Determining Effective Methods of Presenting Bayesian Problems to a General Audience

Stephen Harrison Dewitt

Submitted in partial fulfillment of the requirements of the Degree of Doctor of Philosophy.

Risk and Information Management Research Group

Electronic Engineering and Computer Science

Queen Mary University of London

United Kingdom

11th July 2016

To Alice, without whom it would have been impossible.

Declaration

I, Stephen Dewitt confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Stephen Dewitt.

Abstract

The thesis presents six experiments designed to further understanding of effective methods of presenting Bayesian problems to a general audience. The first four experiments (Part I) focus on general Bayesian reasoning. The final two experiments (Part II) focus specifically on the legal domain.

Experiment one compares two leading theories for Bayesian presentation: Macchi's (2000) 'nested sets' approach, and Krynski and Tenenbaum's (2007) 'causal' approach. It also uses a think aloud protocol, requiring thought-process recording during solution. A nested sets framing effect is found, but no causal framing effect. From the think aloud data, a five-stage solution process (the 'nested sets' process), modal among successful individuals, is found. In experiment two, Macchi's approach is tested on a problem with greater ecological validity. An increase in accuracy is still seen. Experiment two also finds that conversion of the problem to integers by participants is highly associated with accuracy. Experiment three confirms the null causal finding of experiment one and finds that the think aloud protocol itself increases accuracy. Experiment four experimentally tests whether prompting problem conversion to integers, 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.

Experiment five tested whether statistically untrained individuals can undertake accurate Bayesian reasoning of a legal case including necessary forensic error rates (Fenton et al., 2014). No single individual is found to provide the normative answer. Instead a range of heuristics are found. Building upon this, experiment six compares two approaches to presenting the Bayesian output of a legal case: the popular event

tree diagram, and the Bayesian network diagram recommended by (Fenton et al., 2014). Without inclusion of false positives and negatives the event-tree diagram was rated more trust worthy and easy to understand than the Bayesian network diagram. However, including these error types, this pattern reversed.

Acknowledgements

First and foremost I would like to thank my primary supervisor Anne Hsu for her incredible guidance and advice in all areas as well as her unending patience throughout and her tireless and generous feedback. I would like to thank my secondary supervisor Norman Fenton for his invaluable experience and guidance on the project and insightful critiques of my work. I would also like to thank David Lagnado for his extremely generous collaborative advice, guidance and feedback on so many elements of this work. Finally I would like to thank Martin Neil and Magda Osman for their feedback at a number of crucial stages.

I would further like to thank Julian Tillmann and Eva Kagan for all the advice they have given me on my thesis and all things academic.

Finally, I would like to thank Alice Alberici. While completing a doctorate of your own, you have tirelessly read, advised, guided and helped me in all areas of my thesis into the ungodly hours on far too many occasions. I would not be here without you.

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.

Contents

List of Tables

List of Figures

1 General Introduction

1.1 Overview

The work presented herein represents an attempt to build upon and improve current methods for presenting Bayesian statistics to the general public to increase the proportion of individuals accurately solving such problems and reduce fallacious reasoning. The thesis begins with a general introduction to the area, tracing the history of Bayesian statistics as well as attempts to reconcile Bayesian principles with human reasoning. Following this, the thesis has a two part structure.

Part one presents four experiments aimed at improving the presentation of Bayesian statistics in a general sense, and the recommendations which arise from the findings of these four experiments apply to the vast majority of situations in which the general public encounter Bayesian statistics. In general, part one finds little evidence for an approach based upon enhancing the causal framing of such problems, and instead across several experiments find evidence for an approach involving a shift in perspective / presentation from a single individual (the 'inside' view) to a group focus (the 'outside' view). It also proposes a 5-stage process model by which individuals are thought to solve such Bayesian problems, regardless of specific framing. This model is found through exploratory analysis in experiments on to three, and tested experimentally in experiment four, showing a successful increase in accuracy. In experiment three, a clear increase in accuracy is also found for the mere use of a protocol which requires participants to type their reasoning process while solving the problem.

Across two experiments, part two attempts to provide a solution to an anomalous

situation (the legal trial), in which the approaches advocated in part one are not applicable. This is due, firstly, to the impossibility of framing such a situation from the 'outside' view, and secondly, due to the high complexity of the typical legal case. In this situation, it is confirmed in the first experiment that untrained individuals cannot successfully solve the Bayesian statistics involved in a legal trial. Given this inability, it is argued that computational methods such as Bayesian Networks should be employed to undertake the calculations, and legal professionals / jurors should focus instead on defining the input to these programs, and on interpreting the output. Two methods for presenting the output (Event Trees and Bayesian Networks) are compared in the final experiment, with the conclusion that when the necessary complexity of the legal trial is taken into account, the Bayesian network is more trusted and understood than the event tree.

In a final concluding section, the findings of the two parts are drawn together and ramifications for theory and recommendations for practise are provided. An overall analysis of common errors on Bayesian problems from both parts are analysed, finding evidence for a theory which proposes that semantic confusion of the text of Bayesian problems is a major source of error in such problems and against a long-held theory that certain information is simply ignored / under-weighted.

1.2 Bayesian Inference

Bayesian Inference is a statistical approach to updating the probabilistic belief of a hypothesis being true in the light of new data (Sedlmeier, 1997). Through a turbulent history with fluctuating popularity, it has now emerged as a cornerstone of modern probability theory (Box and Tiao, 2011) and has been described by Jeffreys (1973) as "to the theory of probability what the Pythagorean Theorem is to geometry". The approach has now found application in genetics (Beaumont and Rannala, 2004), linguistics (Frank and Goodman, 2012), image processing (Geman and Geman, 1993), cosmology (Hobson, 2010), machine learning (Bishop, 2006), epidemiology (Lawson, 2013), forensic science (Taroni and Aitken, 2006), ecology

(Parent and Rivot, 2012) and many more fields of enquiry.

With the increasing presentation of risk communication in numerical form, the ability to make simple Bayesian inferences from given statistics is rapidly becoming a necessary skill even for individuals with no statistical training (Meder and Gigerenzer, 2014; Gigerenzer, 2015). Two key areas where the general public frequently encounter statistical problems requiring Bayesian inference include medicine and law (Forrest, 2003; Gigerenzer and Edwards, 2003; Meder et al., 2009; Barrett and McKenna, 2011; Fenton et al., 2014). In medicine, both doctors and patients need Bayesian inference to make accurate assessments of the risk of the patient having a specific condition following diagnostic testing (Meder et al., 2009; Barrett and McKenna, 2011; Wegwarth et al., 2012). In law, the increasing use of statistical forensic evidence combined with the discouragement of forensic experts from presenting Bayesian analyses (Donnelly, 2005; Redmayne et al., 2011) has left both lawyers and jurors to make the necessary calculations for themselves ((Donnelly, 2005; Fenton et al., 2014). Ineffective presentations of such statistics greatly increase error rates in comprehension. The consequences have been shown to include poor patient decisions (Gigerenzer and Edwards, 2003; Navarrete et al., 2014) and miscarriages of justice (e.g. Forrest, 2003; Mehlum, 2009).

The actual process of Bayesian inference was first described and formalised by Thomas Bayes and was submitted posthumously in a paper to the Royal Society by Richard Price (Bayes and Price, 1763). In this paper, Bayes and Price provided the first formulation of how to mathematically incorporate new data into pre-existing beliefs to arrive at new belief levels. The same process was independently rediscovered by Pierre-Simon LaPlace (see Laplace and Simon, 1951) who was the first to publish the general modern formulation of what is now known as Bayes‘ rule:

$$\mathbf{P(H|D) = \frac{P(D|H) \times P(H)}{P(D|H) \times P(H) + P(D|\text{-}H) \times P(\text{-}H)}}$$

Fig. 1.1: Bayes‘ formula (1763), developed by LaPlace (1814 / 1951)

The output of the formula, P(H|D), commonly known as the 'posterior' belief level, is the probability of the hypothesis under question (H) being true, given the new piece of data observed (D). P(D|H) is the probability of that data occurring if H were true, and P(D|-H) is the probability of that data occurring if H were not true. P(H) is the probability that the hypothesis was true before the new data was observed, and P(-H) is the probability the hypothesis was not true before that data was observed.

1.3 Rational Man

1.3.1 The Enlightenment View of Human Rationality

In a similar pattern to the varying popularity of the Bayesian approach to statistical inference itself, the degree to which human reasoning on probability-updating tasks has been believed to conform to Bayes' rule has risen and fallen on several occasions. Classical probability theorists of the enlightenment such as Laplace as well as Poisson and Condorcet saw probability theories like Bayes' rule as being embedded in the common sense of educated persons and as an accurate representation of human judgement (Daston, 1995). Laplace stated that "the theory of probability is at bottom nothing more than good sense reduced to a calculus which evaluates that which good minds know by a sort of instinct, without being able to explain how with precision" (Laplace and Simon, 1951, pp. 196).

1.3.2 1944: The Economic Model

This view of the human faculties persisted for over two centuries, and was even further entrenched within the field of economics following the publication of the 'Expected Utility Theory' of Neumann et al. (1944). This model treated human agents as entirely rational decision-makers who would follow the laws of probability precisely when reasoning or making decisions. While initially theoretical, the theory swiftly became the standard assumption when modelling human behaviour within

economics (Thaler, 1980).

1.4 Base Rate Neglect

1.4.1 Early Findings

1955: Meehl and Rosen

Criticism of Expected Utility Theory (Neumann et al., 1944) as a predictor of human behaviour came swiftly following its publication. After a decade of growing discontent, two seminal papers by Edwards (1954) and Simon (1956) demonstrated the failure of this theory to account for actual human behaviour in a range of situations. Simon in particular criticised the implicit assumptions of the model that when faced with a decision or problem humans have unlimited time and processing power with which to make the decision, two ideas that Simon demonstrated to be clearly impossible. Simon's ideas later came to be known as 'Bounded Rationality' (Simon, 1972) and have been hugely influential in the decision making field (Gigerenzer, 1996).

In regards to Bayes' rule in particular, a major blow to the entirely-rational view of the human faculties arrived when Meehl and Rosen (1955) wrote a paper to the Psychological Bulletin expressing concern over the diagnostic tests being used by fellow psychologists. They also expressed reservations as to the theoretical accounts being put forward in the literature on how these tests should be interpreted. Meehl and Rosen believed that, when determining whether a patient had a given disorder, these psychologists were relying entirely on the outcome of their diagnostic test as well as the 'true positive rate' (the frequency with which that test correctly identifies the illness), and were overlooking other key information, such as the incidence of the diagnosis in the population.

An assessment of whether an individual has a given clinical diagnosis is one which requires inference in accordance with Bayes' rule: the diagnostician's prior belief (P[H]) in the probability of the patient having the disease before the diagnostic

test must be combined with the test result itself (P[D|H]) to arrive at a posterior, or updated belief in the likelihood that the patient has that disease (P[H|D]). However, the psychologists that Meehl and Rosen (1955) observed, in focusing entirely on the true positive rate of their tests, were entirely ignoring other information known prior to conducting the test such as the frequency of the illness in the population (otherwise known as the base rate). This was a violation of normative rationality and of probability and Bayesian theory, and stood in stark contrast to the picture of human faculties previously adhered to.

1966: Edwards and colleagues

The earliest empirical attempts to test whether human reasoning conformed to Bayes' rule were undertaken by Rouanet (1961), Phillips and Edwards (1966) and Edwards (1968). In a typical problem, Phillips and Edwards (1966) presented participants with a problem in which 10 bags each contained 100 red and blue poker chips and were told that the experimenter would soon choose one of the bags randomly. In a certain number of bags (the figure varied between conditions and was provided to the participant), red chips were more common, and in the remaining bags, blue were more common. This figure served as the Bayesian 'prior' (P[H]) and participants were asked to judge the probability that a predominantly red bag or blue bag was chosen on this data alone. They were then told that in the predominantly red bag the ratio of red to blue chips was 70:30, and the inverse ratio held for the blue bags (P[D|H]).

Participants were then shown a sequence of twenty chips (with replacement) from a randomly chosen bag and were asked to estimate the probability that each sequence of chips came from a predominantly red, or blue, bag (P[H|D]). This estimate was made by the participants by distributing 100 metal washers over two pegs, one indicating a red bag, one blue. With this design, Phillips and Edwards (1966) as well as Rouanet (1961) and Edwards (1968) found what they called 'Bayesian Conservatism'. By this they meant that participants' estimates, on the whole, were

lower, or less extreme, than was afforded by the data when calculated via Bayes' rule. However, it was also noted that all participants integrated the two pieces of data to some extent, providing an answer somewhere between the two values representing the distribution of bags, and the distribution of chips.

1972: Kahneman and Tversky

Several years later, Kahneman and Tversky (1972) presented participants with the now-classic Bayesian 'taxi-cab' problem:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city.
>
> You are given the following data: 85% of the cabs in the city are Green [P(-H)] and 15% are Blue [P(H)].
>
> A witness identified the cab as Blue [Data]. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colours 80% [P(D|H)] of the time and failed 20% of the time [P(-D|H)].
>
> What is the probability that the cab involved in the accident was Blue rather than Green [P(H|D)]?

In this problem participants are provided with the prior (P[H]) and it's compliment (P[-H]), as well as a single piece of diagnostic data which serves both as P(D|-H) and P(D|-H). Their task is to give the Bayesian posterior, P(H|D). This requires integration of all of these figures. However, the authors found that the majority of their participants' estimations of P(H|D) relied entirely on the eye witness accuracy rate [P(D|H)] - the modal answer was '80%' These participants, they noted, were entirely neglecting the distributional, or 'base rate' data of the proportions of cabs in the city (P[H]), which ultimately led to an over-estimation compared to the Bayesian norm. This pattern of responses was entirely unlike Phillips and Edwards (1966) finding of under-estimation relative to the Bayesian norm but highly analogous to the narrow focus of Meehls and Rosen's (1955) colleagues on their diagnostic tests to the exclusion of disorder prevalence base rates.

The problem used by Kahneman and Tversky (1972) has similarities to that

used by Phillips and Edwards (1966): a 'prior' distribution of green and blue cabs is given (equivalent to the red-predominant and blue-predominant bags distribution (and which both serve as P[H])), as well as further information of the reliability of the witness observing those cabs (equivalent to the distribution of chips within the bags (and which both serve as P[D|H])). There are however substantial differences in the presentation of these variables. These two pieces of information in Kahneman and Tversky's (1972) paper were quite dissimilar (the distributional data of the cabs in a city vs the reliability of an eye witness) whereas those of Phillips and Edwards were highly similar (distributional data of bags vs distributional data of chips within those bags). This may provide some explanation as to the greater amount of 'integration' of the two pieces of data in Phillips and Edwards study: perhaps it was much clearer to participants that the two pieces of data in that experiment could in fact be integrated mathematically. Another substantial difference in methodology is the way in which the data was presented to, and the responses were requested from, participants. While participants in Phillips and Edwards' study directly sampled the chip distributions and responded via "intuitive estimates" (Phillips and Edwards, 1966, pp.3) by placing metal washers on a peg, Kahneman and Tversky's participants received the data via word-based statistics and responded precisely and formally by writing down a numerical percentage. Given the greater realism of Phillips and Edwards approach, this might suggest that Kahneman and Tversky's finding was an artefact of abstract word problems only.

However, the extreme similarity of Kahneman and Tversky's (1972) findings to those observed naturalistically by Meehl (1955) make this simple interpretation less tenable. The form of the problem presented by Kahneman and Tversky was highly similar to the real-life problem investigated by Meehl. Both involved the combination of distributional data (the cab company proportions / the base rate of disorders in a population (P[H])) with diagnostic data (eye witness reliability / diagnostic test reliability (P[D|H])). In both cases, the subjects of the analysis were found to entirely ignore the distributional data and to focus entirely upon the

diagnostic data. The fact that this was observed by Meehl under natural conditions suggested that Kahneman and Tversky's finding was not a mere artefact of study design and may reflect a deeper psychological principle. Indeed, the phenomenon discovered by Meehl and Kahneman and Tversky has come to be known as 'Base Rate Neglect', and questions of when, under what circumstances, for whom, and why, it occurs, have formed the core of a large field of enquiry which persists to this day (Fischhoff et al., 1979; Bar-Hillel, 1980; Casscells et al., 1978; Eddy, 1982; Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996; Macchi, 2000; Evans et al., 2000; Sloman et al., 2003; Krynski and Tenenbaum, 2007; Welsh and Navarro, 2012; Johnson and Tubau, 2013; McNair and Feeney, 2014a).

1.4.2 1973 - 1980: Replications and Extensions

Many replications followed Kahneman and Tversky's (1972) finding. In the following year, Hammerton (1973), in a similar design to the original, but using a Bayesian 'Medical' problem in which the probability of a person having a disease [P(H|D)] was estimated from an imperfect diagnostic test with a given true positive rate [P(D|H)]. Hammerton found that when participants were asked to estimate the Bayesian posterior but weren't provided the prior (P[H]) informaion at all, no participants showed an awareness that a vital piece of information for solving the problem was missing.

Liu (1975) furthered Hammerton's work, presenting the same medical problem but including a prior and with a range of different values for the true positive rate [P(D|H)] and the prior [P(H)]. In two departures from previous designs, Liu first presented the base rate value after the diagnostic information (in all previous experiments the prior had always been given first), and secondly removed participants who had indicated they were 'just guessing' on the problem. Liu still found the modal answer in 11/12 conditions to be equal to P(D|H), indicating complete 'base rate neglect' as the most common response. Further, on additional questioning, almost one third of participants indicated the sentence containing P(D|H) as the most

important for answering the question.

Building more directly upon Kahneman and Tversky's (1972) work, Lyon and Slovic (1976) gave participants the same taxi-cab problem and then presented them with three reasoned arguments for the 'correct answer' which amounted to the statements that (a) only the base rate (P[H]) was needed , (b) only the diagnostic information (P[D|H]) was needed and (c) both pieces of information were needed. Fifty percent of their subjects chose (b), despite (c) being the correct argument.

Six years after Kahneman and Tversky's (1972) work, in another study which has since become a classic, Casscells et al. (1978) provided a worrying empirical demonstration of the 'base rate neglect' effect within medical experts. They gave the following 'disease problem' to 95 students and staff at Harvard Medical School:

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

This problem provided an even closer analogy to Meehl and Rosen's (1955) observations of psychologists' diagnoses. The problem provides as the base rate or prior the prevalence of the disease (P[H]: 1/1000) and the 'diagnostic' or 'new' information is presented via the positive test result and its false positive rate (P[D|-H]: 5%). Despite the relative simplicity in terms of number of variables, the authors found that only 18% of participants were able to give the correct answer of '2%' [P(H|D)]. Forty five percent of participants however, provided the answer '95%' which is generated by subtracting the 5% false positive rate from 100% [1-P(D|-H)]. This was the modal answer in the study, and again, relies entirely on the 'diagnostic information', completely neglecting the base rate information.

This problem introduced to the literature a new form of 'base rate neglect'. While all previous work had found that the majority of these responses involved giving P(D|H) as the answer, Casscells et al (1978) found the modal response to be 1-P(D|-H). A possible reason for this is that previous work provided problem solvers with P(D|H), while Casscells et al provided P(D|-H).

In a theoretical development, Ajzen (1977) proposed that a causal link between the base rate and the event of interest (e.g. the car accident) was necessary to avoid base rate neglect. This theory was later validated using two versions of the taxi-cab problem (Tversky and Kahneman, 1980). In the original version (Kahneman and Tversky, 1972), the base rate was based on the distribution of taxi cabs in the city. This, Tversky and Kahneman claimed, was not perceived by problem solvers as being 'causally related' to the accident, and was instead seen as 'mere statistical data'. In this new version (Tversky and Kahneman, 1980), the size of the two taxi-cab companies were made equal but a causal element was added: it was stated that 85% of accidents were caused by the green taxi company, while only 15% were caused by the blue (P[H]). This new framing, the authors thought, provided a strong causal schema: green taxi drivers were 'dangerous' drivers and were causing accidents much more often than blue taxi cabs. In this version only 18% of participants neglected the base rate entirely, and a full 60% of participants provided an answer which in some way combined both base rate and diagnostic information.

Bar-Hillel (1980) added to Ajzen's (1977) and Tversky and Kahneman's (1980) 'causal' necessity for base rates to be incorporated by problem solves with a further necessity: that of 'specificity'. Specificity was defined as the degree to which the reference class of the base rate was specific to the event of interest (e.g. the car accident). If the base-rate concerned an entire population it would be deemed as less-relevant to the event of interest. However if it concerned a sub-group of the population, and the target was within that sub-group it would be deemed more relevant. The more specific the sub-group, the more relevant the data would be considered to be, and less likely it would be to be neglected by problem solvers. To demonstrate the importance of this principle independent of the causality principle, Bar-Hillel presented a new version of the taxi-cab problem, where the 'diagnostic' witness information (P[D|H]) was replaced with non-causal distributional data similar to the base rate (P[H]). While both pieces of information were now equally 'causally' related to the accident, the diagnostic information was altered to be more

'specific' to the estimation requested. It was stated that in the neighbourhood in which the accident occurred "80% of all taxis are blue and 20% are green". In this situation, where both base-rate and diagnostic data were equally non-causal and 'distributional', but the diagnostic information was more 'specific' to the situation being judged (because it was 'in the area of the accident' rather than for the city as a whole) high levels of base rate neglect were again seen. This demonstrated, according to the author, that specificity was another important determinant in predicting whether problem solvers would neglect the base rate or not.

Bar-Hillel (1980) combined 'causality' and 'specificity' into the concept of 'relevance'. If people thought the base rate was both causally and specifically 'relevant' to the problem at hand, they would use it, and if they did not, they would neglect it. Bar-Hillel also went further: she proposed that more relevant information tended to 'dominate' less relevant information to an extent that was not justifiable: instead of simply down-weighting less-relevant information, as the Bayesian approach would advocate, many problem solvers simply disregarded it entirely. Bar-Hillel demonstrated this in another modification of the original taxi-cab problem where the diagnostic information was made both equally causal and equally specific to the base rate data. The base rate statement itself was not changed from Kahneman and Tversky's (1972) original. However, by downgrading the 'relevance' of the diagnostic information, Bar-Hillel (1980) found a dramatic drop in base rate neglect (12% of participants): now that the two pieces of information were equally 'relevant' neither dominated the other and both were integrated by the majority of participants to arrive at the final answer.

1.5 1981: New Paradigms

1.5.1 Paradigm Criticisms

Shortly after Bar-Hillel's work, the first in a long line of criticisms of the paradigm being used following Kahneman and Tversky's (1972) work began. The first came

from Cohen (1981), who claimed that the reasoning behind the development of the normative standards (the answer considered to be 'correct') for such problems was questionable. He stated that the problems being used in fact offered many interpretations, some of which produced answers in line with responses commonly being classed as fallacies, such as base rate neglect. One key issue Cohen focused on, echoing Bar-Hillel (1980) was the degree to which the reference class of the base rate was specific enough to apply to the problem at hand. Bar-Hillel had demonstrated that people were sensitive to this, which Cohen stated showed a good level of understanding in the majority of solvers. Cohen argued that depending on the level of non-specificity, and on the circumstances, a weighting of zero (and therefore base rate neglect) may in fact be an appropriate and rational approach, in line with Bayes' rule. Another issue raised by Cohen was that solvers may be using moral principles as a reason to avoid the base rate. For example, in the taxi cab problem, the solver is effectively in the place of a juror, and should therefore assume that the defendant is innocent until proven guilty. This may provide a moral imperative to ignore the distributional information, as is common in law. If this were the case Cohen pointed out, it would not be reasonable to conclude this person was succumbing to a fallacy: instead they may be well aware of what the mathematically 'correct' answer is, but be choosing to provide a different answer on ethical grounds. Continuing Cohen's line of thought, Niiniluoto (1981) claimed that if participants did deem the base rate to be non-relevant, for whatever reason, a non-informative Bayesian prior of .50 would yield .80 (equal to the most common answer in the taxi-cab problem), and equal to P[E|H]) as the normative answer.

A further criticism came two years later from Birnbaum and Mellers (1983). Birnbaum also questioned whether the normative values being used for the problems in the field were appropriate. The simplistic norms typically used in the taxi-cab problem relied, Birnbaum stated, upon the questionable assumption that the accuracy rate of the witness in the scenario would be entirely independent from the distribution of the cabs in the city. There are many reasons to suppose this may be

a faulty assumption. For example, the witness is likely to have been exposed to this distribution of cabs by being a resident of the city, which will have some impact on their accuracy in detecting both colours. The wide range of possible dependency values for these two factors according to 'signal detection theory' (e.g. Swets, 1964) is not stated in the problem, and is therefore, according to Birnbaum, free for problem solvers to infer. This range of 'dependency' values produces a similarly wide range of potential 'normative' posteriors, which, Birnbaum claimed, included the '.80' answer typically classified as base rate neglect.

Criticisms of the early paradigm continue to the present day. Over a decade after Cohen (1981) and Birnbaum's (1983) challenges, Koehler (1996), and even later, Welsh and Navarro (2012) offered fresh criticism. Koehler firstly agreed with Cohen and Birnbaum that the normative standards used to provide the 'correct answer' were asserted with far greater confidence than was warranted by established theory - many legitimate interpretations of the problems were possible other than the single 'acceptable' standard assumed by experiments. Further, echoing Cohen, Koehler criticized this blind focus on numerical normative answers without acknowledging or attempting to investigate potential individual differences in solvers' task goals, values and assumptions. He also criticized the ambiguity and lack of clarity of much of the language used in several of the classic problems, in particular Casscells et al. (1978). Concluding, Koehler lamented the lack of ecological validity or 'realism' in the field, and called for a paradigmatic change.

McKenzie (2003) echoed the discontent surrounding the ecological validity of the word problems being used in the field and the use of certain concrete language such as 'rational', 'irrational' and 'fallacy'. McKenzie proposed that the supposed 'correct' answers to the problems being used should be considered mere 'norms' and should cease to be used as hard 'standards' next to which people's rationality or irrationality should be judged.

Later, Welsh and Navarro (2012) developed these ideas, drawing attention to the fact that all problem solvers will bring some 'prior knowledge' of the problem

they are being asked to solve into the laboratory with them. Each problem solvers' prior may be different, and importantly, may also be different to the experimenter's prior. Bayesian inference allows for this, the authors noted, and it was perfectly 'rational' for participants to do this. Welsh and Navarro went on to experimentally investigate one area of 'prior experience': that any data source is never perfectly reliable and should never be entirely trusted. In the taxi-cab problem, for example, participants might justifiably wonder how recently the data regarding the taxi-cab-colour proportions was taken, and whether it is still accurate, as well as the extent to which the testing conditions for the witness were really "similar" enough to the night of the crash. It is exceedingly unlikely in fact that either of these pieces of data is perfectly valid, and so it is in fact more 'rational' for participants to distrust this data to some extent than to trust it entirely. Thus, on this view, the experimenter's 'correct' answer is in fact erroneous to some unknown extent as it assumes the data is perfectly accurate. Welsh and Navarro believed that data distrust might partly explain base rate neglect. The authors looked at four possible sources of data distrust and measured their effect on participants' answers to a Bayesian problem. They looked at age, location, source and the sample size of the data and found that participants' answers were closer to the Bayesian norm when data was stated as being recent, from a local location, from a trustworthy source and when the sample size was large. Clearly then, participants appeared to have a fairly strong understanding of principles of data validity, further calling into question assumptions that they are acting 'irrationally' in the taxi-cab and disease problems.

Interestingly however, while Welsh and Navarro found that 'high trust' data increased the frequency of normative responses over 'low trust' data, a third 'no statement' scenario, in which the authors did not even state any of the 'trust' details regarding the age, location, source or sample size of the data, in fact produced an even higher frequency of normative Bayesian responses than the 'high trust' scenario. The authors believed this was due to Gricean (Grice, 1975) principles of conversation: participants appeared to assume that the data is trustworthy unless

stated otherwise. Even the scenario in which the data is stated as being highly trustworthy may inherently imply that there is some reason to think the data might not be trustworthy. Therefore, while illuminating an important point that all future experiments should keep in mind, Welsh and Navarro's findings in this regard actually suggested that data distrust is unlikely to be a component of the reason for base rate neglect in the early studies. The original taxi-cab and disease problems had inadvertently in fact left unstated the trustworthiness of the data, which according to Welsh and Navarro's findings, should have in fact ensured higher trust in the data than if they had attempted to state how trustworthy it was.

A further note of interest in Welsh and Navarro's (2012) findings is the indirect evidence it provides for Bar-Hillel's (1980) 'dominance' theory. Welsh and Navarro gave participants 'prior' and 'new information' on the prevalence of various animals and plants on a fictional 'alien' planet. Both prior and new information were the same type of data (surveys of the area), however they varied on sample age, location, source and size. In Bar-Hillel's language, each of these factors alters the 'relevance' of the data. Participants were then asked to estimate the frequency of the given plant or animal in a given location at the present time.

On average, participants' answers were found to vary in a 'reasonable' way in response to each of these variables. As stated above, older samples from further away which were taken by a less trust-worthy source and with a smaller sample size were all weighted less in calculations. When examining the averages, these changes in estimates appeared to respond 'smoothly' to changes in the trustworthiness of the data, suggesting that participants were slowly 'down-weighting' the data in an appropriate manner. However, when examining participants' responses on an individual level, a completely different picture emerged. At the individual level it became clear that when data was 'low trust' a large number of participants were entirely neglecting the base rate information. Further, as the data moved from 'low trust' to 'high trust', participants were 'flipping' from providing complete base rate neglect to complete base rate conservatism (providing only the base rate as

the answer). This is precisely what was predicted by Bar-Hillel's (1980) dominance theory: it appeared that the 'less relevant' information, on an individual level, instead of simply being down-weighted, was, at a certain threshold, being entirely neglected.

1.5.2 The Confusion Hypothesis

From the 1980's a group of researchers also became increasingly focused on understanding the psychological mechanisms behind base rate neglect.

Eddy (1982), in a highly comparable paper to Meehl and Rosen's (1955) paper on psychological diagnostic tests, expressed concern both with the discussion of medical diagnostic tests in the literature as well as the actual practice of interpreting those tests by medical professionals. In contrast to Meehl and Rosen's assessment however, Eddy believed that many medical professionals may be confusing P(D|H) with P(H|D) and quoted one author in a 1972 issue of Surgery, Gynaecology and Obstetrics stating that:

> "In women with proved carcinoma of the breast, in whom mammograms are performed, there is no X-ray evidence of malignant disease in approximately one out of five patients examined [P(-E|H)]. If then on the basis of a negative mammogram, we are to defer biopsy of a solid lesion of the breast, there is a one in five chance that we are deferring biopsy of a malignant lesion [P(H|-E)]." (Eddy, 1982 pp.254)

The author here clearly indicates that the value for P(H|-D) should be equal to the value for P(-D|H. In a further study, reported informally in Eddy, doctors were presented with a 'medical diagnosis' problem based on this example, which provided participants with the incidence of the disease in the population (the prior, (P[H]), and the true positive rate of the mammogram test [P(D|H)]. They were told a woman had got a positive result on this test, and were asked to estimate the likelihood she actually had cancer [P(H|D)]. Ninety five out of 100 doctors were stated to confuse the false positive rate [P(D|H)] with this answer. Dawes (1986) suggested that this confusion may be the cause behind much of the base rate neglect phenomenon.

Pollatsek et al. (1987) demonstrated that the majority of people do in fact have the capacity to distinguish between two inverse conditionals e.g. 'the probability that a person who has a fever is sick' (P[F|S]) and 'the probability that a person who is sick has fever' (P[S|F]). However, Hamm (1988) later suggested that the way in which these were presented in Pollatsek et al made the distinction much clearer than in the paradigmatic presentation of Bayesian word problems in the field at large, such as in the taxi-cab and disease problems. Dawes (1986) echoed this view that verbal misunderstanding might be the issue, stating that "words are poor vehicles for discussing inverse probabilities" (Dawes, 1986, pp.80). Later, Macchi (1995) concurred with this, stating that other work (Bar-Hillel, 1990; Thüring and Jungermann, 1990) had also demonstrated that people could distinguish between conditional probabilities in other contexts. Macchi's conclusion, similar to Dawes, was that "this confusion may not rest so much on a natural tendency to err but, rather, on an ambiguous transmission of information produced by the structure of the text problem" (Macchi, 1990, pp.198).

Hamm (1988) also reiterated and expanded the 'Confusion Hypothesis' from Dawes (1986), stating that the data so far collected in the field suggests that the 'integrative' base rate neglect theory of Kahneman and Tversky (1972) and Bar-Hillel (1980), is wrong. In this theory, solvers recognise the problem as one in which two pieces of information need integration, but see the base rate as irrelevant to the problem at hand. The confusion hypothesis on the other hand, hypothesises that solvers first confuse the true positive rate (P[E|H]) or the inverse of the false positive rate (1-P[E|-H]) with the requested answer (P[H|E]), and then, (correctly, allowing for the initial confusion), conclude that this is the only piece of information necessary, and ignore the redundant 'base rate'. A series of additional studies (Hamm, 1988; Hamm and Miller, 1990) using various process-trace and verbalising protocols seemed to confirm this view: participants did not appear to immediately consider the base rate irrelevant, but did frequently appear to confuse the two conditional probabilities. This general picture has been advocated by several other

authors (Braine and Connell, 1990; Hamm, 1994; Wolfe, 1995; Lewis and Keren, 1999; Mellers and McGraw, 1999; Fiedler et al., 2000; Macchi, 2000).

Wolfe (1995) developed the confusion hypothesis further by drawing on fuzzy trace theory (Brainerd and Reyna, 1990; Reyna and Brainerd, 1991). Fuzzy trace theory states that when faced with problems such as Bayesian word problems, solvers maintain multiple representations of the variables involved, at varying levels of precision. At the lowest levels of precision, the conditional probability P(H|E) will be encoded as a simple 'gist' representation where the solver merely has awareness that 'H's and 'E's tend to be related, or co-occur. If solvers' are operating at this level of 'gist' processing when attempting to solve the problem, P(H|E) and P(E|H) will both be encoded identically (the co-occurrence of H's and E's), and may be psychologically indistinguishable.

