9%	1	0.2%	1	0.2%	0	0%
Age								
Minimum	20	-	18	-	19	-	18	-
Maximum	66	-	71	-	75	-	67	-
Mean	33.1	-	34.2	-	36.7	-	35.1	-
Standard Dev.	10.0	-	11.6	-	12.3	-	10.8	-
Education								
High School	31	27.4%	157	30.1%	141	32.9%	138	37.9%
Bachelor’s Degree	55	48.7%	267	51.6%	199	46.4%	172	47.3%
Master’s Degree	22	19.5%	67	10.9%	63	14.7%	36	9.9%
Doctoral Degree	2	1.8%	12	2.3%	13	3.0%	4	1.1%
Other	3	2.7%	26	5%	13	3.0%	14	3.8%
Occupation								
Professional / Managerial	42	37.2%	218	41.8%	162	37.7%	130	35.7%
Labour / Service	35	31.0%	107	20.5%	129	30.1%	115	31.6%
Student	5	4.4%	65	12.5%	23	5.3%	4	10.2%
Unemployed	16	14.2%	70	13.4%	69	16.1%	5	12.4%
Other	15	13.3%	61	11.7%	46	10.7%	4	10.2%
First Language								
English	110	97.3%	517	97.7%	422	98.3%	-	-
Other	3	2.7%	4	2.3%	7	1.7%	-	-

Design. Experiment one employed a 2 (nested vs non-nested) x 2 (causal vs non-causal) within-subjects design resulting in four ‘conditions’ which all 113 participants undertook: ‘basic’, ‘nested’, ‘causal’ and ‘nested-causal’. Four different ‘scenarios’ were also created: ’Mammogram’, ‘College’, ‘Library’ and ‘Gotham’, totalling sixteen possible condition-scenario combinations. Each participant only saw four of these: each participant responded to every condition, and saw every scenario, but exactly which four combinations of these they saw was randomly determined.

Given the focus on individual process a within-participants design was chosen in this experiment principally in order to ensure interpretative clarity of the combined nested-causal condition. If a between-participants design was used, and the nested-causal condition outperformed the nested and causal conditions separately, it still could not be concluded that the combination of the two was beneficial on the individual level: if individual differences existed as to which of the two approaches people found helpful, an alternative explanation could be that some of the participants in the combined condition found the nested aspect helpful, while a different set found the causal aspect helpful, creating a higher average. With a within-participants design, however, if the nested-causal condition was higher than either nested or causal conditions, it would be possible to infer, and to confirm on an individual level, that the combination of nested and causal prompts is more assistive than either alone.

The study also employed a mixed quantitative-qualitative ‘think aloud’ method, drawing on Ericsson and Simon (1998) and Gigerenzer and Hoffrage’s (1995) approach. In that study’s design, participants wrote on paper as they worked out the problem, and these writings were analysed. In the present study, this method was adapted for computer-based experiments by asking participants to write their thought process while working out the problem in an open-ended text box. Crucially, they undertook this before having access to the next page where they could then enter their numerical answer.

Materials. The study was an online-survey conducted through Amazon MTurk, and which participants therefore accessed through their own computers. Colour-blind safe colours were used where colour was necessary, which were sampled from www.colourbrewer.org.

We used a version of the mammogram problem which was an amalgam of Gigerenzer & Hoffrage (1995), Krynski & Tenenbaum (2007) and Macchi (2000). A modified version of the college entrance exam problem (e.g. Brase, 2008) was also used. Two further problems were created for the study, one based on a ‘Macedonian Library’ and another based on crime rates in ‘Gotham city’. Problem for all sixteen conditions can be seen in the supplementary materials and the basic mammogram problem can also be seen below:

> Every year the government advises women to take part in routine mammography screening using an X-ray machine to determine if they have breast cancer. 200 out of every 1,000 women at age forty who participate in this routine

> screening have breast cancer, while 800 do not. If a woman has breast cancer, she will always get a positive mammography. If a woman does not have breast cancer, there is still a 10% chance that she will get a positive mammography.
>
> A woman in this age group had a positive mammography in routine screening. What is the percentage chance that she actually has breast cancer?

Each scenario had the same mathematical / logical structure but was otherwise designed to be as different as possible. This was done in order to reduce the likelihood of framing effects confounding the experiment and in combination with the use of multiple scenarios should therefore have made the study design more robust and the results more generalizable to other Bayesian problems. The scenarios varied in word-length and while ’Mammogram’ and ’College’ were both problems about humans, ’Library’ and ’Gotham’ were about objects (books and crime reports, respectively). Finally the actual numbers and population values used differed considerably across the scenarios.

The design of the conditions was based on Macchi (2000) and Krynski & Tenenbaum (2007). In all conditions, the population and two base rates were given as frequencies e.g. ‘200 out of every 1,000 women at forty who participate in this routine screening have breast cancer, while 800 do not.’ As can be seen in this statement, the frequency of ‘no-cancer’ (and equivalents in other scenarios) was also given, which is a departure from Macchi’s design. This was done to reduce difficulty in order to ensure that no floor effect was seen in the basic condition. This was considered a possibility as Macchi found 6% accuracy in the basic condition in undergraduate students, and the present experiment was conducted within the general population, which may have lower numeracy (Salthouse, 1996; Brase et al., 2006). This difference also departs from Krynski and Tenenbaum who gave the base rate of the first cause (e.g. cancer) as a percentage.

The nested sets manipulation (nested and nested-causal conditions) was produced as follows: in the basic and causal conditions, the '100% true positive' statement for the first cause was given from the perspective of an individual e.g. 'If a woman has breast cancer, she will always get a positive mammography', while in the nested and nested-causal conditions, this was given from the perspective of a group e.g. 'All of the women who have breast cancer will get a positive result on the mammography.' This was also the case for the false positive rate: in the basic and causal conditions this was given as 'If a women does not have cancer, there is still a 10% chance that she will get a positive mammography', while in the nested and nested-causal conditions this was given as 'Out of all those women who do not have breast cancer, 10% will also get a positive mammography.' Finally, in the basic and causal conditions, the question was also framed from an individual point of view e.g. 'A woman in this age group had a positive mammography in routine screening. What is the percentage chance that she actually has breast cancer?' whereas in the nested conditions it was framed from a group perspective e.g. 'What percentage of women who get a positive mammography in routine screening actually have breast cancer?'

The causal manipulation (causal and nested-causal conditions) change was more subtle. In the basic and nested conditions, no explanation was given for why the effect was still observed (e.g. a positive result) even when the first cause (cancer) was not present. This was in line with Krynski and Tenenbaum's (2007) original design, who proposed that in such cases, readers were not able to form a complete causal mental model. While Krynski and Tenenbaum used the mammogram problem they did not use the other three scenarios presented in this paper. The scenarios were therefore designed to ensure that in the basic and nested conditions the second 'hidden cause' would not be obvious to the reader (see supplementary materials for all four scenarios). In the causal and nested-causal conditions, an additional statement was given in order to provide this cause. In the mammogram problem the 'data' was a positive test result and the hidden cause was 'harmless cysts'. In the college scenario the data was entrance into the college and the hidden cause was that students with exceptional high school grades were also admitted even if they failed the exam. In the Library scenario the data was the presence of the book in the library and the hidden cause was the similarity of the Greek and Macedonian languages. Finally, in the Gotham scenario, the data was the presence of a crime in the 'other' folder and the hidden cause was a 'cover-up' of murder rates. There were no other differences between the conditions in the study.

Participants also completed the 7-item Berlin Numeracy Test, which has been shown to have good reliability and validity, and to be less subject to ceiling effects than some other numeracy tests used in the field (Cokely et al., 2012).

Procedure. Participants were recruited through Amazon MTurk. Participants were presented with the consent form, and then the instructions for the study, which included an extensive section on the ‘think aloud’ instructions, including an example (see supplementary materials). Participants then were assigned sequentially to four of the sixteen problems such that they saw each scenario and each condition only once. For each problem they were presented with the problem text and question itself and were asked to write their thought processes while they worked out the problem in a ‘think aloud’ open-ended text-box. Once this was complete they were able to give their actual numerical answer on the next page. Once participants had completed all four of their problems they were presented with the Berlin Numeracy test. Finally they answered the demographic questions and a final question regarding whether they had undertaken any of the problems before.

Data Analysis. In line with Gigerenzer & Hoffrage (1995) a dual criteria was used when determining if participants had given the correct answer. The correct numerical answer was not enough to get a point for each problem:

participants think aloud protocol was also analysed in order to detect whether they had used an ‘unacceptable process’. An unacceptable process was one which coincidentally led to the normative numerical answer on the particular problem but would not have if the numbers in the problem were different. If the participant provided the correct numerical answer but the process used was not clear, they were assumed to be ’correct’. In each problem there was at least one common error which gave the same numerical answer as the normative Bayesian answer, making this distinction important. Confusion between correct answers and numerically coincident but incorrect answers has blighted many previous papers in the field (e.g. see Evans et al. (2000) for a discussion of a similar issue with the paper by Cosmides and Tooby (1996)).

Additionally, certain answers which did not give the correct numerical answer were also accepted. This was the case when the think aloud data indicated that they had undertaken an acceptable process (one which would lead to the normative Bayesian answer), but had made an ’uninteresting error’. Uninteresting errors came in two types in the present study. Firstly, arithmetical errors, which in fact only occurred in 5 cases out of 452. Secondly, when participants gave the wrong ’cause’ as the answer (e.g. the percentage of women without cancer, instead of the percentage with cancer) but had again undertaken an acceptable process, their answer was also accepted as correct. This also occurred in 5 cases. These were both accepted because this study was not interested in improving these types of mistakes, but in whether participants could undertake accurate Bayesian reasoning.

All qualitative analysis of the think aloud protocol was undertaken blind to the participant’s condition. Analysis was first undertaken by the first author. In the first analysis phase the first author began by reading all transcripts looking for potential common themes between participants’ approaches to the problem. Once a set of common approaches to the problem were outlined, the second phase of analysis began wherein the first author reread all answers, coding them as to which approach the participant had used, and which of the steps in those processes they demonstrated. No new approaches were discovered in the second phase. In the third phase, the fourth author was provided definitions of each of the codes by the first author and was asked to assign all participants’ answers to whichever codes they deemed appropriate while blind to both participant condition and to the first authors’ original coding. Inter-rater reliability between the first and fourth authors was above 90%. In the fourth stage of analysis, any discrepancies between first and fourth authors were resolved by the seventh author who also possessed qualitative analysis experience. The seventh author was given the discrepant participants’ think aloud answers and the coding categories. The seventh author’s decision was taken as the final result for each of these participants.

Quantitative analysis was undertaken in IBM SPSS for Windows, Version 22. For the main analyses, generalised linear modelling was used. To implement this, the ’generalised linear model’ function of SPSS was employed, which uses regression modelling but allows for binary outcome data (correct vs incorrect answer to the problem) and for mixed analysis of both between-subject and within-subject variables.

Results

Quantitative. Combining all four conditions, 31.9% of all cases gave the Bayesian normative answer. No significant differences in accuracy were found between any two demographic sub-groups (gender, education or occupation). Below in Figure 3 the percentage of participants giving the Bayesian normative answer can be seen for all four conditions. It is immediately apparent that both the basic and causal conditions performed at a similar, and lower, level than the nested and nested-causal conditions. Confirming this difference, a repeated-measures generalised linear model (GLM) with binomial distribution and logit function found a significant main effect of nested sets framing (Wald = 7.358, p=.007), but no significant main effect of causal framing (Wald = .834, p=.361) and no significant interaction (Wald =.237, p=.626).

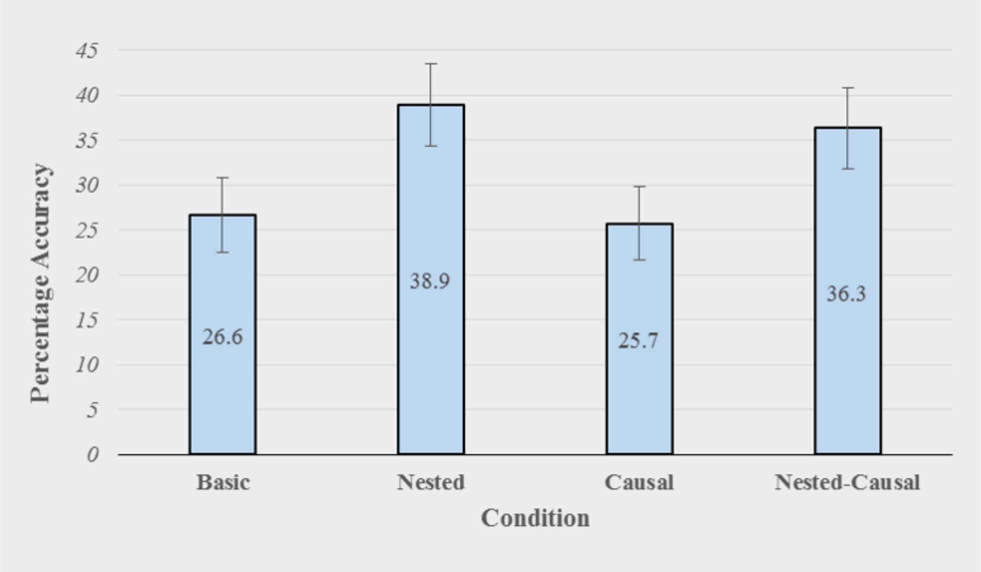


Figure 3. The percentage of participants giving the Bayesian normative answer for basic, nested, causal and nested-causal conditions. Error bars represent one standard error.

To test the two hypotheses that a main effect of the causal framing and an interactive effect between the nested sets and causal framings would be seen in the high-numerate sub-group, two different high-numerate sub-groups were created. Overall Berlin Numeracy Test scores showed a mean of 3.89 (SD = 1.85) and a median of 4. The first sub-group was created by splitting the sample at the median (as the median was also modal, these participants were included in the high numerate group to increase sample size (n = 70)). Mean numeracy score was 5.10 (SD = 1.10) for this group. The second high numerate sub-group was made by following the same method used by McNair and Feeney (2014b) to allow for direct comparison. McNair and Feeney had to remove question 7 on the numeracy test due to a printing error. Replicating this removal, scores were then split at the median from McNair and Feeney's experiment (5), creating a high numerate group (n=35) with similar numeracy levels to theirs (M=5.97, SD = 0.78). A non-significant effect was seen for the causal framing in the high numerate sub-group (Wald = 3.246, p=.072), with the causal group actually performing below the non-causal group (causal accuracy = 23.6%, non-causal accuracy = 31.4%). No significant effect was seen in the low numerate sub-group either (Wald = 2.843, p=.098). Further, no interaction between nested and causal was seen (Wald = .974, p=.324). Overall, low numerates were accurate 13.95% of the time, while high numerates were accurate 42.87% of the time, which was a significant difference (Wald = 26.03, p<.001). Finally, no main effect was seen in the 'McNair and Feeney' high numerate sub-group (Wald = 1.567, p=.211) and no interaction was seen here either (Wald = .167, p=.682).

The effect of the nested sets framing was also examined within the same high (n = 70, M = 5.10, SD = 1.10) and low numerate sub-groups (N = 43, M = 1.95, SD = 0.97) as for the causal framing. A significant effect of the nested sets framing was still seen within the low numerate sub-group (Wald = 5.359, p=.021) and a borderline significant effect was seen within the high numerate sub-group (Wald = 3.159, p=.076).

Qualitative. ***Key to section.*** The following analysis applies to all four experimental conditions. The variables in the problems are assigned the code below for this analysis:

1. H: The number or proportion of all units corresponding to the baserate for the first hypothesis presented (e.g. cancer).
2. -H: The number or proportion of all units corresponding to the base rate for the second hypothesis presented (e.g. no cancer).
3. D: The number or proportion of all units corresponding to the 'data' type requested in the question (e.g. a positive test result).
4. -D: The number or proportion of all units corresponding to the data type not requested (e.g. a negative test result). This variable was generally not used but is included for completeness.
5. P(H&D) / P(-H&D): The number or proportion of all units (e.g. women) with the given hypothesis (e.g. cancer / no cancer) and the requested data (e.g. a positive test result).
6. P(D|H) / P(D|-H): The number of units who have the requested data (e.g. a positive test result) P(D|H) as a proportion of all those units who correspond to a given hypothesis (e.g. cancer / no-cancer). P(D|H) was 1 in all four problems, while P(D|-H) varied between problems.

Successful Participants. Qualitative analysis of the think aloud data for successful participants revealed a 5-step process which comprised two representations of the problem and three computational steps. This was, except for 9 cases out of 452 (discussed below), the only process identified for successful individuals, and furthermore was identified in all four experimental conditions.

Step one / representation one: The hypothesis-focused representation. The first step in this process entailed the presentation of what is here called the un-populated ‘hypothesis-focused representation’, which can be seen below in Figure 4. This representation of the problem will be highly familiar from illustrations given in previous work (e.g. Gigerenzer and Hoffrage, 1995) and earlier in this paper. It begins by sub-dividing a sample of units into the ‘hypotheses’ (e.g. cancer / no cancer) and then further sub-dividing these by the ‘data’ (e.g. positive / negative). Notably, it does not include actual values in the lower-most nodes but instead represents the conceptual structure only through written word.

