Measures of Bayesian Reasoning Performance on 'Normal' and 'Natural' Frequency Tasks

Rosemary Stock, University of West London

John E. Fisk, University of Central Lancashire

Catharine Montgomery, Liverpool John Moores University

Correspondence to be addressed to:

Rosemary Stock

School of Psychology, Social Work and Human Sciences

Boston Manor Road

Brentford Middlesex

TW8 9GA

Tel: 0208 209 4458

Email: Rosey.Stock@uwl.ac.uk

Abstract

While the majority of similar studies examining Bayesian reasoning investigate how participants avoid common errors such as base rate neglect, the current research also examines whether different formats (frequency and probability) lead to a difference in levels of absolute accuracy.

In Study One older (≥60 years) and younger (18 to 29 years) participants completed tasks in the probability and normalised frequency formats. In Study 2, older and younger participants completed tasks in probability and natural frequency formats.

Findings are that frequencies lead to less over-estimation, particularly in natural frequency tasks, which also reveal an interaction between age and task format whereby older adults seem unaffected by format (this interaction is mediated by information processing speed). There was no association found between format and the avoidance of errors such as base rate neglect. Findings are discussed in the light of dual and multi process theories of reasoning, having failed to support the theory that frequency formats elicit System 2 reasoning processes.

Key Words: Bayesian, natural frequency, probability, cognitive aging

Introduction

Reasoning and decision making is a subject of great and growing interest within not just the academic study of psychology but also to a wider audience, as evidenced by the popularity of Kahneman's 'Thinking, Fast and Slow' (2011). As Kahneman's work details, some of the more fascinating things to be learnt about our decision making, and the processes underlying our decisions, stem from the fact that many of our decisions are ultimately contradictory, and fail to follow the logical, and what has been termed the rational, path (see, for instance, Kahneman & Frederick, 2002).

One method that has been used extensively to investigate reasoning is the Bayesian probability task. To achieve the correct or normative value of a Bayesian probability, the following formula is appropriate:

$$P(E|A) = \frac{P(E)\times P(A|E)}{P(E)\times P(A|E) + P(\text{not } E)\times P(A|\text{not } E)}$$

where P(E|A) refers to the probability of some event E *given* the evidence, A.

Such reasoning is required whenever there is a need to consider how likely any given event may be, given some previously established evidence. One area where such decisions can have wide ranging implications for those involved is in clinical diagnostics, for example, in cases where a clinician obtains a test result and must decide on the probability of their patient having the condition in question given that test result. There is a large body of evidence that practitioners do not properly consider such probabilities (Gigerenzer, 1996; Croskerry, 2009;

Anderson, Gigerenzer, Parker & Schulkin, 2014), nor do patients, or journalists and politicians (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2008).

Utilising the formula above when making Bayesian judgements would clearly be a complex and cognitively demanding process. Fisk (2005) points out that even if one is not attempting to solve the exact equation, an individual who is attempting to solve a Bayesian task through calculation would need to manipulate a considerable amount of information about A and E, and their relationship to each other, in order to come up with any kind of solution which approached the normatively correct one. Birnbaum (2004) states that over 80% of student participants will fail to respond with the normative answer, and in a study by Gigerenzer and Hoffrage (1995), half of all participants failed to use any form of reasoning that could be described as 'Bayesian', even when the frequency format, a way of presenting the data believed to facilitate performance, was used (see below). Despite evidence that participants have processed the relevant base rate information, the most common response to such tasks is to respond by producing an estimate based on the likelihood of the additional evidence, a tendency known as base rate neglect (Johansen, Fouquet & Shanks, 2007; De Neys & Glumicic, 2008).

Using the following example of the cab problem devised by Tversky and Kahneman (1980), Birnbaum (2004) identifies three main non-normative ways of responding to such tasks:

"A cab was involved in a hit and run accident at night. There are two cab companies in the city, with 85% of cabs being green and the other 15% Blue cabs. A witness testified that the cab in the accident was "Blue." The witness was tested for ability to discriminate Green from

Blue cabs and was found to be correct 80% of the time. What is the probability that the cab in the accident was Blue as the witness testified?"