To test this theory, Wolfe (1995) devised an experiment in which one group of participants were provided training on the 'Transposed Conditional Fallacy', another name for the confusion of P(E|H) with P(H|E). In a paradigm where participants 'requested' the numbers for the variables they wanted to use in their calculations, Wolfe found that participants in the training group showed greater interest in the base rate number (83% versus 50%) and the group showed a mean estimate significantly closer to the Bayesian norm than the non-training group. Frequencies of answers correct were not given however.

1.5.3 1995: Natural Frequencies and Nested Sets

Gigerenzer and Hoffrage

From a historical perspective, one of the most key and pivotal moments in the field examining Bayesian word problems came in 1995 with the publication of a paper by Gigerenzer and Hoffrage. This paper and its successors drew attention initially away from the confusion hypothesis and focused instead on the format of the numerical values being used in the problems.

All previous work up to this point had used percentages and probabilities to

describe the likelihood of both the base rate and the diagnostic information in the Bayesian problems presented. Gigerenzer and Hoffrage (1995) argued that this usage may in fact lie behind the low accuracy and high base rate neglect being seen. While Bar-Hillel's (1980) work had succeeded in reducing complete base rate neglect, the normative answer was still being seen in less than 20% of participant answers. Gigerenzer and Hoffrage produced an argument from evolution to explain why success rates were so low in that previous work: if people did possess a mental capacity for Bayesian inference, they stated, it would be very unlikely to be designed to operate on probabilities or percentages, which were invented extremely recently in human history. The type of data humans would have experienced in their past, and therefore should be adapted to process, would have been in the form of the counting of instances over time (i.e. direct sampling experience of frequencies). Gigerenzer and Hoffrage called this process 'natural sampling' 'and the figures that come out of this process 'natural frequencies'. They presented a new version of the medical diagnosis problem (Eddy, 1982) as an example:

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer.
>
> 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography.
>
> Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer?
>
> __ out of __

Unlike in Casscells et al's (1978) similar disease problem, where the base rate of the disease and the diagnostic information are presented as probabilities, percentages or normalized frequencies, Gigerenzer and Hoffrage's (1995) scenario presents 'natural frequencies'. These are the data you arrive at if you naturally sample the population, one by one, counting whether each individual possesses the hypothesis (H) of interest (e.g. the disease) and the effect (D) of interest (e.g. the symptom).

Gigerenzer and Hoffrage (1995) converted several of the previous problems used, including the above medical diagnosis and also the taxi-cab problems, into natural

frequency formats, and found a large increase in success rates from around 16% with probability formats (typical of previous work) to around 46% with natural frequency formats. This success rate was far greater than any previously seen either though Bar-Hillel's (1980) 'relevance' changes or Wolfe's (1995) 'confusion' changes. Furthermore, 'base rate neglect' dropped dramatically, to around 13%.

Cosmides and Tooby

Much controversy followed this work, particularly surrounding the psychological mechanisms and causes behind the increased accuracy with natural frequency formats. In the following year, Cosmides and Tooby (1996) focused even more acutely on the evolutionary argument, claiming that humans possessed a 'mental module' designed for processing Bayesian problems but which only worked on frequencies. They produced a large number of experiments demonstrating the benefit of frequency formats on Bayesian inference. Cosmides and Tooby's methods however caused great confusion in the following decade due the fact that, in contrast to Gigerenzer and Hoffrage (1995), they in fact used normalised frequencies (with a fixed or 'normalized' denominator: natural frequencies have a denominator which is determined by the natural sampling process and thus conveys additional base rate information), but still achieved high accuracy. This was a surprising result as Gigerenzer and Hoffrage (1995) had in fact predicted that normalised frequencies would not enhance accuracy. However, subsequent work (Evans et al., 2000) showed that the particular numbers and format used in Cosmides and Tooby's paper allowed participants to interpret the problem as natural frequencies. The two answers produced, depending on whether the problem was interpreted as natural, or normalised frequencies were extremely similar, and Tooby and Cosmides' (1996) analysis failed to distinguish between these two interpretations.

When this confound was removed, Evans et al. (2000) found that normalised frequencies did not outperform percentage formats. Gigerenzer and Hoffrage (1995) had in fact also already demonstrated in their rarely-cited second experiment in

that paper, that normalised frequencies did not improve accuracy above probability and percentage formats. The impact of this confusion in the field lingers to the present day. Hoffrage et al. (2002) were forced to issue a clarification letter detailing studies which had confused natural frequencies with normalised frequencies (e.g. Evans et al., 2000; Macchi, 2000; Johnson-Laird et al., 1999; Girotto and Gonzalez, 2001; Macchi and Mosconi, 1998). However the confusion was not fully abated by this, and later, Sloman et al. (2003) also committed the same error.

Fortuitously, this confusion may in fact have had a beneficial outcome. Many of the authors mentioned above, in determining that normalised frequencies, outside of the confounded design used by Cosmides and Tooby (1996), did not in fact improve accuracy, began to search for what factor, if not frequency, was the cause of the facilitation. One clear answer, already expressed in Gigerenzer and Hoffrage's (1995) paper, was that natural frequency Bayesian problems (and the 'pseudo-natural' frequencies used in Cosmides and Tooby (1996) were computationally simpler to solve than probability or percentage-based problems. This was due to the fact that natural frequencies provide the conjunctions of the hypothesis (e.g. cancer or no cancer) with the effect of interest (e.g. positive result or no positive result), which are a necessary step in calculating the final product and which need to be calculated from first principles in probability formats. This difference however does not explain the entire effect, as demonstrated by Gigerenzer and Hoffrage's (1995) own 'short menu' experiment which provided the same conjunctions for the probability format also, making each computationally identical. Natural frequency formats, even with this control, still outperformed probability formats.

The remaining capacity of natural frequency formats to increase accuracy on Bayesian problems has come to be regarded by many authors in the field (e.g. Evans et al., 2000; Macchi, 2000; Sloman et al., 2003; Barbey and Sloman, 2007) as due to the fact that they reveal the 'nested sets structure' of the problem (See Figure 1.2 below). This again was expressed in Gigerenzer and Hoffrage's (1995) paper, but was not considered to be separable or independent from the natural frequency format

itself. The nested sets structure of a Bayesian problem is demonstrated effectively with a parent / child node diagram. The diagram below is based upon the medical diagnosis problem and adapted from Gigerenzer and Hoffrage (1995). It depicts the total population in the problem as the parent node, which is subdivided into child nodes corresponding to those women with and without breast cancer. These are each further sub-divided into those women with and without a positive test result. It is the revelation of this sub-divided structure which the above authors believe is the largest factor in the success of natural frequencies rather than the particular unit (e.g. frequencies) used.

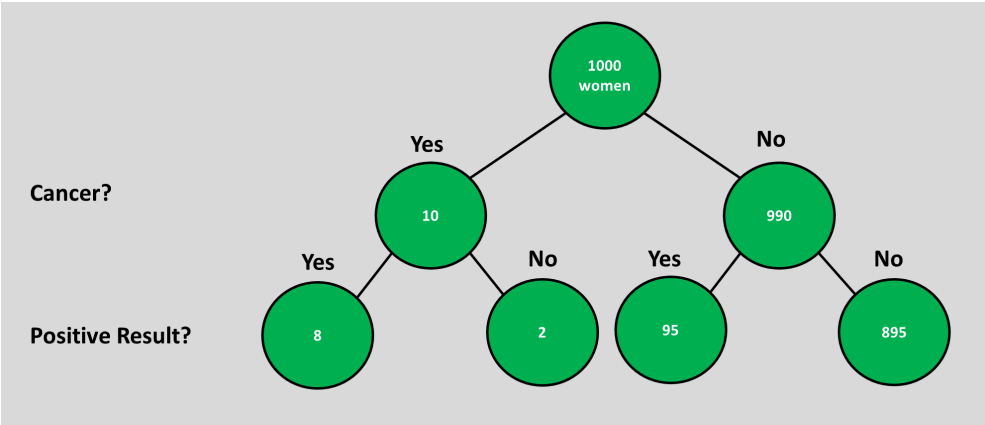


Fig. 1.2: A depiction of a natural frequencies, or nested sets representation of the mammogram problem, adapted from Gigerenzer and Hoffrage (1995). The model contains the information that women with and without breast cancer are subgroups of all women and that these two sub-groups can be further subdivided into those women with positive and negative test results

Macchi

Despite Gigerenzer and Hoffrage's (1995) belief that the nested-sets-revelation feature of natural frequency formats was inseparable from the format itself, Macchi (2000), developing on previous work by Macchi (1995) and Macchi and Mosconi (1998) found an approach which succeeded in separating them experimentally. The method relied upon expressing both of the statistics in the problem (the base rate and the diagnostic information) as well as the question from an 'outside' or 'group' perspective rather than from an 'inside' or 'individual' perspective, as was typical in

previous work. Importantly, this could be achieved with percentages as equally as it could for frequencies. For example, with this paradigm, the false positive rate in the medical diagnosis problem would be expressed from the group perspective as 'Out of those women who do not have cancer, 15% would still receive a positive result' rather than from the individual perspective as 'A woman who does not have cancer still has a 15% chance of receiving a positive result'. Such an approach was in fact proposed for various other problems prior to Macchi's work. For example, Tversky and Kahneman (1983) found that the 'outside perspective' focused on a group of individuals reduced the number of participants committing the conjunction fallacy compared to the 'inside perspective' focused on an individual. The outside perspective had also previously been adopted to reduce the planning fallacy (Griffin and Buehler, 1999) and overconfidence (Griffin and Tversky, 1992). However Macchi was the first to apply it to a Bayesian problem.

Macchi (2000) compared these two percentage formats to each other and also to a natural frequency format. Outside-framed percentages and natural frequencies were found to be non-significantly different from each other in terms of participant accuracy, while both formats significantly outperformed inside-framed percentages. In the same year, Fiedler et al. (2000) also compared 'outside / group' percentages to natural frequencies, but unlike Macchi presented the question in the same, inside-view / probability format in both conditions. Fiedler et al. (2000) used the mean-deviation method (average distance of answers from the normative answer) of comparing the two formats, as opposed to a simple count of accurate answers, and also found no significant difference. A significant difference however was found between both formats and a normalized frequency format.

Two years after Macchi's study, Hoffrage et al. (2002) claimed that Macchi's (2000) and Fiedler et al's (2000) participants may actually have solved the problem in those studies by constructing a natural frequency version of the problem for themselves. It may be, the authors claimed, that the 'outside-framed' percentage view simply encourages solvers to construct natural frequency versions to a greater

extent than the 'inside view' and that the difference between these two conditions can be fully accounted for by an increase in construction of natural frequency formats by participants during solution. This was thought all the more likely as Macchi (2000) in fact provided the base rate figures as integers rather than percentages. The question of whether participants solve the problem this way can be most easily answered by a process-tracing method. Unfortunately, while Macchi (2000) did in fact record participants solution processes and analysed these, they were only used to provide counts of different types of responses and no analysis of the process individuals undertook was provided. This interpretation of her work by Hoffrage et al therefore cannot be ruled out. However, the work does incontrovertibly provide evidence that the computational simplicity feature of natural frequencies does not play a major role in increased accuracy.

Girotto and Gonzalez

While the nested sets approach to assisting Bayesian reasoning flourished in the new millennium, subsequent work did not in fact build upon Macchi's (2000) approach, despite it's success in separating the nested sets format from the natural frequency format. Instead, in the following year, Girotto and Gonzalez (2001) took a drastically different approach. The authors wished to demonstrate that the assistive effect of natural frequencies was due in majority to two features: the fact that the body of the text in the problem possessed a 'partitioned structure' (similar to Macchi's proposal) and the fact that the question structure, for example in the medical diagnosis problem, encourages individuals to firstly calculate the total number of positive results, and to then divide this by the conjunction of cancer with positive results, which provides the correct answer. They claimed that these two features were not in fact unique to natural frequencies and could be replicated in other formats. The format they chose to demonstrate this was 'chances' expressed as natural numbers. They produced the below version of the 'disease' problem (e.g. Casscells et al, 1978):

> A screening test of an infection is being studied. Here is the information about the infection and the test results. A person who was tested had 4 chances out of 100 of having the infection.
>
> 3 of the 4 chances of having the infection were associated with a positive reaction to the test.
>
> 12 of the remaining 96 chances of not having the infection were also associated with a positive reaction to the test.

Girotto and Gonzalez (2001) compared this version to a natural frequencies version using individual 'people' instead of 'chances'. They also crossed this design with '1-step' versus '2-step' question formats. In the 1-step question format, individuals were asked (in the chance version) 'If Pierre has a positive reaction, there will be ___ chance(s) out of ___ that the infection is associated with his positive reaction' while in the 2-step version they were asked 'Imagine that Pierre is tested now. Out of a total of 100 chances, Pierre has ___ chances of having a positive reaction, ___ of which will be associated with having the infection.'

The authors firstly found no difference between the chances format and the natural frequency format. Both elicited similar levels of accuracy. This was a surprising result, as the chances version of the problem was in fact focused on the 'inside view', which was found to be inferior in Macchi's (2000) work. It concerns only one individual (Pierre) and his abstract 'chance' of having the disease. However the information format is still 'partitive' or possessing a 'nested sets' or 'class inclusion' structure. This study therefore gave evidence firstly that natural frequencies are not a privileged format and secondly that the target of the task does not even need to be presented from the 'outside view' in order to increase accuracy. The results suggest it is simply a partitive or nested information structure that is required to elicit high levels of reasoning accuracy on Bayesian problems. However, there may be an issue with this final interpretation. While the 'chances' task did take the 'inside view' in regards to 'Pierre' in that it focused on him as an individual rather than focusing on a group of individuals similar to Pierre as in Gigerenzer and Hoffrage's (1995) and Macchi's (2000) work, it may be argued that the format still provides an outside view, but simply using the units of 'chance', instead of people. The reader

is effectively encouraged to imagine a 'population' or 'group' of 100 chances and in the end is asked to calculate the number of chances with a positive result which also relate to having cancer. In effect, the wording of the problem shifts the perspective from the subject Pierre entirely to the abstract concept of 'chance'. Pierre is indeed being considered as an individual, unlike in Macchi's approach, but he is no longer the subject of analysis. The object of analysis is the unit of chance, and these are being presented from the outside view. Therefore, while this result may not have completely overturned previous work, it did provide a novel new way of presenting Bayesian problems without having to focus on a group of individuals.

Girotto and Gonzalez's (2001) comparison of 1-step to 2-step questions also contributed an important finding to the field. Their 2-step question forced participants to compute the total number of positive results, D (in the medical diagnosis problem) prior to computing the final product. This change hugely increased accuracy from 18% to around 50% of participants and suggested that this 'computation of total positive results' might be an important step in the process by which individuals complete Bayesian problems.

Following the results of Macchi (2000); Fiedler et al. (2000); Girotto and Gonzalez (2001),, debate continued within the field as to whether the natural frequency effect was, or was not, any more than the 'nested sets revelation' effect plus the 'computational simplicity' effect. Hoffrage et al (2002) maintained that there was a difference. Those authors' opinion was that the revelation of the nested sets structure was just one inseparable element of the natural frequency format. They claimed that Girotto and Gonzalez's (2001) chances-format was just "natural frequencies disguised as probabilities" (Hoffrage et al, 2002, pp.350). This criticism however missed the point. Girotto and Gonzalez's chance format did take elements from Gigerenzer and Hoffrage's (1995) natural frequency format, but it also left behind an important feature: the focus on concrete, non-divisible units (e.g. people) as opposed to abstract 'chances'. Their work showed that this feature which many authors within the frequentist approach (in particular Cosmides and Tooby, 1996;

and later, Brase (2007)) claimed was a necessary component for assisting Bayesian reasoning, was in fact superfluous.

Some later work by Brase (2008) also called into question the distinction between Girotto and Gonzalez (2001) 'chances' format and the natural frequency format. Brase was interested in how participants were interpreting the 'chances' format, and presented his participants, in a series of experiments, with a range of different 'chances' and 'natural frequency' formats. Brase then also gave his participants two definitions, one for natural numbers (i.e. frequencies) and one for probability. Participants were asked to choose which of these definitions most closely resembled their interpretation of the problem. Brase stated his experimental reasoning as "If the crucial aspect is only the existence of non-normalized set relations, single-event probability interpretations of chances should not affect performance on tasks, so long as that non-normalised set structure is in place." (Brase, 2008, pp.285). Brase found that those participants who interpreted the chances format as natural numbers outperformed those who interpreted them in a probabilistic sense. Brase concluded that this provided evidence that, similar to Hoffrage et al's (2002) complaint regarding other 'nested sets' approaches, that natural frequencies was also the true underlying reason for the increased accuracy in Girotto and Gonzalez's study. In a clear and important communication of the modern frequentist position, Brase concluded his paper with the claim that natural frequencies should still be considered a 'privileged format', and stating that:

> "This claim does not imply that the human mind is incapable of utilising other information formats, either in other settings or even within a particular setting; that would be an exclusive, rather than a privileged representational format. This type of privileged, nonexclusive situation has been descried by Sperber (1994) in terms of a proper domain (the content with which a cognitive system was actually designed to work) and a larger actual domain (the content with which a cognitive system can be persuaded to work). In the assessment of natural frequencies as a cognitively privileged representational format, there have been some confusions across these proper and actual domains. Some tasks claimed to be within the envelope of the proper domain conditions have, in fact, violated those bounds (e.g. by rote conversion of numbers into frequencies, rather than using natural frequencies [Here Brase refers to studies

> such as Macchi (2000) and Fiedler et al (2000)]). Conversely, other tasks claimed to be outside the envelope of the proper domain conditions (e.g. using chances as purported non-frequencies [Here Brase refers to Girotto and Gonzalez, 2001]) have, in fact, actually been within those bounds." (Brase, 2008, pp.288)

Diagrammatic Approaches

In the subsequent years, several studies continued to attempt to demonstrate that methods other than natural frequencies could reveal the nested sets structure of a Bayesian problem. Many used diagrammatic approaches to do this. A range of diagrams were used including the Euler diagram(Sloman et al., 2003), a novel 'roulette-wheel' diagram (Yamagishi, 2003), and later the Venn diagram (Micallef et al., 2012), all of which ultimately reveal the 'nested sets' or 'sub-divided' nature of the problem. Each of these generally showed increased accuracy on Bayesian problems. Again much debate surrounded the degree to which these studies effectively ruled out natural frequencies as the cause of increased accuracy. Brase (2009) demonstrated that diagrams based on the natural frequencies concept (such as frequency grid diagrams) significantly outperformed other diagram types such as the Venn diagram, and in fact also found no difference between providing the problem with a Venn diagram and no diagram. One potential issue with this finding however is that in the frequency grid diagram it was possible for participants to literally 'count' the number of individuals on the diagram, potentially allowing them to answer the question through a secondary means. This was not possible in the Venn diagram condition and must therefore be considered a potential confound.

Furthermore, diagram-based studies are limited in their value for resolving this issue. While they do demonstrate methods that can be used to help people understand Bayesian inference, they contribute little to any theoretical attempt to isolate the nested sets effect. A visual diagram must always be considered potentially more engaging to a participant than a block of text, and this greater engagement will act as a potential confound on any diagram vs text study. It is also in general difficult to be certain what other factors a diagram is introducing other than the nested

sets revelation. More intriguing however, are text-diagram-cross studies such as Garcia-Retamero and Hoffrage (2013), who crossed text format (inside-percentage vs natural frequency) with diagram type (no diagram vs frequency grid). They found that frequency grids were assistive to Bayesian accuracy when compared to inside-percentage-text but showed no significant increase when compared to natural frequency texts. This results suggests that such diagrams may be 'working' via the same mechanism as the natural frequency text format. This finding is not exclusive to the natural frequency realm. In a similar approach but using a 'nested sets approach', Sloman et al. (2003) found in another crossed design, that a Euler diagram produced a greater increase in accuracy compared to their non-nested-sets 'inside' percentage text than their nested-sets 'outside' percentage text.

1.5.4 2007: The Causal Approach

In 2007, in a third major directional-shift in the field, Krynski and Tenenbaum introduced a drastically new approach to presenting Bayesian problems. Their approach built upon developments in computer science that drew upon 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 the early work in the field (e.g. Kahneman and Tversky, 1972; Casscells et al., 1978; Bar-Hillel, 1980) as well as the natural frequencies and nested sets (e.g. Gigerenzer & Hoffrage, 1995; Fiedler et al, 2000; Macchi, 2000; Girotto & Gonzalez, 2001) approaches, had been proven ineffective by computer models in more realistic, everyday reasoning scenarios due to the complexity and low levels of information-certainty in those situations. Since humans also operate and evolved in these noisy environments, they claimed that purely statistical models were unlikely to be good descriptive models of human reasoning. They therefore theorised that people in fact normally approach Bayesian reasoning problems by 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 'diagnostic' or 'new' data (the hit rate of the test). This, Krynski and Tenenbaum argued, prevented solvers from constructing a causal model and thus from solving the problem in their normal way, resulting in the low accuracy seen and the high levels of base rate neglect. An adapted version of Krynski and Tenenbaum's core model for the medical diagnosis problem can be seen below in 1.3, which embodies the information that both breast cancer and harmless cysts (the 'cause' which they added) are possible causes of positive mammogram test results.

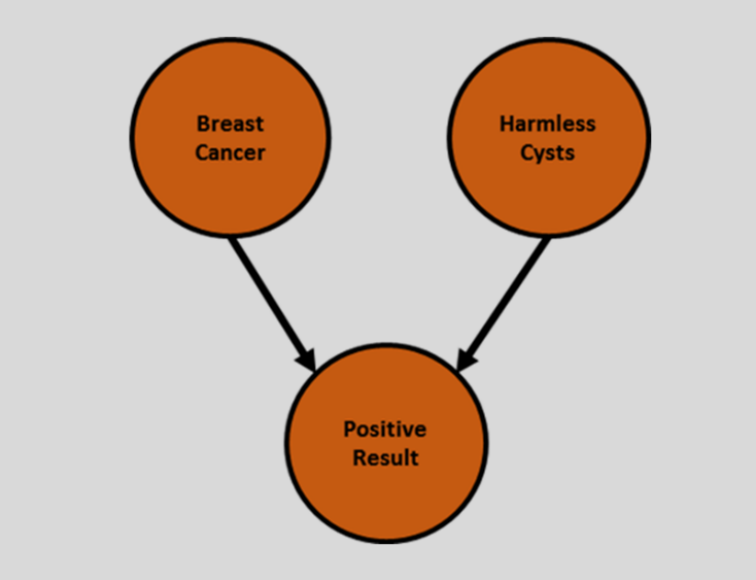


Fig. 1.3: A simple representation of Krynski and Tenenbaum's causal model of the mammogram problem. The model contains the information that both breast cancer and harmless cysts are possible causes of positive mammogram test results

The authors found in two separate experiments that by simply adding a single sentence providing this cause, accuracy increased from around 20% to around 40%. Two further experiments drawing on the Taxi-Cab problem (Kahneman and Tversky, 1972) and one with a novel 2x2 design, both also showed a beneficial effect for providing a clear causal structure. Some replicative success has followed this paper in subsequent years, with one study finding an improvement with the causal format

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 the high-numerate sub-group (McNair and Feeney, 2014b).

Part I

REASONING ON BAYESIAN WORD PROBLEMS

2 The Need for Reform

In the previous few years a growing discontent with the lack of progress being made in the cognitive science field of Bayesian reasoning has emerged. In a recent appeal, McNair (2015) expressed concern that despite a long history and a large amount of experimental data, recent work with both the 'nested sets', 'natural frequency' and 'causal' approaches were still showing high variability in accuracy rates across experiments even with similar methodologies. This, McNair claimed, demonstrated a lack of fundamental understanding of why some individuals succeed and others fail. For this reason McNair advocated greater use of the 'think aloud' protocol (Ericsson and Simon, 1980; Gigerenzer and Hoffrage, 1995) and individual differences measures in order to understand the thought processes that both successful, and unsuccessful individuals go through when attempting to solve Bayesian problems.

This general view of the need to improve a stagnating field was expressed in another recent review by Johnson and Tubau (2015). They believed that even the best methods in use had not achieved particularly impressive accuracy levels, demonstrating a lack of fundamental understanding of the psychological mechanisms underpinning Bayesian reasoning. They also lamented the field for being too focused on purely numerical outcome measures. The authors stated that "Moving forward, we believe there is a need to shift perspective from the facilitators of Bayesian reasoning to more process-oriented measures aimed at uncovering the strategies evoked by successful and unsuccessful reasoners, and the stages in the problem solving process at which these differences emerge." (Johnson and Tubau, 2015, pp.14).

A further issue in the field, but by no means limited to it, and not touched upon by either of the above authors is the widespread use of undergraduate populations as

participants. This latter issue is important because other work has suggested 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, Fiddick, and Harries (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. McNair and Feeney (2014b) made a similar argument in regards to the causal approach of Krynski and Tenenbaum (2007), and suspected that the likely high-numerate sample used by Krynski and Tenenbaum (Massachusetts Institute of Technology students) may have contributed to their finding of a causal effect within their entire sample. McNair and Feeney found in their sample that the causal approach only produced increased accuracy on Bayesian problems within the high-numerate sub-group of their sample. In further support of this argument, Micallef et al. (2012), who did study 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).

It is clear therefore that the field as it stands suffers from a range of issues. On the whole, it lacks information on which individuals fail to solve Bayesian problems, why those individuals fail, and at what point in the solution process. Even within undergraduate populations, accuracy remains relatively low (40% or less). Further, these low accuracy rates are likely to still be an overestimate compared to the general population. As it is the general population, and not just undergraduate students who must face Bayesian problems in real life situations such as in medical decision-making situations, future research must firstly ensure that the success of current approaches such as the nested sets and causal approaches also apply to the

general population. Further, it must be determined whether those individuals with lower numeracy also benefit from these approaches, or whether new approaches are needed. Finally the underlying cause of the increased accuracy of these approaches must be understood more deeply than currently, and this must be determined on an individual basis, in order to find ways to improve these methods in order to achieve the maximum possible number of individuals, across the numerical ability spectrum, having the ability to understand and solve Bayesian problems.

3 Experiment One

3.1 Introduction

The first aim of the present study is to replicate the natural frequency / nested sets and causal approaches to improving Bayesian accuracy in large general population samples. Macchi's (2000) 'outside-framed' approach (which presents percentages framed as proportions of groups, rather than as the abstract concept of chance) to revealing nested sets structure will be used to represent the former as it has achieved a greater reduction in interpretative confounds than other approaches such as natural frequencies themselves (computational confound), Girotto and Gonzalez's (2001) chances approach (according to Brase [2007] there is debate over whether these are distinct from natural frequencies or not) or diagrams (potential engagement confound). This approach will be compared to Krynski and Tenenbaum's (2007) causal approach. Within this aim it is hypothesised that a significant main effect of the nested sets framing will be found. It is further hypothesised, based on McNair and Feeney (2014b) and the fact that the present sample is likely to have lower numeracy levels than previous, that in this sample no significant main effect of the causal framing will be found in the sample as a whole. However, it is hypothesised that the high-numerate sub-group, split at the median, will show a significant main effect of causal framing.

The second aim is to combine the nested sets and causal approaches in one condition and to compare all four conditions to determine if the two effects are additive. It is again hypothesised that no significant interactive effect between the two conditions will be seen in the sample as a whole, but a significant and positive

interactive effect will be seen in the high-numerate sub-group.

The third aim of this study is to heed McNair (2015) and Johnson and Tubau's (2015) appeal to examine problem-solving processes and individual differences by using a 'think aloud' methodology alongside a numeracy measure 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' processes people undertake when approaching these problems as well as how the nested sets and causal framings affect these.

3.2 Method

3.2.1 Participants

The final sample size for the experiment 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 this experiment as well as experiments 2, 3 and 4 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.

Tab. 3.1: Demographics for experiments one, two, three and four

	Experiment One		Experiment Two		Experiment Three		Experiment Four	
	Numeric	Percent	Numeric	Percent	Numeric	Percent	Numeric	Percent
Total Sample	113	100%	521	100%	429	100%	364	100%
Gender								
Male	51	45.10%	232	44.50%	220	51.30%	212	58.20%
Female	61	54.00%	288	55.30%	208	48.50%	152	41.80%
Other	1	0.90%	1	0.20%	1	0.20%	0	0%
Age								
Minimum	20	-	18	-	19	-	18	-
Maximum	66	-	71	-	75	-	67	-
Mean	33.1	-	34.2	-	36.7	-	35.1	-
Std Dev.	10	-	11.6	-	12.3	-	10.8	-
Education								
High School	31	27.40%	157	30.10%	141	32.90%	138	37.90%
Bachelor's Degree	55	48.70%	267	51.60%	199	46.40%	172	47.30%
Master's Degree	22	19.50%	67	10.90%	63	14.70%	36	9.90%
Doctoral Degree	2	1.80%	12	2.30%	13	3.00%	4	1.10%
Other	3	2.70%	26	5%	13	3.00%	14	3.80%
Occupation								
Professional / Managerial	42	37.20%	218	41.80%	162	37.70%	130	35.70%
Labour / Service	35	31.00%	107	20.50%	129	30.10%	115	31.60%
Student	5	4.40%	65	12.50%	23	5.30%	4	10.20%
Unemployed	16	14.20%	70	13.40%	69	16.10%	5	12.40%
Other	15	13.30%	61	11.70%	46	10.70%	4	10.20%
First Language								
English	110	97.30%	517	97.70%	422	98.30%	-	-
Other	3	2.70%	4	2.30%	7	1.70	-	-

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

3.2.2 Design

The study employed a fully crossed 2 (nested vs non-nested) x 2 (causal vs non-causal) within-participants 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.

A within-participants design was chosen 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 there were individual differences as to which of the two framings people find 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 workings-out 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. They undertook this before having access to the next page where they could then enter their numerical answer.

3.2.3 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'. The basic-Mammogram, nested-College, causal-Library and nested-causal-Gotham problems can be seen in Appendix A 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
>
> 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 generalisable 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 Appendix A 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).

3.2.4 Procedure

Participants were recruited through the Amazon MTurk outsourcing service. 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 Appendix B. 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.

3.2.5 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 lead to the normative numerical answer on this one particular problem but would not have if the numbers in the problem were different. 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 classic paper by Cosmides and Tooby (1996)). Finally this method avoids the issue that many previous studies have faced of having to decide whether to accept only precise answers (McNair & Feeney, 2014a), accept those within a certain percentage of the correct answer (Krynski & Tenenbaum, 2007), or conduct a distance-from-correct calculation (Micallef et al., 2012), 2012), all of which will face issues with false positives or false negatives. A mixed method has greater power to detect effects as it is able to reduce statistical noise by reducing false positives and false negatives in categorising

answers as correct or incorrect.

Additionally, certain answers which did not give the correct numerical answer were also accepted. This was the case when their think aloud data indicated that they had undertaken an acceptable process, but had made an 'uninteresting error'. Uninteresting errors came in two types in the present study. Firstly, when participants had clearly done everything correct except for an arithmetical error, their answer was accepted as correct. In fact only 5 cases out of 452 included an arithmetic error. Secondly, when participants gave the wrong 'cause' as the answer (e.g. percentage of women without cancer, instead of percentage with cancer) but had again undertaken the correct Bayesian 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 author. In the first analysis phase the 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 author re-read 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, a volunteer coder with prior qualitative analysis experience was provided definitions of each of the codes by the author and was asked to assign all participants' answers to whichever codes they deemed appropriate while blind to both participant condition and to the authors' original coding. Inter-rater reliability between the author and this volunteer coder was ninety percent. In the fourth stage of analysis, any discrepancies between author and first volunteer coding assignments were resolved by an additional volunteer, also with previous qualitative analysis experience. This second volunteer was given the discrepant participants' think aloud answers and the

coding categories. The second volunteer'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, regression 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 simultaneous analysis of both between-subject and within-subject variables.

3.3 Results

3.3.1 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 and occupation). Below, in figure 3.1 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, but lower level, than the nested and nested-causal conditions. Confirming this difference, a repeated-measures generalised linear model with binomial distribution and logit function found a significant main effect of nested sets framing (χ^2 = 7.358, p=.007), but no significant main effect of causal framing (χ^2 = .834, p=.361) and no significant interaction (χ^2 =.237, p=.626).

To test the two hypotheses that a main effect of the causal framing and an interactive effect between 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 sub-group was made by following the same method used by

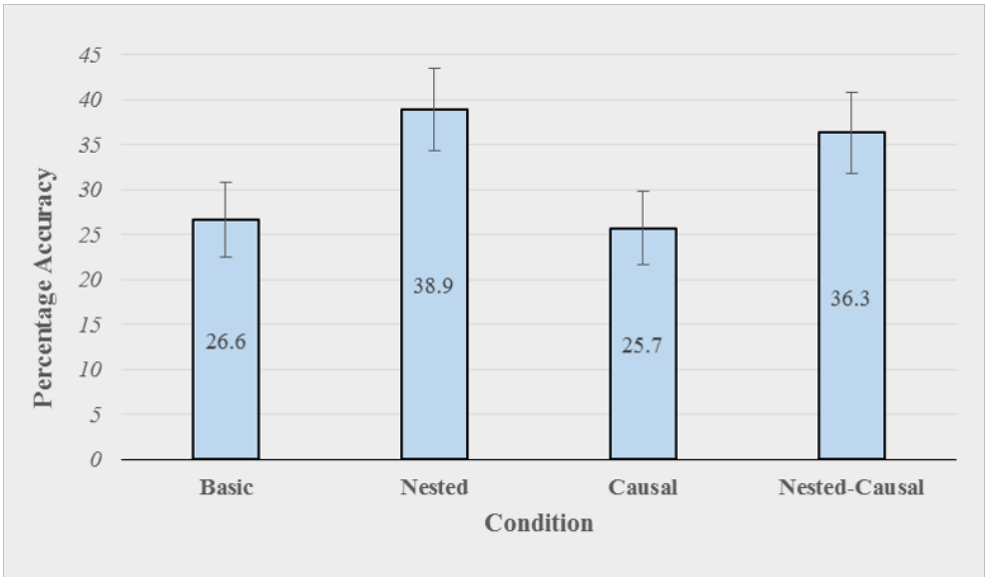


Fig. 3.1: Percentage of participants giving the Bayesian normative answer for basic, nested, causal and nested-causal conditions. Error bars represent one standard error.