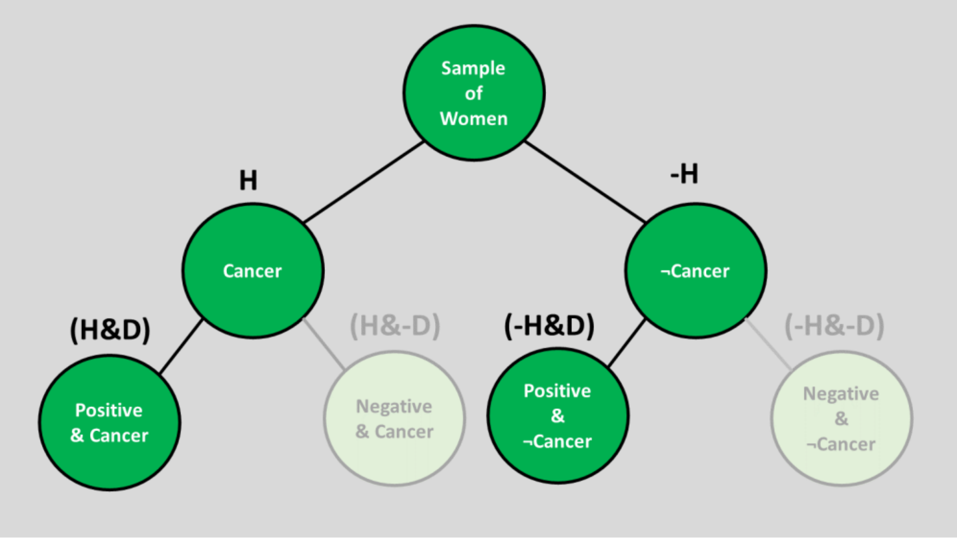


Figure 4. A visual depiction of the un-populated hypothesis-focused representation.

The requirement for this classification was a word-based subdivision of the two hypotheses (e.g. cancer / no cancer) into the data requested (e.g. positive results). A mathematical formula was neither necessary nor sufficient for this classification. An example of this classification can be seen in P40 who said in the nested sets condition: ‘So 200 women will definitely have a positive. 800 do not, but 10% of them will still get a positive.’ A further example comes from P5 who said in the ‘basic’ condition: ‘10% of the 800 women without breast cancer get a positive mammography result.’

In the basic and causal conditions, 9 cases were also detected where an ‘Individual’, ‘Inside’ or ‘Chance’ structure was portrayed. For example, P26 said, on the Gotham problem: ‘150 murders with a 40% chance of being filed as other means 60 murders were filed as others.’

Even within those two conditions however, a greater number (36) of cases presented the hypothesis-focused representation in their think aloud data than chance structure.

Step two / computation one: Populating the hypothesis-focused representation. Following the construction of the hypothesis-focused representation, successful participants subsequently undertook the calculations necessary to populate the bottom ‘D’ nodes representing the conjunctions P(H&D) and P(-H&D) in the hypothesis-focused representation diagram. P(H&D) was calculated by multiplying H by P(D|H). No single participant calculated the '-D' nodes (P(H&-D) and P(-H&-D)), presumably as these were not necessary to solve the problem.

Step three / representation two: The data-focused representation. Following the computation of the two positive conjunctions P(H&D) and P(-H&D), the next step in the process entailed laying out what is here called the un-populated 'data-focused representation'. A diagram depicting this can be seen below in Figure 5. This representation, instead of using the hypotheses as the mid-level nodes (e.g. cancer / no cancer), uses the data (e.g. positive / negative test result). Again, as the problem is inherently focused on 'D' (e.g. positive results) the '-D' of this diagram was neglected (not mentioned by participants) and the original sample (top node) was also typically neglected as neither were required to solve the problem from this point.

The requirement for the data-focused representation classification was a word-based indication that P(H&D) and P(-H&D) are subsets of D (see Figure 5). Again, a mathematical formula was not necessary or sufficient for this classification. Two broad sub-categories of this classification were identified:

Bottom up: i.e. by first defining the subsets P(H&D) and P(-H&D) and then demonstrating that they are in fact subsets of D e.g. P36 who said in response to the college scenario: '98000 fail and 1% of them get in - so that is 980 students

[H&D] add that to the 2000 [-H&D] who passed the test means 2980 [D] students total got in.'

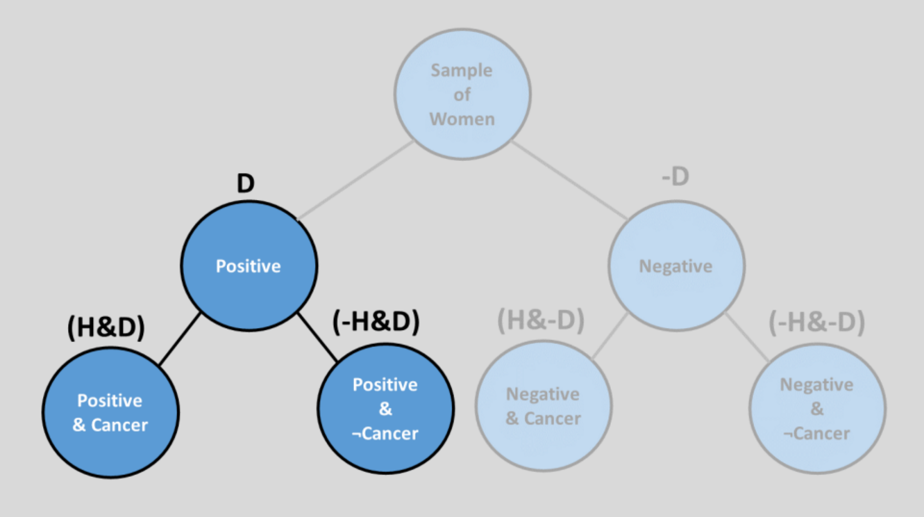


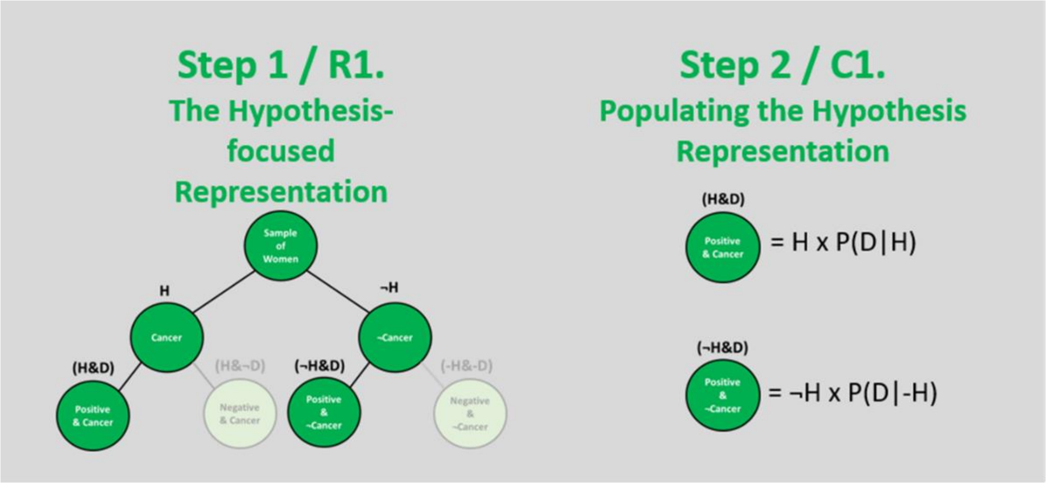
Figure 5. A visual depiction of the un-populated data-focused representation.

Top down: i.e. by first defining D and then demonstrating the subsets e.g. P15 in response to the medical scenario: 'So 280 [D] women will get a positive test. 200 [P(H&D)] actually have the cancer.' Typically this subcategory only mentioned P(H&D) as P(-H&D) was at this point not necessary to solve the problem.

Step four / computation two: Populating the data-focused representation. Either simultaneously with, or immediately following, the laying out of the data-focused representation, participants mathematically summed P(H&D) and P(-H&D) to obtain the total D (e.g. total positive results).

Step five / computation three: The computation of the final product P(H|D). Following the laying out and population of the data-focused representation, successful participants then divided the conjunction P(H&D) by the total number of positive results (D) to compute the normative Bayesian answer, P(H|D).

The nested sets process model. In line with the majority of recent work (e.g. Johnson & Tubau, 2015), we have called this process model the 'Nested Sets Process Model' as both representations of the problem (hypothesis and data) are inherently based upon the identification of certain sets of units in the problem being nested within others. This identification of the nested sets structure of the problem is indeed the key requirement for the classification of both representations. The remainder of the process model consists of populating each of these representations (C1 and C2), and then calculating the final Bayesian product (C3). A depiction of this entire model can be seen below in Figure 6. Apart from 9 cases, the Nested Sets Process Model was the only approach identified in the think aloud data of successful participants, even in the non-nested sets conditions. The entire process model, including each representation and computational step, was found in 11.5% of cases in the basic condition, 24.8% of cases in the nested condition, 15.9% of cases in the causal condition and 21.2% of cases in the nested-causal condition.

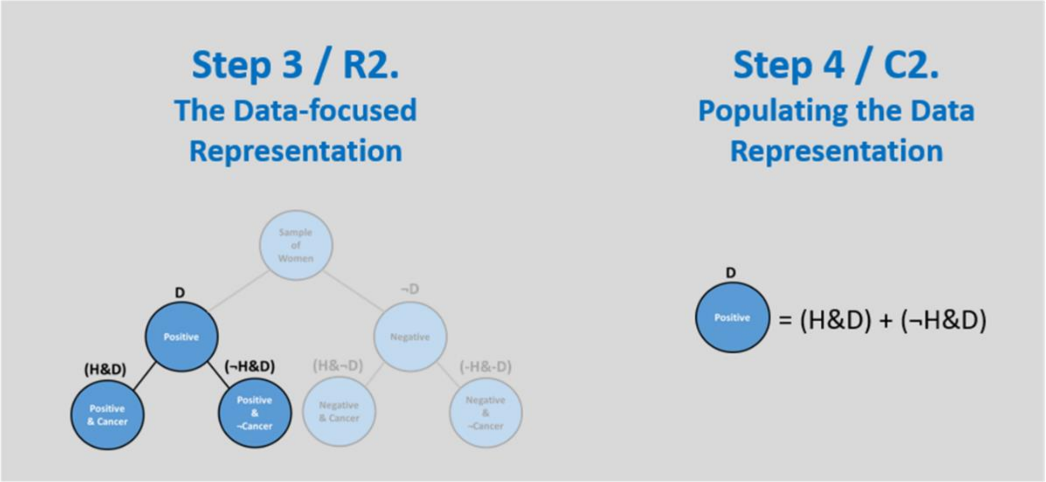


Step 1 / R1.
The Hypothesis-focused Representation

Step 2 / C1.
Populating the Hypothesis Representation

(H&D) = H x P(D|H)

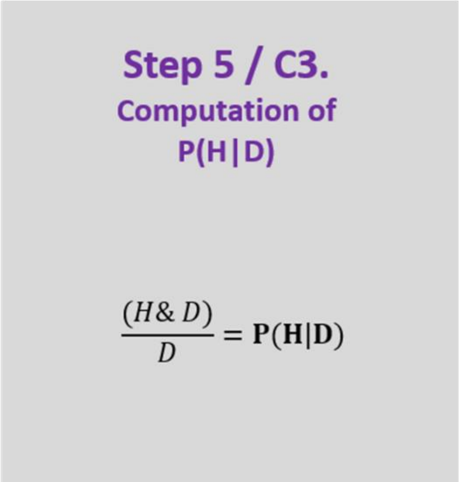
(¬H&D) = ¬H x P(D|-H)



Step 3 / R2.
The Data-focused Representation

Step 4 / C2.
Populating the Data Representation

D = (H&D) + (¬H&D)



Step 5 / C3.
Computation of P(H|D)

$$\frac{(H\&\,D)}{D} = \mathbf{P(H|D)}$$

Figure 6. The Nested Sets Process Model.

It should be noted that it is likely that the two representations are under-detected in the present study (more participants may have mentally formed these representations than was detected by the methodology). This is because many participants simply used mathematical formulae in their think aloud protocol and so could not be assigned either the hypothesis-focused representation or data-focused representation classifications, which used the stricter criterion of a word-based explanation.

The relationship between the proposed process model and previous work. The present model shows clear similarities to previous theoretical work (e.g. Evans et al., 2000; Gigerenzer and Hoffrage, 1995; Girotto and Gonzalez, 2001; Johnson-Laird et al., 1999; Lewis and Keren, 1999; Macchi and Mosconi, 1998; Macchi, 1995; Mellers and McGraw, 1999; Sloman et al., 2003). However, to the best of our knowledge, it is the first time that the entire process has been presented. Indeed, much previous work advocating the nested sets / partitive / subsets / natural frequency formats has attributed the value of the format to either the revelation of the hypothesis-focused representation alone (e.g. Macchi, 2000) or to the data-focused representation alone (e.g. Johnson and Tubau, 2015; Johnson-Laird et al., 1999; Mellers and McGraw, 1999; Sloman et al., 2003) and even when both have been referenced in a single paper (Evans et al., 2000; Girotto and

Gonzalez, 2001) no formal distinction between the two has been made. It is hoped that this explicitness of the differences between these two representations and their role in successful solution will bring additional clarity to the understanding of reasoning on Bayesian problems.

Nested sets framing and the process model. The nested sets framing used in this experiment and taken from Macchi (2000) consists of two changes to the problem: those to the text body and those to the question form. As is typical of Bayesian word problems, the body of the text contains the information relating to the hypothesis-focused representation (H, -H, P[H|D], P[H|-D]), whereas the question contains the information relating to the data-focused representation (D, P(H&D)). This framing may therefore may be expected to impact on both representations. An effect of the nested sets framing was seen on the frequency of hypothesis-focused representations produced (nested sets 26.7% vs non-nested sets 22.1%: Wald = 14.115, p<.001). An effect was also seen on frequency of data-focused representations (nested sets 27.4% vs non-nested sets 16.4%: Wald = 7.957, p=.005. However, when examining only those individuals who constructed the hypothesis-focused representation, no effect of the nested sets framing was seen on frequency of data-focused representations (Wald = 0.193, p=.661).

Similarly, an effect of the nested sets framing was seen on computational step one (nested sets 54.8% vs non-nested sets 42.0%: Wald = 8.625, p=.003), step two (nested sets 35.0% vs non-nested sets 25.2%: Wald = 5.154, p=.023) and step three (nested sets 35.4% vs non-nested sets 24.3%: Wald = 6.593, p=.010). However, when examining only those participants who correctly completed step one, no effect of the nested sets framing was seen on step two (Wald = .501, p=.479) or step three (Wald = 1.270, p=.260).

Causal framing and the process model. The causal framing (Krynski & Tenenbaum, 2007) in fact only made a single change to the problem: the addition of a single sentence in the body of the text providing a ‘cause’ for the false positive rate. No effect of causal framing was detected on frequency of the hypothesis-focused representation (Wald = .367, p=.547) or the data-focused representation (Wald = .020, p=.889). Further, no effect of the causal framing was detected on computational step one (Wald = 0.013, p=.908), step two (Wald = .069, p=.793) or step three (Wald = 0.778, p=.378). Finally, no single causal representation was identified in the think aloud protocol in any condition.

Drop off rates and numeracy. The percentage of people presenting each of the five successive stages in the process for both the computational steps and the representations, can be found in the upper graphs of Figure 7 and Figure 8 below. In the bottom graph of each figure, the average numeracy levels for the computational steps and the representations can also be seen. These graphs are cumulative: each successive step in the graph gives the percentage of participants achieving (or numeracy level) both that step as well as all previous steps. This method allows drop-off to be estimated more precisely. The computational steps and the representations are presented on separate figures due to the fact that, as mentioned previously, the think aloud process is very likely to under-detect the representations in comparison to the computational steps. A combined graph would therefore give the unjustified conclusion that the majority of ‘drop off’ occurs at the representations, when this may in fact just be due to the methodology used.

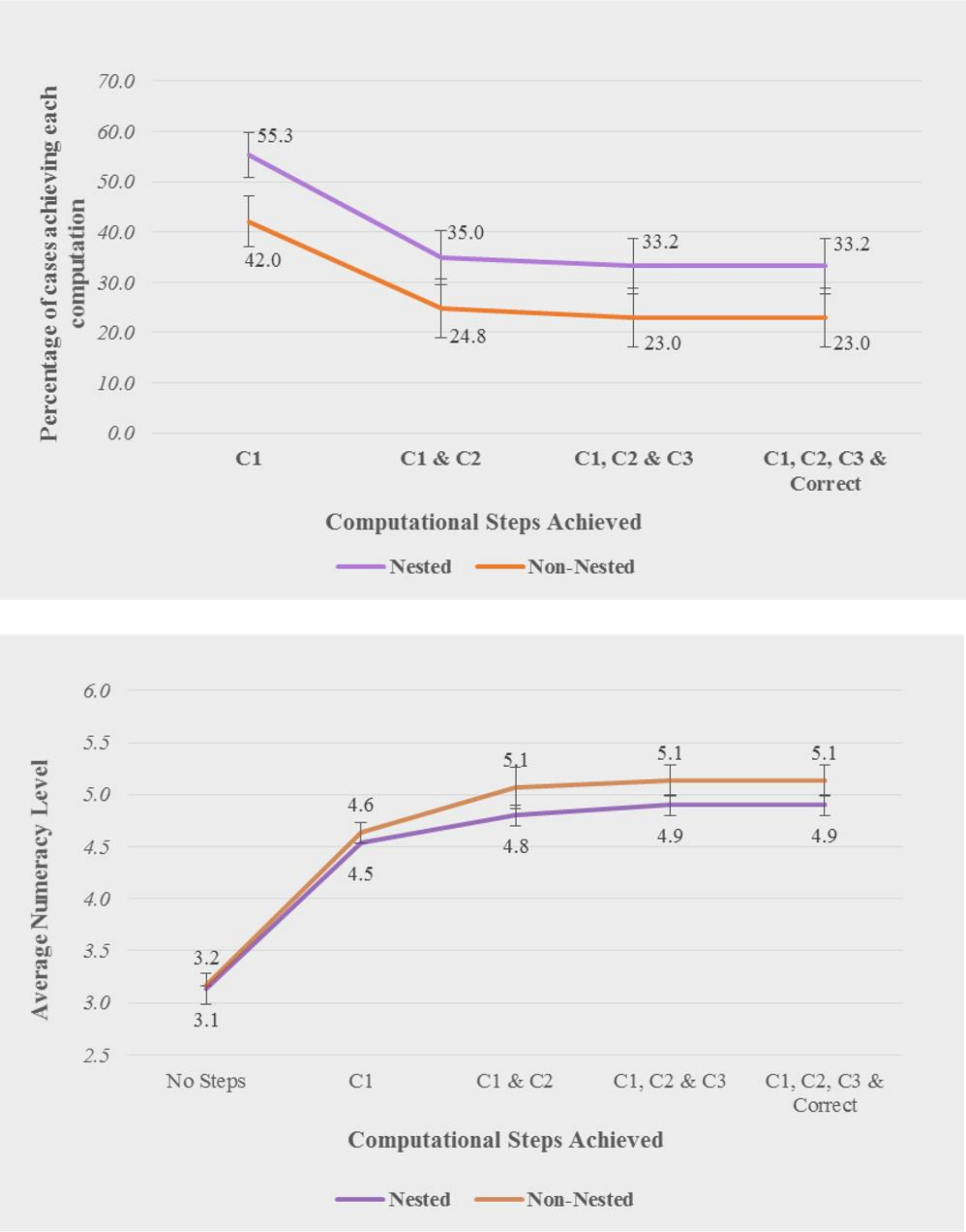


Figure 7. The cumulative percentages (top) and average numeracy levels (bottom) for participants achieving the three computational steps. Error bars indicate one standard error.