Birnbaum (2004) states that the majority of participants, 60%, show base rate neglect by giving the witness reliability rate of 80% as their answer (a finding replicated by Hinsz, Tindale & Nagao, 2008), with a further 20% responding with the base rate only – that is, the 15% rate of blue cabs. The latter error has been called the reverse base rate fallacy (Teigen & Keren, 2007), also found by Philips and Edwards, (1966), who referred to this as conservatism, and is a result of the base rate being relied upon while the information specific to the current case, i.e. the reliability of the witness, is ignored entirely (see also Achtiger, Alos-Ferrer, Hugelschafer & Steinhauser, 2014). Birnbaum also identified a group of participants as multiplying base rate by witness accuracy, to get 12%, and states that very few participants ever give the correct answer of 41%. This being the case, it is the *type* of answer that is of interest in the current study, rather than the apparent inability to come up with the normative response.

Frequency Formats

Gigerenzer and Hoffrage (1995) have demonstrated that there is a 'frequency effect' whereby participants answer Bayesian problems more accurately when the relevant information is presented in terms of frequencies, rather than probabilities (see also Gigerenzer & Galesic, 2012; Cosmides & Tooby, 1996). When expressed as probabilities a typical task, taken in this case from Evans, Handley, Perham, Over and Thompson, (2000), reads:

One out of every 1000 people has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out as positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, 5% of healthy people test positive for the disease. Imagine that we selected a random sample of 1000 people. Given the information above:

On average, how many people who test positive for the disease will actually have the disease? ***___%***

While a frequency version removes the reference to percentages, to become (with changes highlighted here):

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

On average, how many people who test positive for the disease will actually have the disease? ***___ out of ___***

This effect found by Cosmides and Tooby (1996) was so robust that even when the previous information is presented in a probability format, requesting the answer in terms of frequency (that is, as a number of cases rather than a probability) can produce a significant difference between that and the same problem with the same background information but the question worded so as to elicit a response based on single event probability. Gigerenzer and Hoffrage (1995) found that 76% of subjects found the correct answer to a frequency version of the problem, without additional prompting and clarification of the concepts involved, and a 92% accuracy level when told to work the problem out pictorially. They therefore concluded that previous literature in this area may have been too negative about the human ability to reason in accordance with Bayes theorem.

In accordance with this, there is evidence that young children can make probability judgements that are consistently better than chance. Gonazlez and Girotto (2011) examined children aged six to 10 years, asking them to look at a simple, single property (e.g. colour) and to give their answer as an approximation – that is, responding to 'what is more likely', instead of being required to estimate, or calculate a 'how likely' response. They found that the children could reason surprisingly well, with 10 year olds reaching levels of performance which were equivalent to the adult participants.

In their own examination of children's reasoning, Téglás et al. (2011) used a form of probabilistic reasoning with 'dynamic' scenes. Twelve month old babies saw films of different coloured bouncing objects inside a container which had one opening. The view of these objects was then briefly hidden from them, and one of the objects would fall through this opening. There were two factors which indicated which object should be most likely to fall. First, of the four objects three were the same shape and one was different, so the most

common object would normally be the one most likely to fall. Second, the objects were initially shown in motion in an arrangement such that the falling object could have been seen (before they were hidden) as being 'near' the opening or 'far' from the opening. Babies were surprised – as demonstrated by their longer gaze time – when they saw the least common object go through the opening, and when they saw the 'far' object do so. This indicates their understanding that each of this factors – the unusual shape and further distance from the opening – both indicated a low probability, or high surprise value, of the event occurring. Further to this, the surprise caused by the distance factor was reduced when the brief hiding period was extended from .04 to one, and then to two seconds, indicating a still more sophisticated understanding based on the fact that during that longer hiding period the 'far' object may well have bounced round to the opening. Téglás et al. (2011) suggest that babies are showing intuitive reasoning which does closely fit a Bayseian model, incorporating data regarding the hiding period, the proximity to the opening, and the numerosity of the object involved.

For both Téglás et al. (2011) and Gonzalez and Girotto (2010), the emphasis is on the recognition of unlikely outcomes, rather than an estimation of how likely an outcome may be. Whatever form of cognition is occurring in such young children, it is clear that preverbal infants are not using the full formula set out above. Instead, Téglás et al. (2011) suggest a form of intuition, or 'pure reasoning' is being conducted. This form of reasoning is also conducted by adults, as addressed by dual process theories of reasoning.

Dual Process Theories of Reasoning

There are three key dual process theories of reasoning; Evans and Over's Dual-Process Theory (1996; Evans, 2009;2010), Sloman's Dual-System Theory (2002) and Stanovich and West's Two-Systems Theory (2000; Stanovich 2004). Each of these suggest that we have two systems, or sets of systems, for reasoning, one of these being rapid and automatic, the other being slower and more deliberate.