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 (χ^2 = 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 (χ^2 = 2.843, p=.098). Further, no interaction between nested and causal was seen (χ^2 =.974, p=.324). Further no main effect was seen in the 'McNair and Feeney' high numerate sub-group (χ^2 =1.567, p=.211) and no interaction was seen here either (χ^2 =.167, p=.682).

The effect of the nested sets framing was also examined within the high (n = 70, M = 5.10, SD = 1.10) and low numerate sub-groups (N = 43, M = 1.95, SD = 0.97). Numeracy was split at the median (4), and these were included in the high numerate sub-group, in line with the analysis for the causal framing. A significant effect of the nested sets framing was still seen within the low numerate sub-group (χ^2 = 5.359,

p=.021) and a borderline significant effect was seen within the high numerate subgroup (χ^2 = 3.159, p=.076). Overall, low numerates were accurate 13.95% of the time, while high numerates were accurate 42.87% of the time, a significant difference (χ^2 = 26.03, p<.001).

To determine whether this lack of a difference for the causal manipulation was due to an issue with the problems created for this study, accuracy across all 16 scenario-condition combinations was analysed. Overall causal and basic conditions were consistently similar across all four problems and there was no clear evidence of a bias in any direction. While there was a difference on the original mammogram problem between causal (37.0%) and basic (25.0%), and this difference was larger than any of the other three problems, this difference was far from significant (2 = 0.93, p = .33). Furthermore, on the same problem, the nested-causal condition (51.6%) in fact produced lower accuracy than the nested condition (57.7%).

3.3.2 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 base rate 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. (H&D) / (-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 unpopulated 'Hypothesis-focused Representation', which can be seen below in Figure 3.2. This representation of the problem should be highly familiar from illustrations given in previous work (e.g. Gigerenzer and Hoffrage, 1995). It begins by subdividing 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.

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 did not suffice for this classification. An example of this representation 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

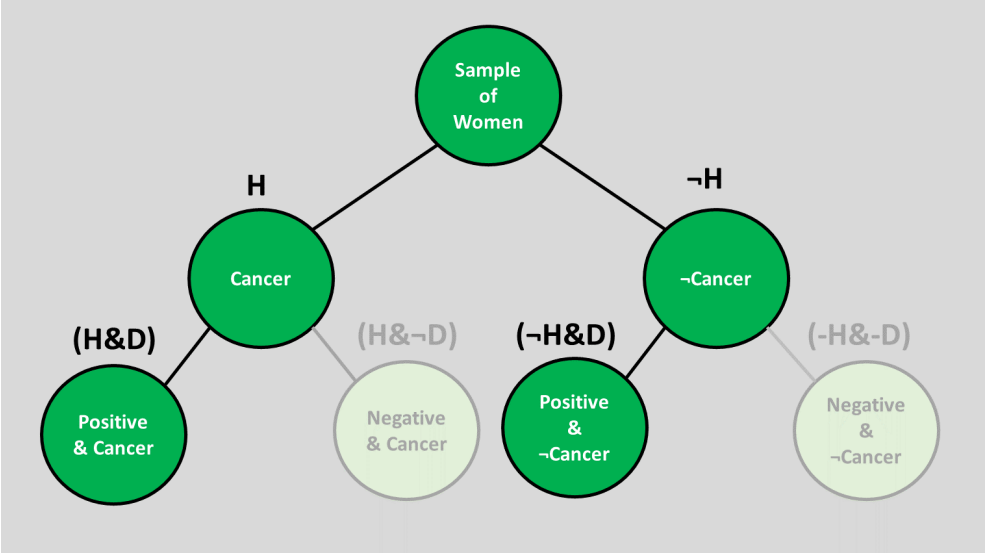


Fig. 3.2: An event-tree depiction of the un-populated Hypothesis-focused Representation.

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' 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 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 compute the bottom 'D' nodes representing the conjunctions H&D and -H&D, in the Hypothesis-focused Representation diagram. H&D was calculated by multiplying H by P(D|H). No single participant calculated the '-D' nodes (H&-D and -H&-D) 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 H&D and -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. This representation, instead of using the hypotheses as the mid-level nodes (e.g. cancer / no cancer), uses the data (e.g. positive / negative 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 neglected as neither were required to solve the problem from this point.

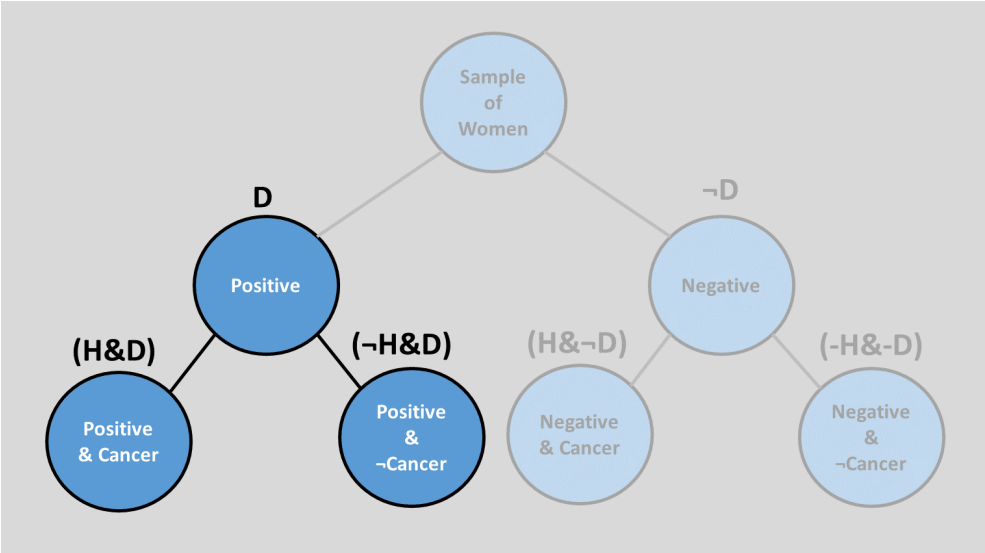


Fig. 3.3: An event-tree depiction of the un-populated Data-focused Representation.

The requirement for the Data-focused Representation classification was a word-based indication that H&D and -H&D are subsets of D (see Figure 3.3). Again, a mathematical formula did not suffice for this classification. Two broad sub-categories of this classification were identified:

Bottom up: i.e. by first defining the subsets H&D and -H&D and then demonstrating that they are in fact subsets of D e.g. P36 who said '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.'

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 [H&D] actually have the cancer.' Typically this sub-category only mentioned H&D as -H&D was no longer needed to solve the problem.

Step 4 / Computation Two: Populating the Data-focused Representation

Either simultaneously with, or immediately following, the laying out of the Data-focused Representation, participants mathematically summed H&D and -H&D to obtain the total D (e.g. total positive results).

Step 5 / Computation 3: 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 H&D by the total number of positive results (D) to compute the normative Bayesian answer, P(H|D).

The Think Aloud Protocol and the Representations

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 Nested Sets Process Model

In line with the majority of recent work (e.g. Johnson & Tubau, 2015), this process model is named the 'Nested Sets Process Model' as both representation of the problem (both 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 of these 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 3.4. Apart from 9 cases, the Nested Sets Process Model was the only approach taken to solving the problem by 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. In regards to those participants who did not follow the process model, this thesis will wait to present an analysis of the causes behind the most common erroneous approaches until experiment 2, where a greater range of problem types, and therefore a richer analysis, is available.

Fig. 3.4: A depiction of the proposed Nested Sets Process Model including the un-populated Hypothesis-focused Representation (R1), the population of this (C1), the un-populated Data-focused Representation (R2), the population of this (C2) and the final computation (C3)

The Relationship Between the Proposed Process Model and Previous Work

The present model clearly draws heavily on 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 the author's knowledge, it is the first time that the entire process has been presented. Further, much previous work presenting either 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 will bring greater clarity to the theoretical underpinnings of the field, and allow it to progress more swiftly.

Nested Sets Framing and the Process Model

As stated above, the nested sets framing taken from Macchi (2000) consists of two changes to the problem: those to the text body and those to the question form. 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, H&D). It would therefore be expected that the nested sets framing format would impact on both of these. 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%: $\chi^2 = 14.115$, $p<.001$). An effect was also seen on frequency of Data-focused Representations (nested sets 27.4% vs non-nested sets 16.4%: $\chi^2 = 7.957$, $p=.005$. 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 ($\chi^2 = 0.193$, p=.661).

Further, an effect of the nested sets framing was seen on computational step one (nested sets 54.8% vs non-nested sets 42.0%: χ^2 =8.625, p=.003), step two (nested sets 35.0% vs non-nested sets 25.2%:$\chi^2 = 5.154$, p=.023) and step three (nested sets 35.4% vs non-nested sets 24.3%:$\chi^2 = 6.593$, 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 ($\chi^2 = .501$, p=.479) or step three ($\chi^2 = 1.270$, 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 was detected on frequency of the Hypothesis-focused Representation ($\chi^2 = 0.367$, p=0.547) or the Data-focused Representation ($\chi^2 = .020$, p=.889). Further, no effect of the causal framing was detected on Computational step one ($\chi^2 = 0.013$, p=.908), step two ($\chi^2 = .069$, p=.793) or step three ($\chi^2 = 0.778$, p=.378). Finally, no single causal representation was detected in the think aloud protocol in any condition.

Drop off Rates and Numeracy

The drop-off rates of the percentage of people giving each successive stage in the process for both the computational steps and the representations, can be found below, in Figures 3.5 and 3.6 , respectively. In 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 both that step as well as all previous steps. Using this method drop off can 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.

In regards to the computation drop-off diagram (Figure 3.5a below), 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 3.5b) shows a similar, but inverse, pattern. Individuals who achieve step one 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 steps 1 and 2, and no further increase is seen either at step three, 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 Figures 3.6a and 3.6b 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.

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 got the answer correct.

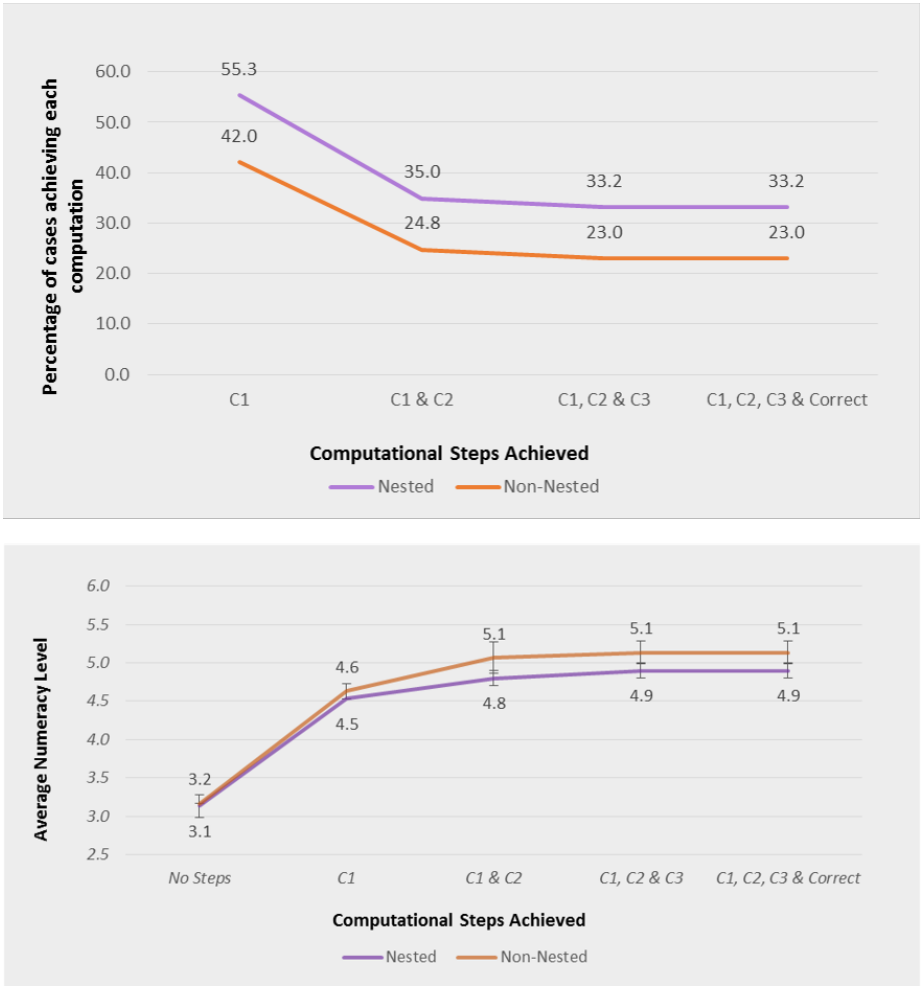


Fig. 3.5: The 'drop off' rates (top) and average numeracy levels (bottom) for the three computational steps. Percentage values are given as a proportion of the number of cases for the nested (n = 226) and non-nested (n = 226) condition

3.4 Discussion

3.4.1 Aims and Hypotheses

The aims of this experiment were three-fold. First, to verify in the general population the work of Macchi (2000) and Krynski and Tenenbaum (2007), which investigated student-population performance on Bayesian problems using nested sets and causal framings, respectively. Second, to examine whether simultaneous presentation of both nested sets and causal framings would produce higher accuracy than either alone. Third, to supplement the experimental paradigm with think aloud and numeracy data to gain a greater understanding of the processes that individuals

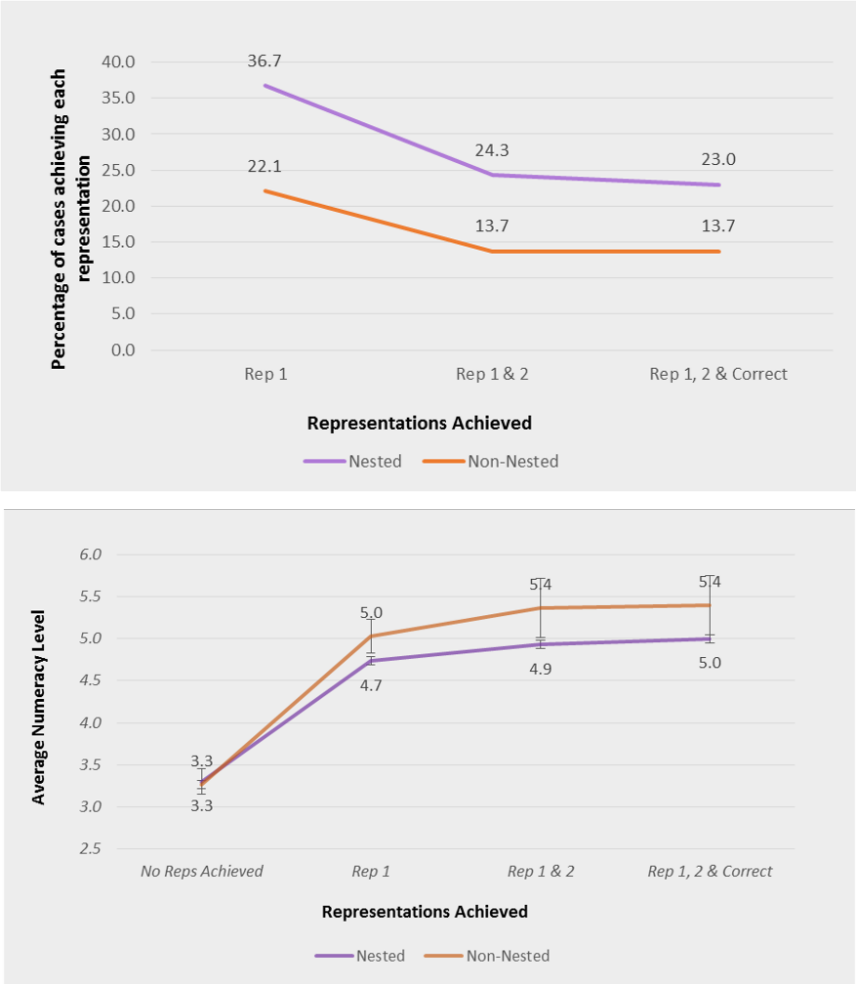


Fig. 3.6: The 'drop off' rates (top) and average numeracy levels (bottom) for the two representational steps. Percentage values are given as a proportion of the number of cases for the nested (n = 226) and non-nested (n = 226) conditions

undertake when solving simple Bayesian problems.

Within the first 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. This result again stands in contrast to Krynski and Tenenbaum (2007) and subsequent replications with the exception of McNair and Feeney (2014a) who also did not find an effect of the causal framing.

Within the third aim, the results provide a more cohesive and systematic model of the processes people use to solve simple Bayesian problems 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 format' and 'causal format'), 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. This stands in contrast 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).

3.4.2 Nested Sets versus 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). This may also have encouraged them to convert the problem into integers (or natural frequencies), and it is possible that without this feature, individuals may work through the process model without converting from percentages to frequencies.

3.4.3 Nested Sets Effect: Text Body change 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 2 and 3, 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 3.6 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 Figure 3.5. 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 step one but does not succeed in improving the frequency of the Data-focused Representation and step two other than indirectly via that increase in the Hypothesis-focused Representation and step one.

3.4.4 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 also removed for the analysis. It is speculated therefore that this may indicate that the second step, and 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 may need greater numerical ability to achieve the middle steps of the process than the first steps. In contrast, the final computational step three, appears to be trivial once one has achieved step two.

3.4.5 Causal Null Finding Explanation

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 sub-group 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 assumes 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' for 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 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, makes this explanation less likely.

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 contained a nested sets prompt to some extent.

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.

4 Experiment Two

4.1 Introduction

In experiment one, Macchi's 'outside-framed percentage' approach to increasing accuracy on Bayesian problems was replicated and further, found to be efficacious in the general population. Second, the nested sets process model outlined in that paper, and hypothesised by previous nested sets authors, 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. However, the problems used in that study were relatively simple 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, it must be tested without these fictitious simplifications. Further, as the process model was found within the data using exploratory analysis in experiment one, the present experiment will serve to verify the existence of the model. It will also serve as a test for whether the model is present in more ecologically valid contexts.

First, the 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, Macchi (2000) and Fiedler et al. (2000) all used integers 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 much previous work (e.g. (Casscells et al., 1978; Eddy, 1982; Kahneman & Tversky, 1982; Krynski & 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 integer 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). It was similarly theorised in experiment one that this use of integers may have 'got participants started' in following the nested sets process particularly as that process begins with the simulation of a target population which is then sub-divided. Without the provision of this, it is therefore again possible that the nested sets framing effect may be reduced and it may also be possible 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 (Brase, 2007; Cosmides & Tooby, 1998; Cosmides & 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. While this fraction-effect may be weaker with the percentage values used in the present experiment, previous work (experiment one;

Hoffrage et al, 2002) has suggested that a large proportion of individuals are likely to convert the problem in to integers, or 'natural frequencies'. If this is the case in the present experiment, it may be possible to simultaneously test whether these fractional values impact on the nested sets framing effect, and the percentage of people following the nested sets process with both percentages and integers.

Finally, the simplicity of problems used in the first experiment (due to the above discussed absence of false negatives) made the analysis of errors of limited value. Due to the potential lack of ecological validity of the problem used, it is quite possible that entirely different errors may be made in more realistic situations. The present experiment, with a greater range of problem formats, will expand the examination of think aloud protocols to unsuccessful individuals in an to attempt to elucidate any themes in underlying cognitive reasoning patterns that lead to errors. This will be examined through think aloud data to build on previous work by Gigerenzer and Hoffrage (1995) on the causes of common errors on Bayesian problems. Gigerenzer and Hoffrage promoted the use of methods which could elucidate solver processes in order to understand the cause of various errors, most notably that type previously named 'base rate neglect'. They noted that the entirety of work at the time had focused only on the numerical outcome of solvers' answers to determine the type of error and stated that "No comprehensive theory of why and when people neglect base rate information has yet been found." (Gigerenzer and Hoffrage, 1995, pp.29) This issue remains to this day. In previous work only one study (Macchi, 2000) has used a 'think aloud' method to elucidate the processes behind errors. While Macchi produced descriptive statistics for the most common error types she did not attempt to examine the processes which lead to these errors. The processes participants undertake when they arrive at these erroneous answers will therefore be examined and analysed for common themes to determine if one or several underlying causes can be determined.

It is 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 decimal and complex conditions. Further, based on the findings of experiment one, a mediation analysis will be conducted with the hypothesis that the hypothesis centred and data centred representations will mediate the relationship between the nested sets framing and accuracy.

4.2 Method

4.2.1 Participants

The final sample for the experiment was 521. From an original sample of 528, 7 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.

4.2.2 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 complex) 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.

4.2.3 Materials

The mammogram 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 complex 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. The nested-decimal-complex 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?

4.2.4 Procedure

Participants were recruited through the Amazon MTurk outsourcing service. 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 Appendix B). 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.

4.2.5 Data Analysis

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

4.3 Results

4.3.1 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 4.1 below, accuracy for all eight conditions can be seen.

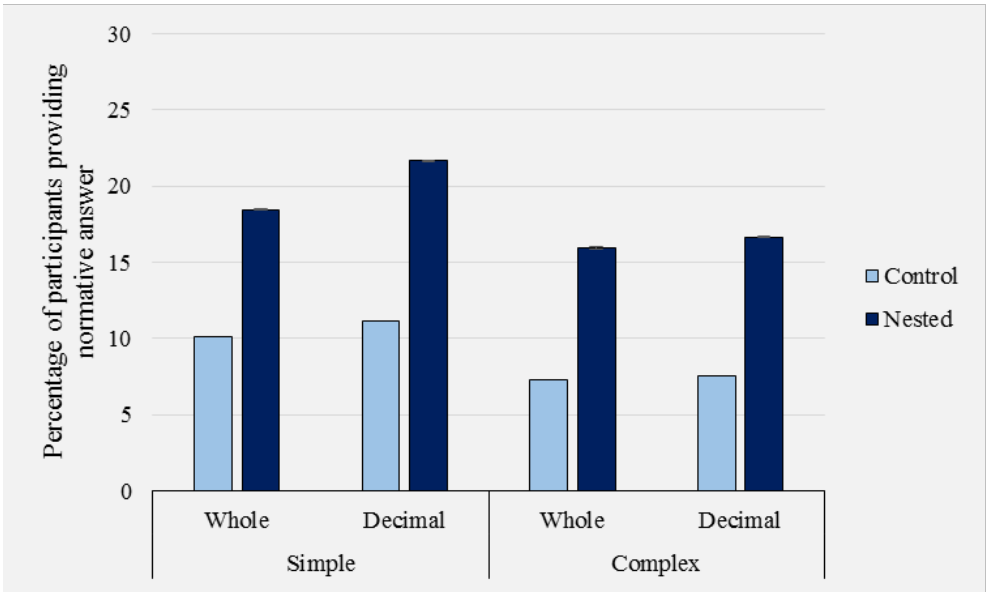


Fig. 4.1: The percentage of correct answers across all eight conditions

A binary logistic regression, using 'score' as the dependent variable and the three condition-comparisons (non-nested vs nested; whole vs decimal; simple vs complex) as independent variables found a highly significant main effect for the

non-nested-nested comparison ($\chi^2 = 8.984$, p=.003), no main effect for the whole-decimal comparison ($\chi^2 = 0.184$, p=.668) and no main effect for the simple-complex comparison ($\chi^2 = 1.350$, p=.245).

To determine if the effect of the nested sets framing was significantly present within the four 'decimal' conditions, a generalised linear model was run on this group only. A main effect of the non-nested-nested comparison was found ($\chi^2 = 4.821$, p=.028). Similarly, to determine if the nested sets effect was present within the four 'complex' conditions, a generalised linear model was run on this group only and a main effect of the non-nested-nested comparison was found here also ($\chi^2 = 4.784$, p=.029).

4.3.2 Qualitative

Process Model

In Figure 4.2 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.

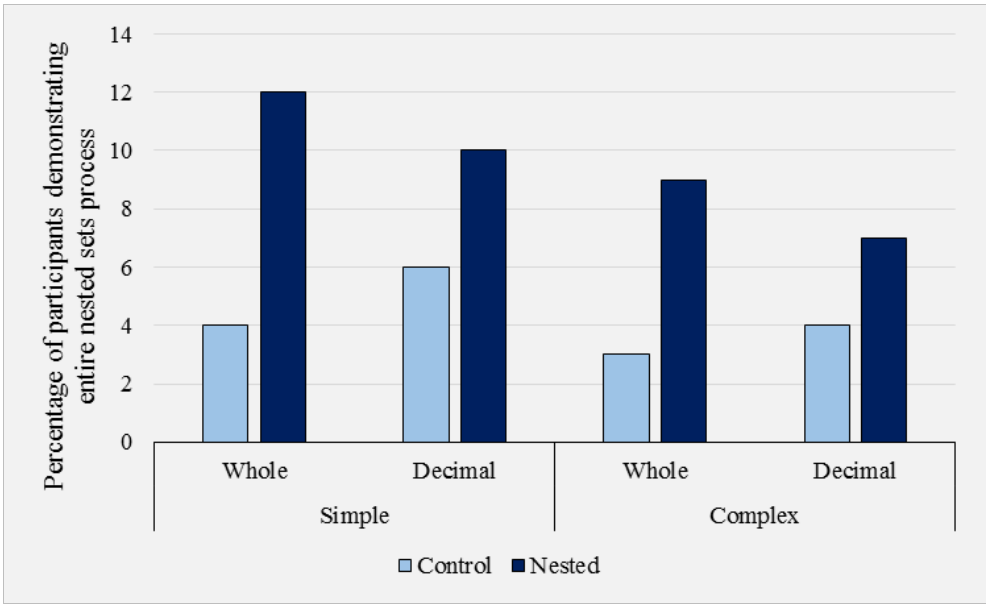


Fig. 4.2: The percentage of individuals achieving all steps of the nested sets process model for all eight conditions

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) and the dependent variable and the nested sets vs non-nested, simple vs complex and whole vs decimal variables as independent variables. A highly significant effect of the nested vs non-nested variable was seen (χ^2 = 9.728, p=.002), while no significant effect was seen for the simple vs complex (p=.161) or for the whole vs decimal condition (p=.816).

In Figure 4.3 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 / Step one, with further substantial but smaller drop off between these and Representation two / Step two, and no substantial subsequent drop off between these and step three or accurate completion of the problem. Further, highly 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.

Confirming the results of Figure 4.3, 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 (χ^2 = 0.04, 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 (χ^2 = 0.019, p=.891).

Conversion to integers

Eighty eight participants (16.9%) converted the problem from percentages into integers before attempting to solve. For this classification, a 'sample' or 'population' of women with a integer 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 into a population of 100 women, and 9 converted to a population of 1000.

Conversion of the problem into integers 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 ($\chi^2 = 980.6$, p<.001).

Conversion of the problem was also associated with a greater frequency of nested sets and data centred representations. Binary logistic regressions showed a significant main effect of conversion on the hypothesis centred representation ($\chi^2 = 93.1$, p<.001) and on the data centred representation ($\chi^2 = 75.4$, p<.001).

A general linear model using logit function examined the prevalence of conversion between conditions. A significant main effect of the nested sets condition was found ($\chi^2 = 7.233$, 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 ($\chi^2 = 0.7$, p=.412) or the simple-complex comparison ($\chi^2 = 0.3$, 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 completely the problem and followed the nested sets process 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

Eight out of the nine individuals who converted the problem to a sample of 1000 were in one of the decimal conditions. Converting to a sample of 1000 turned 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 2 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 centred representations were found in the decimal conditions as the whole-number conditions. These results suggest that apart from 3 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

In order to ensure that the results of the analysis of errors is valuable for future research, the analysis will look only at the most ecologically valid conditions: the

complex-decimal conditions. Given that the present work has shown accuracy is equal in these conditions, future research should be looking to employ these more complex problems , rather than the more simple problems often used in order to ensure any improvements can be implemented in real situations. Brief comparisons to the overall rates will also be given.

Non-Nested Sets The most common answer within the non-nested complex-decimal condition was one which did not achieve either mathematical step. This 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 in the non-nested conditions overall.

The think aloud data was coded by the candidate and second coded by independent researchers with no knowledge of study hypotheses 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%.'. In each of these cases, the first value given (15% or 10%) is P(D|-H) in the problem text. However, the value is being expressed in the place of P(-H|D). This faulty leap in reasoning can perhaps be most clearly seen 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 just gave 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-complex condition, making it the second most common answer. The two most common answers in this condition were the correct answer, and the conjunction (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 (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. The error is to provide this as the final product instead of dividing it by the sum of itself, (H&D) with the opposite conjunction (-H&D), which provides the correct answer.

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 suggests a similar misunderstanding of the question 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 integers and secondly

by the nested sets and data centred 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 (χ^2 = 4.1, p=.042) and a large significant effect of conversion was found (χ^2 = 76.4, p<.001).

In the second model, the nested condition variable and the hypothesis centred and data centred representations were included as independent variables. In this model, the nested sets condition variable was a non-significant predictor of accuracy (χ^2 = 0.0, p=.933) while both hypothesis centred (χ^2 = 32.6, p=<.001) and data centred (χ^2 = 19.9, p=<.001) representations were large significant predictors.

In the third model, all four factors were included as independent variables. In this model, the nested condition variable was not a significant predictor (χ^2 = 0.0, p=.922) and the conversion variable was much reduced in predictive power (χ^2 = 3.1, p=.079). The hypothesis centred representation (χ^2 = 28.1, p=<.001) and the data centred representation (χ^2 = 16.9, p=<.001) representations remained highly significant predictors and were little-reduced in predictive power.

4.4 Discussion

4.4.1 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 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-complex comparison was found. The nested sets framing effect was also found separately within the four 'decimal' conditions and within the four 'complex' 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-complex comparison was observed. This confirms the first hypothesis.

4.4.2 Mediation Analysis

A mediation analysis compiling all conditions found, in two separate analyses, found partial mediation of the nested sets effect on accuracy by conversion from percentages to integers, and full mediation of the nested sets effect on accuracy by the hypothesis centred and data centred representations of the process model. This latter finding confirms the second hypotheses of the study. Further, a mediation analysis including all four variables showed a mediation of the effect of conversion on accuracy by the hypothesis and data centred representations. These results suggest the following solution narrative: Individuals in the nested sets conditions are more likely than those in the non-nested sets conditions to convert the problem from percentages to integers. The individuals who convert are more likely to follow the nested sets process, and the individuals who follow the process are more likely to succeed on the problem.

4.4.3 Nested Sets versus Natural Frequencies

The present results suggest that a nested sets format increases accuracy on Bayesian problems even with percentage base rates. In regards to whether this increase is equal in power, a direct comparison cannot be made as percentage and integer base rates were not directly compared in the present experiment. However, some comparisons may be 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 exper-

iment one, however, accuracy in the complex-whole-nested condition (15.9%) was also lower than the comparable condition in Macchi (2000), who found 33.3% and used a between subjects design. Given that these two studies used 'integer' base rates instead of percentages this may suggest that the provision of these improves accuracy on Bayesian problems above the nested sets format alone.

Previously, Hoffrage et al (2002) hypothesized that the presence of 'integer' base rates / population figures in previous problems (Macchi, 2000; Fiedler et al., 2000) may have been a factor in encouraging individuals to create 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 results demonstrate that the majority of successful participants (60%) in this experiment, when presented with percentage base rates, firstly constructed for themselves a 'integer' version of the problem. This phenomena was also noted by Cosmides and Tooby (1996), who briefly examined participants' workings out but did not conduct a systematic analysis. This conversion process was also highly associated with success. This may provide valuable insight into solution process given that the 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. Further, results also showed that the relationship between conversion and success was partially mediated by the hypothesis and Data-focused Representations. It is therefore possible that this conversion made it more likely that participants would follow the nested sets process, ultimately leading to success on the problem.

However, it must not be ignored that a substantial portion of 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 'integer' population, or the use of integers at all. Therefore, in temperance of Hoffrage et al.'s (2002) conjecture, the construction of a natural frequency format may not in fact be necessary for all

participants to solve Bayesian problems, or even to follow the nested sets process, but instead may simply increase the likelihood that they will do so.

4.4.4 Decimal Values

A further finding of note was that overall, individuals dealt with decimal values in exactly the same way as whole numbers. Even within those individuals who converted the problem into integers no difference between decimal and whole-number conditions was seen. This was considered plausible as there may be a psychological difference between '12.5% of women' and '12.5 women', with the latter being a metaphysical impossibility. Indeed 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.

4.4.5 Errors

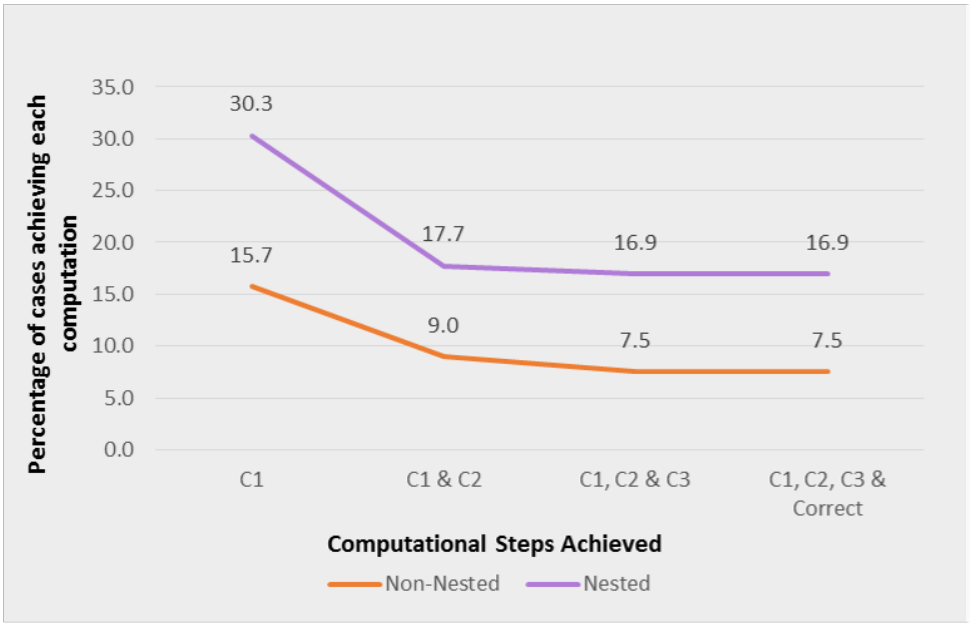
1-P(D|-H) Error

The analysis of errors also produced information which may be valuable to future work. The most common error in the non-nested-decimal-complex condition when presenting with an individual or 'chance' framing was 1-P(D|-H). This was also the second most common error in the complex-decimal-nested 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 given by 2.9% of participants in experiment one and was also found by Macchi (2000). 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 with fits with previous work advocating the 'confusion hypothesis (e.g. Hamm, 1987; Hamm & Miller, 1988; Wolfe, 1995; Macchi, 1995).