In regards to the computation drop-off diagram (Figure 7), highly similar drop-off curves for both nested and non-nested conditions can be seen, while the nested condition shows higher overall rates at each step. Further, it is clear that the vast majority of drop off occurs at C1 and further substantial drop off occurs between C1 and C2. Drop off is very slight between C2 and C3, and 100% of individuals presenting C3 also give the correct answer.

The average numeracy level diagram (Figure 7) shows a similar, but inverse, pattern. Individuals who achieve C1 have considerably higher average numeracy than those who achieve no steps, and this is the case for both nested and non-nested conditions. A further, but smaller increase in average numeracy is seen between C1 and C2, and no further increase is seen between C2 and C3, or in terms of those who also provided the correct answer.

A highly similar pattern to the computational steps diagrams is seen for the representation diagrams in Figure 8 below. In terms of the drop-off graph, highly similar curves are again seen for nested and non-nested conditions. Again, the vast majority of drop-off occurs at representation one, with further substantial drop off between one and two, and only very slight drop off after this, with the vast majority of individuals who construct representation two, also providing the correct answer.

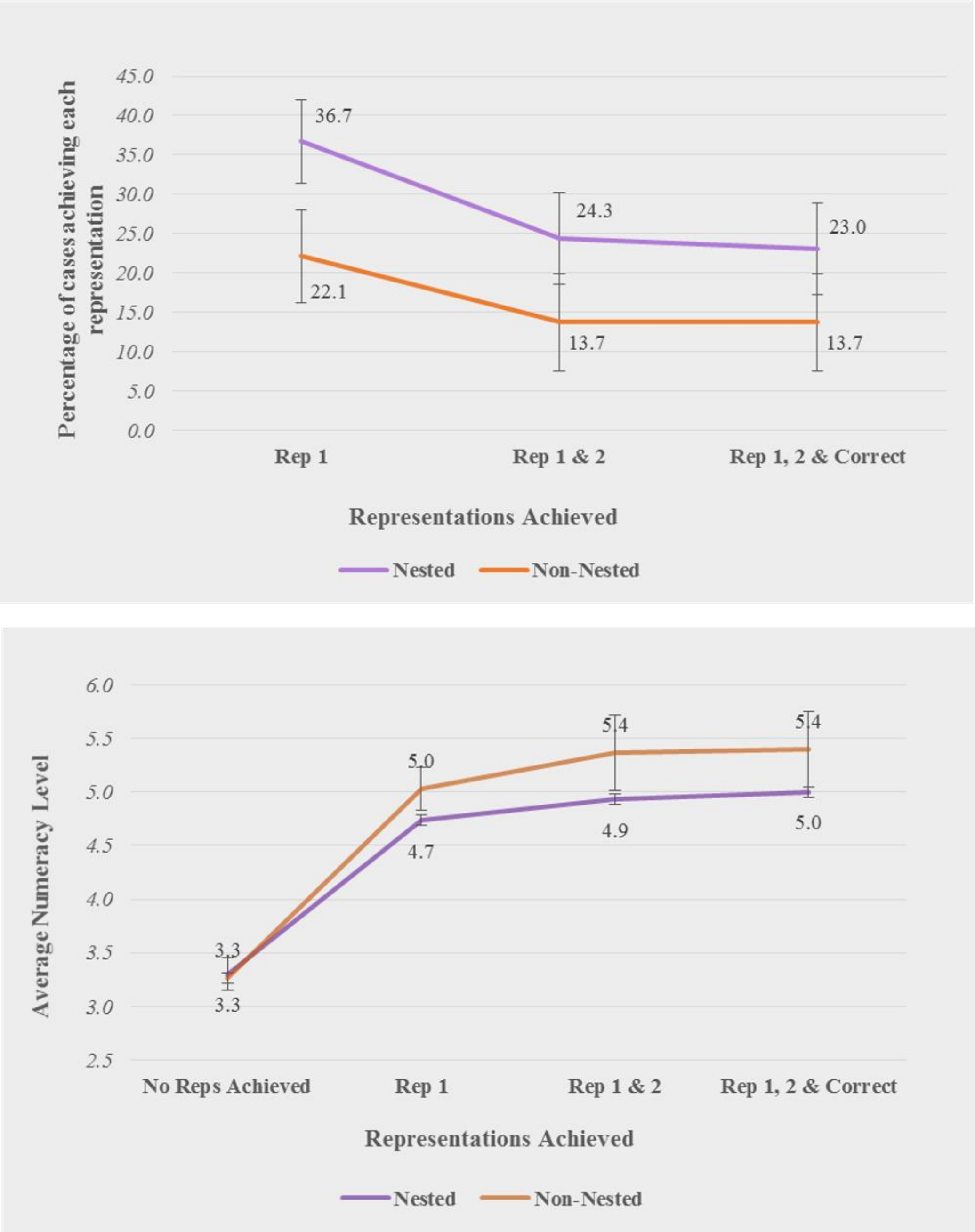


Figure 8. The cumulative percentages (top) and average numeracy levels (bottom) for participants achieving the two representational steps. Error bars indicate one standard error.

In terms of the numeracy graph, the biggest increase in average numeracy is again seen from no representations to representation one, and a further, but smaller increase is seen from one to two. No substantial change is seen when looking additionally at those who gave the correct answer.

Errors. Think aloud data was also analysed for individuals who did not provide the correct answer to determine why and where failure occurred. Twenty point five percent of all errors could not be categorised as the solution method was not clear and the numerical value did not point to a clear error type. Error type was assigned based on a two-step criteria. Firstly, the appropriate values for each scenario for all errors reported in previous work (e.g. Gigerenzer and Hoffrage, 1995; Macchi, 2000) were calculated and all responses which provided this value were preliminarily assigned that error categorisation. Following this, the think aloud data for every case was checked to determine if the method used disconfirmed the error type assigned. For example, if the participant gave the answer of 10% in the medical diagnosis problem, the ‘P(D|-H)’ error (see below) was firstly assumed. The participant’s think aloud data, when examined, may however have revealed that they used a different method to arrive at the numerical answer of 10%. For example they may have divided -H by the population and then added the P(D|-H) rate. If no other method was apparent, or the method could not be discerned, the error type was left unchanged (i.e. determined solely by the numerical response). This approach was used due to the large number of undiscernible responses which would have left many errors uncategorised. The six most numerous errors can be seen in Table 2 below.

Table 2. Percentage of cases for the six most common error types by condition, scenario and Berlin numeracy score.

		Condition				Scenario			
	Total	Basic	Nested	Causal	Nested-Causal	Medical	College	Library	Gotham
Total (Frequency)	**308**	**83**	**69**	**84**	**72**	**65**	**56**	**85**	**102**
Error Type									
P(D|-H)	16.9	15.7	17.4	16.7	18.1	7.7	28.6	8.2	23.5
P(D|-H)/H	10.7	7.2	14.5	9.5	12.5	0.0	0.0	2.4	30.4
P(-H&D) / Pop	7.8	14.5	8.7	8.3	4.2	0.0	8.9	27.1	0.0
-H / Pop	7.5	7.2	10.1	9.5	4.2	3.1	12.5	16.5	1.0
1-P(D|-H)	5.5	6.0	4.3	8.3	2.8	21.5	0.0	1.2	2.0
H / Pop	5.2	3.6	2.9	3.6	11.1	21.5	0.0	0.0	2.0

	Berlin Score							
	0	1	2	3	4	5	6	7
Total (Frequency)	**12**	**42**	**40**	**54**	**81**	**43**	**22**	**14**
Error Type								
P(D|-H)	1.6	35.7	17.5	16.7	13.6	7.0	4.5	7.1
P(D|-H)/H	0.0	2.4	2.5	5.6	18.5	23.3	13.6	0.0
P(-H&D) / Pop	0.0	4.8	7.5	14.8	11.1	11.6	0.0	7.1
-H / Pop	0.0	0.0	2.5	9.3	8.6	20.9	4.5	7.1
1-P(D|-H)	0.0	4.8	5.0	9.3	8.6	2.3	0.0	0.0
H / Pop	16.7	4.8	7.5	3.7	7.4	0.0	4.5	0.0

All errors were also assigned a score based on how many computational steps were involved in the calculation. Errors where the participant simply provided one of the figures in the problem (e.g. the false positive rate) were assigned a score of zero, while those involving a single computational step (e.g. the first base rate figure divided by the population) were assigned a score of one, and so on. Overall, 35.1% of codable errors were classified as involving zero computational steps, 48.9% as one step, 10.9% as two steps, 5.2% as three steps, 2.9% as four steps and 0.57% as five steps.

No significant predictive effect of either Nested (Wald = 2.286, p=.131) or Causal (Wald = .034, p=.854) variables on the number of computational steps involved in errors made was found, however a significant effect of Berlin numeracy score was found (Wald (7) = 37.022, p = .000). A graph depicting this relationship between computational steps of errors and the numeracy level of the participant making that error can be seen in Figure 9 below.

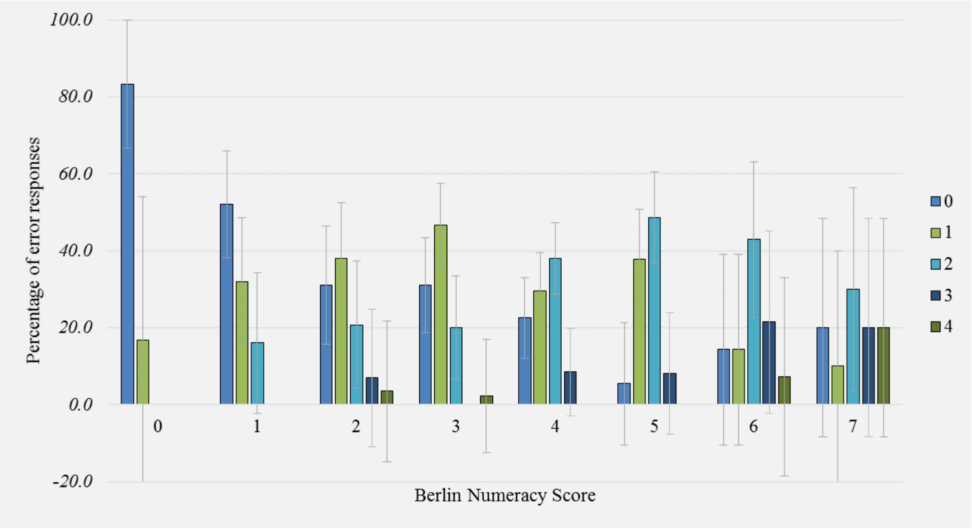


Figure 9. The percentage of errors which achieved 0, 1, 2, 3 or 4 computational steps for each score on the Berlin numeracy test. Error bars indicate one standard error.

It is clear from this graph that errors with zero computational steps (notably, the FP / P(D|-H) error) peak at the zero numeracy score levels. Similarly, errors with one computational step peak in the three numeracy score level, two computational steps peaks at the five numeracy level, three steps at the six numeracy level and four steps at the seven numeracy level. A general trend towards greater computational complexity of errors with increasing numeracy of the participant making that error is therefore clear.

A further exploratory analysis of the think aloud data was undertaken for the most common error, the false positive rate (P[D|-H]), to determine if any common thought processes leading to the error could be discerned. Exploratory analysis of the H/Pop and –H/Pop errors was also undertaken to determine if base rate conservatism logic could be detected. The think aloud data was coded by the first author and blind-second coded by the seventh author.

Out of the 52 instances of the P(D|-H) error, 14 were classified as ‘unknown’ as no clear process could be divined. Thirty two were merely classed as ‘P(D|-H)’, as these participants simply stated the false positive rate in their answer, either as a mathematical figure, or in word form, for example:

> “I believe the answer to this would be 1%. It says it in the last paragraph.” [P32, Scenario 2]

> “There is a 20% chance because it is clearly stated, no math needs to be done.” [P31, Scenario 3]

> “The answer is there 40%” [P86, Scenario 4]

Five instances clearly showed a version of the ’Inverse fallacy’, or a confusion between the false positive rate (P[D|-H]) and P(-H|D):

> “40%, it says 40% of murders are accidently labeled other. Making 40% of the others folder actually murders” [P28, Scenario 4]

> “This seems like a silly question. If it’s asking what percentage of students are ACCEPTED actually FAILED the exam, it states in the problem, ”1% of applicants who FAIL the entrance exam are also ACCEPTED into the university.” So the answer would be 1%.” [P2, Scenario 2]

> “there are 100000 applicants and 2000 pass the exam while 98000 fail the exam. when someone fails they have a 1% chance of still getting admitted. That means that the percentage chance that they actually failed the entrance exam but was still accepted should be 1%” [P21, Scenario 2]

In terms of the H/Pop and –H/Pop errors, the vast majority of participants simply stated that this was the answer to the problem, providing little insight into their underlying logic, e.g. P58 who stated that “200 [H] of 1000 [Pop] is 20%, so there is a 20% chance”. However, two participants did potentially indicate base rate conservatism logic (under-valuing of false positive / true positive figures) in stating that “the stats on false positives mean nothing in this question” (P34), and “trick question, still 20% [H/Pop]” (P28).

Discussion

Aims and hypotheses. Within the first study aim, the hypothesis that a significant main effect of the nested sets framing would be found has been confirmed, providing evidence of the student-population findings of Macchi (2000) in the general population. It was also hypothesised that no main effect of the causal framing would be found in the whole sample but a main effect would be seen in the higher numerate sub-group, in line with McNair and Feeney’s (2014b) findings. However, no main causal effect was found in either the whole sample or in any high numerate sub-group. This finding stands in contrast to Krynski and Tenenbaum’s (2007) original work, as well as subsequent replications by Hayes et al. (2013) who found a whole-sample effect, and McNair and Feeney (2014b) who found a causal effect in high numerates. The finding may however be in line with

McNair and Feeney (2014a) who found no causal effect in a whole-sample analysis.

Within the second aim, it was hypothesised that no significant interaction would be found between nested sets and causal framings in the whole sample, but that such an effect would be found in the high-numerate sub-group. In fact again no significant interaction was found in either the whole sample or any sub-groups.

Within the third aim, the results provide a more cohesive and systematic model of the processes people use to solve simple Bayesian problems (whether successfully or unsuccessfully) than has been presented previously within the nested sets literature, providing particular emphasis on the separation of the hypothesis and data-focused representations. Further, the qualitative data has made it possible to determine that this process model is followed by the vast majority of successful individuals in three major framing types explored in the literature (‘inside percentages’, ‘nested/partitive/subset’ and ‘causal’ formats), suggesting that this may be the preferred method of solution for individuals even in the absence of a specific attempt to encourage it (the basic condition) and even in the presence of a specific attempt to encourage a different process (the causal condition). This also suggests that the model is not simply a regurgitation of the nested sets framing, but is spontaneously produced by solvers in the absence of any prompt. Given that this was the modal response in this paper, this may provide exception to comments made by both Tversky and Kahneman (1983) and Sloman et al. (2003) who claimed that the ‘default’ problem-solving perspective was the ‘inside’ point of view unless an explicit cue was given to adopt the ‘outside’ perspective. In the present paper, a large number of participants adopted this outside perspective, and succeeded in solving the problem, even in the absence of any such cue (i.e. in the basic and causal conditions).

Nested sets and natural frequencies. There has been much previous debate in the literature in regards to the relative distinctiveness of the nested sets / partitive / subset / outside-framed approach to improving Bayesian reasoning from the natural frequencies approach. Hoffrage et al. (2002) claimed that Macchi’s (2000) outside-framed approach to improving Bayesian reasoning, as used in the present paper, works by encouraging individuals to construct a natural frequencies version of the problem for themselves, which is then thought to be the true cause of the increase in accuracy. This possibility continues to plague modern work such as Sirota et al. (2015) and can only be resolved by a protocol such as a think aloud analysis which can record solver processes. It is also given some evidential backing by the present experiment as ‘populating the hypothesis focused representation’ as it is termed in this experiment, could also be considered a ‘conversion’ to natural frequencies. However, as noted by Hoffrage et al. (2002), Macchi (2000) and the present paper used real-number values for the whole sample (e.g. 1000 women) as well as for the two base rates (e.g. 800 women have cancer, 200 women do not have cancer), which is likely to encourage participants to work with real numbers throughout the problem. It is possible that without this methodological feature, individuals may work through the process model without converting from percentages to frequencies.

Nested sets effect: text body versus question format. In the present paper an overall effect of the nested sets framing was found on accuracy rate. The nested sets framing used in this experiment and taken from Macchi (2000) contains changes to the body of the text and the question form. The body of the text contains the information related to the hypothesis-focused representation and computational step one, and any changes here may be expected to affect these primarily. The question contains the information related to the data hypothesis and computations two and three, and any changes here may be expected to affect these primarily.