The systems are similarly defined in each of the theories, with system 1 requiring less conscious effort in each case, being both less conscious (or entirely unconscious) and rapid, while system 2 is more deliberate, more time consuming and more commonly described as being 'rule based'. It has been suggested that it is System 1 that is in use when any of the reasoning heuristics are being utilised, and can be seen to be using intuition, while System 2 is held to utilise, and be limited by, working memory resources to make more 'rational choices. It is also clear that while heuristics are, by definition, short cuts in reasoning, this does not mean that they lead to inaccurate or unhelpful judgements in all situations (see for instance Hogarth & Karelaia, 2006; 2007; Evans, 2014, Mandel 2014).

There is also evidence in support of a tri, or multi-system theory, which allows for the existence of some form of reflective mind, which influences whether system 1 or 2 is used at any time. This process may be conscious and explicit (e.g. Stanovich 2009) or more implicit (Thompson 2009).

It should be noted that the expression 'dual process theory' in this paper should not be taken to imply that the perspective adopted is exclusively, or even primarily, that of Evans and Over's Dual-Process Theory, nor is it intended to exclude the growing support for each

system containing a number of processes (Stanovich, 2004: Evans, 2009, 2010), or the existence of a third 'reflective' process (Stanovich, 2009: Thompson, 2009). Instead the phrase is being used in general terms to describe any such theory, with emphasis on their commonalities, that System 1 operates on intuition, without using formal rules of logical calculation or a heavy cognitive load, while System 2 is heavily cognition dependent, rather than their differences.

Cognitive Aging and Reasoning

It is well established that many cognitive skills do deteriorate throughout adulthood, with working memory in particular showing consistent age related decline (Salthouse, 1998; Salthouse & Babcock, 1991; Ghisletta, Rabbitt, Lunn & Lindenberger, 2012).

As such, it is reasonable to expect that reasoning performance will also show age related decline, but research to date shows mixed results, with any evident decline being far from universal on all tasks. There is evidence that every day problem solving and decision making effectiveness (EPSE) does deteriorate with age, as found by Thornton and Dumke (2005) in a meta-analysis including a total of 4,482 participants. While EPSE is suggested by the authors to be quite different from what they call 'traditional' problem solving, of the sort addressed in this paper, whereby the problem and its solution are often more stylised, and designed by the researcher to examine particular phenomena.

Research by Yam, Gross, Prindle and Marsiske (2014) finds that inductive reasoning, measured through tasks involving pattern completion, is actually strong predictor of abilities

in more ecologically valid 'every day cognition' tasks within older (65 years and above) sample. Mutter, (2000) has suggested that age related detriments in reasoning are often elicited by the addition of cognitive load tasks, while Chasseigne *et al.*, (2004), Chasseigne, Mullet and Steward (1997), Mutter (2000), Mutter, Haggbloom, Plumlee and Schirmer (2006) and Mutter and Williams (2004) have all established that tasks which require participants to respond to multiple cues are disproportionately difficult for older participants, with age decrements on reasoning tasks often only becoming apparent when the tasks are cognitively demanding.

Fisk (2005) found no clear detrimental effect of age upon Bayesian reasoning performance, with participants' scores actually indicating that the older participants were giving estimates that were closer to the normative than were their younger counterparts. While Fisk (2005) did not find an age effect on Bayesian reasoning, despite the more complex nature of the tasks, this current study aims to further investigate any effect of age by looking not only at the magnitude of error made by participants, but also at the apparent underlying processes. This will be done by examining whether incorrect answers can be attributed to the participants focusing only on one cue within the task, or by substituting the required, complex, calculation for one that has (to them) some face validity but is a much simpler calculation. For instance, simply summing or multiplying two or more cue values.

With natural sampling giving participants clearly stated values that already contain base rate information, this may reduce the cognitive load on older participants. To use the framework of dual process theories of thinking and reasoning, this reduced cognitive load would enable the use of the deliberate and analytical system 2 processes (which are thought to be primed by the frequency format, Kahneman & Frederick, 2002). In a review of the literature on aging

and decision making Peters and Bruine de Bruin (2012) conclude that older adults are more likely to be swayed by how tasks are presented, and also that they put in less cognitive effort when solving such tasks, two things that would also suggest that older adults should find the natural frequency presentations of tasks to be particularly advantageous. Enabling older participants to make better decisions where possible is important not just because it provides them with increased independence (as suggested by Chen and Sun, 2003) but also due to the fact that society assigns important decision making roles to older adults, within fields such as law, politics and medicine (suggested by Peters & Bruine de Bruin, 2012).