(H&D) Error

The most common error within the complex-decimal-nested sets condition was to provide the conjunction (H&D). This answer was also the second most common error in Gigerenzer and Hoffrage's (1995) first experiment, 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 reasoning error behind this mistake. 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 with a positive result who actually have breast cancer'.



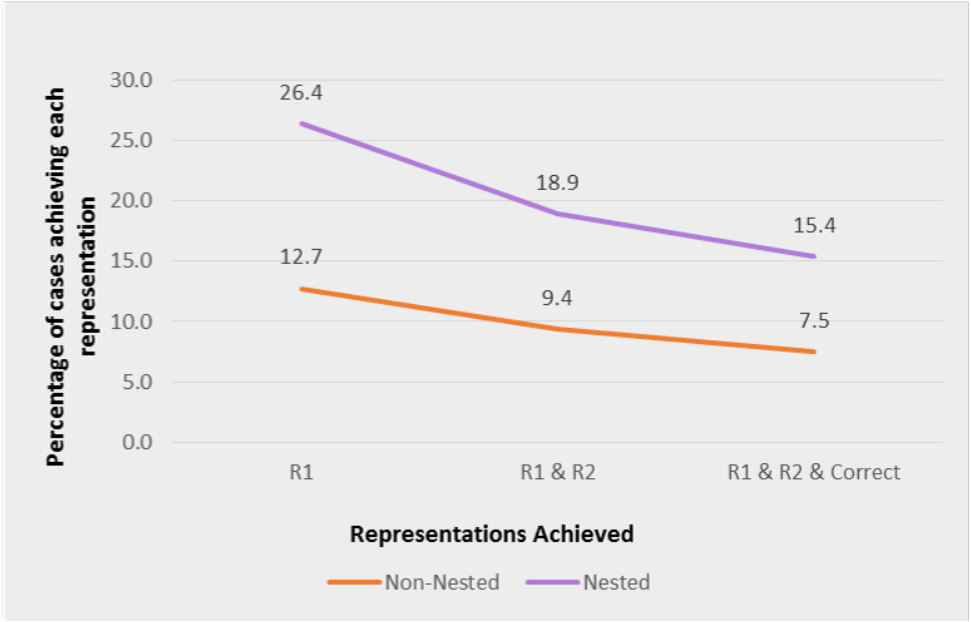


Fig. 4.3: Drop off graphs for each computational and representational step

5 Experiment Three

5.1 Introduction

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 which were discussed within that experiment. 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 potentially confounding factors. Four conditions were analysed. Each had percentage base rates similar to those used by Krynski and Tenenbaum (2007), which was an 'outside percentage' in that experiment, meaning that their original experiment in fact used a mixed outside / inside percentage format. Further, to test the theory that a think aloud protocol may have affected the result (as theorised in experiment one) one causal and one control condition used a think aloud protocol, while one causal and one control condition did not. Based on experiment one it was hypothesised that in a logistic regression analysis, a think aloud effect would be seen but no causal effect would be. It was however 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.

5.2 Method

5.2.1 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.

5.2.2 Design

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

5.2.3 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, and the causal version 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. 20% of women at age forty who participate in this routine screening have breast cancer. If a woman has breast cancer, she will always get a positive mammography result. If a woman does not have breast cancer, there is still a 10% chance that she will also get a positive mammography result. This can happen if she has a dense but harmless cyst which causes a positive result because it looks like a cancer to the X-ray machine.
>
> A woman in this age group had a positive mammography in routine screening. What is the percentage chance that she actually has breast cancer?

5.2.4 Procedure

Participants were shown the consent form for the experiment, 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 Appendix B 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.

5.2.5 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 experiment one and experiment 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.

5.3 Results

5.3.1 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 5.1.

It is clear that the think aloud conditions produced substantially more numerically correct answers than the non-think aloud conditions. However the causal

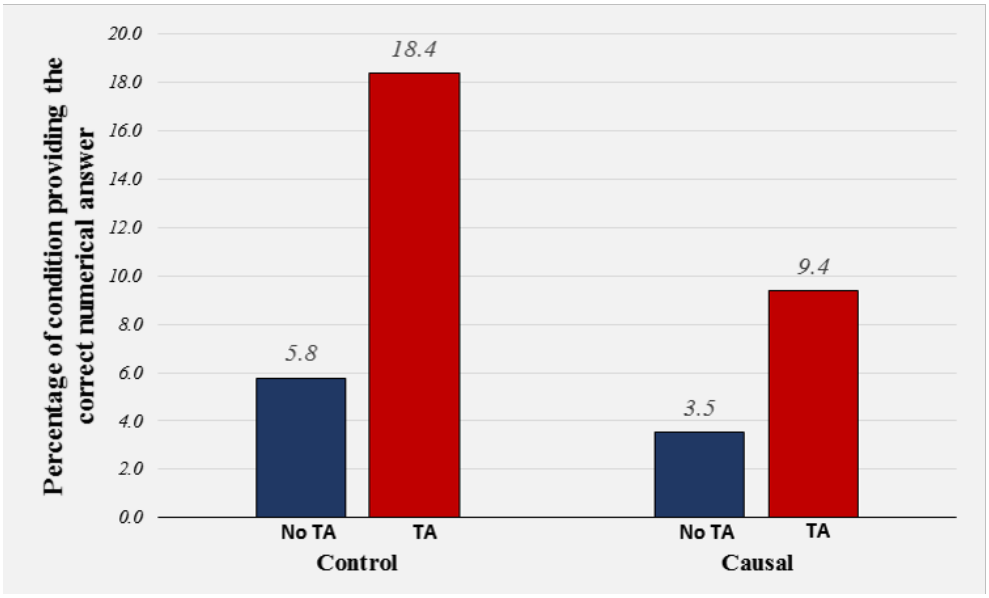


Fig. 5.1: Percentage of participants providing the correct numerical response across all four conditions in Experiment three

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 ($\chi^2 = 9.065$, p=.003), but no significant effect for the causal variable ($\chi^2 = 2.774$, p =.096) and no interaction effect ($\chi^2 = .114$, 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 9 out of 9 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).

5.3.2 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 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 got the correct method, 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.'.

5.4 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 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 integer base rates or the use of a think aloud protocol. The results therefore give further evidence against Krynski and Tenenbaum's 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.

6 Experiment Four

6.1 Introduction

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 'integer' 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 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 and 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 two and three found that a large proportion of participants who were not provided with integer population figures constructed such a figure for themselves in the early stages of solution of the problem. This 'conversion' of the problem into integers 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 compare the provision of a real-number population figure to no figure to determine if solver accuracy is affected.

Previous work (Experiment one; Experiment two; Experiment three) has also discovered 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. Cosmides and Tooby (1996), in one experiment in their paper, provided participants with leading questions which, in the language of the present studies, encouraged individuals to complete what have here been named 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. Their analysis however included a very low sample size and so is very likely to have lacked the power to detect an effect of this size or smaller. The second aim of the present experiment is therefore two 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 (step one and step two) would increase accuracy alone and that both questions combined would increase accuracy further.

6.2 Method

6.2.1 Participants

From an original sample of 419 participants, 15 were removed because they stated they had undertaken the same problem in the past. 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 364. Participant demographics can be found

in Table 1.

6.2.2 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: no-population-no-steps; no-population-step-1; no-population-step-2; no-population-both-steps; population-no-steps; population-step-1;population-step-2;population-both-steps. 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?
>
> 2. What percentage of women have no cancer but still received a positive result?
>
> 3. What percentage of women receive a positive result in total?
>
> 4. What percentage of women at age forty who get a positive mammography in routine screening actually have breast cancer?

Participants in the no-leading questions conditions were only presented with question 4; those in the step one conditions were presented with questions 1, 2 and 4; those in the step two conditions were presented with question 3 and 4. 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.

6.2.3 Procedure

Participants were recruited through the Amazon MTurk outsourcing service. 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.

6.2.4 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.

6.3 Results

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

In Figure 6.1 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, and is also included for the condition from experiment two which used an identical scenario to condition one (it did not include any leading questions) but included a think aloud protocol. 16.7% accuracy was seen in that previous work, while 5.1% accuracy was seen in condition one in the present study, which is a significant difference (= 5.900, p=.015). In further comparison to condition one, condition two, which included H&D and H&D questions produced 7.6% accuracy (= .530, p=.467), while condition three produced 6.5% accuracy (= .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 (= .4114, p=.043). A further analysis of the step one and step two question separately across all four conditions demonstrated a trend towards significance for the step one questions (= 2.774, p=.096) while the step two questions did not show a significant result (= 1.615, p=.204).

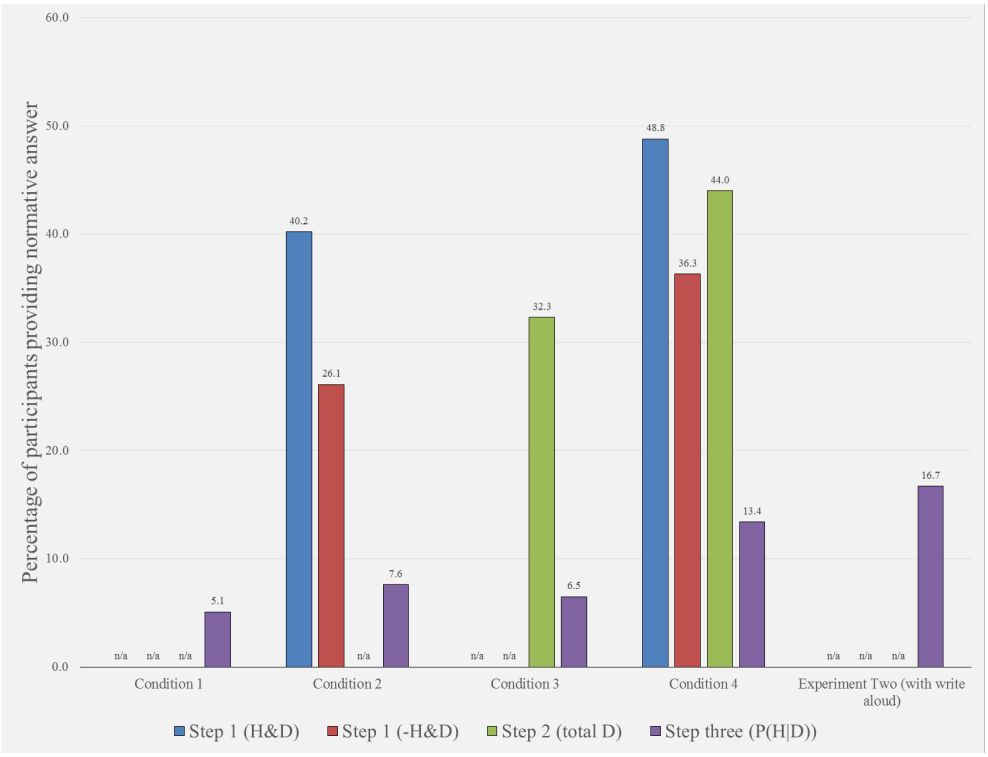


Fig. 6.1: Percentage accuracy for all questions asked for all four conditions in the present study plus equivalent experiment two condition

It is also immediately clear from Figure 6.1 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 (= 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 (= 7.657, p=.006).

Similarly for the step one H&D and 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 (= 15.9, p<.001) and condition four (χ^2 = 11.1, p=.001).

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

Further, in condition four, when H&D, -H&D and step two accuracy were all included as independent factors in a model predicting step three accuracy, H&D did not act as a significant predictor (χ^2 = .870, p=.351), -H&D did not act a significant

predictor (χ^2 = .172, p=.678), while step two did act as a significant predictor (χ^2 = 4.837, p=.028).

6.4 Discussion

6.4.1 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 between the two 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 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).

6.4.2 Population Prompt

The present work found no relationship between the provision of a '100' population sample and accuracy on the problem. This difference was firstly hypothesised because previous work (Hoffrage et al, 2002) had theorised that the provision of integer populations and base rates was behind a large proportion of the success of the nested sets approach. Further however, experiment two found a strong relationship between the construction of a population sample in their 'think aloud' data and success on the problem. Based principally on these correlational results and a mediation analysis the causal theory was proposed in which individuals who convert the problem in to integers are more likely to be able to construct the hypothesis and Data-focused Representations which are themselves highly predictive of success.

6.4.3 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 questions can be an efficacious way to increase accuracy on Bayesian problems. 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 again appears in contradiction to the correlational findings of experiment two which found that the successful completion of step one (Questions one and two) and step two near-guaranteed success on the P(H|D) question. A notable major difference between these two works is that the result from experiment two is correlational while the present was experimental. This may shed further light on the nature of the process being undertaken by participants. It has been theorised above, in regards to the finding that the population figure does not facilitate performance, that participants may therefore have a broad conception of the two representations of the problem before attempting a mathematical solution to the problem. The present result may suggest that the same is true for the computational steps also, and by extension, the entire process. In experiment two successful participants may have had a rough or conceptual idea of the entire process they would go through to solve the problem before making any actual calculation. Therefore participants who completed the computational steps or the two representations in experiment two did so in general because they were likely to have done so because they were working towards a pre-conceived successful solution

procedure. However in the present study, participants were forced to answer the computational step questions regardless of whether they had the correct solution procedure in mind or not. Therefore some participants who did not know how to solve the problem would still have been able to solve the earlier computational questions (which were easier, according to the results of the present experiment). This suggests that the conjecture in experiment two that the successful construction of the Hypothesis-focused Representation of the problem is not the near-guarantor of success that it was claimed to be. In fact encouraging individuals to construct this representation (by forcing them to make the step one calculations) did not increase accuracy significantly at all. One issue with this conclusion is that participants were not provided an overview of the steps they needed to undertake to solve the problem. 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, and may not have seen the connection at all. Given the previous conclusion that participants appear to be 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 in combination with Data-focused Representation diagrams (which no previous work

has attempted).

7 Part I Discussion

7.1 The Nested Sets Approach

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

7.2 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 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. This conjecture was given further tentative evidence by the results of experiment four, which showed higher accuracy on earlier stage leading questions (e.g. step one) compared to later stage (e.g. step three).

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 2 and 3 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 is having 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, which themselves increase the likelihood of individuals achieving the later steps. This possibility is given further evidence by the result of experiment two demonstrating that in the nested sets conditions, the most common error appeared to be, according to the think aloud data, due to a confusion in regards to the question being asked.

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 non-significant 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, however, 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 first calculate D separately (i.e. the total number of positive results) and only then to calculation P(H|D).

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 for Macchi's outside-framed approach 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, 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 at the correct time in the solution process in order to convert a greater percentage of those individuals successfully completing the Hypothesis-focused Representation into individuals providing the full normative Bayesian answer.

7.3 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 3 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 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 was 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 because they encourage individuals to follow the same process which successful individuals independently adopt even in the absence of such approaches.

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 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 an overall (but far from total) preference for integers 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).

In opposition to this view, 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 integers) 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 more able to solve the problem by converting to integers and some participants able to solve with percentages. However, high numeracy levels and familiarity with percentages are proposed as plausible factors for future work to consider.

7.4 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. step 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. 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.

7.5 Nested Sets versus Natural Frequencies

While a direct contrast was not possible, a comparison between experiment one and Macchi (2000), which both used integer base rates, and experiment two, which used percentage, suggests there may be a further increase in accuracy when integers are used. Further evidence for this comes from the fact that 60% of successful individuals in experiment two, and 50% in experiment three, converted the problem from percentage form to integer form. This 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 integers 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 beneficial for everyone (Cosmides and Tooby, 1996). However, it is not as obviously 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, it should be noted that it is difficult to distinguish between participants' preference for integers due to a greater familiarity with them, and the frequentist proposal that our evolutionary history has designed our brain to deal with them more effectively.

This finding fits well with some 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. integers) were more successful than those who interpreted the problem as probabilities. The present work suggests that these frequency-interpreting individuals may have constructed a 'integer' sample for themselves, and that this may have led to their greater increase in accuracy in that study by encouraging them to follow the nested sets process.

In temperance of this view however, an experimental attempt to increase accuracy by providing participants with a integer population rate failed in experiment four. This finding casts doubt on the causal account developed above in which conversion of the problem leads to participants following the nested sets process and ultimately, to the normative answer. Instead, the results of experiment four tentatively suggest the possibility that the direction of the causal relationship may be the opposite: it may be that those participants who perceive the hypothesis and Data-focused Representations are subsequently more likely to construct a integer sample. Since conversion of the problem tends to come 'before' the presentation of either of these representations in the think aloud data, for this theory to be true, we must make the further assumption that successful participants, when faced with the problem, firstly construct a rough mental solution procedure (in terms of the steps of the nested sets process) and may then use their preferred unit to work through the problem (with the majority preferring integers).

This theory, if accurate, would suggest that the use of integers is not causally linked to success on the problem at all, in contrast to Hoffrage et al's (2002) conjecture, and much previous theorising by frequentist theorists (e.g. Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Brase, 2007). It suggests that correct application of the nested sets process are the main determinants of success and that the construction of a integer population merely helps some individuals (60% in experiment two) work through the subsequent mathematics.

One final assumption of this theory required to account for the relationship be-

tween conversion to integers and success in experiment two is that participants who approach the problem in erroneous ways are less likely to convert to integers to solve the problem. This may be a reasonable assumption as the vast majority of incorrect answers (such as the two most common P(D|H) and 1-(P[D|-H) errors) are mathematically simpler than the correct answer and so may be more achievable with out conversion, if the function of conversion, as theorised above, is to allow individuals to work through difficult mathematical steps. This assumption is given some evidence by the fact that integer conversion was not significantly related to success on the problem after controlling for the hypothesis and Data-focused Representations in experiment two.

7.6 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 integer 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 the author does 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, (from which Krynski and Tenen-

baum'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 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 the variables increase and the 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.

7.7 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 who 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 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 is not possible to directly test whether the think aloud protocol changes the nature of the cognitive processes individuals follow, as Ericsson and Simon (1998) have maintained, it does appear to change outcome in this case.

The think aloud protocol requires individuals to work through the problem before providing their numerical answer. The fact that this process increases accuracy so substantially therefore may suggest that some part of inaccuracy on these problems may be due to a lack of sufficient engagement on the part of participants in non-write-aloud experiments. This fits with some findings from experiment two who 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 think aloud protocol may provide the opportunity to avoid such confusion. It 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 the problem showed greater predictive power for accuracy on Bayesian problems than any other measure including cognitive ability. This analysis leads to two conclusions. The first is that understanding of simple Bayesian problems in real situations, such as where a patient is attempting to understand their chance of having a disease after routine screening, is likely to be higher than the rates found in experiments such as the present due to the presumed greater engagement such a patient would have for understanding their risk. The second is that overall engagement is a promising target for future interventions to increase Bayesian accuracy.

7.8 The Confusion Hypothesis and Base Rate Neglect

In experiment two, think aloud data revealed the 1-P(D|-H) error as the most frequently observed among incorrect answers in the inside-frame conditions. This answer, along with P(D|H), has been the most frequent answer labelled 'base rate neglect' by previous work, as it only utilises the 'new data' and does not incorporate the 'prior' or base-rate in its calculation. (Tversky & Kahneman, 1982; Casscells et al, 1978; Bar-Hillel, 1980; Cosmides & Tooby, 1996; Evans et al, 2000; Barbey & Sloman, 2007). Barbey and Sloman (2007) theorized that this type of error was caused by an over-reliance on 'associative' processing systems (e.g. Sloman et al, 2003) as opposed to 'rule-based' processing systems. However, the analysis of the most prominent identified reason for this error does not fit well with this view. The predominant identified error in reasoning in the process was a misunderstanding of the meaning of the false positive rate. Participants committing this error took this rate to mean 'the total number of positive results which were in fact false', a 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). Importantly, the participants' reasoning following this misunderstanding was in fact sound: 100% minus this rate, as interpreted and in a problem with only two possible causes of positive results, would indeed give the percentage of positive results which were true, which is the answer to the question. There is no evidence therefore that this approach to the problem was less 'rule-based' (Barbey & Sloman, 2007) than the correct answer. Allowing for the single misunderstanding of the false positive rate, the reasoning process was sound. This also does not fit well with Bar-Hillel's (1980) dominance view of base rate neglect. Bar-Hillel theorised that the less 'relevant' piece of data in a Bayesian problem was 'dominated' by the more relevant piece, and was 'discarded' entirely. Conversely this finding suggests that the 'base rate neglect' answer is ar-

rived at through an erroneous reasoning process, rather than a simple discarding of one piece of data. It further suggests that that title is not an appropriate description of the error, as others have previously noted (e.g. Hamm, 1987; Hamm & Miller, 1988; Wolfe, 1995).

The finding fits more closely with previous authors who have theorized that semantic misunderstanding of problem texts lies behind many errors on Bayesian 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 finding confirms 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 (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 be searching for '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 perspective again fits closely with previous authors who have theorized that

semantic misunderstanding lies behind many errors on Bayesian word problems (e.g. Gigerenzer, 1996; Macchi, 1995; Macchi & Mosconi, 1998; Macchi, 2000; Welsh and Navarro, 2012). 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.

7.9 Conclusion, Impact and Future Work

The present section 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 integer base rates and with whole number and decimal values. It 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, including in medical contexts.

The present section has also demonstrated that Macchi's (2000) outside-framed approach could be improved further. Analysis over two 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 question form, or attempting to improving 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 section 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 and comprised five stages. An intervention experiment encouraging individuals to follow these steps prior to giving their final answer also found an increase in 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 in some entirely different manner. Further, these interventions should be focused principally on encouraging the later steps in this process, as these appear to be more difficult and require greater numerical ability, to achieve.

In the longstanding debate over the relative distinctiveness of the nested sets and natural frequency approaches to increasing accuracy on Bayesian problems, the present section also contributes valuable findings and theoretical developments. In both the nested sets and natural frequency literature, the same underlying process has been hinted at, which has here been formally outlined for the first time as the 'nested sets process'. It has been named this, rather than the 'natural frequency process' because evidence from experiment two and three, while indeed suggesting a slim majority preference for integers (60% in experiment two, 50% in experiment three), has shown that many individuals are fully capable of undertaking the same basic process with non-integers. 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 integers, or even integers at all. This suggests that while there may be some preference for particular units of analysis, this is considerably 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, experiment two has contributed several valuable findings. It was determined through analysis of the think aloud data that the two of the most common causes of error were due to misunderstanding of, firstly, the meaning of the false positive rate and secondly, the figure that the questions was requesting. This finding 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). In addition, the present section has found that the mere addition of a think aloud protocol can increase accuracy considerably. This process forces individually to engage with the problem before they can provide a numerical answer. This finding may therefore suggest that engagement is a major factor in inaccuracy on Bayesian problems, and that this may be a sensible target for accuracy improvement interventions. 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.

Part II

BAYESIAN REASONING IN LEGAL CASES

8 Literature Review

8.1 Bayes and Law

With the unceasing rise in technological sophistication available to modern forensic science, the frequency with which jurors, lawyers and judges in legal trials are encountering evidence presented in a probabilistic format is rapidly increasing (Walsh et al., 2004; Saks and Koehler, 2005). Unfortunately, this has not been met with the necessary countervailing increase in education in how to incorporate this type of evidence into legal assessments (Gigerenzer and Edwards, 2003). This discrepancy has already lead to disastrous outcomes, with incorrect presentation assessment of probabilistic evidence leading to many miscarriages of justice (e.g. Forrest, 2003; Mehlum, 2009; Donnelly, 2005).

Typically, when forensic evidence of a match between a defendant and some aspect of the crime scene (such as a DNA sample, or a foot or finger print) is presented in court it will be in the form of a statement of the 'random match probability' (RMP) of the match (Koehler et al., 1995). This is the probability of picking a random member of the population and finding that they match the sample, typically presented either as a percentage / probability (e.g. .0000005%) or as a frequency (e.g. 1 in 200 million). It can also be viewed as the conditional probability of the match occurring if the defendant was not actually the source of the sample (often denoted as P[M|-S]).

While the RMP is most often provided in a match scenario, the necessary figure for a jury member to assess the impact of the evidence is the probability that the defendant is the source of the sample. Calculating this figure from the RMP is

another instance of a situation often classed as a 'Bayesian' problem (Bayes and Price, 1763). Taking the legal example above, with just the single piece of 'match' evidence, Bayes' rule states that the probability of the defendant being the source of a sample found at a crime scene given that a match has been found between them and that sample is:

$$P(S|M) = \frac{P(M|S) * P(S)}{P(M|S) * P(S) + P(M|\neg S) * P(\neg S)}$$

Fig. 8.1: Bayes rule (Bayes and Price, 1763), developed by Laplace and Simon (1951)

There are several key terms to elucidate upon here (the following makes the simplistic assumption that the testing of sources and suspects by the forensic teams is error-free). P(S|M), also known as the 'posterior' is the probability that the defendant is the source of the tested sample (e.g. the hair fibre / fingerprint / footprint etc.) P(S), also known as the 'prior' is the probability that the defendant is the source without considering, or 'before' the match evidence. While this is in practise often highly difficult to calculate, in theory, given that we presume the innocence of the defendant before any evidence is presented against them, they are considered in the eyes of law equally likely to have committed the crime as all other people who could feasibly have committed the crime (sometimes considered to be the number (X) of people in a given area such as a city), or 1/X. Assuming only two hypotheses, P(-S), the probability that the defendant is not the source of the sample, is the negative of this figure or 1-(1/X). The probability of the match assumingthat the defendant is the source (P[M|S]) is equal to 1. The probability of the match if the defendant is not the source (P[M|-S]) is equal to the RMP, discussed above.

8.2 Cognitive Science and Legal Literatures

Unfortunately, outside of the legal realm and within the cognitive science literature, a large body of experimental research accrued over the previous 40 years has

demonstrated that non-statisticians, when faced with even the most simple Bayesian problems, typically make large errors in their estimates (Phillips and Edwards, 1966; Kahneman and Tversky, 1972; Lyon and Slovic, 1976; Ajzen, 1977; Casscells et al., 1978; Bar-Hillel, 1980; Eddy, 1982; Gigerenzer and Hoffrage, 1995; Evans et al., 2000; Macchi, 2000; Fiedler et al., 2000; Girotto and Gonzalez, 2001; Sloman et al., 2003; Krynski and Tenenbaum, 2007; Hill and Brase, 2012; Welsh and Navarro, 2012; Johnson and Tubau, 2013; McNair and Feeney, 2014a). One of the most common errors routinely reported throughout this literature, in the language of the legal problem above, is to substitute the probability of a match if the defendant was not the source (P[M|-S]) with the probability of not being the source, given that a match has occurred (P[-S|M]). This is known as the transposed conditional fallacy within that literature, but this same error in reasoning has become notorious within the legal literature under another name, the 'Prosecutor's Fallacy' (Thompson and Schumann, 1987). This error has been instrumental in many miscarriages of justice, including the infamous case of Sally Clark (see Forrest, 2003), and has also been widely documented within the literature studying legal reasoning of match evidence, which has developed in the previous 20 years following the rise of the routine use of DNA evidence in legal trials and the statistical presentations which it has brought with it (Koehler et al., 1995).

Within the cognitive science literature, developments have been made in terms of the best method for presenting Bayesian problems to elicit accurate reasoning and reducing errors such as the prosecutor's fallacy. Accuracy has been increased from as low as 18% (Casscells et al., 1978) to around 40% with the natural frequency (Gigerenzer & Hoffrage, 1995) and nested sets (e.g. Macchi, 2000) approaches. In the first part of this thesis, evidence was provided over a series of experiments that this nested sets approach using Macchi's outside-framed percentage format could improve accuracy on Bayesian problems, and the underlying reasons for this change were elucidated. Such an approach to presenting statistics in court has never been tested within the legal literature and so may appear to be a fruitful

avenue for reducing instances of the prosecutor's fallacy in court rooms. However, unfortunately the inherent structure of a typical legal situation does not lend itself to this approach. This is because Macchi's approach relies on changing the perspective of the problem from a single individual (e.g. a medical patient) to a reference group of highly similar individuals (e.g. other medical patients of the same age, etc) and thus harnessing the problem solver's greater capacity for thinking in terms of groups and sets. The same process in law (the use of a reference class to infer the probability that a defendant committed a crime) has long been considered prejudicial and unacceptable and therefore cannot be considered a fruitful avenue for change. For this situation therefore, new methods will need to be devised.

While the cognitive science and legal literatures studying Bayesian reasoning have coexisted for several decades, there has been little overlap in authorship. Perhaps partly due to this lack of overlap, there are substantial differences in the experimental paradigms employed. One important difference is that the cognitive science studies cited above have typically presented participants with Bayesian problems which are 'fully quantified'. By this it is meant that every variable in the Bayesian formula above is presented to the participant in numerical format, or is derivable from other numerical statistics presented. Within the legal literature however, every experiment to date has presented problem solvers with a mixture of both quantified and non-quantified evidence (Thompson and Schumann, 1987; Faigman and Baglioni, 1988; Koehler et al., 1995; Smith, 1996; Taroni and Aitken, 1998; Schklar and Diamond, 1999; Nance and Morris, 2002, 2005; Kaye et al., 2007; Pozzulo et al., 2009; Dartnall and Goodman-Delahunty, 2006; Mnookin et al., 2011; Thompson et al., 2013). There are advantages and disadvantages to both approaches. The fully-quantified method has frequently been criticised as lacking ecological validity (e.g. Welsh and Navarro, 2012), as Bayesian problems outside the laboratory are rarely fully quantified. There are few better examples of this than legal cases, where jurors are frequently asked to synthesise quantified evidence (such as a DNA match with an RMP) with non-quantified evidence (such as an alibi, eye-witness testimony,

etc.). However, the fully-quantified method employed in cognitive science does allow the researcher, ideally, to assign a 'normative standard' to the problem, according to Bayes' rule, and compare solvers' answers to this standard to determine accuracy with greater precision. The legal literature on the other hand, has typically only been able to demonstrate whether changes in presentation format increase or decrease estimates of guilt / conviction rates relative to each other, with no ability to compare to a normative standard. Furthermore, the simplicity of the fully-quantified format lends itself to a methodology known within that literature as a 'think aloud' protocol (e.g. Ericsson and Simon, 1980; Gigerenzer and Hoffrage, 1995; Macchi, 2000). With this methodology the problem solver is requested to record their thought processes during solution. It has not to date been employed within the legal literature, which may be because the number and complexity of interpretations of mixed-evidence formats is much higher than fully-quantified formats, making for greater variety in data and lower ease of interpretation through qualitative analysis. Overall, the fully quantified paradigm with think aloud protocol, while possibly less ecologically valid, does provide a more concrete answer to the simple question of whether the solver is able to properly undertake Bayesian inference.

8.3 The Need for Errors

A further area which urgently requires investigation is the effect of testing errors during forensic analysis on legal Bayesian reasoning. Typically in legal trials, while the RMP is given during forensic reporting of a match, far less frequently, the probability of there being an error in the analysis is reported (Schklar and Diamond, 1999). While some forensic scientists have argued in court that it is impossible to obtain a false match using current DNA technology (Koehler, 1993) the few published proficiency tests present a very different picture, showing that error probabilities ranging between one in several hundred to one in several thousand are likely (Koehler et al., 1995). Both error rates and the RMP each represent ways in which the match between the defendant and the source could have occurred if the defendant was in fact

not the source: due to coincidence (the RMP) or due to a mistake (the error rate), respectively. While both types of match errors are important to a court case, in practice, the probative value of forensic error (which is more common with a typical error rate of 1 in 1000) will eclipse the value of a random match (which is very uncommon with a typical random match probability of 1 in 2,000,000). Thus forensic scientists are currently frequently presenting the coincidental match probability (the RMP) only and conveying this as highly improbable, while failing to reveal that there is a far more probable way in which such a match may have occurred without the defendant being the source (through a false positive: see Thompson et al., 2003). This is very troubling, and has even been called 'prejudicial' by Koehler et al. (1995).

It is clear therefore that any test of people's ability to undertake a legal Bayesian assessment must also include testing errors in the problem. Unfortunately while some previous work has included testing errors in the problem they presented to participants (e.g. Dartnall and Goodman-Delahunty, 2006; Koehler et al., 1995; Schklar and Diamond, 1999), some major and important simplifications of the presentations of those errors have been made. Firstly, all three of the above studies simply presented a 'labaratory error rate', without making it clear that there are two types of error that can be made: a false positive, and a false negative. It may not have been clear to participants which type of error the problem was referring to. Secondly, especially in work such as Schklar and Diamond (1999), the analysis of participant responses has made the implicit assumption that the RMP and forensic testing errors can be dealt with in an identical mathematical way. This would be the case if errors were only possible when testing the forensic sample from the defendant, however, error is equally possible when testing the sample obtained from the crime scene. This dual-role of the error rate adds a great deal of additional complexity to the problem, and precludes a simple use of the error rate as assumed by the above work. If error rates are to be presented in legal cases, false positives and false negatives must be separated (they cannot be presented as a single figure),

and the impact of that error rate must be taken into account on all tests made, not only those on the suspect. Finally, none of these studies used a fully-quantified approach, and were therefore unable to observe the impact of each on the entirety of an individual's Bayesian reasoning process. Therefore the extent to which statistically untrained individuals are capable of undertaking a Bayesian analysis of a match case when errors are included is currently largely unknown. Successful analysis may seem unlikely given the above outline of the true complexity of the addition of forensic error (Fenton and Neil, 2011).