When the steps of the process model were examined independently, it was found that there was an effect of the nested sets framing on the frequency of both the hypothesis-focused representation and the data-focused representation. However, when examining only those who successfully constructed the hypothesis-focused representation, no effect of the nested sets framing was seen on the data-focused representation. This can be seen in Figure 8, wherein a substantially larger amount of participants achieve the hypothesis-focused representation in the nested sets framing (36.7%) than the non-nested framing (22.1%), but subsequently, a very similar percentage of those participants (66.2%

in the nested condition vs 62.0% in the basic condition) go on to construct the data-focused representation.

When examining only those individuals who completed computational step one, there was also no effect of the nested sets framing on computational steps 2 or 3. This can be seen clearly in 7. A substantially larger number of nested sets cases achieved C1 (55.3% vs 42.0%) but out of these individuals a highly similar proportion of individuals (63.3% in the nested condition vs 59.0% in the non-nested conditions) achieved computational step two and step three.

Overall, this analysis suggests that Macchi’s outside-framed approach to improving Bayesian reasoning succeeds in improving the frequency of the hypothesis-focused representation and computational step one but may not succeed in additionally improving the frequency of the data-focused representation and steps two or three.

Numeracy and the process. This experiment also demonstrated that the average numeracy levels of those individuals completing each computational and representational step in the process increases largely from no steps to both computational step one, and also to the hypothesis-focused representation. A further, smaller increase is seen between these and step two, and the data-focused representation. No further increase in numeracy is seen for individuals achieving further steps after these. This progression is not, it should be noted, due to a cumulative effect of arithmetic errors, as only 5 of these were detected in the entire study, and they were removed for this analysis.

It is speculated therefore that this may indicate that the second step, and thus the second representation of the solution process may be more difficult to achieve than the first step, or representation. By this it is meant that individuals, even if they correctly initially perceive the problem in terms of the hypothesis representation, may need greater numerical ability to subsequently achieve the middle steps of the process. In contrast, the final computational step three, appears to be trivial once step two has been achieved.

Causal Null Finding. In regards to the null finding for the causal framing, several methodological issues must be considered in regards to the present experiment and some of which may require further experimentation to rule out. Firstly, given that McNair and Feeney (2014b) only found a causal effect in their high-numerate sub-group, it was considered possible that the numeracy level of the present sample might be the reason for the null finding. Indeed, the present sample did, in fact, have a lower median numeracy level than McNair and Feeney (when correcting for their ‘missing’ question the present study had a median of 4 while McNair and Feeney had a median of 5). While no high-numerate sub-group, including one constructed to have the same parameters as McNair and Feeney’s high numerate group showed a causal effect either, this subgroup was quite small in size (n = 35) and so may have lacked power to detect the effect.

Another potential limitation could be that the three scenarios which were invented for the study (College, Library and Gotham) may not have been designed adequately to test the causal framing. The causal manipulation rests on the assumption that in the basic condition the ‘cause’ of the false positive rate is not only not stated (which is simple to ensure), but further, not ‘easily inventible’ by the solver either. If an obvious second cause springs to mind for the solver then they would be able to create a causal mental model equally well in both conditions and no difference would be predicted between the two conditions. The three new scenarios were therefore all carefully designed to ensure that in the basic version the cause would not be obvious. However, it is possible that the cause was more obvious than in the original mammogram problem used by Krynski and Tenenbaum (2007) which may have weakened the overall effect of the causal framing in this study. Some evidence for this comes from the fact that the causal condition outperformed the basic condition in the mammogram scenario to a greater extent than in the other scenarios. However, even this difference was very far from significant and the combined nested-causal condition in fact under-performed in comparison to the nested condition even in the mammogram scenario. Further, the fact that no single participant even referenced a causal structure in their think aloud data, even in the causal condition, casts further doubt on this explanation.

A further possibility is that the fact that the Total Population and H and -H were given as sub-divided frequencies, rather than the percentages used in Krynski and Tenenbaum's (2007) study may have 'got participants started' with constructing a nested sets representation, precluding them from taking a causal approach. This fact is also an alternative explanation for why the nested sets representation was modal in all four conditions. In effect, it is possible that all conditions all contained a mild nested sets prompt.

A further possibility is that the repeated measures nature of the study may have led to practise effects, which, if present, would have the effect of making the accuracy of all conditions more similar to each other, and thus reducing the effect size of both the nested sets and causal framings. However, the causal condition actually produced slightly lower accuracy than the basic framing and the nested-causal framing also produced slightly lower accuracy than the nested sets framing alone. Practise effects could only reduce effect sizes and could not reverse their direction, suggesting that this is not a good explanatory candidate for the null finding.

One final possibility is that the introduction of the 'think aloud' process reduced the effect of the causal prompt. Given that participants were asked to write down their thought processes prior to giving a numerical answer, it is plausible that this also encouraged them to think more deeply and for longer about the problem than in previous experiments not containing this feature (such as Krynski and Tenenbaum, 2007). This has been suggested previously by Ericsson and Simon (1998). Depending on the mode by which the causal framing facilitates reasoning on the problem, this additional thinking time may have compensated for its absence in the basic condition.

Errors. The present experiment found little variety in error types between conditions but great variety between different scenarios and different numeracy levels. The medical diagnosis scenario was notably unique in producing a large number of 1-P(D|H) errors, while the most common error overall, P(D|-H) was mostly made within the College and Gotham scenarios. Both of these were also common errors in Gigerenzer and Hoffrage (1995) who used a version of the medical diagnosis scenario and was there named the ’false alarm component’ and ’Fisherian’ responses, respectively. Other common errors included the provision of the conjunction of P(H&-D) and the individual base rates H and -H expressed as a percentage of the overall population. Overall, errors tended towards being more simplistic in terms of the number of computational steps employed, with zero and one-step solutions comprising 84.0% of all errors. This tendency can be seen in the six most common errors, which broadly involve either the provision of a figure provided in the problem text (e.g. P[D|-H]), or the division of such a figure by one other, such as the population.

Examination of the think aloud protocols for participants providing the most common error, P(D|-H), found five instances of individuals showing clear confusion between P(D|-H) and P(-H|D). The majority of the remaining participants merely stated that P(D|-H) was the answer to the problem, which could be attributed to either the previous confusion of conditional probabilities, a misunderstanding of the question asked in the problem, or neglect of the importance of the base rate.

Experiment two

In experiment one, Macchi's 'outside-framed percentage' approach to increasing accuracy on Bayesian problems was replicated and found to be efficacious in the general population. Second, the nested sets process model outlined in that paper was found to be modal in all conditions, regardless of specific framing. Third, the increase in accuracy provided by the nested sets framing was found to coincide with a greater number of individuals following the nested sets process model than with other framing types.

These findings suggest that Macchi's approach could have widespread social value in situations such as medicine and law, where the general public are frequently exposed to Bayesian problems. It also suggests that successful individuals typically solve such problems via a single five-step process. However, the problems used in that study were relatively simplistic in a number of ways and may have suffered from a lack of realism, or ecological validity. If the nested sets approach is to be used and recommended for real situations and the nested sets process model advocated as a general solution procedure for Bayesian problems, both must be tested without these fictitious simplifications.

The following elements will be altered in the current experiment to increase ecological validity. First, the original problems used had a 0% ‘false negative’ rate (e.g. ‘All women who have cancer receive a positive result’), which simplified the problem but is impossible with any real test. The present study will add a non-zero false negative rate (e.g. ‘Out of all the women who have cancer, 80% receive a positive result’). While Macchi (2000) in fact did include this complication, that paper did not publish the solution processes of their participants. This added complication will necessarily make the nested sets process itself more complex and could therefore feasibly deter participants from perceiving, or following it. This may result in a weakening of the nested sets effect, and potentially in participants using a different process to solve the problem.

Second, experiment one, similarly to Macchi (2000) and Fiedler et al. (2000), used real numbers for the total population and / or the base rates e.g. ’200 out of 1000 women have cancer’. This again may not be the case in all real settings and most previous work (Casscells et al., 1978; Eddy, 1982; Kahneman and Tversky, 1972; Krynski and Tenenbaum, 2007) has generally used percentage base rates and not explicitly included a population figure at all (e.g. ‘20% of women have cancer’). Hoffrage et al. (2002) also theorised that the real number format might encourage individuals to construct a ‘natural frequency’ version of the problem for themselves, which may be the cause of the facilitation in Macchi (2000) and Fiedler et al (2000). Thus it is possible that in experiment one this use of real numbers may have ‘got participants started’ in following the nested sets process as that process begins with the simulation of a target population which is then sub-divided. Without the provision of this, it is possible that the nested sets framing effect may be reduced and that participants will use a different process to solve the problem.

Third, the particular numbers used in experiment one and in Macchi (2000) also allowed participants to work solely with whole numbers throughout the entire solution procedure, until the final product. This would also be very uncommon in a real setting. The nested sets process relies upon mental simulation of units and their sub-divisions and previous work (e.g. Brase, 2007; Cosmides and Tooby, 1996) has suggested that individuals may have difficulty mentally simulating fractions of units. Therefore, fractional values may deter individuals from following the nested sets process model. This ‘fractional values’ effect may be weaker with the percentage values used in the present experiment than if real numbers were used. However, previous work (Hoffrage et al, 2002) and our results from experiment one suggest that a significant proportion of individuals are likely to convert the problem in to real numbers, or ’natural frequencies’ as part of their problem solving process. This will therefore allow observation of whether this ‘fractional values’ effect affects solution process both for participants who solve with percentages and with real numbers.

It is therefore hypothesised that a positive significant overall effect of the nested sets framing on both accuracy and completion of the nested sets process will be seen in comparison to the non-nested condition both in the sample as a whole, and separately within both the ‘hard’ and ‘decimal’ conditions. Further, to test the role of the nested sets process found in experiment one, a mediation analysis will be conducted with the hypothesis that the hypothesis-focused and data-focused representations will mediate the relationship between the nested sets framing and solution accuracy.

Method

Participants. The final sample for the experiment was 521. From an original sample of 528, seven individuals were removed because they stated that they had undertaken the problem presented in the experiment previously. Demographic data can be found in Table 1. Participants were paid an average of $6.00 per hour for taking part.

Design. The study was a between-subjects 2 (non-nested vs nested sets framing) x 2 (whole numbers vs decimals) x 2 (simple problem vs hard) design resulting in eight groups. It also employed the same mixed-methods design using the ‘think aloud’ procedure developed by Ericsson and Simon (1998) and used in experiment one.

Materials. The same medical diagnosis problem used in experiment one was again employed. The simple conditions used a true positive rate of 100%, identical to experiment one (e.g. ‘All women who have cancer receive a positive result’). This allowed participants to use a calculation shortcut in which they substituted H (the number of women with cancer) for (H&D: the number of women with cancer and a positive result) because the former simply needed to be multiplied by 100% (the true positive rate) to obtain the latter, making no change. In the hard condition however, the true positive rate was set at less than 100%, introducing the possibility for false negatives and requiring participants to calculate both conjunctions and therefore increasing solution and representational complexity.

The whole-number conditions used figures which produced whole number products at every stage (i.e. in this condition, the conjunctions were whole numbers) in the process except the final product, which was a decimal in all conditions. The decimal-condition importantly resulted in decimal values for the two ‘conjunctions’ of (H&D: women with cancer and a positive result) and (-H&D: women without cancer and a positive result), the calculation of which were a necessary step to solution of the problem. All condition problems can be seen in the supplementary materials. The nested-decimal-hard condition can be seen below.

> Every year the government advises women to take part in routine mammography screening using an X-ray machine to determine if they have breast cancer. Among women at age forty who participate in this routine screening 10%

have breast cancer, while 90% do not. However the screening test is not always accurate. Specifically, out of those women who have breast cancer, only 76% will actually get a positive mammography. Furthermore, out of all of those women who do not have breast cancer, 15% will also get a positive mammography. What percentage of women at age forty who get a positive mammography in routine screening actually have breast cancer?

Procedure. Participants were recruited through Amazon MTurk. Participants were presented with the consent form, and then the instructions for the study, which included an extensive section on the ‘think aloud’ instructions, including an example (see supplementary materials). Participants were then randomly assigned to one of the eight conditions. For each problem they were presented with the problem text and question itself and were asked to write their thought processes while they worked out the problem in a ‘think aloud’ open-ended text-box. They were also provided with a link to an online calculator wherever required. Once this was complete they were able to give their actual numerical answer on the next page. Finally they answered the demographic questions and a final question regarding whether they had undertaken the problem in the study before.

Data Analysis. The same dual criteria to determine correct answers used in experiment one was again employed in experiment two.

Results

Quantitative. Overall accuracy for the experiment was 13.5% with an average accuracy of 9.0% for the non-nested conditions and 18.1% for the nested conditions. In Figure 10 below, accuracy for all eight conditions can be seen.

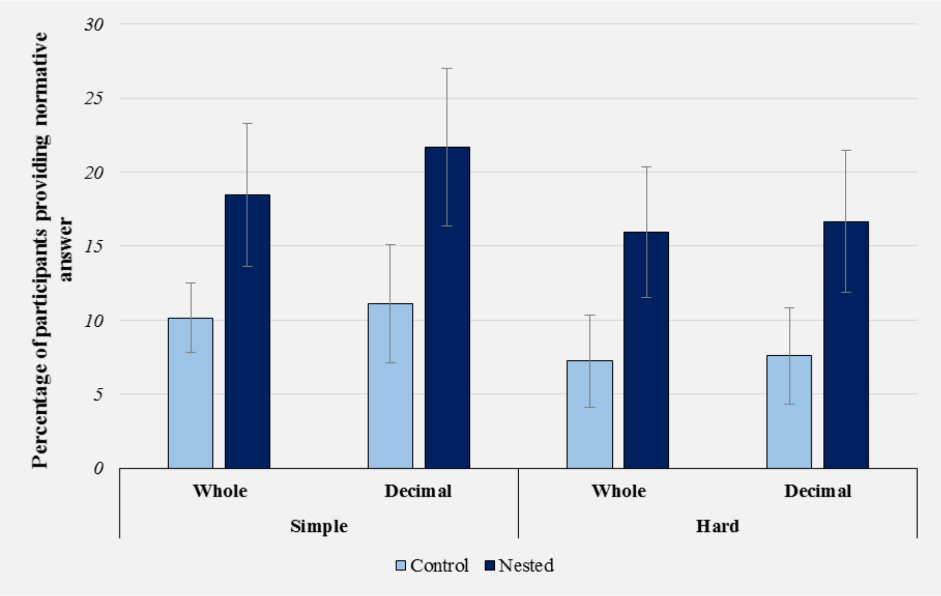


Figure 10. The percentage of participants providing the normative Bayesian answers across all eight conditions. Error bars represent one standard error.

A between subjects GLM with binomial distribution and logit function, using ‘score’ as the dependent variable and the three condition-comparisons (non-nested vs nested; whole vs decimal; simple vs hard) as independent variables found a highly significant main effect for the non-nested-nested comparison (Wald = 8.984, p=.003), no main effect for the whole-decimal comparison (Wald = .184, p=.668) and no main effect for the simple-hard comparison (Wald = 1.350, p=.245).

To determine if the effect of the nested sets framing was significantly present within the four ‘decimal’ conditions, the same GLM model was run on this group only. A main effect of the non-nested-nested comparison was found (Wald =

4.821, p=.028). Similarly, to determine if the nested sets effect was present within the four ‘hard conditions, a GLM was run on this group only and a main effect of the non-nested-nested comparison was found here also (Wald = 4.784, p=.029).

Qualitative. ***Process model.*** In Figure 11 below, the percentage of individuals achieving every step of the process model can be seen for all eight conditions. A highly similar pattern to overall success is immediately apparent, with the nested sets conditions producing more process models in every instance.

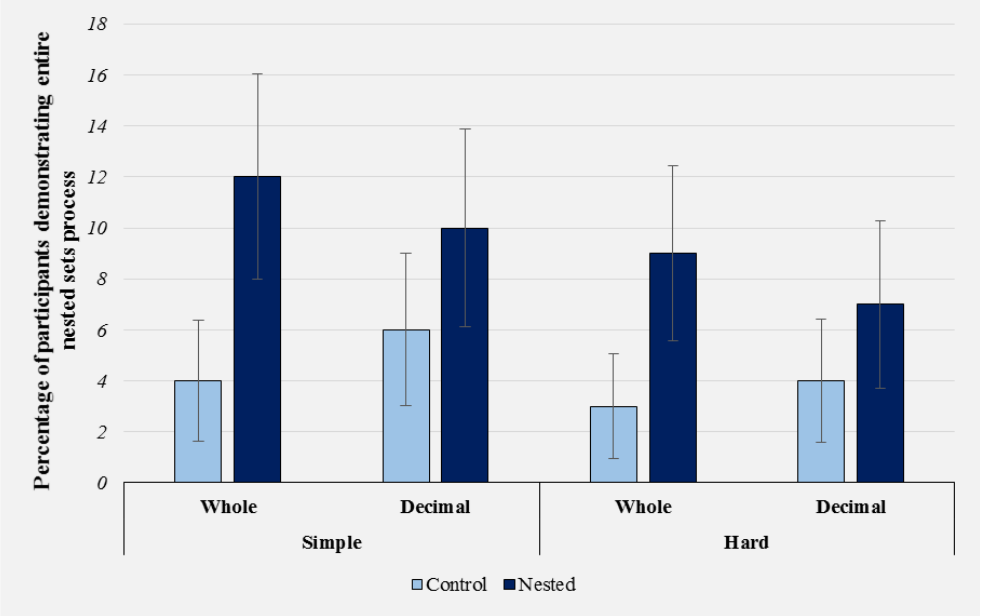


Figure 11. The percentage of individuals achieving all steps of the nested sets process model for all eight conditions. Error bars represent one standard error.