Peters et al. (2000) suggest that the increasing prevalence of base rate neglect among older individuals in 'real life' situations is primarily due to their increased use of heuristics. However, priming the analytic system 2 by making the set structures clear in a natural frequency format, combined with the above fact that cognitive load has been reduced (by including information about the base rates within the frequencies) may lead to more effective analytic reasoning. This could give rise to more considered answers reliant on active reasoning, and/or more accurate answers.

One issue is whether or not older participants will be able to make use of the salient information in the tasks, (see Mutter & Pliske, 1994; Johnson 1993). If it were the case that older participants could not utilise the frequency information effectively, we would not expect older participants to show a great difference in performance across the different task formats, as they would not benefit from the natural frequency information in the same way as the younger participants. However, with System 2 being working memory dependent, the use of frequency formats – presenting the data in a more immediately accessible form – can be

expected to enable System 2, if primed, to work more efficiently and lead to more accurate answers.

Younger participants are also expected to benefit from natural frequencies, as per previous research (for instance Brase, 2008; Gigerenzer & Hoffrage, 1999). However, it is anticipated that this benefit will not be as pronounced as it is for the older participants, with the younger group being more able to deal with the greater cognitive load of the probability versions – given that the frequency versions reduce this load by containing information regarding the base rate throughout (see Gigerenzer & Hoffrage, 1995, and Mellers & McGraw,1999).

It is therefore anticipated that those participants given the tasks worded as natural frequencies will show more evidence of having attempted to manipulate the information to reach an answer, rather than simply responding with one of the values in the task. That is, they will be less likely to give 'base rate only' or 'base rate neglect' type responses.

Two separate studies were conducted. The first compares performance on tasks worded as probabilities with those worded as normalised frequencies. The second compares performance on those tasks worded as probabilities with those worded as natural frequencies. In each, the data were examined in two ways. First, by looking at the estimates of likelihood as values out of a hundred, allowing examination for overall accuracy, and second by looking at estimates in terms of response categories that are apparent in such data (for instance, base rate neglect, and 'base rate only' type responses).

In both cases, it is hypothesised that:
There will be a facilitating effect of frequency format upon accuracy of task.

There will be an interaction between age and format, whereby older participants experience a greater facilitating effect of the frequency format.

There will be an association between the frequency format and attempts to manipulate the data contained within the task. This will be indicated by participants in the frequency condition being more likely to give responses which are not merely re-stating single values from the task, but are clearly indicative of some manipulation of the data.

Study 1

Method

Participants

Seventy-seven older individuals, with a mean age of 70.53 (SD = 7.12) years and a range from 60 to 88 were recruited through the University of the Third Age (U3A). There were 24 males and 53 females in this group. The U3A is a group for older people which provides a range of classes aimed to provide education on a wide range of topics. Hultsch, Hertzog, Small and Dixon (1999) suggest in their paper 'Use it or lose it' that cognitive decline may be caused by (and is certainly related to) cognitive inactivity. As such, it was important that this target population be matched to the younger group of students, in that they were cognitively active and continuing to learn. Potential participants received information about the research at their local U3A meetings, and those that wished to take part were then invited to come in

to the university at a convenient time. They were each given a £10 supermarket gift voucher in lieu of any expenses incurred.

Regarding the younger participants, 139 took part, 20 male and 119 female, ranging from 18 to 29, with a mean age of 19.32 (2.05) years. All of this group were students of Liverpool John Moores University, and the experimental tasks were integrated into their coursework on a research methods module. As such, all students were given the opportunity to take part, resulting in a sample that was larger than could be obtained for the older age group. For ethical considerations students were also free to choose an alternative activity to fulfil coursework requirements, although none chose to do so.

In both cases, participants were randomly assigned to either the probability or the normalised frequency condition, and as far as could be ascertained, none of the participants had ever taken part in any similar research, at this or any other institution, and were naïve as to the issues being studied.