8.4 A New Approach

Due to this increase in complexity and the clear necessity of including testing errors, Fenton and Neil (2011) have claimed that any approach which attempts to get jurors or legal professionals to understand the Bayesian calculations behind a legal trial is fruitless and should be abandoned. Instead, while the quantification of evidence should always be undertaken by humans, the actual Bayesian calculations should be performed by validated computer programmes such as Bayesian Network packages (Edwards, 1991; Kadane and Schum, 1998; Aitken, 1995; Taroni and Aitken, 2006) and trusted as mathematical fact in the same way that a calculator would be, and has been, trusted to produce the correct answer for a large multiplication or division (e.g. Donnelly, 2005).

The necessity of this new approach to presenting statistical formulations of legal cases suggested by Fenton and Neil (2011) explicitly rests on the assumption that jurors are incapable of understanding the calculations themselves. There are examples of previous legal cases where attempts to explain complex calculations have been made, (such as R v Denis: Donnelly, 2005). However this was not considered successful by the statistician recruited to provide the explanation.

While several strands of evidence therefore suggest it unlikely that untrained statisticians will be able to undertake accurate Bayesian reasoning of a legal match case with forensic errors, it remains an open empirical question. Further, if jurors are

incapable of this, a greater understanding of the reasoning processes they undertake are necessary to understand why they struggle, in order to inform the design of assistive interventions. For this purpose, a think aloud protocol will be a valuable tool (Ericsson and Simon, 1998). Finally, if jurors are not capable of undertaking these calculations themselves, the best method for presenting the results of these calculations, as proposed by Fenton et al. (2014), must be trialled and compared.

9 Experiment Five

9.1 Introduction

To answer the question of whether statistically untrained individuals are able to undertake Bayesian reasoning of a legal match case including forensic testing errors, the present study firstly aimed to provide participants with a fully quantified Bayesian match problem. To further understanding of the processes individuals undertake when attempting to reason in these situations, and the reasons those who fail to produce the normative answer do so, the present study firstly requested participants to provide estimates of guilt after each new piece of evidence was presented. For this reason a 'think aloud' protocol (Ericsson and Simon, 1980; Gigerenzer and Hoffrage, 1995) was also employed, which required participants to record their thought processes while attempting to solve the problem. The combination of these two features allows precise mapping of the change in estimates participants undertake in response to each new piece of information as well as the cognitive processes involved in bringing about these changes.

9.1.1 Research Question 1a

Will participants over-weight, under-weight, or appropriately-weight the random match probability in their estimates in comparison to the Bayesian norm?

The present work is exploratory in nature, aiming to provide a first impression of the capacity of the general public to reason through a quantified legal Bayesian problem including forensic testing errors. When asking participants to reason about quantified pieces of 'match' evidence previous work has found a range of responses,

including over-, under- and appropriate-weighting, with under-weighting being reported as most common (Faigman and Baglioni, 1988; Smith, 1996; Taroni and Aitken, 1998; Schklar and Diamond, 1999).

9.1.2 Research Question 1b

What reasoning processes lead to this weighting of the random match probability?

Further, following the detailing of the accuracy of participants in the simple version of the problem (without errors), the think aloud data will be explored and analysed in order to determine the reasoning process behind participants' numerical answers.

9.1.3 Research Question 2a

Will participants over-weight, under-weight, or appropriately-weight the false positive and false negative error rates in their estimates in comparison to the Bayesian norm, and will they do this to a greater or lesser extent than for the RMP?

While the weighting of the error rate has never been previously compared to a fully quantified normative standard, expectation for the relative weighting of the error rates compared tothe RMP can be drawn from three papers: Koehler et al. (1995); Schklar and Diamond (1999); Dartnall and Goodman-Delahunty (2006). Koehler et al. (1995) and Dartnall and Goodman-Delahunty (2006) both found that the inclusion of the RMP in forensic testimony increased conviction rates. However, the inclusion of the possibility for error rates, while, as discussed previously, was many magnitudes more quantitatively important than the RMP, made no significant difference to conviction rates. Further, Schklar and Diamond's (1999) work suggested this may be due to an inherent greater importance attached to the RMP. They presented participants in one condition with an RMP of '1 in a billion' and an

error rate of '1 in 200', and in a second condition, with the values reversed. Comparing these two conditions, participants in fact convicted more frequently when the RMP was presented with the greater '1 in 200' value, than when the error rate was presented with that value, suggesting that participants were in fact more affected in their judgements by the RMP than error rates when controlling for numerical value. In the present study, the RMP is in fact more quantitatively valuable (1 in 20) than the error rates (1 in 100, or 2 in 100). It is therefore expected that greater weight will be given to the RMP than to the error rates. This is also expected to be the case when numerical differences are accounted for (i.e. the proportional weighting of the RMP relative to its true value will be greater than that for error rates).

9.1.4 Research Question 2b

What reasoning processes lead to this weighting of the error rates?

The reasoning processes individuals undertake when attempting to incorporate forensic error is not known. The present study therefore aims to contribute to this through analysis of participants' 'think aloud' data.

9.2 Method

9.2.1 Participants

Eighty six participants comprised the final sample of the experiment. Two participants were removed from an original 88 due to a clear lack of engagement with the experiment. Participants were recruited from the Amazon MTurk outsourcing service and were required to be in the United States of America and have a 95% HIT approval rating. The demographics for experiments five and six can be seen in Table 2 below.

Tab. 9.1: Demographics for experiments five and six.

	Experiment Five		Experiment Six	
	Numerical	Percentage	Numerical	Percentage
Total Sample	88	100%	364	100%
Gender				
Male	44	51.2%	193	53.0%
Female	42	48.8%	171	47.0%
Other	-	-	0	0.0%
Age				
Minimum	20	-	20	-
Maximum	76	-	71	-
Mean	37.1	-	36.9	-
Std Dev	12.4	-	10.7	-
Education				
High School	-	-	114	31.3%
Bachelor's Degree	-	-	175	48.1%
Master's Degree	-	-	50	13.7%
Doctoral Degree	-	-	6	1.6%
Other	-	-	19	5.2%
Occupation				
Professional / Managerial	-	-	161	44.2%
Labour / Service	-	-	101	27.7%
Student	-	-	16	4.4%
Unemployed	-	-	40	11.0%
Other	-	-	46	12.6%
First Language				
English	86	100.0%	358	98.4%
Other	0	0.0%	6	1.6%

9.2.2 Design and Materials

The experiment comprised three 'conditions'. However, these were designed to counterbalance presentations / rule out confounding factors, rather than for direct experimental comparison. The first two conditions contained four 'sections' in the main body of the experiment, corresponding to the four pieces of information participants were required to integrate: the prior, the random match probability, the false positive rate and the false negative rate. In each section, participants were presented with the information, then requested to record their thought process regarding how to use that information before being able to record their numerical answer on the subsequent page. The scenario detailed an attack on a man by a gang. A member of that gang had been arrested and had been found to match a footprint on the body of the victim. The participant's task was to combine all four pieces of numerical information to calculate the probability that the defendant was the person who stamped on the man. The full text provided can be seen below:

The introduction and prior:

> A mob of 100 hoody-wearing gang members walking through a city centre pass an elderly man. One of the gang members in an unprovoked attack punches the man and then stamps on him after he falls to the floor. Upon realising the man is seriously injured (he later dies), all of the gang members run off in different directions. A policeman in a nearby street who hears the commotion randomly grabs one of the fleeing gang members (the defendant) and arrests him on suspicion of being involved in a criminal act. The police are unable to find any of the remaining gang members.

The random match probability section:

> Forensic evidence from the crime scene has been able to identify the size of the footprint left on the victims body, which was left by the person who stamped on him. The footprint was determined to be a size 12. Forensic databases of the shoe sizes of youths in gangs show that out of a group of 20 gang members, 1 gang member would be expected, on average, to have a size 12 shoe. Finally, the forensic team also determined that the defendant was wearing a size 12 shoe when arrested.

The false positive section:

> Studies show that during both of the tests that the forensics team conduct (measuring the foot size of the print on the victims body as well as

> the shoe size of the defendant), the forensics team will sometimes make a mistake. Sometimes they will record that the size of the footprint or shoe is size 12 but actually they have made a mistake and it really is a different size. This type of mistake is called a false positive. The forensic team make a false positive mistake 2 out of every 100 times they make a measurement.

The false negative section:

> Studies show that there is another type of error that the forensics team sometimes makes. Sometimes they will record that the size of the footprint or shoe is size 12 but actually they have made a mistake and it really is a different size. This type of mistake is called a false positive. The forensic team make a false positive mistake 2 out of every 100 times they make a measurement.

The presentation of the two error rates was counterbalanced such that participants in condition one were presented with the false positive rate as the third piece of information and the false negative rate as the fourth piece of information. Participants in condition two received the opposite order. Due to the fact that each estimate participants gave was based on combining all previous pieces of information, this separation allowed independent comparison of the impact of false positive and false negative rates on participant estimates, as well as the comparison of order on final integration of both rates. However, this format also meant that for the non-error portion of the experiment (sections one and two), participants in conditions one and two experienced an identical presentation.

To ensure that any findings of the present experiment weren't due to the segmented presentation of information, condition three was presented with all four pieces of information and only then asked to record their thought processes and provide a numerical estimate. Participants in conditions one and two therefore made four sequential estimates over the course of the experiment, while participants in condition three only made one estimate.

9.2.3 Procedure

Participants completed the experiment in their own homes via an internet survey program. They were presented with the consent form followed by the instructions

for the experiment. In conditions one and two, participants were then presented with each successive piece of information from the problem each on a separate web page (requiring the participant to press 'Next' to receive each subsequent piece of information). Each piece of information was followed by the think aloud (an open ended text box requesting the participant to record their thought process while they solve the problem) and numerical question ('Please record the percentage chance of the defendant being the individual who stamped on the man below') after each new piece of information. In condition three participants were shown all pieces of information first without being asked any questions. Only once they had seen the final piece of information they were presented with both the think aloud and numerical questions. Finally they were presented with the below question which acted as a manipulation check on the random match probability statistics and finally they were shown the demographic questions.

> Please indicate below the extent to which, when initially presented with this evidence, you believed that 1 out of 20 was the correct figure for the amount of members of this particular gang who wore shoe size 12.
>
> I believed that, in regards to how many wore shoe size 12, for this particular gang:
>
> 1. It was far less than 1 out of 20
> 2. It was considerably less than 1 out of 20
> 3. It was slightly less than 1 out of 20
> 4. 1 out of 20 was correct
> 5. It was slightly more than 1 out of 20
> 6. It was considerably more than 1 out of 20
> 7. It was far more than 1 out of 20
> 8. Don't feel able to answer

9.3 Results

9.3.1 Manipulation Checks and Condition Comparisons

Manipulation checks were placed into the experiment to ensure that participants processed and accepted the prior and the RMP in the way that they were intended

to. Following the presentation of the prior (the fact that there were 100 gang members at the scene) participants were asked the chance of the defendant being the one who stamped on the man, and the vast majority of individuals (87.5%) responded '1 in 100'. At the end of the experiment, participants were asked whether they believed the RMP figure of '1 in 20' was correct for 'this particular gang'. Again the majority (65.1%) responded that it was correct.

Further, in regards to the comparison between conditions one, two and three, little difference was seen. Participants in condition one gave a mean chance that the suspect was the source of 1 in 24.53, condition two gave a mean chance of 1 in 24.49 and condition three, 1 in 24.40. A one-way analysis of variance confirmed no difference between these three groups on this variable, (F = .000 [2, 83], p=1.000).

9.3.2 Research Question 1a

In the basic problem, will participants overweight, appropriately weight, or underweight the Random Match Probability?

In conditions one and two, the second question asked participants to compute the chance of the defendant being guilty in the light of both the prior information ('100 gang members were present') and the footprint evidence ('1 out of 20 gang members have size 12').

The Bayesian normative answer after receiving only the prior information was '1 in 100' (0.01). The mean answer given by participants after the prior information was '1 in 89.75' (0.011) while the modal answer was '1 in 100', with 86.0% (49 out of 57) of participants providing this.

When incorporating both the prior (1 in 100) and the RMP (1 in 20), the Bayesian normative answer for the chance of the defendant being the source of the footprint is 1 in 5.95. This is calculated by dividing the 99 gang members (100 - the 1 defendant) by the random match probability (1/20) to arrive at 4.95 probable matches. This is then added to the defendant to arrive at a most-probable 5.95 total matches. The defendant is one of these 5.95, and no more likely to be the source

than any of them. Therefore the chance of him being the source of the footprint is 1 in 5.95.

However, participants mean prior (1 in 89.75) was deviant from the normative prior (1 in 100), and their answers to question two build upon this prior. Therefore, in order to construct a normative answer following the presentation of the RMP, this prior must be incorporated into the calculations. Given this prior, the Bayesian normative posterior after the RMP would be '1 in 5.45' (18.3%). With the '1 in 100' prior, the normative posterior would be '1 in 5.95' (16.8%).

The mean figure provided by participants was in fact '1 in 14' (7.1%), considerably higher than either normative answer. Proportional changes for the normative and actual mean figures for estimates following the RMP compared to the prior can be seen in Figure 9.1 below. We can see that, using the participants' prior, the Bayesian normative answer represents a 15.45 fold increase in probability that the defendant was the source, while the actual change made by participants represents only a 5.41 fold increase in guilt. This suggests that participants on average under-weighted the RMP in comparison to the Bayesian norm.

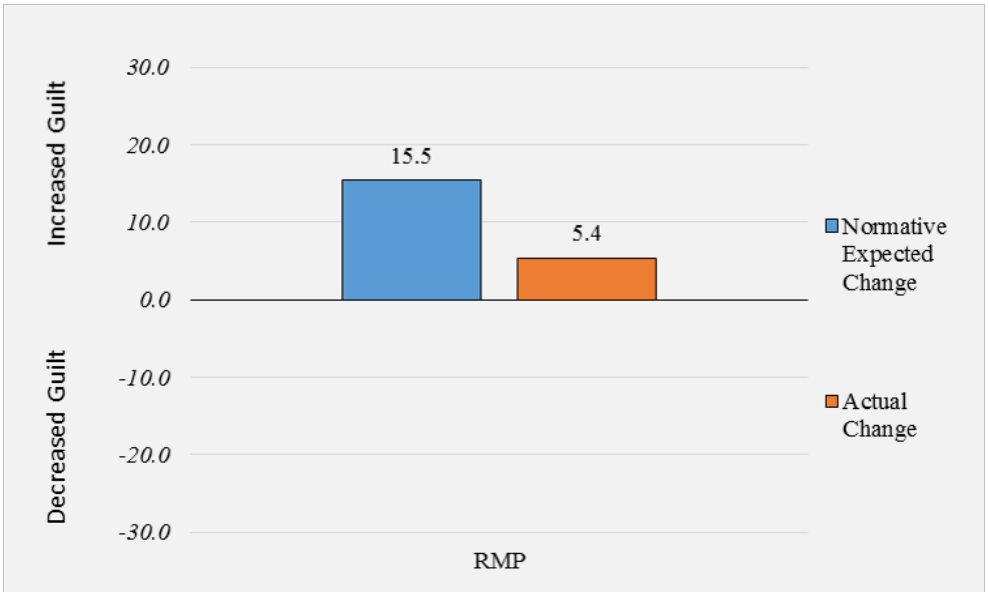


Fig. 9.1: Proportion change figures for the normative and mean values from 'Prior-only' estimates to 'Prior plus RMP' estimates

9.3.3 Research Question 1b

To explore why this under-weighting occurred, an analysis of both the numerical and think aloud data was undertaken to categorise and understand the reason for various response types. The percentage of response types given to the basic problem across all three conditions can be seen below in Figure 9.2.

'1 in 6'

As explained above, the Bayesian-normative answer for question two assuming the correct prior of '1 in 100' was in fact 1 in 5.95 or 1 in 6. No participant gave this answer.

'1 in 5'

Thirty two out of the 56 participants (37.2%) in conditions 1 and 2 gave the answer of '1 in 5', which was accepted as correct under the 'lenient' criteria of ignoring the independent samples rule. These individuals divided all 100 gang members by the '1 / 20' RMP to arrive at 5 individuals with size 12, and determined that this was the correct answer. This answer uses the correct process but ignores the 'Independent Samples' rule: the observation of the defendant being size 12 should is independent from the most likely number of size 12's in the remaining 99 gang members, who's shoe size is still in a probabilistic state (unknown). These individuals therefore needed to divide 99 instead of 100, by 1/20, and then needed to add the defendant on to arrive at the final answer.

Out of the 32 individuals providing this answer, six in fact did not demonstrate appropriate reasoning in their think aloud protocol e.g. P42 who stated 'There are a lot of people who wear a size 12 shoe but since he was caught at the scene and wears the size 12 shoe then it could be him'. The remaining 26 participants (46.4%) demonstrated the (lenient) correct reasoning that 5 gang members (100/20) would have size 12 shoes, and the defendant was one of these 5. For example, P2 stated that 'If 1 in 20 gang members have a size twelve shoe, then 5 of the 100 have that

size shoe. Therefore he is one of the five.' and P9 who stated that 'If 1 in 20 gang members wears a size 12 shoes out of 100 gang members most likely only 5 or so of them have a size 12 making it a 1 in 5 chance he is the right guy.'

'1 in 20'

In response to this second question, 16/56 participants (28.6%) answered '1 in 20'. During qualitative analysis of the justifications given in the 'think aloud' protocol for this answer, three clearly distinct approaches were found which all led to the same numerical answer of '1 in 20' but for very different reasons.

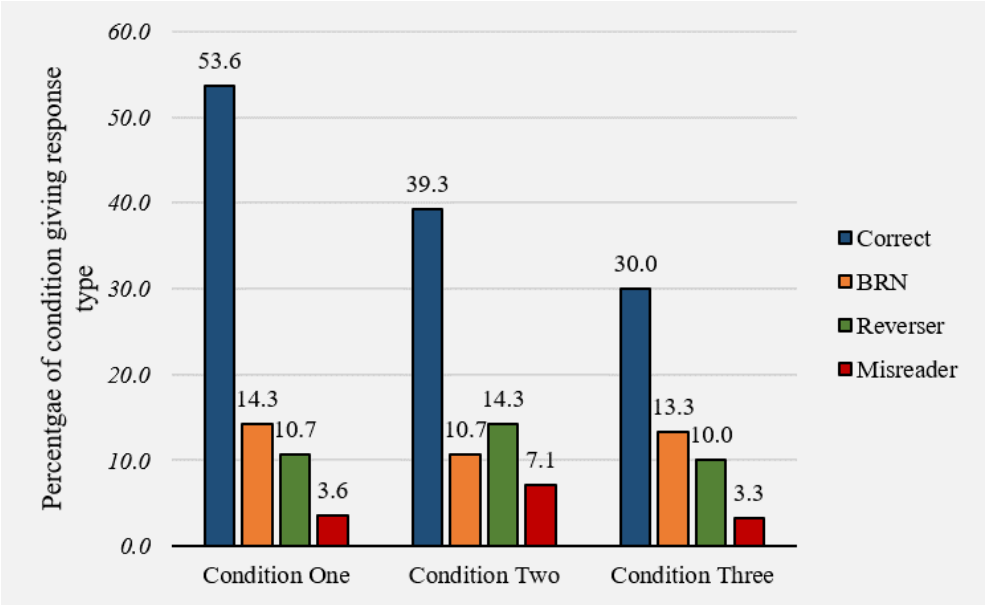


Fig. 9.2: Percentage of response types given to the basic problem in all three conditions.

Reversers Six participants have been classified as 'Reversers'. This name was chosen because these participants began the calculation correctly to arrive at a figure of 1 in 5, but then 'reverse' their work at the end with an extra unnecessary step in their calculation to finally end up back with an answer of 1 in 20. Firstly, each of these participants managed to use both the prior and the RMP to determine that 5 out of the 100 gang members would have this size shoe. The correct next step his now to recognise that the defendant is one of these 5 and so the answer is 1 in 5 (or 1 in 6, ideally). However, what these participants instead did was to

state that 5 out of 100 is the same as 1 in 20, and conclude that this is therefore the answer. They therefore first convert 1 in 20 into 5 in 100, then convert it straight back into 1 in 20. There is little else in the qualitative data to explain exactly why participants do this. Below are their transcripts.

> '1 in 20 gang members have a size 12 shoe. There are 100 gang members, so 5 of them probably have a size 12 shoe. 100 / = 20. 1 in 20 chance.'
>
> '1 in 20 hoodie members have a size 12 according to the stats. Out of 100 hoodie member this would mean 1/5 of the hoodie members would have a size 12. 5/100 is equivalent to 1 in 20.'
>
> '1:20 = 5:100'
>
> 'There were 100 so 5 in 100 have size 12 shoe. Brings it back to 1/20'
>
> 'Out of a group of 20, 1 is expected to have size 12. 5 out of 100 reduces to 1 out of 20.'
>
> 'Based on the stat of 1 in 20 wear size 12, then 5 of 100 or 1 in 20 would be the chance given the numbers, would check for blood spatter consistent with the kind of attack on the shoes worn by the suspect. Since he was apprehended relatively quickly according to the scenario.'

Misreader Two participants were classified as 'Misreaders': these participants seem to have misread the information 'out of a group of 20 gang members, 1 would have size 12.' Their answers seem to indicate that they believed that the information given was that 20 gang members total had shoe size 12. They therefore computed that the defendant was 1 out of these 20 and thus came to the answer of 1 in 20.

> 'If there are only 20 gang members who wear a size 12, then the chance of the suspect being guilty is 1/20'
>
> 'Twenty members of the gang wear a size 12, and he is one of them'

Base Rate Neglecter Seven participants presented a simple conviction that the RMP was the answer to the question, in line with previous work on Base rate neglect (Bar-Hillel, 1980).

> 'The odds of a gang member having a size 12 shoe is 1 in 20. Since the defendant had that size, the same odds apply.'
>
> 'Only 20 members were wearing size 12 shoe and the defendant was wearing size 12 shoes. hence one in 20 was the defendant.'
>
> '1 in 20 wear a size 12 this guy wears a 12.'

> 'Statistically there is a one in 20 chance the suspect would be guilty due to his size 12 shoe. The footprint on the victim was a size 12.'
>
> 'Earlier it was 1 in 100 among the whole crowd of 100, but the new evidence says than among the whole 100 member whoever has the size 12 and among all member of 12 size the probability is 1 out of 20.'
>
> 'Size 12= 1/20 expected'
>
> 'The suspect had a size 12 shoe and the probability is 1 in 20 so that means that this is the same probability that he did it.'

Others

Acceptors and Rejectors At either ends of the spectrum of guilt, 2 participants gave an answer of 1 in 1 while 2 other participants gave an answer of 1 in 100. Both of the first two participants seemed to suggest that they thought the evidence compelling enough to be convinced of the defendant's guilt.

> 'The forensic evidence is totally convincing.'
>
> 'According to statistics, this man was most likely involved in the killing of the man.'

However, the two '1 in 100' participants on the other hand, seemed to find the shoe size evidence unconvincing / unimportant:

> 'The shoe size evidence is too speculative for my tastes so I kept to the original percentage'
>
> 'Although there were findings regarding the shoe size and that 1 in every 20 members is size 12, chances are is that still there is only one in a hundred.'

Unknown One further participant was classified as 'unknown' as they gave little information as to how they formed their conclusion, stating that 'I figure that its lowered to 1 in 20 now.' Four further participants gave the answers of 10, 12, 30 and 50 but their qualitative responses gave little insight into their reasoning or suggested that they were simply guessing.

Condition Three

In condition three, participants were presented with the prior, RMP and two errors before giving their reasoning in a single text box. While this makes it more difficult to pick out the impact of individual pieces of information on participants' reasoning, analysis of the think aloud protocol was still possible, and the exact same patterns of reasoning in regards to the prior and RMP were seen as those in conditions one and two and also in similar proportions (see Figure 9.2 above). Thirty percent demonstrated correct integration of these two variables, while 13.3% demonstrated the BRN response, 10% the Reverser response and 3.3% the Misreader response.

9.3.4 Research Question 2a

Will people under-weight, appropriately weight, or over-weight the false positive and false negative error rates?

In question three, participants in condition one and two were presented with evidence regarding one type of error that the forensic team can make (false positives for condition one; false negatives for condition two). Bayesian normative answers were calculated using the event tree process presented in Fenton et al. (2014) (see 9.3 below). In experiment five the prior (s) was 0.01, the RMP (m) was 0.05, the false positive rate (u) was 0.02 and the false negative rate (v) was 0.01. To calculate the Bayesian posterior the two uppermost branches reflecting scenarios wherein the defendant was the source of the footprint were divided by total of those 2 branches plus the 4 lower branches wherein the defendant was not the source of the footprint. By changing the false positive rate, or false negative rate to zero, answers for question 3 (where only one of these error rates was given) could be calculated in the same manner.

According to this, the Bayesian normative answer for question three, condition one (false positive rate only) was 1 in 10.36, which represents a substantial decrease in the probability that the defendant was the source of the footprint. The false

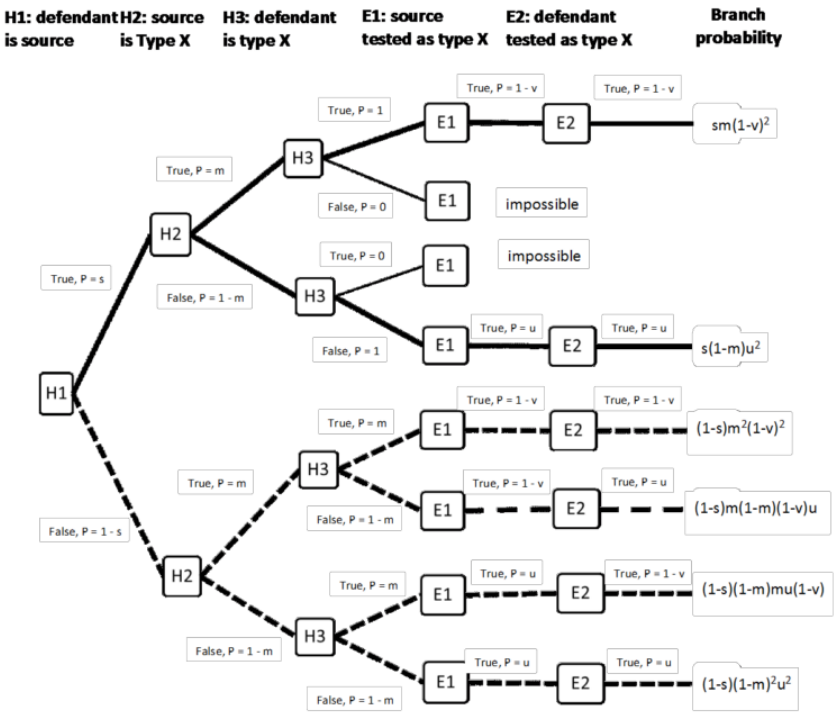


Fig. 9.3: An event tree for a legal match case with errors taken from Fenton et al. (2014). Here s is the prior probability that defendant is source; m is the random match probability for Type X; u is the false positive probability for X and v is the false negative probability for X (so 1- v is the true positive probability that we are interested in). The bold branch is that consistent with the prosecution hypothesis and the dotted branch is that consistent with the defence hypothesis. Cases of E1 and E2 false are not considered.

negative rate alone, (question three, condition two) produced a posterior chance of 1 in 5.95 (no change). Both error rates combined (question four for conditions one and two, and the single question for condition three) produced a Bayesian posterior of 10.41, a small change from the false positive rate only. We can see here therefore that the false negative rate only impacts on the posterior when combined with the false positive rate, and in that case, reduces the chance of the defendant being the source a small amount.

However, similar to in the previous section, we were also able to create Bayesian normative posteriors for the error rates taking into account participants' responses to the previous question. As the error incorporation built upon these previous answers, accurate comparison of participant responses to errors required a comparison to a normative standard which allowed for any inaccuracy in previous calculations. Participants' mean answer in the previous question (combining prior and RMP was)

'1 in 14' (0.0714). Inputting this value into the Bayesian calculation, the normative answer following the incorporation of the false positive rate only (condition one) should have been '1 in 17.2' (0.058) which is a proportional change of probability of -0.19. However, participants' actual mean response to this question was '1 in 21.8' (0.046), a proportional probability change of -0.36.

Similarly for condition two, the introduction of the false negative rate alone should have made no change, retaining the same chance of the suspect being the source of '1 in 14' (0.0714), making no proportional probability change. Participants' actual mean response to this question was '1 in 21.9' (4.6%), a proportional probability change of -0.36.

The incorporation of both errors (question four for conditions one and two) should have yielded a posterior value of '1 in 17.2' (5.8%), a proportional change of zero for condition one, and a proportional change of -0.19 for condition two. Participants' actual mean response was '1 in 24.5' (4.1%), a proportional change of -0.11 for both conditions one and condition two. The proportional changes for both the normative and actual mean responses for conditions one and two can be seen below. The normative change represents a decrease in the chance of the suspect being the source when the false positive rates are incorporated, and a near-zero change when the false negatives are incorporated. However, as can be seen in Figure 9.4, and was demonstrated by an analysis of variance, participants responses in both conditions were identical, regardless of the order of error type presentation.

9.3.5 Research Question 2b

To explore why this erroneous integration of errors occurred, an analysis of the numerical think aloud data was undertaken to categorise and understand the reason for various response types. The percentage of participants corresponding to each categorical type identified can be seen below in Figure 9.5.

In response to the presentation of this first error type , two modal approaches to incorporating the new evidence were seen, here named 'Negligible' and 'Nudge'.

For the 'Negligible' classification, participants needed to demonstrate in their think aloud protocol that they believed the margin of error presented was too small to impact on the problem. For example, P35 said 'I still believe that at least 5 gang members out of 100 had a size 12 shoe. Furthermore, the forensics team is very unlikely at 1 in 100 to have made a mistake.' and P7 said '2 in 100 is a really small chance not enough to change my estimate'. This was uniformly accompanied by a complete absence of change of the solver's numerical response in comparison to the previous question.

The classification of 'Nudge' was given to any participant who, either non-mathematically, or using simple addition or subtraction, 'nudged' their answer from the previous question a small amount in response to the new evidence regarding the error rates. A non-mathematical example comes from P21 (condition one) who altered their answer of '1 in 20' from the previous question (with only prior and RMP) to '1 in 21' and justified this by saying 'The odds of a 'false positive' slightly mitigate the odds of the attacker having a size 12 shoe, but not much.' and P42 (condition two), who raised their previous answer of '1 in 5' to '1 in 10', stating that 'I just figured maybe I ought to give it a little less chance now.' Addition / Subtraction approaches to nudging can be seen in P4 (condition one), who nudged their previous '1 in 5' answer to '1 in 7' and stated that 'The odds of a false positive adds 2 out of 100 to the 5 out of 100 that have a size 12 shoe. Therefore the odds are now 1 in 7.' and P47 (condition two) who nudged their previous answer of '1 in 12' to '1 in 13', stating that 'A 1% error would only increase the possible amount of 100 members by one extra person. So it goes from 1 in 12 to 1 in 13.' Combining both conditions, addition-nudges made up 46.5% of all nudges, subtraction-nudges made up 7% and non-mathematical nudges made up 46.5%.

False Positive Rate

In terms of the Bayesian normative standard for incorporating error rates in this problem, the introduction of false positives has a far larger impact on the proba-

bility that the defendant was the source than the false negative information. The introduction of the false positives decreases this chance from 1 in 5.95, with only the prior and RMP information, to 1 in 10.36. The introduction of the false negative alone makes no change to the chance of guilt. When the false negative is introduced on top of the false positive rate it makes a very small change from 1 in 10.36 to 10.41.

Following the presentation of the false positive rate, only four participants gave an answer within even 3 points of this standard, and in each case their think aloud protocol indicated they had arrived at this figure through the combination of error and coincidence, as opposed to appropriate Bayesian reasoning:

> P18 (1 in 10): 'I guessed randomly.'
>
> P36 (1 in 10): 'There is a 2% chance that they got the wrong shoe size or a false positive. So There is a 2% chance that the person is not it because of a false positive. and since there is a 20% chance he is the right person I did 20% divided by 2% which means there is a 10 percent chance now that the person is correct instead of a 20 percent. so 1 in 10 chance.'
>
> P40 (1 in 10): 'I simply doubled the previous chance to account for an incompetent forensics team.'
>
> P41 (1 in 10): 'I just figured maybe I ought to give it a little less chance now.'

False Negative Rate

For question three in condition two the correct response to the presentation of the false negative error was to make no substantial change to the chance of guilt. This is because the problem deals with a positive result (the 'match' of the defendant) and so although false negatives do have some impact on the chance of guilt, it is minimal. Ten out of 28 individuals (37.5%) made no change to their answer. However the think aloud protocol showed that six of these chose not to change their answer on the basis that the error rate was very small (e.g. P29 who stated 'The error rate is small and the evidence is overwhelming.'), and from one individual it could not be determined if they had this understanding of the false negative rate or not. However three individuals did correctly state that they had not changed their answer because

the problem dealt with a positive result e.g. P53 who stated 'False negatives would not apply to this case because the test was positive.'

Participants in condition one also faced the need to integrate the false negative rate in question four (after having integrated the false positive rate in question three). The correct response again was to make no substantial change to the chances of guilt. Seven out of 28 individuals made no change to their answer, however all of these were classified as giving the 'negligible' response, suggesting the errors were too small to make a difference. None of these referenced the fact that the test was positive as the reason for not changing their answer.

Condition Three

In the first two conditions, the fourth question provided the second piece of error evidence (false negatives for condition one, false positives for condition two). However in the third condition, participants were only asked when question in total - they were presented with all the evidence sequentially and then asked the chance of guilt. 'Negligible' and 'Nudge' were again the most common response types seen. However, in this condition only, a large proportion of participants were also classified as 'ignoring' the errors. This classification was given to individuals who responded to the problem as if only the prior and RMP data were given, and made no mention of the errors at all, and did not include them in their calculations. This classification is distinct from the 'Negligible' classification where participants explicitly state that the errors are too small to be worth considering. For example, P57, omitting any mention of the error information at all, and providing a '1 in 5' answer, stated that 'One in twenty people would wear a size 12 shoe. Since there were 100 gang members, it is reasonable to assume 5 of them would wear a size 12.' The percentage of the sample providing the categorical responses identified can be seen below in Figure 9.6.