A binary logistic regression was run using ‘All Steps’ (a variable which was coded to be ‘1’ if participants completed all steps in the process, and 0 otherwise) as the dependent variable and the nested sets vs non-nested, simple vs causal and whole vs decimal variables as independent variables. A highly significant effect of the nested vs non-nested variable was seen (Wald = 9.728, p = .002), while no significant effect was seen for the simple vs causal (p = .161) or for the whole vs decimal condition (p = .816).

In Figure 12 below, a similar drop-off graph to experiment one for both computational steps and representations can be seen. A similar pattern in both cases is once again apparent, wherein the vast majority of drop off occurs prior to representation one / computational step one, with further substantial but smaller drop off between these and representation two / computational step two, and no substantial subsequent drop off between these and step three or accurate completion of the problem. Further, similar curves were once again seen for both the nested and non-nested conditions. The major difference between the two conditions was the amount of drop off at representation one / step one. After this, and identically to experiment one, highly similar subsequent proportional attrition in both conditions was seen.

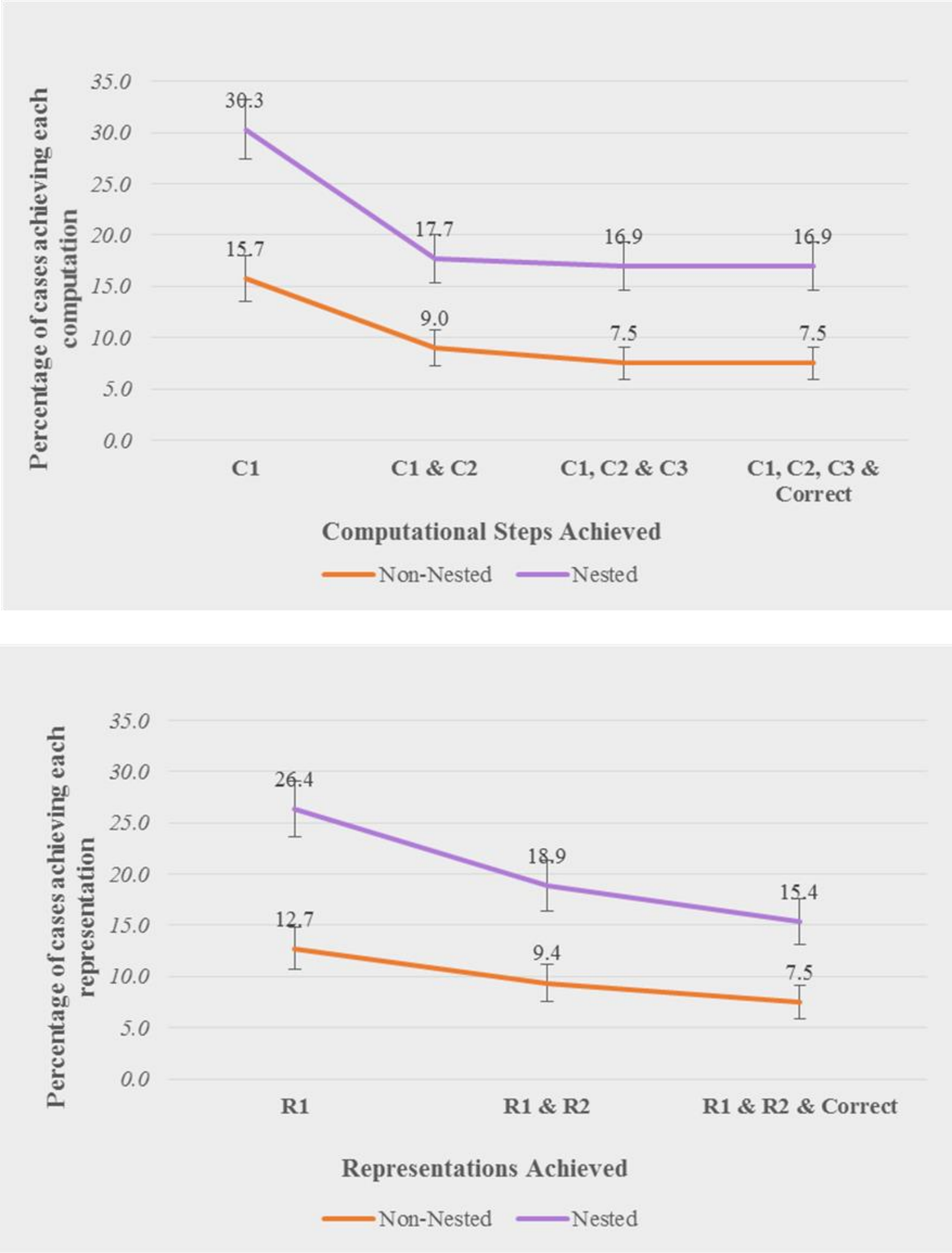


Figure 12. Drop off graphs for each computational and representational step. Error bars represent one standard error.

Confirming the results of Figure 12, a further analysis was conducted to test the finding from experiment one that the nested sets framing effect on the data-focused representation, and computational step two, while significantly predictive alone was non-significant when controlling for the presence of the hypothesis-focused representation, and computational step one, respectively. When examining only those who produced the hypothesis-focused representation, the nested sets framing did not predict the frequency of data-focused representations in this experiment (Wald = .04, p=.841). When examining only those who produced computational step one, the nested sets framing also did not predict the frequency of computational step two (Wald = .019, p=.891).

Conversion to real numbers. Think aloud data indicated that eighty eight participants (16.9%) converted the problem from percentages into real numbers before attempting solution. For this classification, a ‘sample’ or ‘population’ of women with a real number rather than a percentage or probability had to be expressed. For example P105 said ‘To make my math easier, I am going to assume there are 100 women.’ and P186 began ‘Out of 100 women, 10 have breast cancer, while 90 do not.’ Out of the 88 participants who converted the problem to whole numbers, 73 converted to a population of 100 women, and 9 converted to a population of 1000.

Conversion of the problem into real numbers was highly associated with success on the problem. Out of the 434 participants who did not convert the problem, 6.5% provided the normative answer, while out of the 87 individuals who did convert the problem, 48.3% provided the normative answer. This relationship was highly significant (Wald = 80.6, p<.001).

Conversion of the problem was also associated with a greater frequency of nested sets and data-focused representations. Binary logistic regressions showed a significant main effect of conversion on the hypothesis-focused representation (Wald = 93.1, p<.001) and on the data-focused representation (Wald = 75.4, p<.001).

A between subjects GLM with binomial distribution and logit function examined the prevalence of conversion between conditions. A significant main effect of the nested sets condition was found (Wald = 7.233, p=.007) with 12.4% converting in the non-nested condition and 21.3% converting in the nested condition. However no main effect of the whole-decimal comparison (Wald = 0.7, p=.412) or the simple-hard comparison (Wald = 0.3, p=.615) was found on conversion rate.

Finally, it should be stressed that this conversion was not unanimous amongst successful participants, nor amongst those who followed the nested sets process. Of successful individuals, 40.0% did not convert from percentages. For example, P245 correctly solved the problem while entirely using percentages:

> ‘10% have breast cancer, 90% do not - Participants / 76% of the 10% test positive / 15% of 90% test positive / / 76% of 10% is 7.6%. 15% of 90% is 13.5% / / 13.5% + 7.6% = 21.1% / / 13.55%/21.1 = 63.981%’

Dealing with decimals. Out of those individuals who converted the problem to real numbers, nine used a base of 1000 and eight of these were in one of the decimal conditions. Converting to a sample of 1000 simultaneously turned the values in the decimal problem into whole numbers, suggesting this was one strategy that some individuals used to deal with the problem of decimals. However, out of the 34 (13.7%) individuals who achieved step two in the decimal conditions, only two converted to a base of 1000. One more individual converted to a base of 110 which also provided whole numbers in that particular condition. The remaining 31 individuals (91.2%) dealt with the decimal values in a precisely analogous way to the equivalent figures in the whole-number conditions and no single individual attempted to round the decimal values up or down. Further it has already been shown above that an equal amount of nested sets and data-focused representations were found in the decimal conditions as the whole-number conditions. These results suggest that apart from three individuals, successful participants in the decimal conditions dealt with the figures in the precisely same way as individuals in the whole-number conditions. This was not only the case when participants dealt directly with percentages either. Fifteen participants out of the 34 who achieved step two in the decimal conditions converted to a sample of 100 and these individuals frequently mentioned fractions of women. For example, P109 said ‘Out of 100 women, 23.5 women will have test results show positive for cancer’, P482 said ‘so 13.5 women who don’t have breast cancer will also get a positive mammography' and P480 said 'This would mean that 7.6 women out of 10 women who have breast cancer would have a positive mammogram.'

Errors. Error were categorised by type using the same method as experiment one. The five most frequent errors can be seen below.

Error Type	Grand Total	Control	Nested	Whole
Total (Frequency)	451.0	243.0	208.0	237.0
1-P(D|-H)	20.2%	28.8%	10.1%	20.3%
P(H&D) (Simple [H])	7.3%	10.7%	3.4%	5.9%
Hard Only: P(H&D)	6.0%	3.7%	8.7%	6.3%
P(-H&-D)	4.4%	2.9%	6.3%	5.1%
Hard Only: P(D|H)	4.0%	5.8%	1.9%	3.8%

	Decimal	Simple	Hard
Total (Frequency)	214.0	218.0	233.0
1-P(D|-H)	20.1%	20.6%	19.7%
P(H&D) (Simple [H])	8.9%	9.2%	5.6%
Hard Only: P(H&D)	5.6%	0.0%	11.6%
P(-H&-D)	3.7%	9.2%	0.0%
Hard Only: P(D|H)	4.2%	0.0%	7.7%

The think aloud data for the most common errors for both basic and nested conditions was also explored to determine if any underlying cognitive processes could be detected which could provide understanding of why the error occurred. To complement the analysis in experiment one and provide the most valuable information for future research to build upon, the present analysis examined only the most ecologically valid conditions: the decimal-hard conditions. Brief comparisons to the overall rates will also be given.

Non-nested sets. The most common answer within the non-nested decimal-hard condition used zero computational steps and was to provide the complement of the false positive rate, (1-P[D|H]). This answer was given by 25.8% of all participants in that condition. It was also the most common error in the non-nested conditions overall.

The think aloud data was coded by the first author and second coded by the seventh author for further insight into common reasoning that lead to this mistake and a single piece of reasoning was found to be highly prominent (45.8% of cases). This was the confusion of the false positive rate (the rate at which women without cancer still get a positive test result) with the percentage of all positives that were in fact false. Following this confusion, the subsequent accurate deduction was made that 100% minus this value would give the percentage of positives which were correct, which is the answer to the question. This is a confusion of P(D|-H) with P(-H|D). For example, P228 said ‘The fact that 15% of positive mammographies are invalid means that 85% are valid. She therefore has an 85% chance of actually having breast cancer’, P20 who said ‘I guess since 10% of positive tests are inaccurate, that means there’s a 90% chance of her having cancer’ and P133 who said ‘Also of all the women who get a positive mammogram, 15% will not have breast cancer, so I think it is 85%.’. Each of these participants produced the faulty logic that if 1-P(D|-H) = X, then P(H|D) = 1-X, heavily implying a confusion between 1-P(D|-H) and 1-P(-H|D). In some cases a direct confusion between these two was explicitly stated as in P177 who said ‘But there is a 10 percent chance that a woman without breast cancer will get a positive mammogram [true, P(D|-H)], so 10 percent of the positive mammograms are not accurate [false, P(-H|D)]’. In the remainder of participants’ think aloud data, the reasoning could not be extracted from the data. For example, many participants simply provided mathematical notation.

Nested sets. The 1-P(D|-H) answer, while the most common in the non-nested conditions, was in fact only given by 6.7% of all participants in the nested sets decimal-hard condition, making it the second most common answer. The two most common answers in this condition were the correct answer, and the conjunction P(H&D): the percentage, or number, of women with both breast cancer and a positive test result. Each of these was given by 16.7% of participants in this condition. Again these results were mirrored in the overall nested sets conditions. The P(H&D) answer is obtained by multiplying the base rate for cancer with the true positive rate. Its calculation is part of the first step to answering the question correctly.

A single reasoning process behind this error proved more difficult to extract by the coders. However, out of the total 31 individuals who made this error, six clearly stated that they were aiming to find the ‘percentage of women with a positive result and breast cancer', suggesting a potential confusion in the reading of the question. For example, P420 concluded by saying 'so it would be 8% that have positive screens and actual breast cancer.' Similarly, a further 18 people simply stated that 10% of women had breast cancer and X% would get a positive result, then provided the product of these as the answer. This may suggest a similar misunderstanding of the aim of the problem to the six people who articulated this more explicitly. For example, P418 said 'So if 10% of women actually have breast cancer and only 80% of those will actually have received a positive result. So 10% of 100 is 10 and 80% of 10 is 8.' From the remaining seven individuals, no process could be divined.

Mediation analysis. A mediation analysis was carried out to test if the effect of the nested-sets framing on the 'score' variable was mediated firstly by conversion to real numbers and secondly by the nested sets and data-focused representations of the process model proposed in experiment one.

In the first model, the non-nested-nested comparison variable and the conversion variable were used as independent factors in a binary logistic model with score as the dependent variable. A borderline significant effect of the nested sets condition variable was found (Wald = 4.1, p = .042) and a large significant effect of conversion was found (Wald = 76.4, p<.001).

In the second model, the nested condition variable and the hypothesis-focused and data-focused representations were included as independent variables.

In this model, the nested sets condition variable was a non-significant predictor of accuracy (Wald = 0.0, p=.933) while both hypothesis-focused (Wald = 32.6, p=<.001) and data-focused (Wald = 19.9, p=<.001) representations were large significant predictors.

The pattern of these results was confirmed by a series of Sobel tests. A significant mediation effect of conversion on the relationship between the nested-non-nested variable and score was found (z = 2.57, p=.010). Further, mediation of the relationship between the nested-non-nested variable and score was found by the hypothesis representation variable (z = 3.47, p=0.00) and also by the data representation variable (z=3.03, p=0.00).

Discussion

Aims and Hypotheses. The present study aimed to determine if the nested sets framing effect (Macchi, 2000) on accuracy would remain with three methodological departures from experiment one; using a percentage base rate instead of frequency, using decimal values instead of whole numbers, and using a more complex problem than previous, which included false negatives. All three of these departures were intended to increase ecological validity of the problem.

Overall, a main effect of the nested sets condition variable on accuracy was found. No main effect of the whole-decimal comparison or the simple-hard comparison was found. The nested sets framing effect was also found separately within the four ‘decimal’ conditions and within the four ‘hard’ conditions. A significant relationship between the nested sets condition and completion of the nested sets process was also observed while no main effect of the whole-decimal comparison or simple-hard comparison was observed. This confirms the first hypothesis.

A mediation analysis compiling all conditions found full mediation of the nested sets effect on accuracy by the hypothesis-focused and data-focused representations of the process model. This latter finding confirms the second hypotheses of the study suggesting that the increase in accuracy obtained by employing a nested sets framing occurs by encouraging more individuals to follow the outlined nested sets process.

Nested Sets and Natural Frequencies. The present results suggest that a nested sets format also increases accuracy on Bayesian problems when percentage base rates are used. In regards to whether this increase is equal in effect size to when real number base rates are used, a direct comparison cannot be made as percentage and real number base rates were not directly compared in the present experiment. However, some comparisons may prove informative. Overall accuracy in the simple-whole-nested condition (18.5%) was lower than the comparable condition in experiment one, which found 38.9% accuracy. This may be due to the repeated-measures nature of experiment one, however, accuracy in the nested-whole-hard condition (15.9%) was also lower than the comparable condition in Macchi (2000), who found 33.3% and used a between subjects design. Thus, this suggests that the use of ‘real number’ base rates instead of percentages may improve accuracy on Bayesian problems beyond the nested sets format alone. However, direct comparison would be beneficial.

Previously, Hoffrage et al (2002) hypothesized that the success of the nested sets approach (Macchi, 2000; Fiedler et al., 2000) may have been due to participants constructing a ‘natural frequency’ version of the problem for themselves, thus providing the increased accuracy seen in those problems. In partial support of this theory, the present results demonstrate that the majority of successful participants (60%), when presented with percentage base rates, first constructed a ‘real number’ version of the problem. This phenomena was also noted by Cosmides and Tooby (1996), who briefly examined participants’ workings out in their experiments but did not conduct a systematic analysis. This conversion process was also highly associated with success. Furthermore, mediation analysis demonstrated a partial mediation effect of conversion of the problem to real numbers on the relationship between the nested sets framing and solution accuracy. This confirms that while the nested sets framing did encourage more individuals to convert the problem to real numbers, and this led to greater success, a substantial portion of the increased success of the nested sets framing cannot be attributed to conversion. A substantial portion of successful participants (40% of correct answers) were content to follow the nested sets process in percentage form, and without any mention of a population or subsequent creation of a ‘natural frequency’ version of the problem (Hoffrage et al, 2002). This suggests that this process does not necessarily require the simulation of a ‘real number’ population, or the use of real numbers at all. Therefore, in temperance of Hoffrage et al.’s (2002) conjecture, our results indicate that the construction of a natural frequency format does not appear necessary for solving Bayesian problems or to follow the five-step ‘nested sets process’, but does appear to increase the likelihood that solvers will do so.

The reason for the benefit of converting the problem from percentages to real numbers is unclear on the present data. A majority of participants who converted the problem did so from a base of 100% to a base of 100 women, making no mathematical change to the problem and suggesting some other benefit was conferred. The source of this benefit from real number presentations would be another fruitful direction for future research.

Decimal Values. A further finding was that overall, individuals dealt with decimal values in exactly the same way as whole numbers. Even amongst those individuals who converted the problem into real numbers, no difference in accuracy was seen between decimal and whole-number conditions. A difference may have been plausible as there may be a psychological difference between conceptualizing ‘12.5% of women’ and ‘12.5 women’, with the latter being a metaphysical impossibility. However, the think aloud data indicated that no single individual attempted to round the real decimals up or down, and participants appeared to deal with the fractional numbers of women in precisely the same way as their counterparts in the whole-number conditions.