Materials

Participants in each condition were presented with two Bayesian reasoning problems.
The first of these, the 'disease X' problem, was based on one used by Evans, Handley, Perham, Over and Thompson (2000)

The probability version read as follows, with bold added here to emphasise the key differences between the two formats:

One out of every 1000 people has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out as positive. But sometimes the test also comes out positive when it is given to a

person who is completely healthy. Specifically, ***5% of healthy people*** *test positive for the disease. Imagine that we selected a random sample of 1000 people. Given the information above:*

On average, how many people who test positive for the disease will actually have the disease? ***___%***

The frequency version removed the references to percentages, in both the vignette and the response prompt, with the bold again for emphasis:

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

On average, how many people who test positive for the disease will actually have the disease? ***___ out of ___***

In this way, participants were free to think of any number of instances, rather than being led to think of a proportion out of one hundred, as implied by the probability version's use of a percentage response.

The second task was the non-causal version of the cab problem, based on that used by

Tversky and Kahneman (1980), with the tasks again worded to ensure that both groups were given a similar amount of helpful material, with only the prompts to consider percentages being re-worded for the frequency version:

*In a certain city **85% of the taxis are Green and 15% Blue. A witness to an accident in which a taxi was involved identifies the colour of the vehicle as blue.***

*The courts tested the **witness's** ability to identify cabs under the appropriate visibility conditions. **The** witness was presented with a random sample of taxis, (**85%** Green and **15%** Blue). Of the Blue taxis, the witness correctly identified **80%** of them as Blue (while mistakenly identifying **20%** as Green). Of the Green taxis the witness mistakenly identified **20%** as Blue (while correctly identifying **80%** as being Green).*

How likely is it that the taxi involved in the accident was blue? _____%

The frequency version differed as follows:

*In a certain city **85 out of every hundred taxis are green and the other 15 are blue. Over a one year period, there are 100 accidents involving cabs. In each of these 100 incidents, there was a witness available, and each of these witnesses identified the cab involved as a blue cab.***

The courts tested the **witnesses'** ability to identify cabs under the appropriate visibility conditions. **One of the witnesses was chosen to be tested at random and you are to assume that all of the other witnesses exhibited exactly the same degree of accuracy. This** witness was presented with a random sample of **100 taxis** (**85** Green and **15** Blue). Of the blue taxis **in the sample**, the witness correctly identified **12** as being Blue (while

mistakenly identifying **3** as Green). Of the Green taxis the witness mistakenly identified **17** as Blue (while correctly identifying **68** as being Green).

For how many of the 100 accidents where the witness identified the cab as blue was the cab actually blue? _____

Participants also completed the multiple choice Mill Hill Vocabulary Scale (MHVS, found by Raven, Raven & Court, 1998, to have good reliability for adults over 50, at alpha=.9) and Information Processing Speed (IPS) was measured with a paper and pencil version of the letter comparison task (Fisk, 2005). This latter task was developed from that used by Fisk and Sharp (2002), itself based on that by Salthouse and Babcock (1991). In the current version, the participants were required to compare two sets of letters on a page, and asked to decide whether the two sets were the same, or different. Alongside the two sets of letters were a large 'S' and a large 'D'. The participants were required to circle the S if the two sets of letters were the same, and the D if the sets were different. There were three levels of difficulty – sets of three letters, sets of six letters, and sets of nine. In each case, the participants had thirty seconds to complete as many items as they could, as quickly but carefully as possible. This entire process was then repeated, and the participants' score on this task was then the total number of items that they had responded to correctly, across the three levels of difficulty. Before beginning this task, they were given a full page of instructions, as well as five practice items.

Procedure

Younger participants took part in scheduled teaching sessions, with the older participants attending singly or in small groups at times that were convenient to them. The MHVS and

measure of IPS were administered before the Bayesian task. All participants also completed a number of other reasoning tasks as part of a related study.

Study 1 Results

There are two ways of examining the data obtained, the first is to examine the actual responses given as a value out of 100. Participants in the frequency condition were able to choose a denominator other than 100, and in those cases the value was converted by dividing the numerator by the denominator and then multiplying by 100. E.g., an estimate of 40 out of 550 gave (40/550) * 100 = 7.27. Participants' estimates for both age groups in both conditions are reported in Table 1. Due in part to the very uneven sample sizes, for both the disease task and the cab task the data lacked homogeneity of error variance (Levene's showing ps < .01) and no cells showed normal distribution (Kolmogorov-Smirnov ps < .001), although the data did not have significant skew or kurtosis. In order to ensure that ANOVA assumptions could be met, the ANOVA were both recast into a regression with dummy variables, and in STATA the robust command was employed. The resulting effect sizes and significance did not differ up to the first two decimal places, indicating that the original ANOVA are sound. These original ANOVA are therefore reported here.