Finally, in comparing the three experimental conditions, mean estimates between the three conditions were highly similar. Condition one showed a mean of 1 in 24.5

(SE = 6.1), condition two showed a mean of 1 in 24.5 (SE = 5.4) and condition three showed a mean of 1 in 24.4 (SE = 5.0). These were all considerably higher than the normative Bayesian answer of 1 in 10.4.

9.4 Discussion

9.4.1 Manipulation Checks and Condition Comparisons

The figures from the two manipulation checks (of the prior, and the RMP) suggest that on the whole, participants accepted and dealt with these figures in the manner that was intended and that, instead of bringing a large amount of their own world-knowledge to bear on the problem, they dealt with it abstractly, using the numbers given. This result runs in contrast to some previous work (e.g. Welsh & Navarro, 2012) who suggested that failure on these problems may be due to 'distrust' and lack of acceptance of figures such as these.

Further, neither the order of presentation of the false positive or false negative rates, nor the repeat-questioning format of conditions one and two seemed to change final estimates. The first of these points can be seen in the fact that no difference in mean final estimates was seen between condition one (1 in 24.53) and condition two (1 in 24.49). The second point can be seen in the fact that no further difference was seen in comparison to condition three (1 in 24.40), which only asked for an estimate once, after all pieces of information were shown.

9.4.2 Research Questions 1a and 1b

In terms of the basic match problem without errors, the average changes made by participants in response to the RMP statistic were in the same direction as the normative answer (increasing the probability of guilt), suggesting accurate conceptual understanding that this evidence increases the likelihood of the defendant being the source of the footprint. This finding is in keeping with previous work (e.g. Koehler et al., 1995; Schklar and Diamond, 1999; Dartnall and Goodman-Delahunty, 2006)

who all found that the introduction of the RMP increased conviction rates on average.

However, the strength of this change was small relative to the normative standard, suggesting that participants under-weighted this variable, also in line with previous work comparing to a normative standard (Thompson and Schumann, 1987; Faigman and Baglioni, 1988; Goodman, 1992; Smith, 1996; Taroni and Aitken, 1998; Thompson and Newman, 2015).

The reason for this under-weighting in the present study was explored by examining numerical and think aloud responses. No participants were able to incorporate the independent samples rule into their reasoning in order to arrive at the strict Bayesian normative answer of '1 in 5.95' or '1 in 6'. While this does not distort the posterior belief level to a great extent with the present RMP, which is relatively large, the smaller the RMP (such as with DNA evidence, typically several orders of magnitude lower), this could become a larger problem. However this may not be an issue in practise. This is because, at lower RMP levels, the incorrectness of this mistake becomes much clearer. For example, if we had an RMP of 1 in 500 with the present 100 gang members, we would only expect 0.2 matches on average. If the independent samples rule were ignored here the solver would end up with a posterior chance of '1 in 0.2' of guilt, which is a clear impossibility: the correct value is '1 in 1.2'.

Therefore, it may be of greater importance to overlook the failure to apply the independent observations rule, and assess the degree to which participants were able to combine the RMP and the prior to determine the number of matches expected, a key skill in correct Bayesian reasoning for match cases. Out of conditions 1 and 2, 46.4% of participants undertook this process successfully, providing the '1 in 5' answer.

The 'Inversion Fallacy', Base Rate Neglect and the Confusion Hypothesis

In contrast, 28.6% of participants in these two conditions provided the '1 in 20' answer as the chance that the defendant was the one who stamped on the man. Interestingly, this answer has a lot in common with the commonly observed prosecutor's fallacy (Thompson and Schumann, 1987; Smith, 1996; Nance and Morris, 2002, 2005), but is also importantly different. This fallacy was first documented by Nance and Morris (2002) and was named there the 'Inversion Fallacy'. It does not involve confusing the RMP with the chance of 'not being a match' (as in the prosecutor's fallacy), but instead with the chance of 'being a match'.

In previous work, the answer of '1 in 20', or its equivalent, which relies entirely on the 'diagnostic information' (here, the RMP), and ignores the base rate data (here, the population size), has typically been classified as 'base rate neglect' (Bar-Hillel, 1980; Kahneman and Tversky, 1972; Welsh and Navarro, 2012; Obrecht and Chesney, 2013; Pennycook and Thompson, 2012). Typically, however, previous work which has classified these responses in this way has relied entirely on the numerical response from participants. In contrast to this, the present study, by examining the think aloud protocol of solvers, found three distinct processes all of which led to the same numerical value, but only one of which actually involved ignoring the base rate. The first classification which led to the '1 in 20' numerical value was named 'Reverser' as these participants actually integrated the RMP and the base rate together correctly, then added an additional unnecessary and incorrect step to get back to '1 in 20'. The second classification, 'Misreader', interpreted the statement as '20 gang members have size 12', which, if accurate, would legitimately allow the ignoring of the prior. Only the final group, representing 12.5% of the sample actually committed the base rate neglect fallacy, by entirely ignoring the prior and stating that the RMP was the correct value to the problem. This analysis suggests that, behind the numerical values, a large amount of previous work may have overestimated the amount of base rate neglect in their samples by mis-classifying different cognitive processes which coincidentally produce the same numerical value.

Across a range of studies participants have been shown to confuse the Bayesian posterior with the true positive rate (e.g. Eddy, 1982), the false positive rate (e.g. Casscells et al., 1978) and now the random match probability (e.g. Nance and Morris, 2002, and the present study). This provides further evidence for the 'confusion hypothesis' developed within the cognitive science literature mentioned previously (e.g. Braine and Connell, 1990; Cohen, 1981; Dawes, 1986; Eddy, 1982; Hamm and Miller, 1990; Fiedler et al., 2000). It also supports views expressed throughout this paper that participants lack of engagement or lack of deep processing of the problem may contribute to the types of errors being seen. It also provides further evidence for the conjecture also put forward here that problem solvers may interpret the task as one in which one of the figures in the problem text need to be 'picked' to provide the correct answer, rather than interpreting the problem as one in which several figures need to be integrated. If this former interpretation is adopted, participants may then simply look for the figure in the problem which superficially sounds the most like what they are being asked for, and 'pick' that as the answer.

9.4.3 Research Questions 2a and 2b

Participants' mean responses following the presentation of false positive and false negative rates was highly deviant from the Bayesian normative standard. Comparing to the normative standard for change at each stage, the false positive error rate when provided alone was over-weighted in participants' analyses by a factor of 1.89, while the false negative weight when presented alone was over-weighted by an indefinable factor, as it's impact on the outcome probability when presented alone was nil (Figure 9.4. However, in terms of the modal response types, the 'negligible' and 'ignore' response clearly under-weighted both (particularly the false positive rate) and the 'nudge' response under-weighted the false positive rate but over-weighted the false negative rate.

In comparison the RMP on average was under-weighted by a factor of 2.87 (Figure 9.1), demonstrating a greater proportional mis-weighting of the RMP than the

false positive error rate, but with the greatest mis-weighting being of the false negative error rate. This is not fully in line with previous work (e.g. Koehler et al., 1995; Schklar and Diamond, 1999; Dartnall and Goodman-Delahunty, 2006) who found an overall under-weighting of the error rate. However, as stated before two major simplifications of the use of error rates in those works (no distinction between false positive and negative rates, and the ignoring of the impact of forensic error on testing of the source) make comparison difficult.

In regards to why this mis-weighting of the error rates occurred, information can be gathered from both the quantitative and qualitative data. Firstly, Figure 9.4 shows that participants responded on average to both false positive and false negative rates in a remarkably identical manner (a proportional reduction in estimates of 0.19 in each case), suggesting a gross misunderstanding of the nature and meaning of the two error types. Further, the think aloud data suggested that no single participant undertook an appropriate calculation to incorporate the error rates into their calculations. Instead, three categories of responses were record as most common. These were named 'Negligible', 'Nudge' and 'Ignore'. Those participants classified as 'Negligible' typically stated that the error rates were so small that they did not need to be incorporated into their calculations. This was most notably incorrect in regards to the false positive rate, which in fact almost halved the chance that the suspect was the source of the footprint. These participants therefore under-weighted the false positive rate, and appropriately-weighted the false negative rate (but not for a legitimate reason). Another common classification was that of 'Nudge'. The reasoning process of these individuals was highly reminiscent of the 'anchor and adjustment' heuristic first proposed by Tversky and Kahneman (1974). Individuals classified this way typically used the value they had previously calculated combining the prior and the RMP as an anchor and from this 'nudged' their answer, typically towards a lower chance of guilt (the normative direction, if not with the normative magnitude required). This was one of the most common responses for both the false positive and false negative error types. In some cases this involved mathematical

addition / subtraction of the error rate directly from their anchor, while in other cases think aloud data suggested a non-mathematical nudge based upon intuition. In the third condition only, a further common response was seen, which was to simply ignore, or not mention the error rates, and respond to the problem as if only the prior and RMP had been given. This was less possible in conditions one and two as they were directly asked for their revised estimate based on each piece of information. This 'ignore' classification in the third condition again suggests a dismissal of the importance of the error rates similar to the 'negligible' classification.

The combination of these quantitative and qualitative findings suggest that while many statistically untrained individuals can successfully analyse a legal match case when only a prior and RMP are presented, once errors are introduced, the complexity of the problem appears too great and a fundamental misunderstanding of the function of forensic error in a Bayesian calculation is seen. No single individual in our sample approached the analysis of errors in the appropriate Bayesian manner, and all mis-estimated the chance of the suspect being the source of the sample in comparison to the normative Bayesian standard. Finally, the nature of the heuristics participants used to solve the problem in fact suggests that the above quantitative findings in regards to over-weighting and under-weighting would not be stable across different values: the same heuristic might over-weight with some values, and under-weight with others. In general, regardless of the direction of the mistake, it must be strongly concluded that statistically untrained individuals should not be relied upon to make accurate Bayesian inference of legal match cases when forensic testing errors are taken into account.

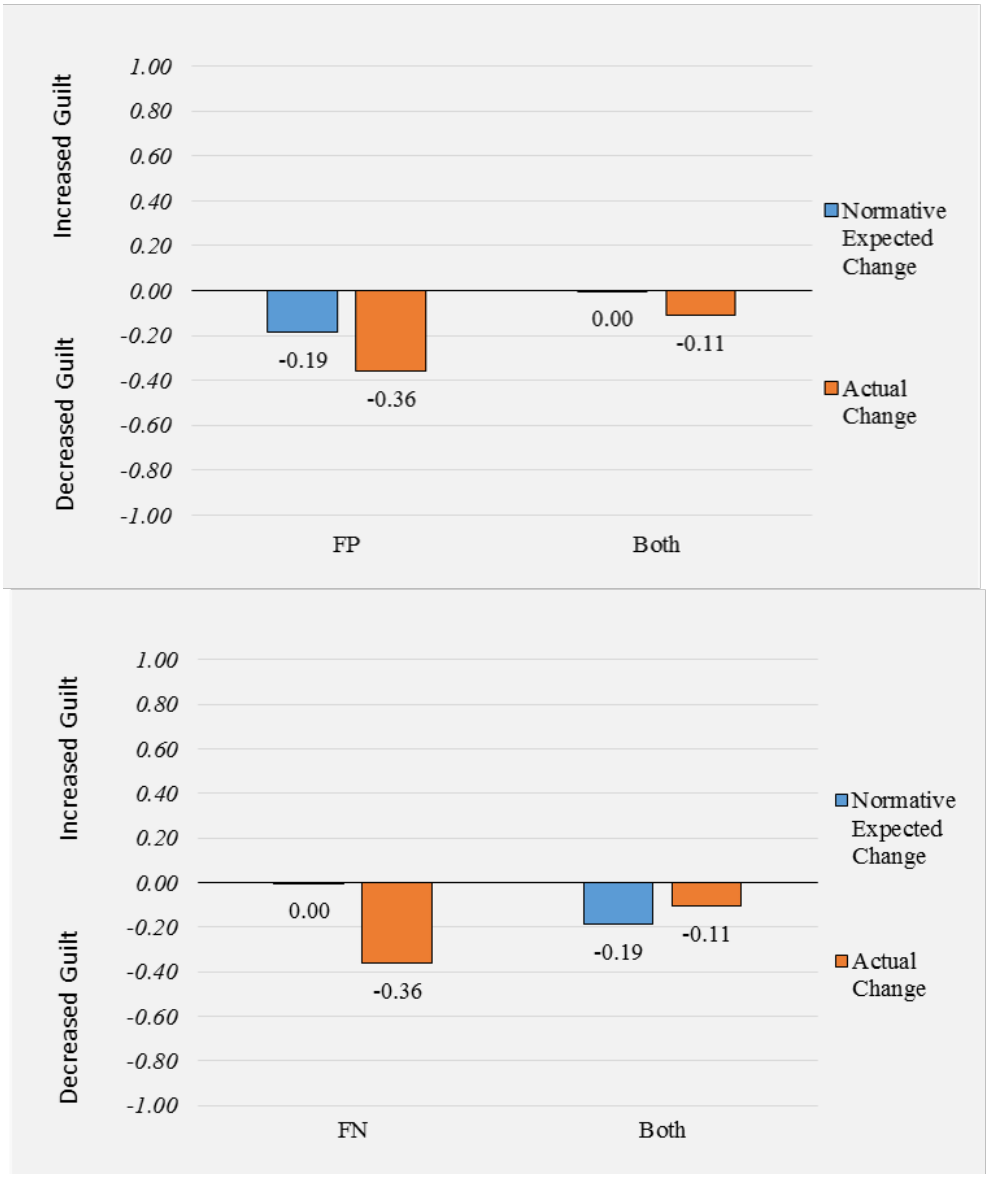


Fig. 9.4: The proportional change in condition three estimates that the defendant is the source from 'Prior plus RMP' to including the false positive rate (top) and false negative rates (bottom) and then to including both false positives and false negatives (both).

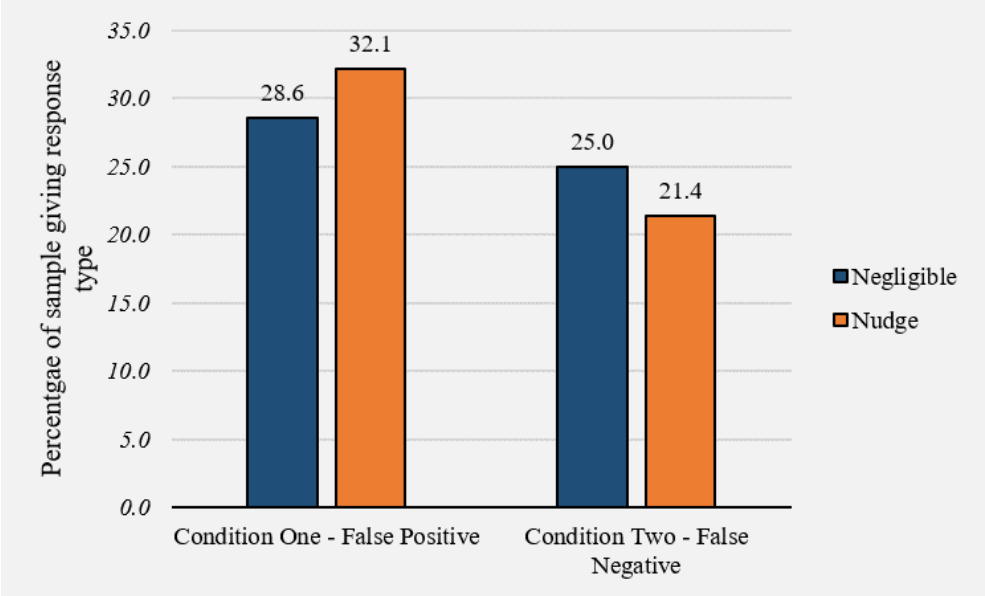


Fig. 9.5: The Percentage of participants coded as providing 'Nudge' and 'Negligible' responses to question three for condition one (false positive) and condition two (false negative).

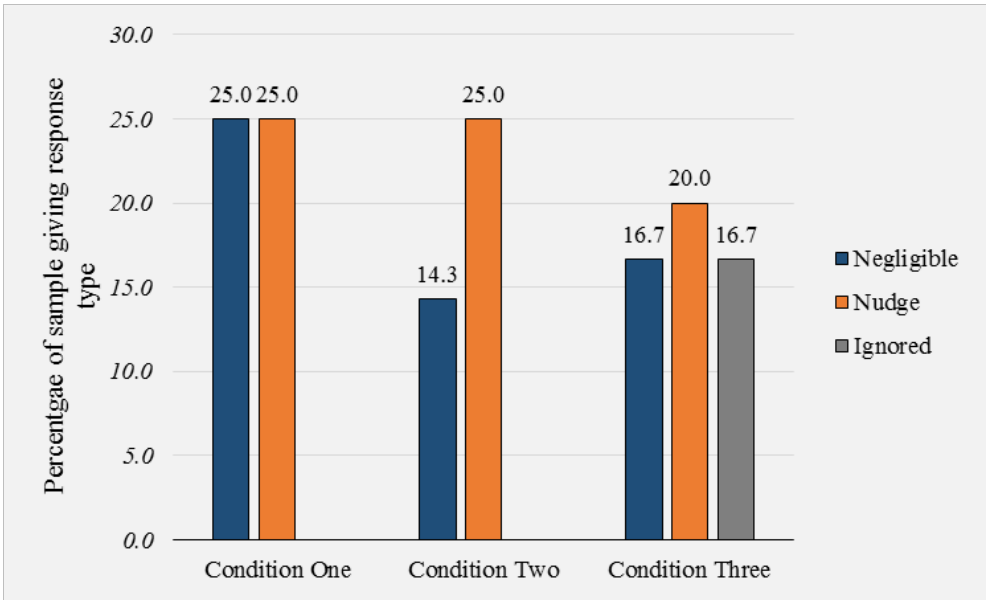


Fig. 9.6: Percentage of sample providing 'Negligible', 'Nudge' and 'Ignored' responses to the presence of both error types (both false positives and false negatives)

10 Experiment Six

10.1 Introduction

Experiment five demonstrated that the capacity of the general public to undertake the calculations required for even the simplest legal match case, when the vital error margins of forensic testing are taken into account is extremely limited. Given that the statistical presentation of DNA in the form of the random match probability is likely to remain a common occurrence within the legal system (Koehler et al., 1995; Schklar and Diamond, 1999), the best approach to presenting the output of Bayesian calculations, removing the need for untrained individuals to undertake the calculations themselves, needs to be determined.

A group of authors (Fenton and Neil, 2011; Fenton et al., 2012, 2014) have recently argued forcefully that any such approach other than a presentation of the Bayesian posterior (i.e. the removal of any need to undertake a calculation by untrained individuals) to these individuals is untenable. This belief was given evidence by the results of experiment five. Fenton and Neil (2011) argue that these calculations should instead be undertaken by Bayesian Network computer programs, a view echoed by many experts in the statistical community (e.g. Dawid et al., 2007; Taroni and Aitken, 2006). However, Fenton et al. (2014) go further and argue that the causally-linked Bayesian network diagram may also make a valuable intuitive tool for presenting those statistics visually to untrained individuals. Fenton et al. (2014) argue that in comparison to event trees (the most common visual form of presentation of such statistics and advocated for legal cases by some authors e.g. Hoffrage et al. (2002)), a Bayesian Network diagram may scale more effectively with

increased complexity. They also stress that even the simplest legal case, when forensic errors are taken into account, presents a level of complexity too great for effective comprehension of any event-tree type depiction.

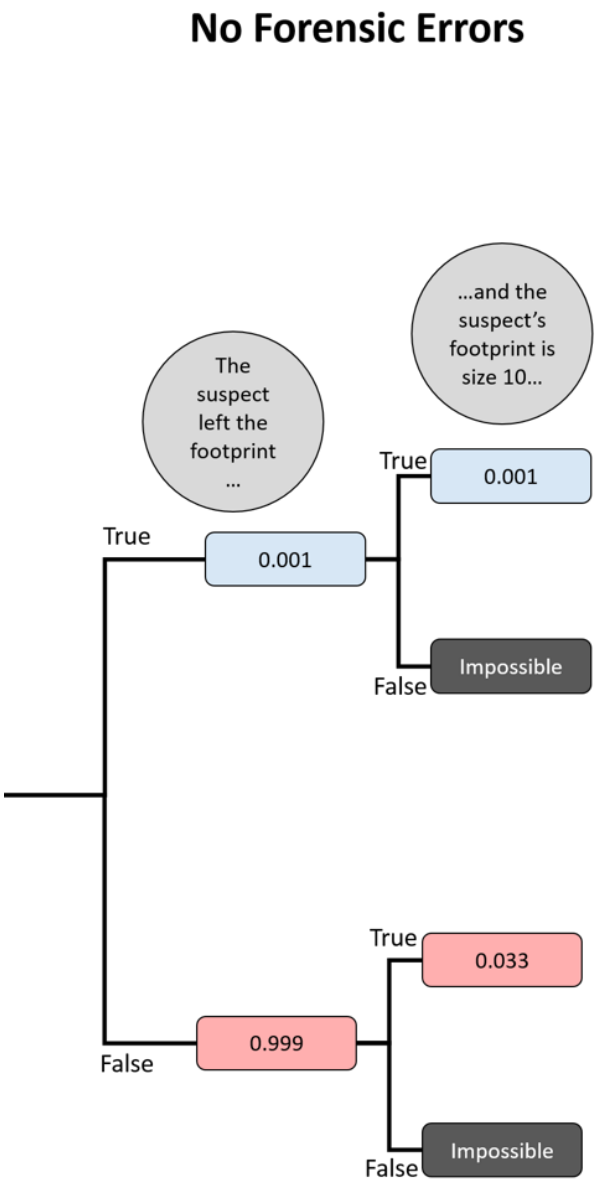


Fig. 10.1: An event-tree diagram depicting the scenario in the experiment with no testing errors: a suspect matches a footprint found at a crime scene.

In Figures 10.1 and 10.2 above the event tree comprises a series of branches each depicting a possible hypothetical scenario that could have led to the available facts in the case , along with the calculated probability of that scenario. The left diagram depicts the simple case, with no testing errors. Here only two scenarios are

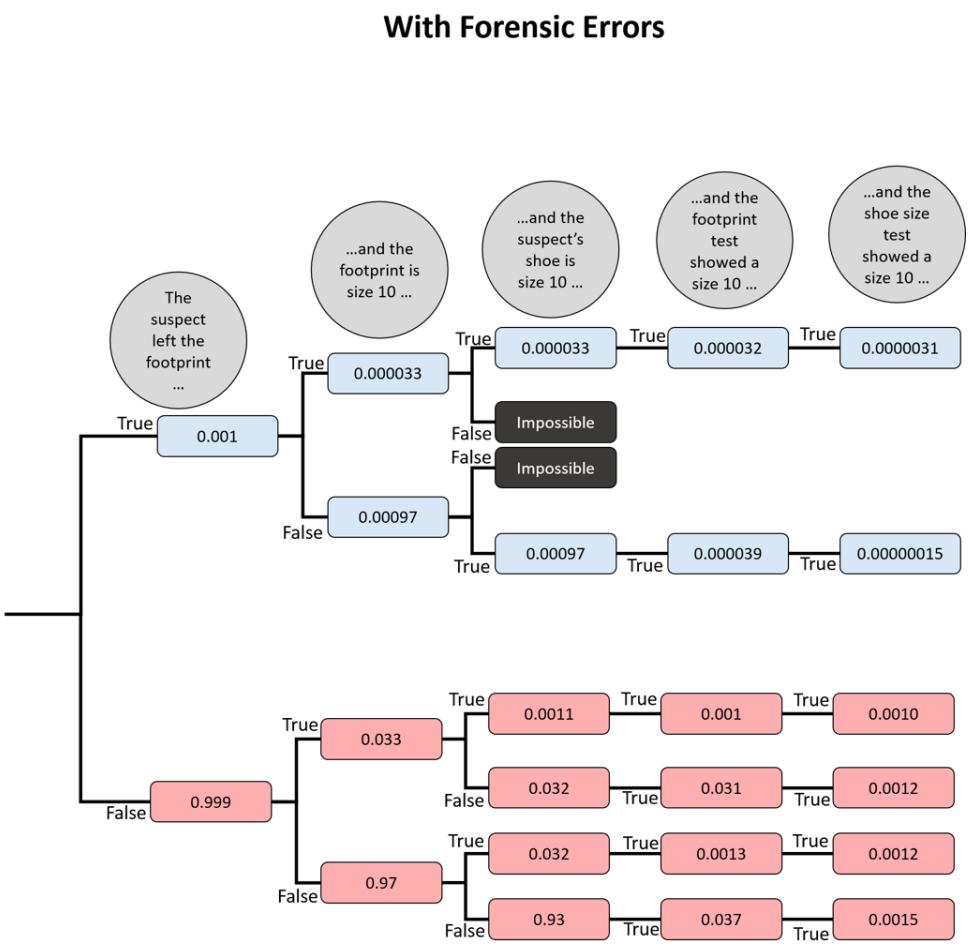


Fig. 10.2: An event-tree diagram depicting the scenario in the experiment with testing errors: a suspect matches a footprint found at a crime scene.

possible ('the defendant committed the crime' or 'the defendant did not commit the crime but coincidentally has the same footprint size'). In the right diagram, when testing errors are taken into account, far more potential scenarios are possible (e.g. the defendant did not commit the crime, does not have the shoe size on the body, and the footprint test on the defendant was a false positive). While the event tree diagram is concise and clear in the simple case (the left diagram), visual complexity is greatly increased when testing errors are included (the right diagram).

Bayesian network diagrams do not suffer from this issue. For example, the Bayes nets in Figures 10.3 and 10.4 depict the scenario used in the present experiment both without forensic errors (left) and with them (right). In the scenario a suspect is found to match a footprint found at a crime scene and the task is to infer the probability that the suspect is actually the source of that footprint. Each 'box'

in the diagram represents a conjecture in the scenario, and denotes the percentage likelihood that that particular conjecture is true. An arrow indicates that a given conjecture (the source of the arrow) has a causal effect on the likelihood of another (the terminus of the arrow) being true. In the left diagram (no errors) we can see that a forensic test showing that the defendant is a size 10 (stated in the scenario) entirely (100%) implies that the defendant really is a size 10. In the scenario with the potential for forensic errors (right), the test still depicts the same factor, but we see that we can no longer be 100% certain that this means the suspect really is size 10 (the forensic team may have made a mistake). The calculations undertaken by a Bayesian Network program utilise these causal relationships between factors to infer the probability of a given conjecture being true (e.g. that the defendant is the source of a footprint) given other facts (e.g. that the defendant is found to match the footprint and that 1/20 individuals in the population also match that footprint). Unlike the event tree diagrams (Figures 10.1 and 10.2), the Bayes net diagrams (see Figures 10.3 and 10.4 below) do not substantially change when errors rates are introduced into the problem. With errors included (right), it remains a 5-node diagram and instead the percentage values in the boxes change to reflect the new information.

The present experiment will seek to test the prediction made by Fenton et al. (2014) that a Bayesian network presentation will result in greater trust and understanding compared to an event tree presentation when errors are taken into account. This will be achieved through direct comparison of the event tree and Bayesian Network presentation methods in both basic and error versions of a legal match case. A further control condition will be added in which a simple statement that 'Computer analysis has calculated the value of the defendant being the source of the footprint to be X%' will also be included. The principal function of this control condition is to ensure that trust in computational analysis is not the reason for any beneficial effect of a Bayesian network presentation. Participants' levels of trust and understanding of each explanation given to them will be measured as the outcome measure. It

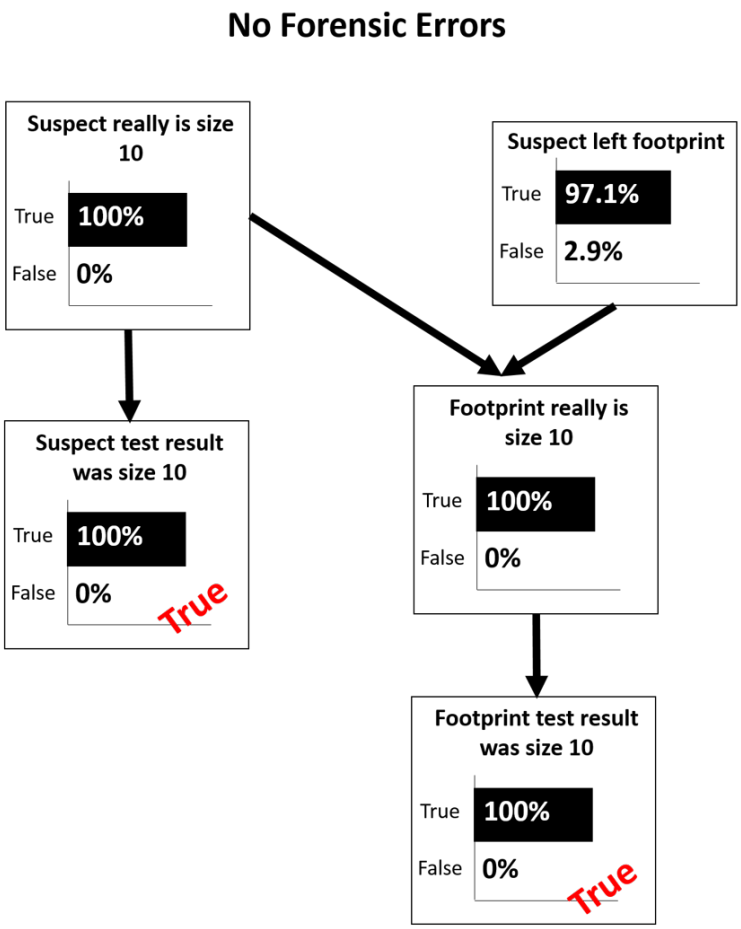


Fig. 10.3: A Bayesian Network diagram depicting the scenario in the experiment with no testing errors: a suspect matches a footprint found at a crime scene.

is hypothesised that the event tree condition will produce equal or greater trust and understanding than the Bayesian condition with the basic, no-errors scenario, however it is further hypothesised that the Bayesian Network condition will produce greater trust and understanding than the event tree condition when forensic error rates are included. No hypothesis is made regarding the control condition in the basic scenario, however it is hypothesised that the Bayesian Network condition will produce greater trust and understanding compared to the control condition in the scenario with errors. Finally, the Berlin numeracy measure (Cokely et al., 2012) will be employed to ensure that any effects found are appropriate for presentation to individuals across the numerical ability spectrum.

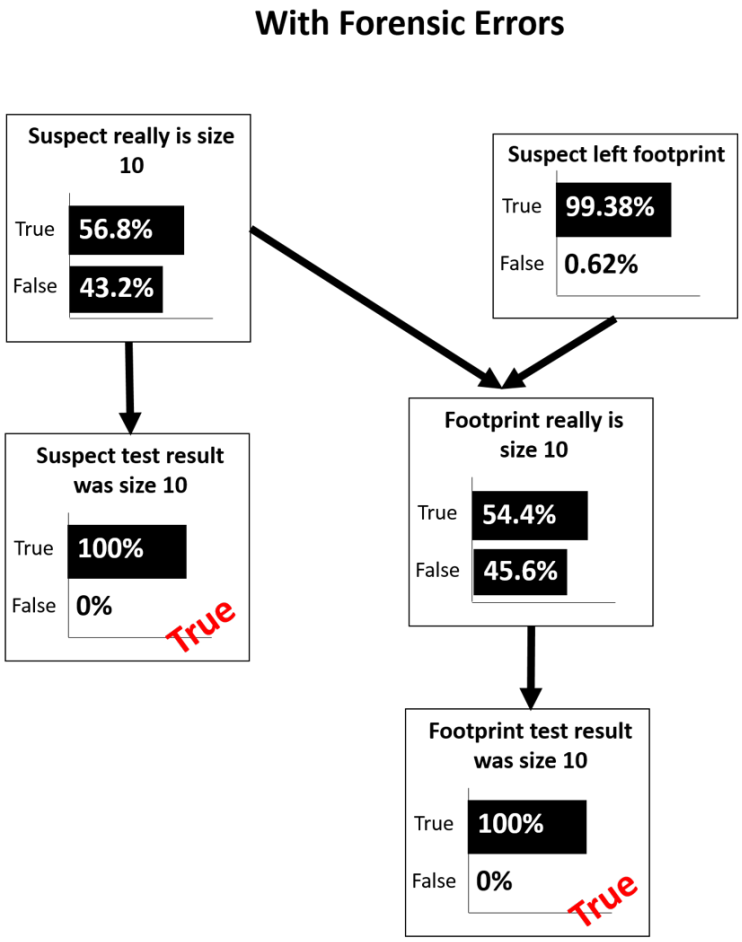


Fig. 10.4: A Bayesian Network diagram depicting the scenario in the experiment with testing errors: a suspect matches a footprint found at a crime scene.

10.2 Method

10.2.1 Participants

Three hundred and sixty four participants took part in the experiment. No participants were removed. Participants were recruited from the Amazon MTurk outsourcing service and were required to be in the United States of America and have a 95% HIT approval rating. The demographics for experiments five and six can be seen in Table 2.

10.2.2 Design

The experiment comprised three conditions to which participants were randomly assigned: the control condition, the event-tree condition and the Bayesian network condition. Each condition comprised two distinct sections, both of which all participants undertook: the basic scenario, with prior and random match probability statistics, and the errors scenario, which extended this by adding false positive and false negative error rates.

10.2.3 Materials

The full visual diagrams used in the experiment can be seen in Appendices C and D. Presentation materials were created through an iterative process involving the production of a series of refinements by the author and feedback from expert statisticians. The scenario used in the problem and presented to participants can be seen below:

> Imagine a remote Island with 1,000 inhabitants. No one ever comes or goes. One day, a man is found dead a couple of miles outside the main village.
>
> The resident law enforcer examines the crime scene and concludes that the man has been murdered. He finds two sets of footprints: the victims and a second set, leading away from the body which must have come from the murderer. He measures these footprints and finds that they are size 10. The law enforcer heads back to his office at the village and checks records from the last 50 years on shoe sizes on the island. He finds that on average only 1 in 30 Islanders have that shoe size. Word gets around the island about the footprints and as the law enforcer leaves his office, he sees a group of Islanders pushing a man in front of them. When they reach the law enforcer, they push the man onto the ground and explain that he has size 10 shoes and so he must be the one who left the footprints. The law enforcer measures the mans feet and agrees they are indeed size 10. He arrests the man as a suspect in the murder, and puts him in a cell pending a public trial.
>
> Now imagine you are a juror at the trial of this man. Given that the shoe size is the only evidence against him, what do you think is the chance that he is the source of the footprints leading away from the body?