Errors. ***1-P(D|-H) error.*** The most common error in the non-nested-decimal-hard condition when presenting with an individual or ‘chance’ framing was 1-P(D|-H). This was also the second most common error in the nested-decimal-hard sets framing. It was named the ‘False Alarm Complement’ in Gigerenzer and Hoffrage (1995), and was given by 3.4% of participants in their second experiment. It was also found by Macchi (2000) as well as by 5.5% of participants in our experiment one. Our think aloud data analysis determined that the most common reasoning process behind this error was to confuse P(D|-H) with P(-H|D), a finding which fits with previous work advocating the ’confusion hypothesis (e.g. Hamm, 1987; Hamm & Miller, 1988; Wolfe, 1995; Macchi, 1995).

P(H&D) error. The most common error within the nested-decimal-hard sets condition was to provide the conjunction P(H&D). This answer was also the second most common error in Gigerenzer and Hoffrage’s (1995) first experiment (there named ’Joint Occurrence’), and even more common than the correct answer in their second experiment. This answer was reported in combination with other ‘Pseudo-Bayesian’ answers in Macchi (2000) which collectively totalled the most common error also. The error was not possible in experiment one as there was no possibility for false negatives.

It proved more difficult to discern a general pattern underlying this ‘Joint Occurrence’ error. However, the most frequent reason identified was a mis-reading of the question, wherein participants seemed to be searching for ‘the percentage of all women with a positive result and breast cancer’, rather than ‘the percentage of women who have breast cancer among those with a positive result’.

Experiment three

Experiment one produced a null finding for the causal framing. However, there were some notable differences between that experiment and previous work which may have contributed to this and were discussed. Experiment one firstly was within-subjects. It also used real-number base rates, a think aloud protocol and multiple scenarios, some of which previous work had not examined. Each of these may have contributed to the null finding for the causal framing.

The aim of present experiment was therefore to test the causal framing effect without these potential confounds. We used percentage base rates similar to those use by Krynski and Tenenbaum (2007), which were ‘outside percentage’ in that experiment while the false positive rate was presented as an ‘inside percentage’. Further, to test the theory that a think aloud protocol may have affected the result (as theorised in experiment one) we also varied whether or not participants were asked to engage in the think aloud protocol. Based on results from experiment one it was hypothesised that in a GLM analysis with logit function a think aloud effect would be seen but no causal effect would be. It was also hypothesised that an interaction between the think aloud and causal effect would be seen, with post-hoc analysis revealing a difference between causal and control within the non-think aloud conditions.

Method

Participants. Twenty seven participants were removed because they stated they had completed the medical diagnosis problem before. The final sample consisted of 429 participants, recruited through Amazon MTurk. A breakdown of the demographics for the experiment can be seen in Table 1.

Design. The experiment comprised a between-subjects 2 (causal vs non-causal) x 2 (think aloud vs non-think aloud) design.

Materials. The experiment was an online survey which participants accessed through their own computers. Colour-blind safe colours were used where colour was necessary, which were sampled from www.colourbrewer.org. The same medical diagnosis problem used in experiments one and two was again used.

Procedure. Participants were shown the consent form for the experiment, and then were randomly assigned to one of the four experimental conditions. Each group was then shown a set of instructions for the experiment which were more extensive for the ‘think aloud’ group (see supplementary materials for think aloud instructions). Participants were then presented with the problem and given the opportunity to respond. Participants in the non-think aloud groups were asked simply to provide a numerical response to the problem, while participants in the think aloud groups were asked to provide a verbal record of their thoughts while working out the problem before being allowed to enter their numerical response on the following page.

Data Analysis. Due to the fact that the main analysis of experiment three compared think aloud conditions to non-think aloud conditions, the same dual criteria analysis of ’correct’ answers employed in experiments one and two could not be used. Instead, in line with previous non-think aloud work (e.g. Krynski and Tenenbaum, 2007; McNair & Feeney, 2014a) answers within 1% of the normative correct answer were categorised as correct.

Results

Quantitative. When comparing the think aloud to the non-think aloud conditions, analysis must rely entirely on the numerical answer to determine which participants were correct. The percentage of participants giving the correct numerical response for each condition can be seen below in Figure 13.

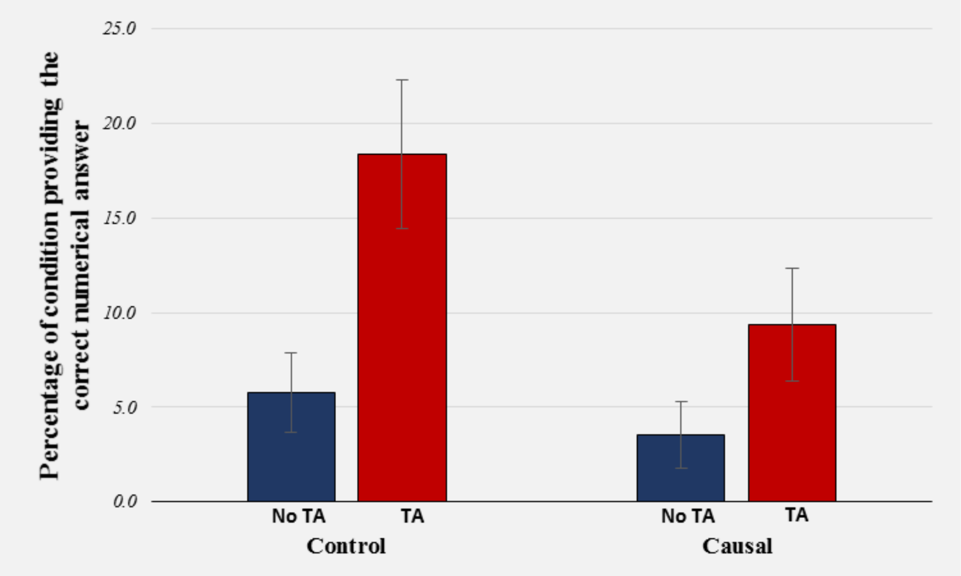


Figure 13. Percentage of participants providing the correct numerical response across all four conditions in experiment three. Error bars represent one standard error.

It is clear that the think aloud conditions produced substantially more numerically correct answers than the non-think aloud conditions. However the causal conditions also clearly showed either equal, or less accuracy than the control condition. A logistic regression model, using the binary ‘correct’ variable as criterion and the ‘TA vs non-TA’ and ‘Causal vs non-Causal’ variables as predictors demonstrated a significant effect for the TA variable (Wald = 9.065, p=.003), but no significant effect for the causal variable (Wald = 2.774, p =.096) and no interaction effect (Wald = .114, p =.735)

Within the TA conditions, think aloud protocols were analysed to ensure that participants who provided the correct numerical responses had also undertaken an appropriate method to arrive at the answer provided. Within the control condition, 17 out of 18 participants who provided the correct answer were also classified as having used an appropriate method. Within the causal condition, all nine out of nine participants were classified this way. A binary logistic regression using ‘correct method’ as criterion variable and ‘causal vs non-causal’ as predictor variable revealed the difference in participant accuracy between these two conditions was non-significant (B = -.707, = 2.583, p=.108).

Qualitative. In terms of the process model outlined in experiment one, 2.6% of participants in experiment three completed every step of the model in the causal condition, and 4.6% in the non-causal condition. Out of those individuals who provided the normative answer, 36.8% demonstrated all five steps of the nested sets process. All other participants failed to demonstrate either one or more steps of the process. Only one approach other than the nested sets process was detected in a single participant: P379 directly used Bayes’ formula by plugging the values in the problem into the appropriate places in the formula and computing the answer. No representation of the problem in terms of a causal structure was detected in the think aloud protocol of any single participant.

From the think aloud protocol it was also possible to replicate the analysis from experiment two of those participants who converted the problem from the original percentage format to a real-number format. Out of the 168 participants who did not get the correct method, no single participant made any numerical conversion of the problem. However, out of the 26 participants in the think aloud conditions who provided the correct Bayesian answer, 14 (50%) converted the problem from percentages to whole numbers. Twelve of these converted to a sample of '100 women', while one converted to a sample of 40 and another converted to a sample of 10. An example of this comes from P5 who began by stating 'Say that 100 women get a mammogram. Then 20 will have positive findings because they have BC.'

It was also possible, using numerical analysis only, to make an estimate of the effect of the think aloud protocol on the three most common errors in this experiment: 1-P(D|-H), P(D|-H) and H. No significant difference in frequency of these errors was seen between non-think aloud and think aloud conditions for the 1-P(D|-H) error (17.0% of all incorrect answers vs 12.6%: Wald =1.47, p=.255), the P(D|-H) error (17.5% of all incorrect answers vs 16.2%: Wald = .118, p=.731) or the H error (26.5% of all incorrect answers vs 22.2%: Wald = .966, p=.966).

Discussion

The aim of the third experiment was to test the null causal finding of the first experiment in a between-subjects design using a full-percentage scenario and without think aloud protocol. A further aim was to test whether an interaction effect existed between the think aloud protocol and the causal framing. An effect for the think aloud protocol was detected, but no effect for the causal framing and no interaction effect. Overall, the causal framing actually produced a non-significant, but lower, level of accuracy. Therefore the first and second hypotheses, of a think-aloud effect and a null overall causal effect, were confirmed. However, the hypothesised interaction effect between think aloud and causal variables was not detected.

The results of this experiment suggest that the null finding for the causal framing in experiment one were not due to the within-subjects nature of that experiment, nor to do with the use of real number base rates or the use of a think aloud protocol. The results therefore give further evidence against Krynski and Tenenbaum’s (2007) theory that, firstly, participants represent simple Bayesian word problems as a causal mental model, and secondly, that providing the second ’hidden’ cause in the medical diagnosis problem can increase accuracy rates.

Experiment four

The majority of previous work using the ‘outside view’ percentage format (e.g. Macchi, 2000; Fiedler et al, 2000; experiment one) has provided participants with either a population value or a base rate in ‘real number’ form in the problem text used. Hoffrage, Gigerenzer, Krauss and Martignon (2002) theorised that this may have encouraged participants to construct a ‘natural frequency’ representation of the problem, and that this may be part of the reason for the effect of the nested sets framing. However, experiment two provided percentage base rates, no real-number figures at all and still found increased accuracy with the nested sets framing. This suggests that the provision of the real-number population / base rates cannot account for the ‘nested sets’ effect. However, it is important to note that accuracy in experiment two was considerably lower than an equivalent condition in experiment one. The only two differences between these conditions is the use of a population figure vs. real-number base rates and the overall study design (within subjects vs between subjects). This suggests that one or both of these factors may increase accuracy on Bayesian word problems. Further, Experiments 2 and 3 found that a large proportion of participants who were not provided with real number population figures constructed such a figure for themselves in the early stages of solution of the problem. This ‘conversion’ of the problem into real numbers was also highly associated with success on the problem, with the participants who converted significantly outperforming those who did not. These two converging pieces of evidence suggest that providing problem solvers with a population figure may increase accuracy. The first aim of the present study is therefore to experimentally test the impact on solver accuracy of provision of a real-number base rate figure in addition to a percentage figure.

Experiments one, two and three also revealed the importance of the ‘nested sets process’ in the successful solution of Bayesian problems. However, these experiments were confined to correlational analyses with no attempt to experimentally test this connection. One experiment in Cosmides and Tooby (1996) provided participants with leading questions which, in the language of the present studies, encouraged individuals to complete computational steps one and two in a ’disease’ problem. They found a 20% greater accuracy with leading questions, but this difference was not significant in their paper. This experiment however included a very small sample size and so is likely to have lacked the power to detect an effect of this size or smaller. The second aim of the present experiment is therefore to prompt participants to make step one and two calculations to determine if this increases their accuracy on Bayesian word problems.

Based on previous work it was hypothesised that the group provided with the population value would show a significantly higher accuracy rate than the group with no population value. It was also hypothesised that each leading question (for step one and then for step two) would increase accuracy alone and that both questions combined would increase accuracy further.

Method

Participants. From an original sample of 419 participants, 15 were removed because they stated they had undertaken the same problem in the past. Prior to any analyses, the 10% of participants with the fastest completion times were also removed as their completion times (<1.5 minutes) were considered to be unlikely to be conducive to an engaged completion of the problem. The final sample was therefore 364 and participant demographics can be found in Table 1.

Design and Materials. The study was a between-subjects 2 (population figure provided vs no figure) x 2 (step one questions vs none) x 2 (step two question vs none) design resulting in eight conditions. The study was an online survey which participants accessed through their own computers. A version of the classic medical diagnosis problem (Eddy 1982; Gigerenzer and Hoffrage, 1995) was used in all eight conditions which can be seen below along with all four questions, and the two phrases inserted for the population and non-population conditions:

> Every year the government advises women to take part in routine mammography screening using an X-ray machine to determine if they have breast cancer.
>
> [Pop: Out of 1,000 women at age forty] [Non-pop: Among women at age forty] who participate in this routine screening 10% have breast cancer, while 90% do not.
>
> However the screening test is not always accurate.
>
> Specifically, out of those women who have breast cancer, only 76% will actually get a positive mammography.
>
> Furthermore, out of all those women who do not have breast cancer, 15% will also get a positive mammography.
>
> 1. What percentage of women have cancer and a positive result (P[H&D])?
> 2. What percentage of women have no cancer but still received a positive result (P[-H&D])?
> 3. What percentage of women receive a positive result in total P(D)?
> 4. What percentage of women at age forty who get a positive mammography in routine screening actually have breast cancer (P[H|D])?

Participants in the ’no leading questions’ conditions were only presented with question four; those in the step one conditions were presented with questions one, two and four; those in the step two conditions were presented with questions three and four. Finally participants in the step one and step two combined conditions were presented with all four questions. Each of these conditions was presented in ’population’ and ’non-population’ versions.

Procedure. Participants were recruited through Amazon MTurk. Participants were presented with the consent form, and then the instructions for the study. Participants were then randomly assigned to one of the eight conditions. For each problem they were presented with the problem text and then were presented with each question on a separate page and were required to click next to access each subsequent question. They were also provided with a link to an online calculator wherever a calculation was required. Finally they answered the demographic questions and a final question regarding whether they had undertaken the problem in the study before.

Data Analysis. Answers within 1% of the correct answer were accepted as correct. Further, the answer corresponding to calculation of the wrong conjunction (i.e. giving the rate of ‘no cancer’ instead of ‘cancer’), which had been identified in experiment one and two was also accepted as it demonstrates accurate Bayesian reasoning while simply failing in misreading which conjunction was required. Answers within 1% of this were also accepted. Three answers of this type were given.

Results

The percentage of solvers providing the normative answer showed no significant difference between those participants who were provided with a ‘100’ population and those who were not (Wald = .182, p=.670).

In Figure 14 below the percentage accuracy can be seen across the four ‘question’ conditions for all four questions asked (the population / non population condition distinction has been collapsed due to the null finding). P(H|D) accuracy is shown in purple. In comparison to condition one, condition two, which included P(H&D) and P(-H&D) questions produced 7.6% accuracy (Wald = .530, p=.467), while condition three produced 6.5% accuracy (Wald = .174, p=.676). Neither of these results were significantly different to condition one.

Condition four however produced 13.4% accuracy, which was significantly higher than the control condition (Wald = .4114, p=.043). A further combined analysis of the step one and step two question separately across all four conditions demonstrated a trend towards significance for the step one questions (Wald = 2.774, p=.096) while the step two questions did not show a significant result (Wald = 1.615, p=.204).

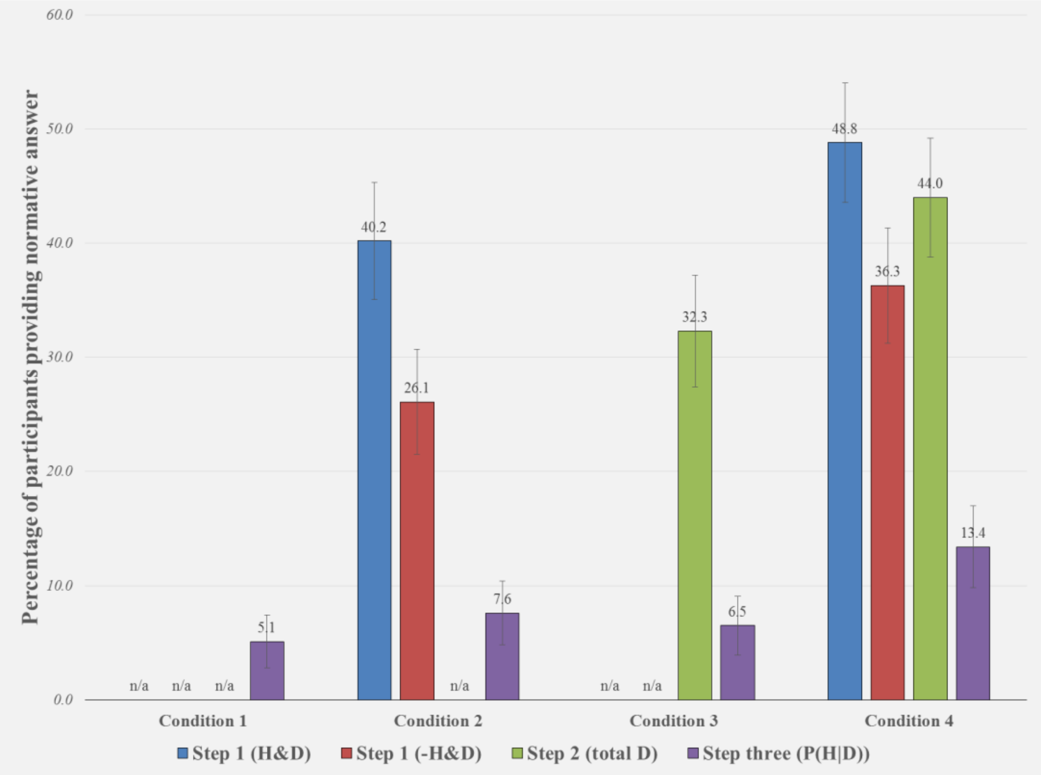


Figure 14. Percentage accuracy for all questions asked for all four conditions in the present study plus equivalent experiment two condition. Error bars indicate one standard error.