< Table 1 around here>

Independent 2 x 2 ANOVA of the disease task data revealed that there was a significant effect of task format, with those in the probability group giving significantly higher estimates than those in the frequency group, $F(1, 209) = 7.74$, $p < .05$, partial $eta^2 = .036$ (see Table 1 for means of each group). Given that the normative answer to the task was 1.96, it can be

seen that those in the frequency group were significantly closer to this correct answer than those in the probability group.

There was no effect of age group, $F(1, 209) = 0.55$, $p > .05$, partial $eta^2 = .003$, and also no significant interaction between age group and task format, $F(1, 209) = 0.60$, $p > .05$, partial $eta^2 = .008$.

<Table 2 around here>

Descriptive statistics for the measures of vocabulary (the MHVS) and Information Processing Speed (IPS) are presented in Table 2. Neither the MHVS nor the IPS correlated significantly with the disease task scores, r = -.04 and -.06 respectively, ps > .05. When examined by task format, for the probability tasks the MHVS coefficient was r = -.10, and the IPS r = .02. For frequency tasks MHVS r = .00, IPS = -.15. ps > .05 in all cases. When analysed by age, the older group also showed no significant relationship, MHSV r = -.18, IPS r = -.04, ps > .05. Only for the young group did either measure show significance, with MHVS r = .11, p > .05 but for the IPS r = -.17, achieving significance at $p < .05$. This indicates that a faster speed is associated with greater over estimates on the task.

When entered into the analysis of variance as a covariate (assumption of homogeneity of regression slopes being sound at ps > .05), IPS was not significantly related to disease task score, $p > .05$, and the existing effect of task format is unaltered, at $F(1, 208) = 7.67$, $p < .01$, partial $eta^2 = .036$. The interaction remains non-significant, $F(1, 208) = 1.41$, $p > .05$, partial $eta^2 = .007$, while the effect of age is largely unaltered with only a slight increase in effect size, $F(1, 208) = 2.92$, $p > .05$, partial $eta^2 = .014$.

The second method of data analysis was to group the responses into categories derived by which cue they were based on, or method by which they appeared to have been generated, based on an approach by Birnbaum (2004).

1. 95, the modal response. May be from taking the false positive rate (5%) away from 100 (despite having been informed that the test is 100% accurate when testing those with the disease). (48 participants)
2. 5, the false positive rate (22 participants)
3. 0.1, the base rate of people within the population who have the disease (24)
4. 1, believed to be indicating some manipulation of both base rate and evidence. (14)
5. 1.96 through to 2, normatively correct answer (23)

There were also a large number of participants (85) who did not fit into any of these groups, but gave responses that seemed instead to depend on error variance, as only a very small number of participants gave each of the other values collected.

Chi square analysis of these groups found a significant association between task format and response category, χ^2 (4,131) = 33.43, $p < .001$. Figure 1 indicates that of those participants in the probability group, many more were giving a response of 95, which as indicated above suggests some manipulation of the data (through taking the false positive rate away from 100) but not an accurate calculation. The frequency group contained a larger proportion of correct answers, but also a greater proportion of 'base rate only' type responses.

< Figure 1 around here>

There was no association between response category and age group, χ^2 (4,131) = 0.79, $p > .05$.

The cab task data are summarised in Table 3. ANOVA revealed that there was a significant effect of task format, with those in the probability group giving significantly higher estimates than those in the frequency group, $F(1, 211) = 14.98$, $p < .001$, partial $eta^2 = .066$. Given that the normative answer in this case was 41, responses in the frequency condition are again indicating less over estimation. As revealed below, however, there was no effect of age group, $F(1, 211) = 0.90$, $p > .05$, partial $eta^2 = .004$, and also no significant interaction between age group and task format, $F(1, 211) = 0.02$, $p > .05$, partial $eta^2 = .000$.