10.2.4 Procedure

Participants accessed the survey from their own computers. They were presented with the consent form followed by the instructions for the experiment. They were then randomly assigned to one of the three conditions. Following this they were presented with the basic scenario, including the backstory, prior and random match probability values. Participants were then shown an explanation of the 'correct' answer to the problem. This explanation, but not the scenario itself, varied between the three conditions.

Following this explanation, participants were asked to provide a rating for two questions (with the following rating scales):

1. To what extent do you trust that the result just given is correct? [Distrust Completely; Distrust Moderately; Distrust Slightly' Neutral; Trust Slightly; Trust Moderately; Trust Completely]

2. To what extent do you understand how the result was calculated? [Don't Understand Completely; Don't Understand Moderately; Don't Understand Slightly; Neutral; Understand Slightly; Understand Moderately; Understand Completely]

Following the first explanation, participants were presented with further explanatory dialogue and the false positive and false negative statistics. They were then again showed a condition-dependent explanation of the correct answer to the problem, taking into account these new values and were again asked the same trust and understanding questions. Following this, participants were presented with the Berlin numeracy scale (Cokely et al., 2012). Finally, participants were presented with the demographic questions and thanked for taking part in the study.

10.3 Results

In Figure 10.5 below, average responses to the trust and understanding questions can be seen for both the basic scenario and including the information about errors. We can see that for average trust levels, the control (4.8) and tree (5.0) conditions are both similar but larger than the Bayes net condition (4.3) in the basic scenario. In the errors scenario however, both control and trees show similar trust levels (control: 4.7; tree: 4.5), while the Bayes net condition shows substantially higher levels (5.2).

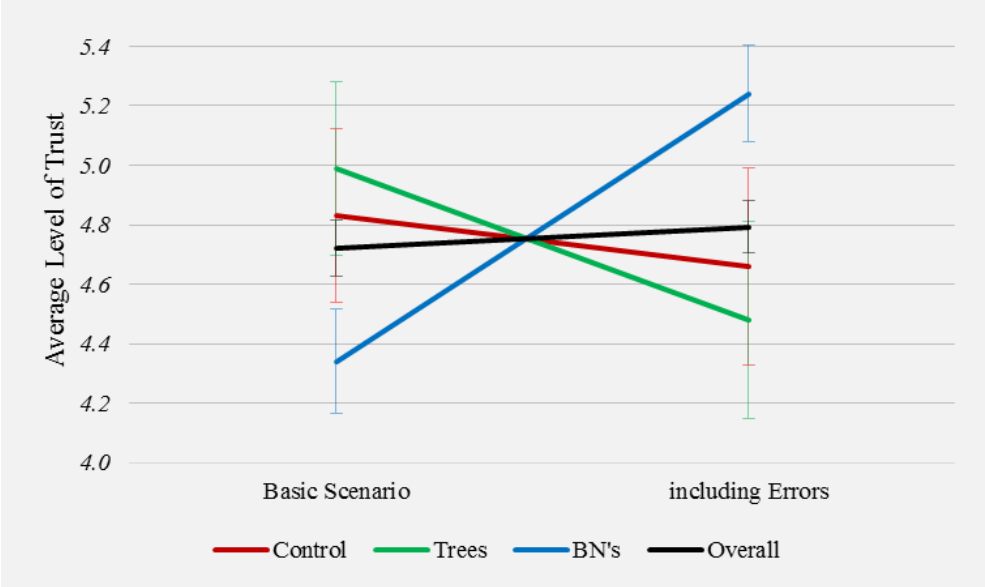


Fig. 10.5: Average level of trust for the basic and 'with errors' scenario across all three conditions and combined

This effect was confirmed with a mixed-effects ANOVA using trust as the dependent variable, scenario as the within subjects variable with 2 levels (Basic vs Errors) and in both ANOVA's Condition as a between subjects factor with 3 levels (Control, Trees, Bayesian Networks) and further, Berlin Numeracy Score as a between subjects factor with 8 levels (each of the possible total scores in the test: 0-7).No linear effect of scenario on trust was found (p=.286), but an interaction between scenario and condition was found ($F = 13.229$ (2, 340), $p<.001$). No interaction between scenario and Berlin score was found (p=.354) or between scenario, Berlin score and condition (p=.548). The average trust scores for individuals high on numeracy (4+) and low(<4) and across all three conditions and both scenarios can be seen below

in Figure 10.6

Paired-samples tests were used as post-hoc tests to analyse change in trust levels for each condition from before the errors were introduced to afterwards. No significant change in trust was seen in the control group (t [1, 126] = 1.016, p=.311), a significant decrease was seen in the event-tree group (t [1, 115] = -2.905, p=.004) and a significant increase was seen in the Bayes' net group (t [1, 120] = 4.357, p<.001).

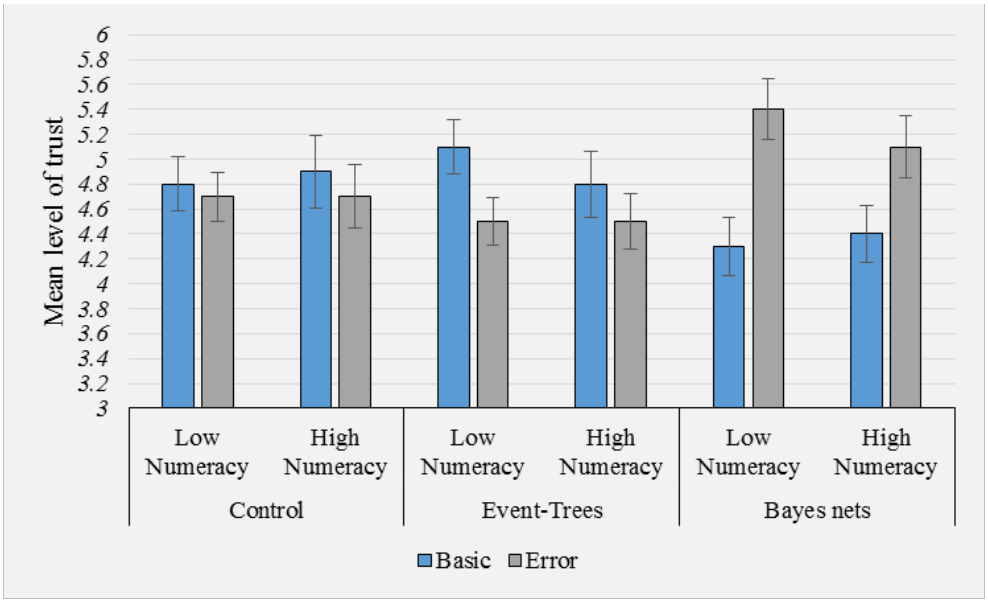


Fig. 10.6: Average level of trust for high and low numerates across all three conditions and both scenarios

Below we can also see that for average understanding levels, all three conditions show very similar understanding levels in the basic scenario (control: 4.7; tree: 4.6; Bayes net: 4.5), however in the scenario including errors, the control (4.1) and tree conditions (3.7) showed a substantial decrease in understanding while the Bayes net condition (4.6) showed a mild increase.

This effect was confirmed with a mixed-effects ANOVA's using understanding as the dependent variable, scenario as the within subjects variable with 2 levels (Basic vs Errors) and in both ANOVA's Condition as a between subjects factor with 3 levels (Control, Trees, Bayesian Networks) and further, Berlin Numeracy Score as a between subjects factor with 8 levels (each of the possible total scores in the test: 0-7). A negative linear effect of scenario on understanding was found (F=10.203 (1, 340), p=.002), and an interaction between scenario and condition (F

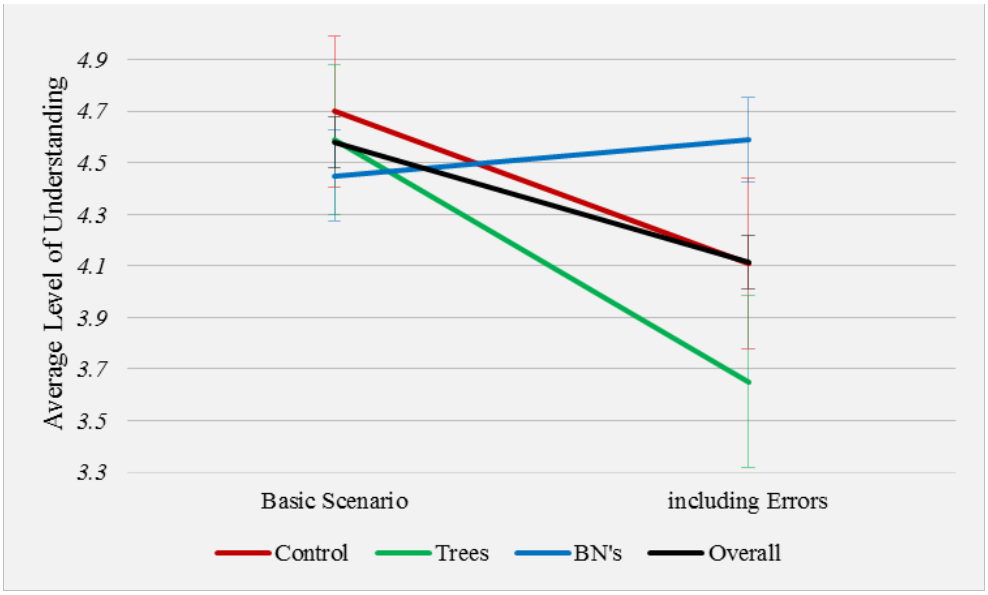


Fig. 10.7: Average level of understanding for the basic and 'with errors' scenario across all three conditions and combined

= 5.018 (2, 340), p=.007). No interaction between scenario and Berlin score was found (p=.841) or between understanding, Berlin score and condition (p=.648). The average understanding scores for individuals high on numeracy (4+) and low(<4) and across all three conditions and both scenarios can be seen below in Figure 10.8.

Paired-samples tests were used as post-hoc tests to analyse change in understanding levels for each condition. A significant decrease in understanding was seen in the control group (t [1, 126] = -3.352, p=.001), a significant decrease was seen in the event-tree group (t [1, 115] = -4.721, p<.001) and no significant change was seen in the Bayes' net group (t [1, 120] = .835, p=.405).

Based on this graph, an exploratory analysis was conducted which separated the sample into high (4+) and low (<4) numeracy levels. In the low numeracy group, a mixed-effects ANOVA revealed an effect of scenario (F = 8.227 (1, 194), p=.004) and an interaction between scenario and condition (F = 3.976 (2, 194), p=0.20).

In the high numeracy condition, a mixed-effects ANOVA also revealed an effect of scenario (F = 12.271 (1, 164), p=.001) and an interaction between scenario and condition (F = 5.633 (2, 164), p=.004).

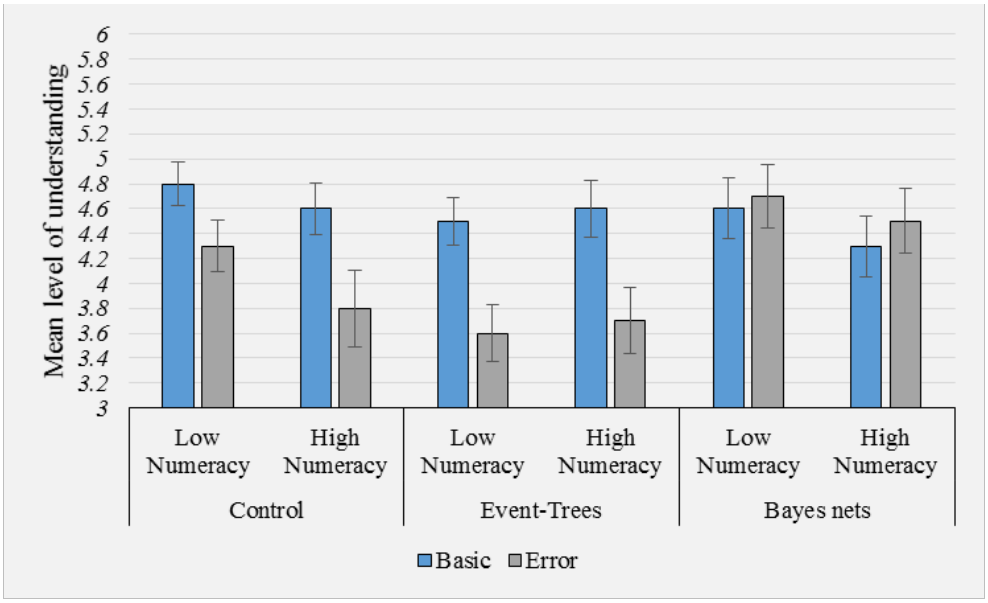


Fig. 10.8: Average level of trust for high and low numerates across all three conditions and both scenarios

10.4 Discussion

The principle aim of the present experiment was to compare the Bayesian network and event tree methods of presenting the statistics involved in a legal match cases to untrained individuals both with and without testing errors, to determine which engendered the greatest trust and understanding. In the basic scenario, both the event tree and control conditions outperformed the Bayes' net condition on trust, while all conditions showed similar values on understanding. However, this trend reversed for the scenario with errors, wherein the Bayes' net condition showed greater trust and understanding than either the event-tree or control conditions. The reversal occurred for trust due to a slight drop in trust for the tree condition, little change in the control condition and a substantial increase for the Bayes' net condition. The reversal occurred for understanding due to a substantial decrease in understanding for the event-tree and control conditions but no change for the Bayes' net condition. This confirms the main hypotheses of this experiment. While without errors the event tree diagram produces greater trust and equal understanding to Bayes nets, when these are included, as they must be in any legal analysis (Fenton et al., 2014), the Bayes net diagram engenders greater understanding and trust in a general au-

dience. Further, the effect of the Bayes net presentation is unlikely to be merely due to trust in computer analyses, as the Bayes' net condition also outperformed the control condition on both trust and understanding in the errors scenario. An analysis of Berlin numeracy scores showed no effect on trust and understanding and no interaction with condition or scenario. Confirming this, a pair of graphs showed remarkable consistency of the main findings within the high and low numeracy sub-groups. Within both groups the control and event-tree conditions showed little change in trust from basic to errors, while the Bayes net condition increased in trust from basic to errors in both numeracy sub-groups. On the understanding measure, a decrease was seen between basic and errors scenarios in the control and event-tree groups in both sub-groups and again no substantial change was seen in the Bayes net condition in either sub-group. This suggests that the findings above in regards to the superiority of the Bayes net presentation for trust and understanding when including errors is not confined to any particular numeracy sub-group but should instead be recommended for all individuals over the event-tree format. When looking at the diagrams presented to participants (see Figures 10.3, 10.4, 10.1 and 10.4) the pattern of understanding and trust levels seen is clearly coherent with the changes in the complexity of the two diagrams from the basic to the errors scenario. The event-tree diagram clearly increases in complexity from the basic (two important branches) to the errors scenario (six important branches). This correlates with the significant decrease in understanding and trust seen in this condition from basic to errors scenarios. Unfortunately this increase in complexity is an inherent feature of the format, and the version used in the present experiment is actually the most minimal version possible with as many unnecessary branches as possible removed for visual simplicity. When looking at the Bayes net presentation, it is clear that no substantial change in complexity to the overall diagram occurs between scenarios. The Bayes net diagram comprises five boxes in both scenarios, and it is only the numbers within those boxes which change. The fact that understanding of the Bayes net explanation did not change between scenarios is therefore perhaps not

surprising. More mysterious however is that trust in the Bayes' net condition in fact significantly increased. The fact that the Bayes' net condition showed lower trust levels even than the control condition may shed some light on this. The Bayes' net presentation included two largely 'superfluous' boxes in the basic scenario (see Appendix D): as there were no testing errors two boxes were redundant and could have been removed. The rationale behind this retention was to create greater coherence between the basic and errors scenario, and so that participants could see that these two boxes 'change' when errors were included. It is theorised therefore that the presence of these boxes may have reduced trust in the basic scenario, and the revelation of their use may have increased trust in the errors scenario. This retention of complexity was not possible with the event-tree format.

11 Part II Discussion

Experiment five demonstrated that the general public are not capable of undertaking accurate statistical reasoning in legal scenarios when vital forensic testing errors are included. Instead, participants use a set of heuristics which mis-estimate the impact of the error rates. Analysis of the think aloud data suggest that no single individual incorporated the error rates into their calculations in a mathematically appropriate manner. This suggests, in line with Fenton et al. (2014), that jurors and legal professionals should not attempt to undertake such calculations themselves, and that this may be best left to trained statisticians employing statistical computer programs such as Bayesian networks.

Building upon these findings in experiment five, experiment six sought to test two methods for presenting the output of such calculations to the general public: the popular event tree format and the Bayesian network diagram approach recommended by Fenton et al. (2014). As hypothesised, in the simple match scenario, without errors, the event tree was either equal or superior to the Bayesian network diagram in terms of the level of trust and understanding in the calculations that it engendered in participants. However, also as hypothesised, when the necessary forensic errors were included in the problem, a complete reversal of this pattern was seen, with the Bayesian network producing higher levels of both trust and understanding than the event tree diagram. This was theorised to be because of the increased visual complexity of the event tree diagram when testing errors are added. The Bayesian network diagram does not suffer from any such increase in complexity, and results demonstrated that participants in fact trusted the Bayesian network diagram significantly more when testing errors were added than when they were

absent.

In line with Fenton and Neil (2011); Fenton et al. (2012, 2014) it is argued that these findings demonstrate that, firstly, jurors should not be asked to undertake Bayesian calculations by hand. This echoes the message of an account of a disastrous attempt to do exactly that by Donnelly (2005) which may have led to a serious miscarriage of justice. Secondly, they suggest that if such calculations must be completed on behalf of the jury then the Bayesian network diagram is recommended for this purpose instead of the currently more-commonly-used event-tree diagram. As well as providing higher levels of both trust and understanding in comparison to the event-tree diagram, this approach has an additional benefit which was not in fact mentioned by Fenton et al. (2014). This is the coherence between the calculations being undertaken on behalf of the jury and the presentation of those calculations to that jury (the presentation is a true representation of the calculations being undertaken). This may reduce issues surrounding the jury being 'misled' by the statistician, a concern which has motivated previous disastrous attempts to have the jury complete calculations by hand such as in the R v Adams case (see Donnelly, 2005).

12 General Summary

This thesis has aimed to determine the most effective methods of presenting the statistics involved in Bayesian problems to a general audience. It has done this within a general framework in part one, and with a more specific focus on presentation in legal trials in part two.

12.1 Part I

In the first experiment in part one, two leading theories from the field focused on presenting Bayesian word problems were, for the first time, compared and combined in a single experiment using a within subjects design. Out of these two approaches, the outside-framed percentage approach of Macchi (2000: used to represent the many different 'natural frequency' and 'nested sets' approaches) was shown to significantly improve accuracy over and above an inside-framed percentage or probability approach. However, the causal approach of Krynski and Tenenbaum (2007) showed no significant increase over the control condition. In a second between-subjects experiment (experiment three) this null finding was replicated when removing several possible confounding factors and with a closer design to Krynski and Tenenbaum's original.

In the first experiment, the design was supplemented with a think aloud protocol, which asked participants to record their thought processes while they solved the Bayesian problem. Analysing this data using qualitative techniques, a five-stage process was discovered to be highly modal among successful participants. While this process was modal amongst successful solvers in all conditions, including the

control and causal conditions, it was found in higher frequency in the nested sets condition. In experiment two, Macchi's (2000) approach, and the presence of the newly-discovered nested sets process were tested in a between subjects design using a full percentage format (experiment one used real-number base rates) and which further crossed problem difficulty (simple versus complex) with the use of decimal numbers (whole number vs decimal). In all conditions, Macchi's approach outperformed the control condition, and this effect again coincided with a greater increase in individuals following the 'nested sets' process. A mediation analysis demonstrated that the nested sets framing effect on accuracy was mediated by the increase in frequency of individuals following the 'nested sets' process. This experiment also found that many successful (but very few unsuccessful) individuals converted the problem from percentages to integers, most often converting from a base of 100% to a base of 100 women, making no mathematical change to the problem. This gave some evidence for a conjecture put forward by previous work (e.g. Hoffrage et al., 2002) that the nested sets approaches (e.g. Evans et al., 2000; Fiedler et al., 2000; Macchi, 2000; Sloman et al., 2003) work only because they encourage more people to convert to 'natural frequencies', and that this is the true reason for the increase in accuracy seen using these formats. However, it is important to note that 40% of successful participants followed the nested sets process without converting from percentages, suggesting that it is by no means an absolute necessity for individuals to do this in order to solve Bayesian problems. While this finding could be interpreted in line with an evolutionary 'frequentist' perspective (Cosmides and Tooby, 1996; Brase, 2007) it may also merely demonstrated a greater familiarity on the part of many individuals with integers. Future work may be valuable in testing whether familiarity with percentages or integers moderates this relationship.

Experiments one and two also demonstrated that the majority of individuals who failed were doing so at the first stage of the nested sets process, with a substantial but smaller number failing in the middle stages, and very few failing in the final stages. They also demonstrated that the benefit from Macchi's outside-framed approach

may be entirely due to the changes made to the body of the problem, which was thought to be assistive at this earlier stage, and that the further changes to question form (which were assistive at later stages) may be superfluous. Further, qualitative analysis of the think aloud data in experiment two demonstrated that the most common error made was due to a confusion of the question being asked, again suggesting that this part of the problem in Macchi's approach could be further refined.

In a fourth experiment the correlational nested sets process and conversion-to-real-number findings of experiments one and two were tested experimentally in a crossed design. The provision of a real-number population, designed to encourage individuals to 'convert' the problem to integers, showed no increase in accuracy. This suggested that the causal story developed following the results of experiment two may be false: instead of conversion leading to success, individuals who are already destined to be successful may subsequently be more likely to convert the problem, possible to aid their working through of the mathematics. The provision of two sets of leading questions designed to encourage individuals to follow the nested sets process did show a significant increase in accuracy. Finally, this experiment as well as experiment three demonstrated that the mere presence of a write aloud protocol also substantially increases accuracy on Bayesian problems.

12.2 Part II

While the 'outside-framed' nested sets approach was widely advocated in Part one of the thesis, in Part two, a situation (the presentation of evidence in law) where the nested sets approach is not possible was analysed. Two factors contribute to the inappropriateness of the approach here. Firstly, the legal trial is inherently focused on a single individual and by design, a reference class cannot be used and is considered prejudicial. Secondly, the legal situation, when properly analysed, has a level of complexity too great for statistically untrained individuals to work with, regardless of framing.

In the first experiment in part two, the question of whether a general population sample could accurately undertake Bayesian inference of a legal match case including necessary forensic errors (see Fenton et al., 2014, for a discussion) was tested. The clear conclusion was that this was not possible, with no single participant making an accurate Bayesian inference when errors were included. Instead, participants were found to use a set of heuristics in response to errors. The first set of these involved stating that the error margins were so negligible that they could be ignored or ignoring them entirely (omitting to mention them). When the error rates were incorporated, the most common approach individuals took was to 'nudge' their answer from their previous calculation by several percentage points. While overall, participants' responses represented an over-weighting of the error statistics, the 'negligible' and 'ignore' response clearly under-weighted both (particularly the false positive rate) and the 'nudge' response under-weighted the false positive rate but over-weighted the false negative rate.

Given the findings of experiment one, experiment two sought to compare two methods of presenting the outcome of a Bayesian analysis of a legal match case with testing errors, precluding the need for individuals to make the calculations themselves. The two methods compared were the popular event-tree diagram and the Bayesian network diagram proposed by Fenton et al. (2014). A further control condition was used to rule out trust in 'computer analysis' as a factor in any success seen for the Bayesian network diagram. When participants were shown the 'simple' case with no forensic testing errors, the Bayesian network diagram was rated as either equal to or less than both the control and event-tree diagram presentations for both subjective trust and understanding ratings. However, when the testing errors were added, the Bayesian network diagram resulted in higher trust and understanding than either control or event-tree diagrams. This suggested that when a legal Bayesian problem takes into account the necessary levels of complexity, a Bayesian network diagram will engender greater trust in the result and subjective feeling of understanding in how it was calculated, than an event tree diagram presentation.

12.3 Overall Conclusion

The present thesis has tested a range of presentation methods across a range of domains where the general public frequently have to solve Bayesian problems. Among the numerous 'nested sets' approaches available, including the use of 'natural frequencies' (Gigerenzer and Hoffrage, 1995), the presentation of the necessary statistics in a Bayesian problem as percentages framed from the point of view of a group (Macchi, 2000), and sub-divisions of that group, has been extensively tested and validated. In part one, this nested sets approach has been advocated as the most reliable method to improve accuracy wherever possible. In Part two, a situation (the presentation of evidence in law) where the nested sets approach is not possible was analysed. Here, particularly due to the complexity of an appropriately-analysed legal case, the approach advocated was to engender trust and understanding in the output from a Bayesian network analysis of the case as manual calculation was demonstrated to not be plausible, with no participants able to achieve a normative result.

Across both parts one and two, through qualitative analysis of think aloud data, the thesis has provided a deeper understanding of the processes individuals go through across a range of problems, and the reasons they both succeed and fail, than has previously been possible. These analyses give strong suggestions for the direction of future work and provide a framework on which improvements to current approaches can be made. Where individuals can undertake Bayesian reasoning with reasonable success, this thesis has demonstrated that the nested sets or natural frequency method is most efficacious. In the legal realm, where the Bayesian problems that individuals undertake are too complex, the thesis has demonstrated that a Bayesian network presentation of the outcome of the Bayesian analysis engenders the greatest trust and understanding in statistically untrained individuals. In experiment five, a similar, but importantly different 'base rate neglect' (Bar-Hillel, 1980) phenomenon was found to that in experiment two. While in experiment two,

participants appeared to 'confuse' P(D|H) with P(H|D), participants in experiment six appeared to confuse the inverse value, P(D|-H) with P(H|D). While at first a surprising result, this may provide insight into participant process, and the base rate neglect phenomenon in general. This is because participants in experiments one, two and three were provided with the figure P(D|H) in the text of the problem while participants in experiment six were provided with the figure P(D|-H). This may therefore be explained by participants simply 'picking' the figure from the problem that appears most like the figure being requested in the question. According to fuzzy processing theory (Wolfe, 1995), statistics like these are represented in the mind in a variety of levels of coarseness, and, the less engagement / time to process the figures, the greater the coarseness of representation. At it's most coarse, participants may represent all of these figures as 'D's and H's go together' (Reyna and Brainerd, 1991). At this level of processing, the value P(D|H) and the value P(D|-H) may be indistinguishable from P(H|D). This explanation fits with the overwhelming finding from part one that the mere implementation of a think aloud protocol (which encouraged individuals to process the problem more deeply and consciously by forcing them to write down their thought process) produced a large increase in accuracy, and reduction of the base rate neglect error. In general therefore, qualitative data from both parts of the thesis give evidence against the traditional base rate neglect view that individuals ignore base rates as they do not believe them important (Bar-Hillel, 1980; Kahneman and Tversky, 1972; Welsh and Navarro, 2012; Obrecht and Chesney, 2013; Pennycook and Thompson, 2012) and in favour of the 'confusion hypothesis' (e.g. Braine and Connell, 1990; Cohen, 1981; Dawes, 1986; Eddy, 1982; Hamm and Miller, 1990; Fiedler et al., 2000).

In terms of individuals' preferred mode of solution, the findings of experiment six echoed and extended two core proposals from part one. Firstly that, when using a simple problem, a 'nested' method of presentation is preferred. Secondly that, while in the simple case a causal approach will not be effective (and will be inferior to a nested sets approach), this pattern may reverse in the complex case. The

first of these was demonstrated in the fact that higher levels of trust were seen for the 'event tree' presentation method than the Bayesian network presentation in the simple case. The second proposal was demonstrated in the fact that, when the full complexity of the legal case was taken into account, this method was found to be less trusted and less understood than a Bayesian network presentation, which is an inherently causal method of presentation.

In general therefore, where a reference class is possible, and where complexity is not too great for manual calculation, a nested sets approach (either with outside-framed percentages or natural frequencies) is advocated by this thesis. The situation in which an individual is faced with understanding their risk of having a disease after routine screening, which the vast majority of people will face at some point in their lives, is a good example of this type of situation. Where complexity is too great for individuals to perform manual calculations, such as when analysing a legal case, the output of computations from a Bayesian network presented in a causal, nodal construction has been shown to engender high levels of trust and understanding in experiment six, and is therefore advocated.

Future work should focus on developing Macchi's (2000) outside-framed approach to presenting 'simple' Bayesian problems. A particular focus should be made on improving the format of the question asked, possibly in line with Girotto and Gonzalez's (2001) approach. Further improvements should be informed by the fact that individuals appear to follow the here-outlined nested sets process when solving Bayesian problems. It may therefore be most sensible to design further presentation methods with an aim to encouraging problem solvers to more easily follow this process, particularly by encouraging them to perceive the Hypothesis-focused representation of the problem, which seems to be the greatest predictor of success: once individuals perceive this representation of the problem, their chance of successful solution is considerably higher.

When presenting highly complex Bayesian problems, future work should focus on extending the Bayesian network diagram approach and taking more detailed

mixed methods quantitative-qualitative data to determine why the Bayesian network diagram is so effective when false positives and false negatives are included, and further, how it can be improved. Visual complexity is proposed as a key possible factor, as the event-tree diagram increases visual complexity substantially when error rates are incorporated, while the Bayesian network diagram does not.

Bibliography

Aitken, C. (1995). *Statistics and the evaluation of evidence for forensic scientists*. Wiley.

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.

Bar-Hillel, M. (1990). Back to base rates. *Insights in decision making: A tribute to Hillel J. ....*

Barbey, A. K. and Sloman, S. a. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and brain sciences*, 30(3):241–54; discussion 255–97.

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

Bayes, M. and Price, M. (1763). An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*.

Beaumont, M. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*.

Birnbaum, M. H. and Mellers, B. a. (1983). Bayesian inference: Combining base

rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, 45(4):792–804.

Bishop, C. (2006). Pattern Recognition. *Machine Learning*.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

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

Brainerd, C. and Reyna, V. (1990). Inclusion illusions: Fuzzy-trace theory and perceptual salience effects in cognitive development. *Developmental Review*.

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

Brase, G. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*.

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. 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.

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.

Dartnall, S. and Goodman-Delahunty, J. (2006). Enhancing Juror Understanding of Probabilistic DNA Evidence. *Australian Journal of Forensic Sciences*, 38(2):85–96.

Daston, L. (1995). *Classical probability in the Enlightenment*. Princeton University Press.

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

Dawid, A. P., Mortera, J., and Vicard, P. (2007). Object-oriented bayesian networks for complex forensic dna profiling problems. *Forensic Science International*, 169(2):195–205.

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.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, 51(4):380–417.

Edwards, W. (1968). Conservatism in human information processing. *Formal representation of human judgment*.

Edwards, W. (1991). Influence Diagrams, Bayesian Imperialism, and the Collins Case: An Appeal to Reason. *Cardozo Law Review*, 13.

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.

Faigman, D. L. and Baglioni, A. J. (1988). Bayes' theorem in the trial process - Instructing jurors on the value of statistical evidence. *Law and Human Behavior*, 12(1):1–17.

Fenton, N. and Neil, M. (2011). Avoiding probabilistic reasoning fallacies in legal practice using bayesian networks. *Austl. J. Leg. Phil.*, 36:114.

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.

Fenton, N., Neil, M., and Lagnado, D. a. (2012). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, 37(1):61–102.

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.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, 23(3):339–359.

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

Frank, M. and Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*.

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.

Geman, S. and Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*. *Journal of Applied Statistics*.

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

Gigerenzer, G. (2015). *Risk savvy: How to make good decisions*. Penguin.

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.

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *Am. J. Trial Advoc.*

Grice, P. (1975). Logic and Conversation. *Syntax and Semantics*, 3:41–58.

Griffin, D. and Buehler, R. (1999). Frequency, probability, and prediction: easy solutions to cognitive illusions? *Cognitive psychology*, 38(1):48–78.

Griffin, D. and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3):411–435.

Hamm, R. (1994). Underweighting of base-rate information reflects important difficulties people have with probabilistic inference. *Psycoloquy*, 5(3).

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.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*.

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.

Hobson, M. P. (2010). *Bayesian methods in cosmology*. Cambridge University Press.

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.

Jeffreys, H. (1973). *Scientific inference*. Cambridge University Press.

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.

Kadane, J. and Schum, D. (1998). A Probabilistic Analysis of the Sacco and Vanzetti Evidence. *Technometrics*.

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

Kaye, D. H., Hans, V. P., Dann, B. M., Farley, E., and Albertson, S. (2007). Statistics in the Jury Box: How Jurors Respond to Mitochondrial DNA Match Probabilities. *Journal of Empirical Legal Studies*, 4(4):797–834.

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.

Koehler, J. (1993). Error and Exaggeration in the Presentation of DNA Evidence at Trial. *Jurimetrics Journal*, pages 21–40.

Koehler, J., Chia, A., and Lindsey, S. (1995). The random match probability (RMP) in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal*, pages 201–220.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(01):1.

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.

Laplace, P. S. and Simon, P. (1951). A philosophical essay on probabilities, translated from the 6th french edition by frederick wilson truscott and frederick lincoln emory.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.

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

Liu, A. Y. (1975). Specific information effect in probability estimation. *Perceptual and Motor Skills*, 41:475–478.

Lyon, D. and Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4):287–298.

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.

McKenzie, C. R. M. (2003). Rational models as theories - Not standards - Of behavior. *Trends in Cognitive Sciences*, 7(9):403–406.

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.

Mnookin, J., Cole, S. A., Dror, I., Fisher, B. A., Houk, M., Inman, K., Kaye, D. H., Koehler, J. J., Langenburg, G., Risinger, D. M., et al. (2011). The need for a research culture in the forensic sciences. *Northwestern Public Law Research Paper*, 11-20.