To further test the impact of a think aloud protocol, condition one in the present experiment was compared to the equivalent condition in experiment two, which was identical in other respect. Sixteen point seven percent accuracy was seen in that condition, while 5.1% accuracy was seen in condition one in the present study, which is a significant difference (Wald = 5.900, p=.015).

It is also immediately clear from Figure 14 that accuracy on step one and two questions was substantially higher than the accuracy for the final question, P(H|D). In condition four, 31.4% of those who answered the step two question correctly also answered the P(H|D) question correctly. Conversely, no one who failed to answer the step two question answered the P(H|D) question correctly. This was a significant difference (Wald = 16.4, p<.001). This pattern also held in condition three, where 16.7% of those who answered the step two question correctly also answered P(H|D) correctly, while only 1.6% of those who failed the step two question answered P(H|D) correctly. This difference was also found to be significant (Wald = 7.657, p = .006).

Similarly for the step one P(H&D) and P(-H&D) questions, in condition two only 1.9% of those who got these questions wrong subsequently answered the final question correctly and in condition four this figure was 2.8%. However out of those who got these questions correct in condition two, 27.3% got the final answer correct and in condition four this figure was 33.3%. This difference in final solution accuracy between those who answered the step one questions correct was significant for both condition two (Wald = 15.9, p<.001) and condition four (Wald = 11.1, p=.001).

Overall, accuracy on the P(H&D) step one question was considerably higher than accuracy on the P(-H&D) question (Wald = 15.2, p<.001). Accuracy on the P(-H&D) question was predictive of step two accuracy in condition four (Wald = 12.017, p<.001) but when P(H&D) accuracy was added to the model, P(-H&D) ceased to be a predictive factor (Wald = 1.428, p=.232), while P(H&D) remained highly significant (Wald = 15.060, p<.001). Similarly, the P(H&D) question was predictive of step three accuracy in conditions 2 (Wald = 4.524, p=.033) and 4 (Wald = 4.657, p=.031), but again ceased to be a significant predictor when P(H&D) was added to the model (condition two [Wald = .414, p=.520]; condition four [Wald = .683, p=.408]) while P(H&D) remained a significant predictor in both conditions (condition two [Wald = 13.016, p<.001]; condition four [Wald = 4.732, p=.030]).

Further, in condition four, when H&D, P(-H&D) and step two accuracy were all included as independent factors in a model predicting step three accuracy, P(H&D) did not act as a significant predictor (Wald = .870, p=.351), P(-H&D) did not act a significant predictor (Wald = .172, p=.678), while step two did act as a significant predictor (Wald = 4.837, p=.028).

Discussion

Think Aloud. The control condition in the present study produced considerably lower accuracy than the equivalent condition in experiment two. The two studies used the same population and exactly the same problem. The only difference was the inclusion of a think aloud protocol in experiment two. The control condition in the present study also performed at a similar level to Micallef et al (2012) (around 6%), who used a natural frequency version of the medical diagnosis problem in the general population (also using Amazon MTurk) but with no think aloud protocol. This suggests that such a protocol increases accuracy (to an equal or greater extent than the provision of all three leading questions in this study).

Population Prompt. The present experiment hypothesised that the provision of a real number base rate would increase accuracy on the problem. This hypothesis was based upon previous work (Hoffrage et al, 2002) which theorised that conversion to natural frequencies was the cause of a large proportion of the success of the nested sets approach. A large number of participants were in fact found to do exactly this in experiment two, which also found a partial mediation of the relationship between nested sets framing and solution accuracy.

No relationship between the provision of a ‘100’ population sample and accuracy on the problem was found however. This suggests that while some individuals do appear to benefit from converting the problem to real numbers, they are not assisted in doing this by the provision of a real number population alongside a percentage figure. However, as a base of ‘100’ was used in the present experiment, conversion of the problem for participants into real numbers in the non-population condition may have been a simple process, allowing those participants who wished to convert the problem to do so easily regardless of which condition they were in. Future work testing the provision of a real number population figure with numbers which do not allow such easy conversion from percentages to real numbers may therefore find differing results.

Leading Questions. A non-significant increase in accuracy was seen for the provision of both step one and step two questions separately. However a significant and substantial increase was seen for the provision of both together. Overall this suggests that encouraging individuals to undertake the steps identified in the nested sets process model in experiment one can increase accuracy but only in combination. This is in line with previous work (Cosmides and Tooby, 1996) who actually found a larger (20%) increase with the provision of the same leading questions, but may have lacked statistical power due to small sample size. This result suggests that the provision of such ‘primer’ questions can be an efficacious way to increase accuracy on Bayesian problems.

However, considerably lower accuracy was seen in this experiment on the final question (P[H|D]) compared to the leading questions. Further, successful completion of the leading questions was, while highly related to success, considerably less so than the comparable computational steps in experiment two. This appears in contradiction to the findings of experiment two which found that the successful completion of step one (questions one and two) and step two (question three) near-guaranteed success on the P(H|D) question.

A notable major difference between these two experiments is that the result from experiment two is correlational while the present was experimental. Further, in the present experiment participants undertook the leading questions without any conception of 'where they were leading'. In contrast, in experiment two, participants may have been more likely to undertake the earlier stages of the nested sets process model if they already have an idea in mind of their final solution process (i.e. as a necessary step towards the end goal). In combination with the fact that the leading questions appear easier than the final question, this means that many participants who are in fact not capable of solving the Bayesian problem without leading questions (or indeed, answering the final question in the present experiment) would have still been able to answer the leading questions, reducing the strength of the relationship between successful solution of early stages / leading questions and the correct final answer to the problem.

Another related possible factor is that participants were not provided an overview of the steps they needed to undertake to solve the problem in the present experiment. They were presented with four questions which they may have considered to be entirely disconnected from each other. They may not have recognised that they were being intentionally drawn through a single solution process and so when they arrived at the final question, many participants may not have even attempted to use the knowledge gained from the previous questions in their answer.

If, when faced with a Bayesian problem, participants are broadly conceptualising a solution procedure before attempting any mathematics (and before writing anything in their think aloud answer), this suggests that leading questions (while beneficial, according to this study) may not be the best approach to increasing accuracy. Instead of forcing participants to blindly compute each step in the hope that they will perceive the correct solution process it may be more beneficial to provide prompts which help individuals perceive the correct solution procedure in rough form. The findings of the present experiment in combination with experiment two suggest that it is this rough conceptualisation of the nested sets process which is the greatest guarantor of success on Bayesian problems. This may be done through hypothesis-focused representation diagrams (which some previous work e.g. Sloman et al [2003] has found success with) or, as we would advocate, in combination with data-focused representation diagrams (which no previous work has attempted).

General Discussion

The Nested Sets Approach

Across experiments one and two altering Bayesian problems to use Macchi's (2000) outside-framed percentages instead of inside-framed percentages has been demonstrated to increase participant accuracy rates for the normative answer. This effect was consistently produced in both simple and hard (including false negatives) problems, with whole numbers and with decimals and with base rate information in both real number and percentage forms. In experiment one it was also found likely to be present within both low and high numeracy groups and across two experiments has been shown to increase accuracy using both within and between subjects designs.

Step One versus Step Two

In a deeper analysis of Macchi’s outside-framed approach, results from experiments one and two have both demonstrated that Macchi’s approach has a large impact on the frequency with which individuals produce the hypothesis-focused representation. As the drop-off curves (e.g. Figure 7 and Figure 8) for both experiments demonstrate, the clear difference between the nested and non-nested conditions is that substantially more individuals in the nested sets condition produce the hypothesis-focused representation and computational step one. However, as the drop-off graphs suggest, and statistical analysis in both experiments further demonstrates, when controlling for the presence of the hypothesis-focused representation (or computational step one), the outside-frame approach does not increase the frequency of data-focused representations (or computational step two). This can be seen visually in the flattening of the drop-off curves after the hypothesis-focused representation / computational step one.

As stated previously, the outside-frame approach makes changes to both the body text and the question format. The information regarding the hypothesis-focused representation / computational step one is contained within the body of the text of the problem, while the data-focused representation / computational steps two and three are contained within the question. Based on the above analysis, Macchi’s approach is successful in improving the frequency of the hypothesis-focused representation and computational step one, but has little or no impact on the frequency of the data-focused representation or computational step two other than indirectly via that increase in the hypothesis-focused representation and computational step one. This possibility is further supported by the pattern of errors in experiment two. Firstly the nested sets condition showed a far smaller rate of 1-P(D|-H) errors, which were predominantly identified as caused by a confusion of the false positive rate. However, the most prominent error in the nested sets conditions was the P(H&D) error, which was identified as being caused by confusion of the question being asked in the problem. This further suggests that while the nested sets approach does effectively enhance clarity around the false positive rate (used in populating the hypothesis-focused representation), it is less effective in enhancing clarity of what the solver is ultimately being asked to compute (the latter stages of the nested sets process).

These results fits with Evans et al. (2000) who found no difference between two question formats very similar to those compared in the present study ('individual percentages' (inside) vs 'proportionate percentages' (outside)). It also fits with results by Fiedler et al (2000) who found a nonsignificant difference between a Bayesian problem with an outside-framed text body but inside-framed question and a natural frequency version of the same problem. It does not as easily fit with work by Girotto & Gonzalez (2001) who found an effect on accuracy by altering the question form only. However, the changes made by Girotto and Gonzalez were somewhat different: they included two 'steps' in their question, asking solvers to firstly calculate D separately (i.e. the total number of positive results) and only then to calculation P(H|D). While experiment four also required participants to compute D before computing P(H|D) and found a limited effect, this was presented as a separate question on a separate page 'before' requesting P(H|D). As speculated previously, this may have reduced participants' recognition of the connection between the two answers. Girotto and Gonzalez's approach instead requested both D and P(H|D) simultaneously as a two-step question, making the connection between the two entirely apparent.

In combination these results suggest that a simple flip of the question form from inside to outside perspective may not be a sufficient intervention to improve accuracy, and tentatively suggest that the same results would have been seen in the present experiment if the question form had not been changed at all. However, a more involved change directly requesting D prior to calculation of the final product (and, making the connection between these two clear), may have impact, as demonstrated by Girotto and Gonzalez (2001). This therefore suggests that future work may benefit from further enhancing the question form, with an aim of encouraging individuals to construct the data-focused representation (and to calculate D) at the correct moment in the solution process.

Process Model

Based upon the think aloud data in experiment one, a five-stage process was proposed and was purported to be used by the vast majority of participants who provided the normative answer. This process involved two representations of the problem and three computational steps. This model built upon and formalised a large amount of theoretical work in the previous literature and also introduced novel contributions, including a clear distinction between the hypothesis and data-focused representations of the problem, previously missing from the literature. This process was present in the data of the vast majority of successful solvers in all four conditions, including the control and causal conditions, suggesting that it is the preferred solution process of successful solvers, regardless of particular framing / prompts. This finding runs counter to a commonly-held view in the field that the ’default’ problem-solving perspective in the absence of any specific prompt is the ’inside’ or ’individual’ perspective (e.g. Tversky & Kahneman, 1983; Sloman et al, 2003). In experiments one, two and three a large proportion of individuals spontaneously adopted the outside perspective and followed the nested sets process in the absence of any specific prompt to do so (e.g. in the ’inside’ and ’causal’ conditions). These findings contribute to the aim of Johnson and Tubau (2015) to gain a greater understanding of why leading facilitative approaches to Bayesian problems achieve their success. The answer, based on this work, is that they do so by encouraging more individuals to follow the nested sets process.

Experiment two furthered this finding by demonstrating that successful individuals also follow the nested sets process when presented with decimal numbers, whole numbers and percentages. Experiment four, cementing the importance of the nested sets process found that providing individuals with step one and step two leading questions increases their accuracy on Bayesian problems by 8.3% (previous work by Cosmides and Tooby (1996) found a 20% increase in accuracy, but likely lacked statistical power).

Previous work (e.g. Girotto and Gonzalez, 2001) has also shown that participants are facilitated by an 'outside view' structure with the abstract units of 'chance' (although see Brase [2013] for a reinterpretation of that work). Further, work by Sirota, Juanchich and Hagmayer (2014) has demonstrated that increased accuracy with an outside-frame approach can be seen when divisible units such as 'mgs of wheat' are used, as opposed to 'whole' units such as a 'bag of wheat', which was predicted would not be the case by Brase (2007). While overall this experiment has found a general preference for real numbers as opposed to percentages, no preference was seen for whole numbers over decimal values. These findings in combination contradict some previous theorising by Gigerenzer and Hoffrage (1995), Brase (2007) and Cosmides and Tooby (1996) who theorised that "if there are [mental] mechanisms which represent frequencies in terms of discrete, countable entities, it should be difficult to think about a tenth of a person, and therefore the level of Bayesian performance should decrease." (Cosmides & Tooby, 1996, pp. 55).

The present findings combined with previous (Girotto & Gonzalez, 2003; Sirota, Juanchich & Hagmayer, 2014) suggest that problem solvers are able to solve Bayesian problems using a vast range of different units (although there may be some preference for particular values, such as real numbers) so long as they are guided to follow the nested sets process model outlined in this paper. These findings emphasise that it is the process which is most integral to solution, and not the particular unit of analysis.

Future work may be beneficial in determining the role that individual differences play in unit preference. The present methodology was not able to determine why some participants were more able to solve the problem by converting to real numbers and some participants able to solve with percentages. Numeracy level and familiarity with percentages are proposed as plausible factors for future work to consider.

Drop Off and Problem Difficulty

Experiments one and two both demonstrated that out of those participants who failed to achieve success on the problem, the majority failed at the representation one / computational step one phase, while a smaller proportion failed at later stages. This finding was echoed by data on numeracy levels from experiment one which showed that the numeracy levels of individuals achieving the later stages of the process was higher than those who only achieved representation one / computational step one, which was in turn higher than those who achieved no steps. These correlational findings were broadly confirmed in experiment four which showed that participants found the earlier questions (e.g.

step one) easier than the later questions (e.g. steps two and three). Overall these findings suggest that the later steps in the nested sets process may be more difficult, and require greater numerical ability to undertake than earlier steps. In combination with the results across several experiments on the typical errors individuals make, these findings go some way towards providing the understanding of the stages at which problem solvers err (Johnson and Tubau, 2015). It also again suggests that future work should focus more on altering the question form (which contains the information for the later steps) than the body of the text in Bayesian problems.

Nested Sets and Natural Frequencies

While a direct contrast was not possible, a comparison between experiment one and Macchi (2000), which both used real number base rates, and experiment two, which used percentages, suggests there may be a further increase in accuracy when real numbers base rates are used. Further evidence for this comes from the fact that 60% of successful individuals in experiment two, and 50% in experiment three felt the need to convert the problem from percentage form to real number form. This also gives some support to Hoffrage et al.'s (2002) conjecture that the nested sets outside-percentage format may work via encouraging individuals to construct a natural frequency version of the problem for themselves. However, while the tendency for people to prefer to work with real numbers appears clear, it is important to recognise that a large proportion of individuals in both experiments correctly solved the problem using the same process, but entirely using percentages. This may not sit well with some evolutionary explanations for the benefits of natural frequency formats that assume such formats should be universally beneficial (Cosmides and Tooby, 1996). However, the finding is not contradictory to theoretical work by Brase (2008) who labelled natural frequencies as a ’privileged’ format, and stated that, while the brain systems designed to process natural frequencies will always prefer that format, they can be persuaded to work on other formats, albeit with lesser efficiency. The present finding, that some individuals were able to work through the problem using percentages, while others felt the need to convert to natural frequencies to solve the problem, fits with this view. However, in terms of the evolutionary argument, it must be noted that the present experiment is unable to determine whether participants’ preference for real numbers is due to a greater familiarity with that format or due to the frequentist proposal that our evolutionary history has designed our brains to deal with them more effectively.

The finding of problem conversion fits well with previous work. Brase (2013) showed participants ambiguous ‘chance’ wording Bayesian problems (adapted from Girotto and Gonzalez, 2001) and found that individuals who interpreted the problems as frequencies (i.e. real numbers) were more successful than those who interpreted the problem as probabilities. The present work suggests that these frequency-interpreting individuals in Brase’s work may have constructed a ‘real number’ sample for themselves, and that this may have led to their greater increase in accuracy in that study. In temperance of this view, an experimental attempt to increase accuracy by providing participants with a real number population rate failed in experiment four. However, as noted previously, conversion of the problem (which used a base of 100) may have been overly simple in this experiment, allowing all participants who wished to convert to do so regardless of which condition they were in. Future work would be beneficial in determining if assisting individuals to convert percentage problems to real numbers in less standardised scenarios would increase accuracy.

Causal Framing

No effect was found across two experiments (one and three) for Krynski and Tenenbaum’s (2007) causal framing. The null finding was found for the medical diagnosis problem and three other novel problems with real number base rates in experiment one for high and low numerates within the general population in a within subjects design. It was also found with a between subjects design with and without a think aloud protocol and with percentage base rates in experiment three.

Several things should be noted about this null finding. Firstly, all of the replications of Krynski and Tenenbaum’s (2007) findings, including the present, have focused on the medical diagnosis problem. While it appears, given the mixed results of previous work, and the present, that this effect does not reliably replicate, it is quite possible that a causal framing would be assistive in other problems, including the other problems included in that original paper. Secondly it should also be strongly noted that we do not consider that the null finding and subsequent theorising in this paper discredits Krynski and Tenenbaum’s (2007) causal account of Bayesian reasoning in general. The present experiments examined relatively simple Bayesian problems which may be very different from the Bayesian situations that people deal with in everyday life (and from which Krynski and Tenenbaum’s theory was derived). If it is, in fact, the case that the proposed process model is the best description of human reasoning in the simple Bayesian word problems in the present study it does not necessarily follow that this will hold in more complex situations. In fact, given that the process that humans take in these simple problems appears to be non-intuitive (Lesage, Navarrete, & De Neys, 2013; Sirota, Juanchich, & Hagmayer, 2014) it seems likely that as variable numbers increase and quality of information decreases (such as in the real life situations discussed by Krynski and Tenenbaum), such a reasoned approach will become impossible and a different approach entirely (e.g. intuitive estimates) will become the dominant approach. It is perfectly plausible that in these more realistic situations a causal model will predict and describe human decision making more accurately. More work in these complex situations would therefore be valuable to provide a more thorough understanding of human Bayesian reasoning.

The Confusion hypothesis and Base Rate Neglect

In experiments one and two, think aloud data revealed the P(D|-H) and 1-P(D|-H) errors as the most frequently observed among incorrect answers. These answers have been the most frequent answers labelled ‘base rate neglect’ by previous work, as they only utilise the ‘new data’ and do not incorporate the ‘prior’ or base-rate in their calculation (Tversky & Kahneman, 1982; Casscells et al, 1978; Bar-Hillel, 1980; Cosmides & Tooby, 1996; Evans et al, 2000; Barbey & Sloman, 2007). No direct evidence for the base rate neglect error was found in the think aloud data of any participant, with underlying logic undiscernible in many cases. However, many participants (45.8%) committing the 1-P(D|-H) error demonstrated a clear confusion of P(D|-H) with P(-H|D). This error in reasoning is also known as the ‘transposed conditional’ fallacy (Foreman et al, 2005), or the ‘prosecutor’s fallacy’ in law (e.g. Fenton & Neil, 2011; Nance & Morris, 2005).

While many participant’s think aloud data for the P(D|-H) error also provided little insight into underlying logic, no direct evidence for base rate neglect was found here either. Five participants demonstrating this error also clearly showed confusion between P(D|-H) and P(-H|D), providing some further evidence for the confusion hypothesis and against base rate neglect. It should be noted that base rate neglect also does not provide a convincing explanation for the pattern of errors observed in this experiment or previous experiments which have closely examined errors (e.g. Gigerenzer and Hoffrage, 1995; Macchi, 2000) as a whole. As can be seen in experiment one, two of the most common errors also include the provision of H and – H (the two base rates) as percentages of the population, a phenomenon known as base rate conservatism (e.g. Gigerenzer and Hoffrage, 1995), or a general undervaluing of the false positive / true positive rates on Bayesian word problems. The simultaneous large rates of errors using both the false positive rate only, and the base rates only, therefore appear to make base rate neglect a poor general explanation of human error on Bayesian word problems. Examination of the think aloud data for the base-rate-only errors revealed little about their underlying logic, but with two participants demonstrating some under-valuing of the false positive rate.

The pattern of findings fit more closely with previous authors who have theorized that semantic misunderstanding of problem texts lies behind many errors on Bayesian word problems (e.g. Hamm, 1988; Hamm and Miller, 1990; Gigerenzer and Hoffrage, 1995; Macchi, 2000; Wolfe, 1995; Macchi and Mosconi, 1998; Macchi, 2000; Fiedler et al., 2000; Welsh and Navarro, 2012). The finding also fits with previous work which has theorised a similar ‘confusion’ hypothesis (Braine and Connell, 1990; Cohen, 1981; Dawes, 1986; Eddy, 1982; Hamm and Miller, 1990; Fiedler et al., 2000) for the other major answer labelled as base rate neglect: providing the variable P(D|H) (the true positive rate). The confusion hypothesis has theorised that this is due to a complete misunderstanding of the difference between P(D|H) and the correct answer, P(H|D). In the present paper a similar confusion was seen, but in this case the confusion was between P(D|-H) and P(-H|D). Evans et al. (2000) did not use a think aloud protocol to analyse errors but also theorised that this error was 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.” (Evans et al. 2000, pp. 199). The present findings confirm this conjecture.

Confusion of an element of the text was also the most prominent reasoning error uncovered in the outside-framed conditions in experiment two. The most common error in those conditions was to provide the conjunction P(H&D), the number of women with cancer and a positive result. It proved more difficult to discern a general reasoning error behind this mistake. However, the most frequent reason identified was a mis-reading of the question, wherein participants seemed to believe the aim of the problem was to provide the ‘percentage of all women with a positive result and breast cancer’, rather than ‘the percentage of women with a positive result who actually have breast cancer’. This suggests that future work looking to reduce this error should focus on making the question clearer, perhaps using Girotto and Gonzalez’s (2001) two-step method.

Think Aloud Protocol

Several strands of evidence from experiments one to four suggest that the use of a think aloud protocol increases accuracy on Bayesian word problems. This was demonstrated firstly in a comparison of the control condition in experiment four to the equivalent condition in experiment two and which used the same participant pool (mTurk workers with the same requirements). In the non-think-aloud experiment four condition, 5.1% accuracy was seen, while in the think-aloud condition in experiment two 16.7% accuracy was seen. This compares closely to the more direct comparison made in experiment three: here large differences between non-think-aloud and think aloud conditions were seen both for the non-causal (5.8% vs 18.4%) and causal (3.5% vs 9.4%) conditions. Combining these three comparisons, the addition of a think aloud protocol appears to provide around a 9.8% increase in individuals providing the correct answer, from 4.8% (16 / 333) to 14.6% (37 / 254).

Despite many theoretical suggestions to this effect previously (e.g. Wilson, 1994) and similar findings in related situations (Kim, 2002), this is the first time to the authors' knowledge that this has been demonstrated empirically on a Bayesian word problem. While it may not be possible to directly test whether the think aloud protocol changes the nature of the cognitive processes individuals employ (as, by definition, these cannot be recorded in a non-think aloud condition), as Ericsson and Simon (1998) have maintained, it does appear to increase accuracy.

The think aloud protocol employed in this experiment required individuals to work through the problem before they were allowed to submit their numerical answer. The increased accuracy seen with this methodology may be due to a lack of sufficient engagement on the part of participants in non-write-aloud experiments, or may be due to the process encouraging them to engage in more rule-based processing of the problem (as theorised by Barbey & Sloman, 2007) by typing / writing out the logical steps. This fits with some findings from experiment two which found that the two most common errors made by individuals who do not provide the normative answer were due to a misunderstanding of either the false positive rate, or the meaning of the question. The forced contemplation provided by the present think aloud protocol may provide the opportunity to avoid such confusion. This theory also fits with previous work by Sirota, Juanchich & Hagmayer (2014) who found that accuracy on the Cognitive Reflection Test (CRT), designed to test the capacity to ignore incorrect intuitive responses and inspect more deeply into a problem showed greater predictive power for accuracy on Bayesian problems than any other measure including cognitive ability. Overall, these findings suggests that increasing engagement / processing time may be a promising target for future interventions to increase Bayesian accuracy.

Conclusion and Future Work

The present paper has demonstrated the efficacy of Macchi's (2000) outside-framed approach to improving accuracy on Bayesian word problems across two experiments with within- and between-subjects designs, with and without the possibility for false negatives, with percentage and real number base rates and with whole number and decimal values. This framing was also found to be efficacious within high and low numeracy groups. Macchi's approach can therefore be recommended for improving the presentation of Bayesian problems to the general public in a wide range of situations and formats, including in medical contexts.

The present paper has also demonstrated that Macchi's (2000) outside-framed approach could be improved further. Analysis over several experiments suggest that the improvement in accuracy seen as a result of using Macchi's framing is due to the changes to the body of the text, and that the changes to the question form may be superfluous. Further however, drawing on other previous work such as Girotto and Gonzalez (2001), it appears that more-extensive alterations to the question form can improve accuracy. Future work is therefore recommended either in combining Macchi's text body with Girotto and Gonzalez's two-step question form, or attempting to improve Macchi's question form to further increase accuracy. Such work should also focus on improving the clarity of that question form to reduce the chance of individuals misinterpreting it to mean the percentage of women with cancer and a positive result (the most common error in the nested sets condition in experiment two).

The present paper has also demonstrated that, regardless of specific problem framing (either inside-frame, outside-frame or causal framing), successful individuals overwhelmingly follow a single solution process. This process was defined in experiment one, comprised five stages and was subsequently found in experiments two and three in both control and causal conditions. The interventionist experiment four also found that encouraging individuals to follow these steps prior to giving their final answer increased accuracy.

Given the ubiquity of this process across framing types, it is suggested that this process may be the preferred approach of the majority of individuals, and therefore future attempts to improve accuracy on Bayesian problems should use it as a framework to guide their design of interventions. Future interventions should be designed to make it as easy as possible for solvers to follow this five-stage process, rather than attempt to encourage them to solve the problem through some entirely different process.

In the longstanding debate over the relative distinctiveness of the nested sets and natural frequency approaches to increasing accuracy on Bayesian problems, the present paper also contributes valuable findings and theoretical developments. In both the nested sets and natural frequency literature, the same underlying process has been alluded to (with varying degrees of explicitness), which has here been formally outlined for the first time as the ’nested sets process’. It has been named this here, rather than the ’natural frequency process’ because evidence from Experiments 2 and 3, while indeed suggesting a preference for real numbers (60% in experiment two, 50% in experiment three), has shown that many individuals are fully capable of undertaking the same basic process with non-real numbers (percentages). Combined with previous findings (e.g. Girotto and Gonzalez, 2011; Sirota, Juanchich & Hagmayer, 2014) there is converging evidence that individuals can solve Bayesian problems using a range of ’units’ other than whole real numbers, or even real numbers at all. This suggests that while there may be a preference for that particular unit of analysis, this is less important than the process itself, which has been the central message of the nested sets approach for several decades (Tversky and Kahneman, 1983; Macchi, 1995; Macchi and Mosconi, 1998; Lewis and Keren, 1999; Mellers and McGraw, 1999; Macchi, 2000; Evans et al., 2000; Girotto and Gonzalez, 2001; Sloman et al., 2003).

Further, in the debate over the cause of error on Bayesian problems, the present paper has contributed several valuable findings. It was determined through analysis of the think aloud data that two of the most common causes of error were due to misunderstanding of, firstly, the meaning of the false positive rate and secondly, the statistic that the question was requesting. This finding, as well as the overall pattern of errors, provides evidence towards the ’Confusion hypothesis’ view of error on Bayesian problems (Cohen, 1981; Eddy, 1982; Dawes, 1986; Hamm, 1988; Hamm and Miller, 1990; Braine and Connell, 1990; Gigerenzer, 1996; Macchi, 1995; Wolfe, 1995; Macchi and Mosconi, 1998; Macchi, 2000; Fiedler et al., 2000; Welsh and Navarro, 2012) and against the base rate neglect view of error (Kahneman and Tversky, 1972; Ajzen, 1977; Casscells et al., 1978; Bar-Hillel, 1980).

Finally, the present paper has found that the mere addition of a think aloud protocol can increase accuracy considerably. This process forces individuals to engage with the problem before they can provide a numerical answer. It also encourages individuals to type (or write) their thought process out, potentially encouraging rule-based thinking over ‘associative’ processing (Barbey & Sloman, 2007). This finding may therefore suggest that either engagement or an over-reliance on intuitive reasoning are factors in inaccuracy on Bayesian problems, and that think-aloud type procedures may be valuable not only methodologically for examining underlying thought processes, but also for encouraging sound reasoning in real contexts. Further, while a think aloud protocol has here been advocated, it has also been noted that the approach can under-detect certain mental processes e.g. when participants only provide mathematical notation and do not explain their thoughts in words. Future work therefore may be valuable in devising methods for extracting the mental processes of solvers with greater fidelity.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35(5):303–314.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3052):211–233.

Barrett, B. and McKenna, P. (2011). Communicating benefits and risks of screening for prostate, colon, and breast cancer.

Braine, M. and Connell, J. (1990). Is the base rate fallacy an instance of asserting the consequent. *Lines of thinking*.

Brase, G. (2002). Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni's (1999) mentalmodel theory of extensional reasoning. *Psychological Review*, 109(4):722– 728.

Brase, G. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, 15(2):284–289.

Brase, G. (2013). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, 26(1):81–97.

Brase, G., Fiddick, L., and Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Quarterly journal of experimental psychology*, 59(5):965–76.

Brase, G. and Hill, W. T. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 6(March):1–9.

Brase, G. L. (2007). The (in) flexibility of evolved frequency representations for statistical reasoning: Cognitive styles and brief prompts do not influence bayesian inference. *Acta Psychologica Sinica*, 39(3):398–405.

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England journal of medicine*, 299(18):999–1001.

Chapman, G. and Liu, J. (2009). Numeracy , frequency , and Bayesian reasoning. *Judgment and Decision Making*, 4:34–40.

Cohen, L. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., and Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7(1):25–47.

Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58:1–73.

Dawes, R. M. (1986). Representative thinking in clinical judgment. *Clinical Psychology Review*, 6(5):425–441.

Donnelly, P. (2005). Appealing statistics. *Significance*, 2(1):46–48.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. *Judgement under uncertainty: Heuristics and biases*, pages 249–267.

Ericsson, K. A. and Simon, H. a. (1980). Verbal reports as data.

Ericsson, K. A. and Simon, H. a. (1998). How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind, Culture, and Activity*, 5(3):178–186.

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., and Thompson, V. a. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77:197–213.

Fenton, N., Neil, M., and Hsu, A. (2014). Calculating and understanding the value of any type of match evidence when there are potential testing errors. *Artificial Intelligence and Law*, 22(September):1–28.

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *Journal of experimental psychology. General*, 129(3):399–418.

Forrest, a. R. (2003). Sally Clark–a lesson for us all. *Science & justice : journal of the Forensic Science Society*, 43:63–64.

Garcia-Retamero, R. and Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social science & medicine (1982)*, 83:27–33.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3):592–596.

Gigerenzer, G. and Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ : British Medical Journal*, 327:741–744.

Gigerenzer, G. and Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction : Frequency Formats. *Psychological Review*, 102(4):684–704.

Girotto, V. and Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78(3):247–276.

Hamm, R. M. (1988). Explanations of the use of reliability information as the response in probabilistic inference word problems. Technical report, DTIC Document.

Hamm, R. M. and Miller, M. A. (1990). Interpretation of conditional probabilities in probabilistic interference word problems. Technical report, DTIC Document.

Hayes, B. K., Newell, B. R., and Hawkins, G. E. (2013). Causal model and sampling approaches to reducing base rate neglect. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*.

Hill, W. T. and Brase, G. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Quarterly journal of experimental psychology (2006)*, 65(12):2343–68.

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, 84:343–352.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290(December).

Johnson, E. D. and Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, 28:34–40.

Johnson, E. D. and Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, 6(July):1–19.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., and Caverni, J. P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological review*, 106(1):62–88.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454.

Kim, H. S. (2002). We talk, therefore we think? a cultural analysis of the effect of talking on thinking. *Journal of personality and social psychology*, 83(4):828.

Krynski, T. R. and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of experimental psychology. General*, 136(3):430–50.

Lewis, C. and Keren, G. (1999). On the difficulties underlying Bayesian reasoning: A comment on Gigerenzer and Hoffrage. *Psychological Review,* 106(2):411–416.

Macchi, L. (1995). Pragmatic Aspects of the Base-rate Fallacy. *The Quarterly Journal of Experimental Psychology Section A*, 48(February 2015):188–207.

Macchi, L. (2000). Partitive Formulation of Information in Probabilistic Problems: Beyond Heuristics and Frequency Format Explanations. *Organizational behavior and human decision processes,* 82(2):217–236.

Macchi, L. and Mosconi, G. (1998). computational features vs frequentist phrasing in the base-rate fallacy. *Swiss Journal of Psychology*, 57(2):79–85.

McNair, S. J. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology,* 6(February):1–3.

McNair, S. J. and Feeney, A. (2014a). When does information about causal structure improve statistical reasoning? *Quarterly journal of experimental psychology (2006)*, 67(4):625–45.

McNair, S. J. and Feeney, A. (2014b). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review,* pages 1–7.

Meder, B. and Gigerenzer, G. (2014). Statistical thinking: No one left behind. In *Probabilistic Thinking*, pages 127–148. Springer.

Meder, B., Mayrhofer, R., and Waldmann, M. R. (2009). A Rational Model of Elemental Diagnostic Inference. *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, pages 2176–2181.

Meehl, P. E. and Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological bulletin*, 52(3):194–216.

Mehlum, H. (2009). The Island Problem Revisited. *The American Statistician*, 63(3):269–273.

Mellers, B. a. and McGraw, a. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, 106:417–424.

Micallef, L., Dragicevic, P., and Fekete, J. D. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18:2536–2545.

Navarrete, G., Correia, R., and Froimovitch, D. (2014). Communicating risk in prenatal screening: the consequences of Bayesian misapprehension. *Frontiers in psychology*, 5.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. *Econometric Theory*, 19(04):675–685.

Salthouse, T. a. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review*, 103(3):403–428.

Sirota, M., Kostoviˇcov´a, L., and Vall´ee-Tourangeau, F. (2015). Now you Bayes, now you don’t: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic bulletin & review*, pages 1465–1473.

Sloman, S. A. and Lagnado, D. A. (2005). Do We ”do”? *Cognitive science*, 29(1):5–39.

Sloman, S. a., Over, D., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91(2):296–309.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315.

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., and Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Annals of Internal Medicine*, 156(5):340–349.

Welsh, M. B. and Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, 119(1):1–14.

Wilson, T. D. (1994). Commentary to feature review: the proper protocol: validity and completeness of verbal reports.

Wolfe, C. R. (1995). Information Seeking on Bayesian Conditional Probability Problems : A Fuzzy-trace Theory Account. *Journal of Behavioral Decision Making*, 8(June):85–109.