Nance, D. a. and Morris, S. B. (2002). An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability. *SSRN Electronic Journal*, pages 403–454.

Nance, D. a. and Morris, S. B. (2005). Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small RandomMatch Probability. *The Journal of Legal Studies*, 34(June 2005):395–444.

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

Neumann, J. v., Morgenstern, O., et al. (1944). *Theory of games and economic behavior*, volume 60. Princeton university press Princeton.

Niiniluoto, I. (1981). LJ Cohen versus Bayesianism. *Behavioral and Brain Sciences*.

Obrecht, N. a. and Chesney, D. L. (2013). Sample representativeness affects whether judgments are influenced by base rate or sample size. *Acta Psychologica*, 142(FEBRUARY):370–382.

Parent, E. and Rivot, E. (2012). *Introduction to hierarchical Bayesian modeling for ecological data*. CRC Press.

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

Pennycook, G. and Thompson, V. a. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic bulletin & review*, 19(3):528–34.

Phillips, L. D. and Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, 72(3):346–354.

Pollatsek, A., Well, A., Konold, C., and Hardiman, P. (1987). Understanding conditional probabilities. *Behavior and Human ....*

Pozzulo, J. D., Lemieux, J. M. T., Wilson, A., Crescini, C., and Girardi, A. (2009). The Influence of identification decision and DNA evidence. *Journal of Applied Social Psychology*, 39:2069–2088.

Redmayne, M., Roberts, P., Aitken, C., and Jackson, G. (2011). Forensic science evidence in question. *Criminal Law Review. 5*.

Reyna, V. and Brainerd, C. (1991). Fuzzytrace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*.

Rouanet, H. (1961). Études de décisions expérimentales et calcul de probabilités. *Colloques internationaux du centre national de la ....*

Saks, M. J. and Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science (New York, N.Y.)*, 309(5736):892–5.

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

Schklar, J. and Diamond, S. S. (1999). Juror Reactions to DNA Evidence: Errors and Expectancies. *Law and human behavior*, 23(APRIL 1999):159–184.

Sedlmeier, P. (1997). BasicBayes: A tutor system for simple Bayesian inference. *Behavior Research Methods, Instruments, & Computers*, 29(3):328–336.

Simon, A. H. (1956). Rational choice and the structure of enviroment. *Psychological Review*, 63:129–138.

Sirota, M., Kostovičová, L., and Vallée-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.

Smith, B. C. (1996). Jurors' use of probabilistic evidence. *Law and Human Behavior*, 20(1):49–82.

Sperber, D. (1994). *The modularity of thought and the epidemiology of representations.* Cambridge University Press.

Swets, J. A. (1964). Signal detection and recognition in human observers: Contemporary readings.

Taroni, F. and Aitken, C. (2006). Bayesian networks for evaluating scientific evidence. *... Probabilistic Inference ...*.

Taroni, F. and Aitken, C. G. (1998). Probabilistic reasoning in the law: Part 1: assessment of probabilities and explanation of the value of dna evidence. *Science & Justice*, 38(3):165–177.

Thaler, R. (1980). Judgement and decision making under uncertainty: what economists can learn from psychology. *A.E. - University of Illinois, Department of Agricultural Economics*.

Thompson, W. C., Kaasa, S. O., and Peterson, T. (2013). Do Jurors Give Appropriate Weight to Forensic Identification Evidence? *Journal of Empirical Legal Studies*, 10(2):359–397.

Thompson, W. C. and Newman, E. J. (2015). Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents. *Law and human behavior*, 39(4):332–349.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy. *Law and human behavior*, 11(3):167–187.

Thompson, W. C., Taroni, F., and Aitken, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of forensic sciences*, 48(1):47–54.

Thüring, M. and Jungermann, H. (1990). The conjunction fallacy: Causality vs. event probability. *Journal of Behavioral Decision ...*.

Tversky, a. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, 185(4157):1124–31.

Tversky, A. and Kahneman, D. (1980). Causal schemas in judgments under uncertainty. *Progress in social psychology*.

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

Walsh, S. J., Ribaux, O., Buckleton, J. S., Ross, A., and Roux, C. (2004). DNA Profiling and Criminal Justice: A Contribution to a Changing Debate. *Australian Journal of Forensic Sciences*, 36(1):34–43.

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.

Yamagishi, K. (2003). Facilitating Normative Judgments of Conditional Probability: Frequency or Nested Sets? *Experimental Psychology*, 50:97–106.

APPENDIX

A The Basic-Mammogram, Nested-College, Causal-Library and Nested-Causal-Gotham Word Problems from Experiment One

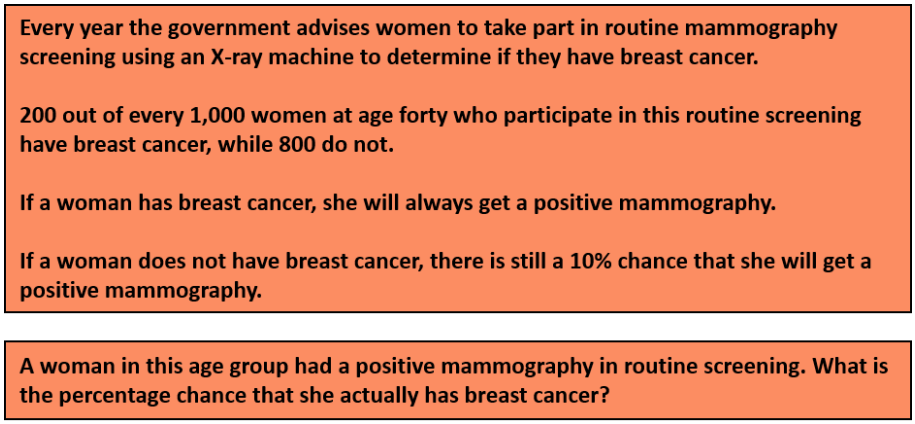


Fig. A.1: The Basic-Mammogram problem from Experiment One

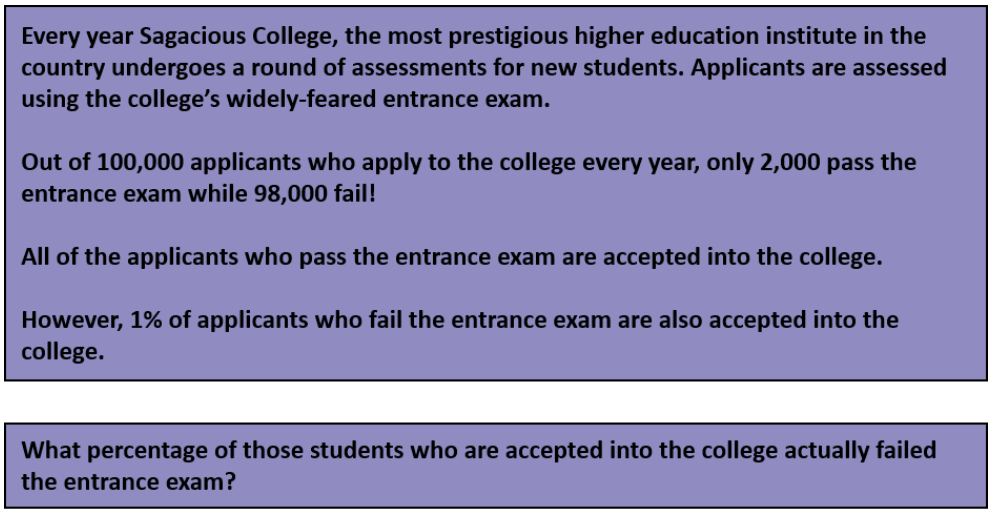


Fig. A.2: The Nested-College problem from Experiment One

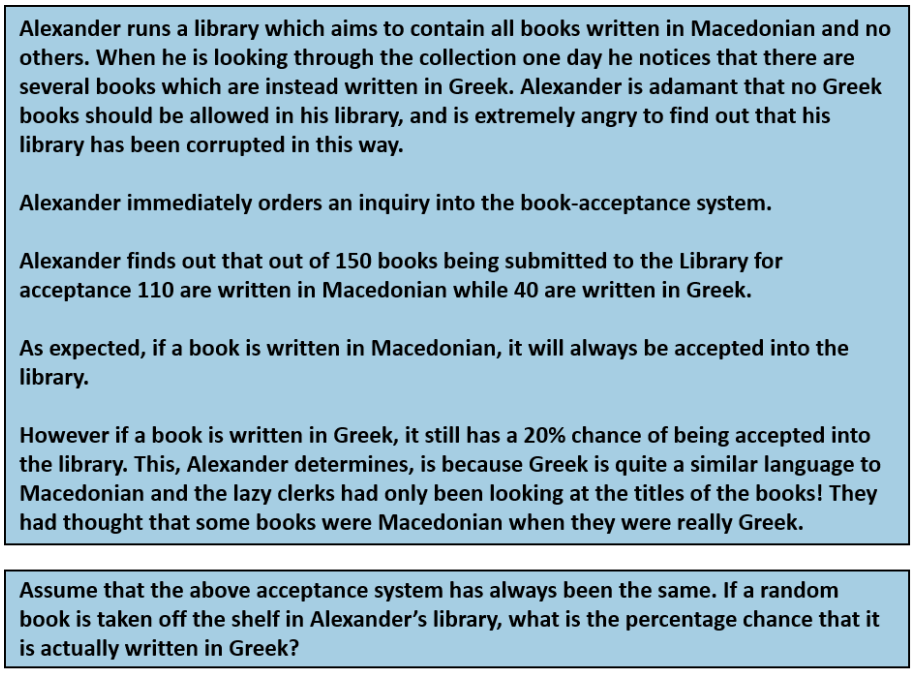


Fig. A.3: The Causal-Library problem from Experiment One

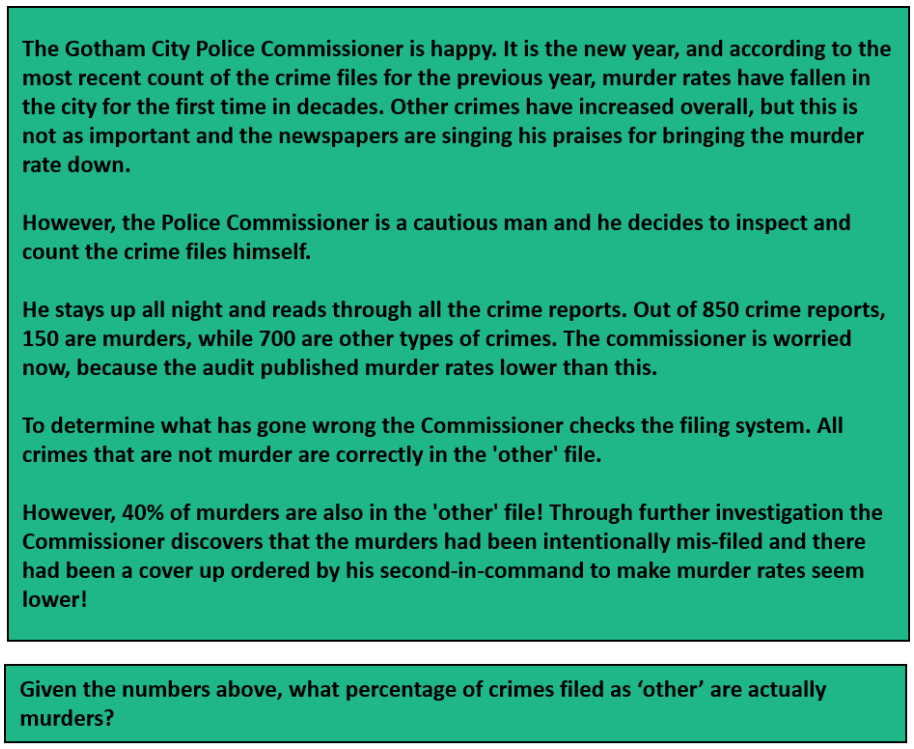
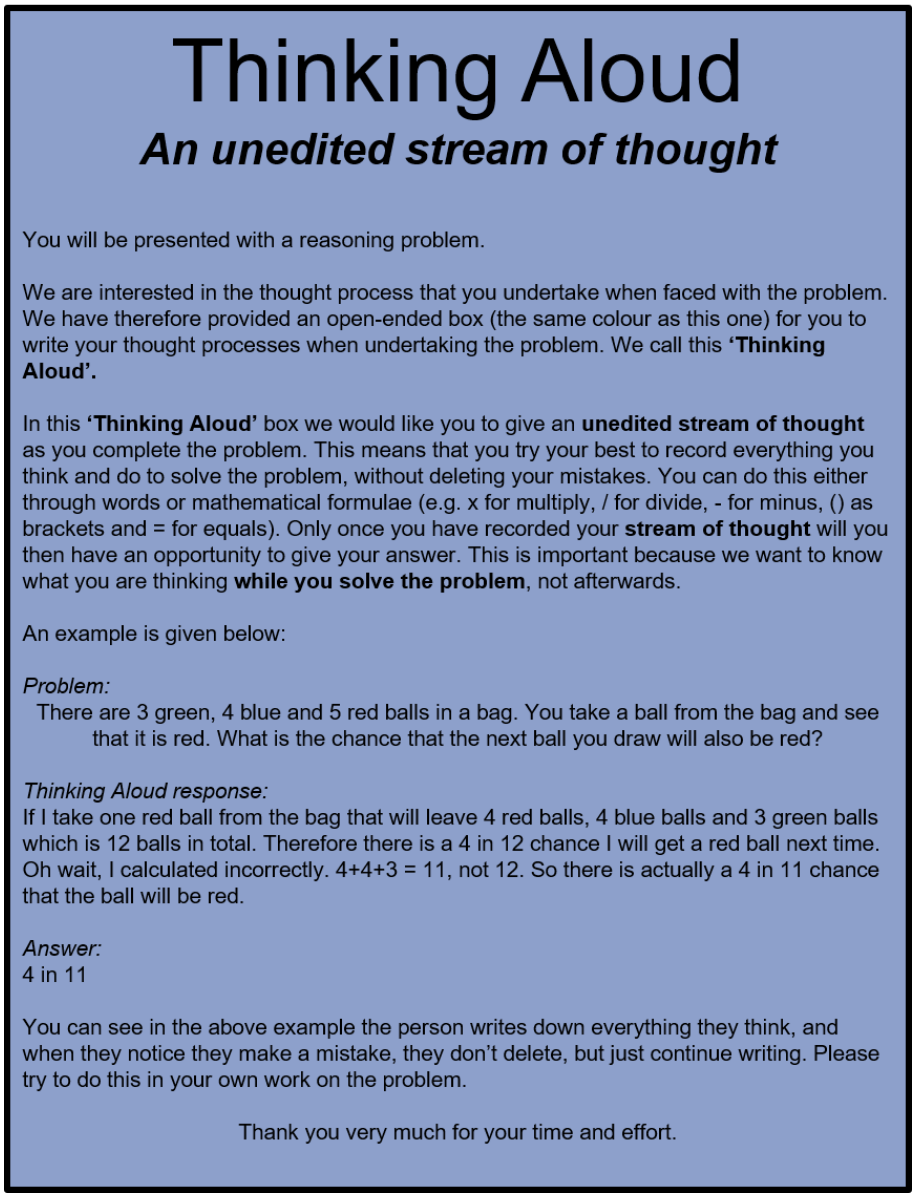


Fig. A.4: The Nested-Causal-Gotham problem from Experiment One

B The Think Aloud Instructions used in Experiments One, Two, Three and Five



Thinking Aloud

An unedited stream of thought

You will be presented with a reasoning problem.

We are interested in the thought process that you undertake when faced with the problem. We have therefore provided an open-ended box (the same colour as this one) for you to write your thought processes when undertaking the problem. We call this **'Thinking Aloud'.**

In this **'Thinking Aloud'** box we would like you to give an **unedited stream of thought** as you complete the problem. This means that you try your best to record everything you think and do to solve the problem, without deleting your mistakes. You can do this either through words or mathematical formulae (e.g. x for multiply, / for divide, - for minus, () as brackets and = for equals). Only once you have recorded your **stream of thought** will you then have an opportunity to give your answer. This is important because we want to know what you are thinking **while you solve the problem**, not afterwards.

An example is given below:

Problem:
There are 3 green, 4 blue and 5 red balls in a bag. You take a ball from the bag and see that it is red. What is the chance that the next ball you draw will also be red?

Thinking Aloud response:
If I take one red ball from the bag that will leave 4 red balls, 4 blue balls and 3 green balls which is 12 balls in total. Therefore there is a 4 in 12 chance I will get a red ball next time. Oh wait, I calculated incorrectly. 4+4+3 = 11, not 12. So there is actually a 4 in 11 chance that the ball will be red.

Answer:
4 in 11

You can see in the above example the person writes down everything they think, and when they notice they make a mistake, they don't delete, but just continue writing. Please try to do this in your own work on the problem.

Thank you very much for your time and effort.

Fig. B.1:

C The Simple Event Tree and Bayesian Network Diagrams Presented to Participants in Experiment Six

Imagine a remote Island with 1,000 inhabitants. No one ever comes or goes. One day, a man is found dead a couple of miles outside the main village.

The resident law enforcer examines the crime scene and concludes that the man has been murdered. He finds two sets of footprints: the victim's and a second set, leading away from the body which must have come from the murderer. He measures these footprints and finds that they are size 10. The law enforcer heads back to his office at the village and checks records from the last 50 years on shoe sizes on the island. He finds that on average only 1 in 30 Islanders have that shoe size.

Word gets around the island about the footprints and as the law enforcer leaves his office, he sees a group of Islanders pushing a man in front of them. When they reach the law enforcer, they push the man onto the ground and explain that he has size 10 shoes and so he must be the one who left the footprints. The law enforcer measures the man's feet and agrees they are indeed size 10. He arrests the man as a suspect in the murder, and puts him in a cell pending a public trial.

Now imagine you are a juror at the trial of this man. Given that the shoe size is the only evidence against him, what do you think is the chance that he is the source of the footprints leading away from the body?

Fig. C.1: The Simple Scenario Presented to Participants in Experiment Six

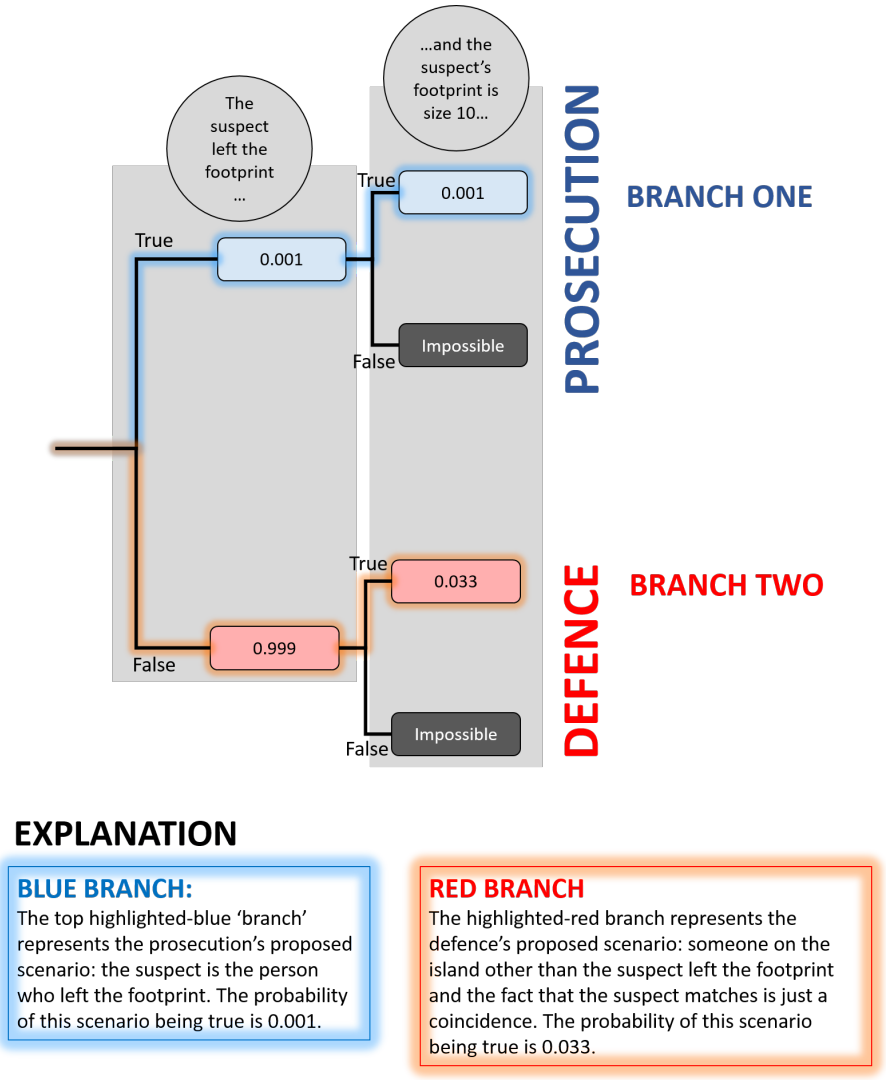


Fig. C.2: The Simple Event Tree Diagram Presented to Participants in Experiment Six

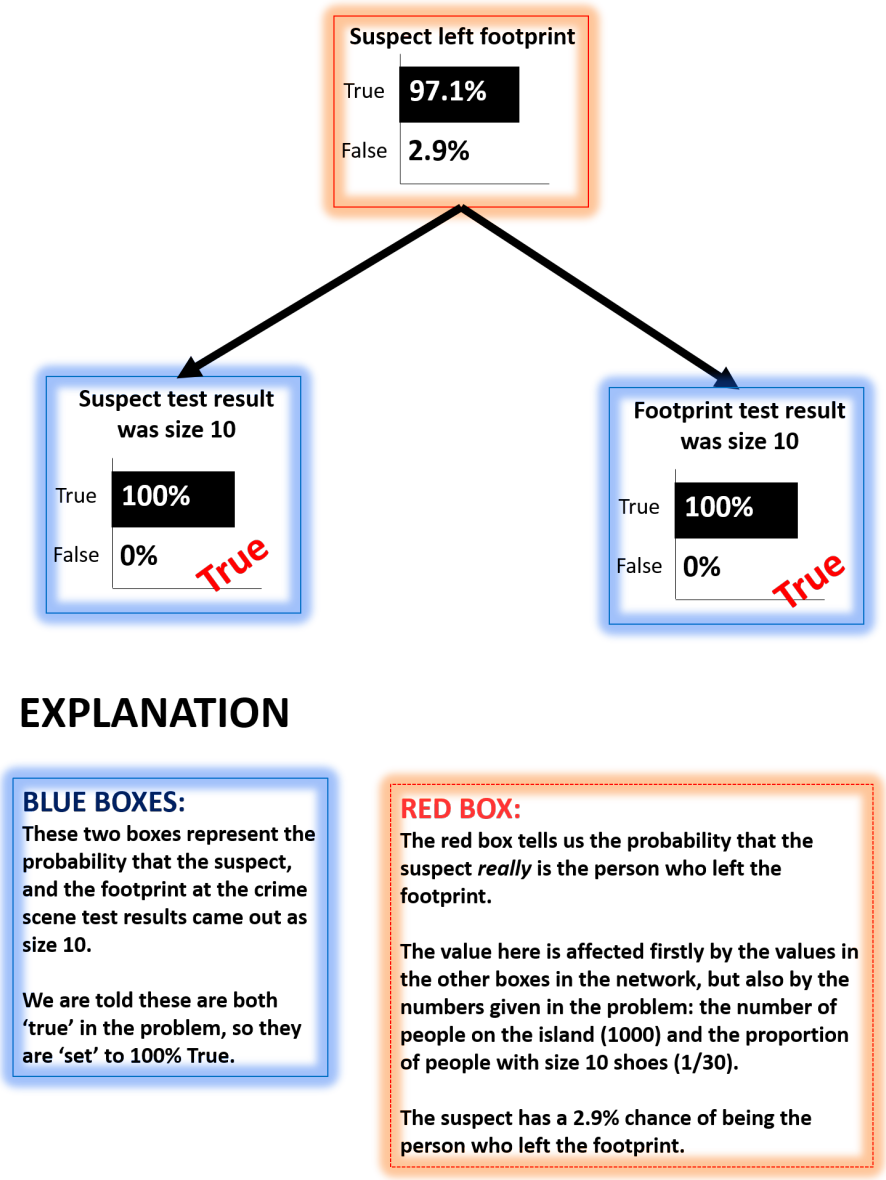


Fig. C.3: The Simple Bayesian Network Diagram Presented to Participants in Experiment Six

D The Additional Errors Information plus Event Tree and Bayesian Network Diagrams with Errors Presented to Participants in Experiment Six

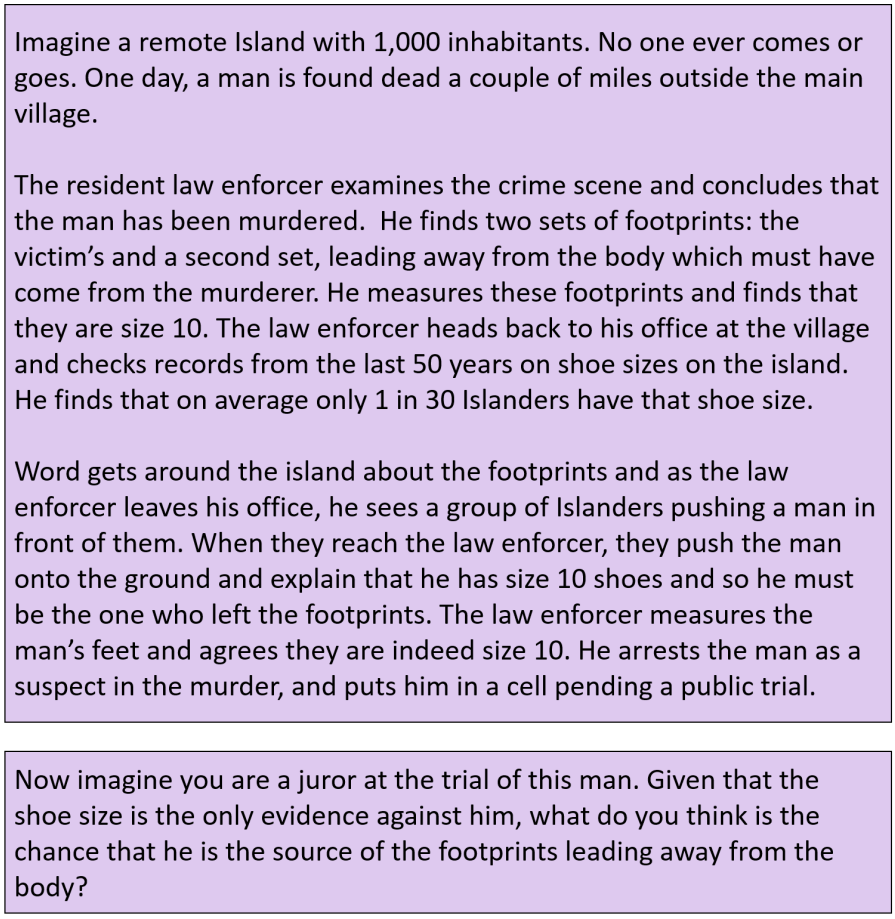
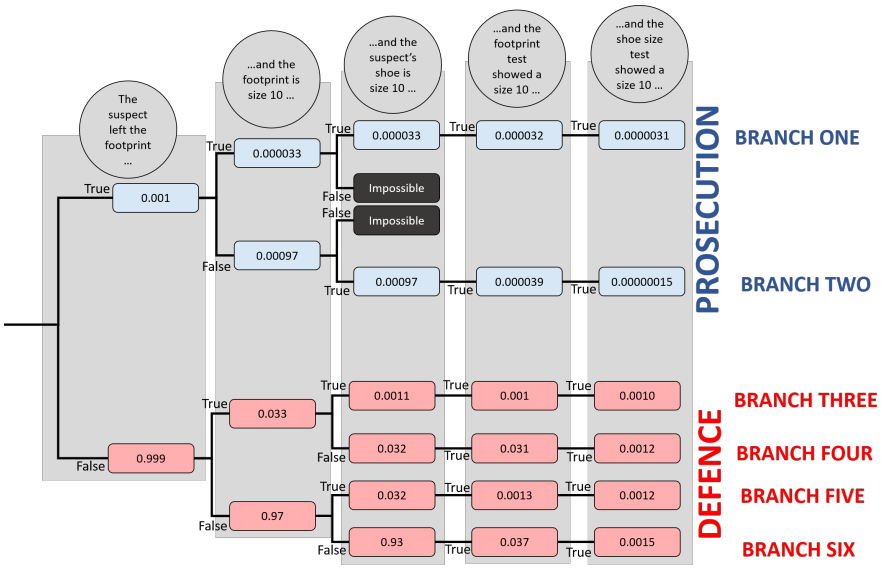


Fig. D.1: The Additional Errors Information Presented to Participants in Experiment Six



BLUE BRANCHES

The blue branches of the tree represent two important scenarios that fit the prosecution's hypothesis (that the suspect left the footprint)

- **BRANCH ONE:** This represents the 'standard' prosecution scenario that both the footprint and shoe size tests were 'true' positives (i.e. the law enforcer didn't make a measurement mistake) and the suspect is the person who left the footprint. The probability of this scenario is 0.00032.
- **BRANCH TWO:** This represents a less common prosecution scenario. Both the footprint and the shoe size tests were false positives, but the suspect's shoe still matches the footprint (e.g. both were size 8, perhaps). The probability of this scenario is 0.00000097.

RED BRANCHES

The red branches of the tree represent four important scenarios that fit the defence hypothesis (that the suspect did not leave the footprint)

- **BRANCH THREE:** This represents the most common defence scenario. Neither footprint nor shoe size test were false positives (the law enforcer didn't make a measurement mistake) but the match between the two is a coincidence (some other size 10 individual left the footprint). The probability of this scenario is 0.0010.
- **BRANCH FOUR:** This represents a scenario which involves a false positive occurring during the shoe size test. The probability of this scenario is 0.00031.
- **BRANCH FIVE:** This represents a scenario which involves a false positive occurring during the footprint test. The probability of this scenario is 0.00031.
- **BRANCH SIX:** This represents the least common defence scenario involving a false positive occurring during both the footprint test and shoe size test. The probability of this scenario is 0.000093.

Fig. D.2: The Event Tree Diagram Including Errors Presented to Participants in Experiment Six

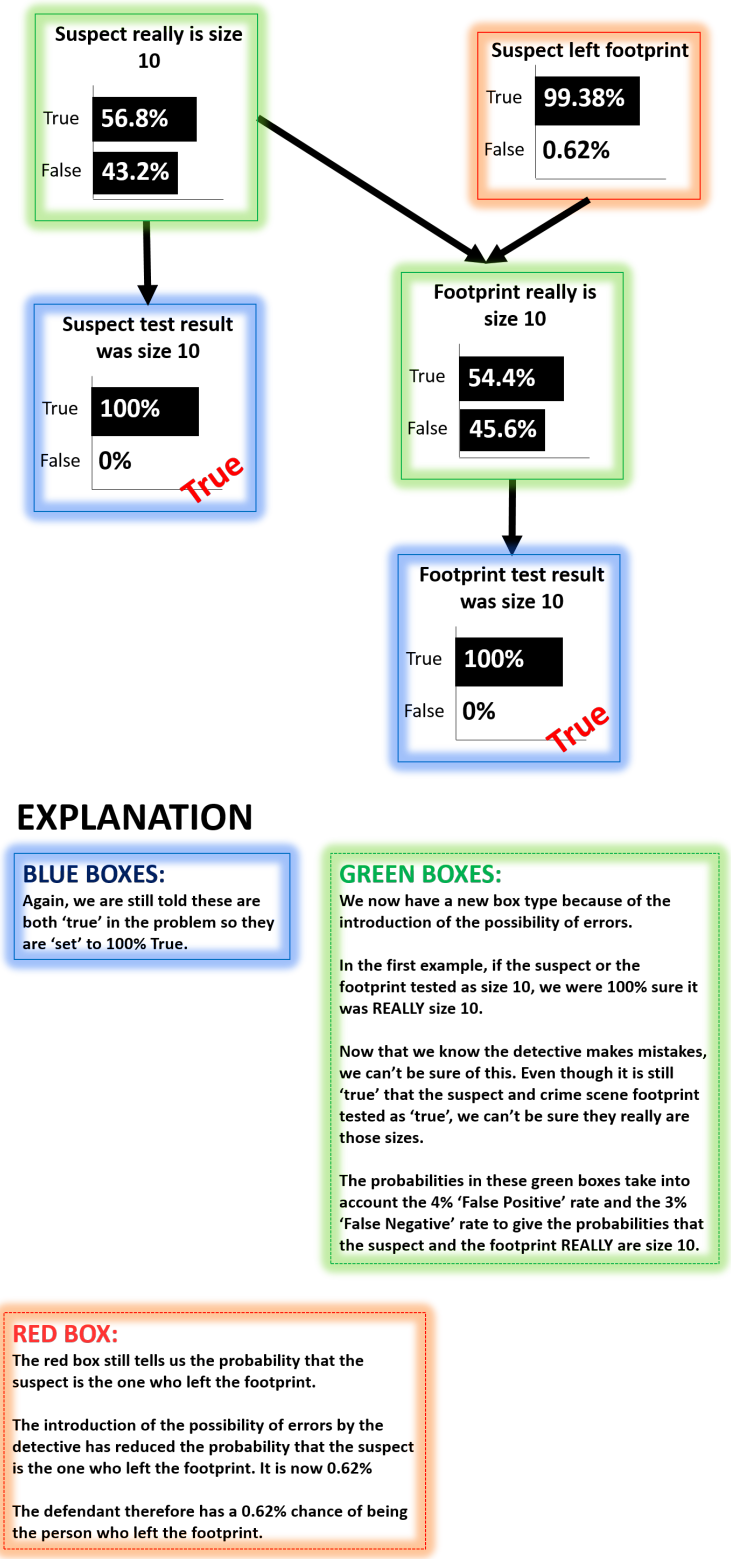


Fig. D.3: The Bayesian Network Diagram Including Errors Presented to Participants in Experiment Six