< Table 3 around here>

Neither the Mill Hill nor the IPS scores correlated strongly (or significantly) with the cab task responses, $r = -.11$ and $r = .05$ respectively, $ps > .05$. No significant correlations were found within any subgroups of the data (e.g. examining the data by format or by age group), $rs < +/- .19$, $ps > .05$

For the cab task, only three values were regularly reported with the remaining participants producing a range of individual responses that could not be readily categorised. The three more common responses were:

1. 80, the modal response and the accuracy of the witness (61 participants fell into this group)
2. 15, the base rate of blue cabs (22 participants)
3. 12, the product of .80 and .15 (29 participants)

No participant gave the normatively correct answer in this case, and again a large number of participants were omitted from the analysis as they did not fit into any clearly defined group.

There was a significant association between format and response category, with all those who gave a response of 12 – indicative of some calculation – being in the frequency group. χ^2 (2, 112) = 39.31, p <.001. In contrast to the data from the disease task, the base rate only response was slightly more likely in the probability format.

<Figure 2 around here>

There was no association between age group and response category, χ^2 (2,112) = 2.51, p > .05 (see Figure 2).

Study 1 Conclusions

Of the three hypotheses, only the first, that there would be a facilitating effect of the frequency format, has been supported. There was no interaction between age and format, and no clear indication that the frequency format was associated with responses indicative of attempts to manipulate more than one of the values present in the task, with only the cab task indicating that the frequency format facilitated performance in this way. Coupled with the lack of main effect of age upon response in either task, this indicates that the older participants did not perform more poorly than their younger counterparts, and the frequency effect was found across both age groups.

There was a clear effect of format in both of the reasoning tasks presented, with those in the probability condition making significantly larger over estimates than those in the normalised frequency condition. It should be noted that (as confirmed by chi square analysis) it is not the case that those in the normalised frequency condition are more likely to obtain the correct

answer, but given that the average response to such tasks is to greatly over estimate likelihoods (by neglecting the base rate) such a reduction in over estimates could be of value in an applied setting.

Study 2

Introduction

Study 1 compared probability tasks with frequency tasks, using a form of frequency task known as 'normalised' frequencies, and with no clear evidence that such a format lead to greater manipulation of the task data, found no evident increase of use of System 2 reasoning processes. Hoffrage, Gigerenzer, Krauss and Martignon (2002) and Gigerener (2011) stress the clear difference between wording tasks as *natural* frequencies, as opposed to *normalised* frequencies:

*"*Natural *frequencies: Out of 1000 patients, 40 are infected. Out of 40 infected patients, 30 will test positive. Out of 960 uninfected patients, 120 will also test positive.*

Normalised *frequencies: Out of each 1000 patients, 40 are infected. Out of 1000 infected patients, 750 will test positive. Out of 1000 uninfected patients, 125 will also test positive."*

Hoffrage, Gigerenzer, Krauss & Martignon, 2002, p. 346

In the first case, it is easier to see that the answer to the question 'how many of those who test positive actually do have the disease?' can be found by totalling the number of people who have tested positive to get 150 (30, who have the disease, and 120, who do not) and using this as a denominator in a simple equation with the number of people who test positive and have

the disease, 30, as the numerator. If the response is asked for as ___ out of ___, this gives the participant scope to express their result as 30 out of 150, with no attempt at calculating the answer of 20%.

Both the standard probability format, and the normalised frequency format (as stated above) potentially require the following calculation:

$$P(E|A) = \frac{P(E)\times P(A|E)}{P(E)\times P(A|E) + P(\text{not } E)\times P(A|\text{not } E)}$$

P(E) being the prior probability of the event, or base rate, while P(E|A) is the posterior probability of the event, *given* the existence of A. Gigerenzer and Hoffrage (1995) apply the same Bayesian expression in order to evaluate the likelihood of H (the hypothesis that the patient has the disease) or not H (does not have the disease) in the context of the probabilities associated with D (the data obtained, in this case the positive test result). As such, the formula becomes:

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

Gigerenzer and Hoffrage (1995) and Mellers and McGraw (1999) illustrate how the natural frequency format renders this full calculation unnecessary, as one of the numbers presented directly corresponds to the numerator of the above expression and the sum of this together with one other presented values constitute the denominator. As such, the necessary

calculations have already been conducted and are embedded in problem format and the formula thus becomes (with lower case letters indicating 'data' and 'hypothesis'):

$$P(H|D) = \frac{d\&h}{d\&h + d\&noth}$$

In other words, the number showing both a positive test result and actually having the disease, is simply divided by those who show positive and have the disease *plus* those who show positive but do not have the disease. When we consider that 'd' represents all of those who tested positive, regardless of whether or not they actually have the disease, this can be expressed even more simply as:

$$P(H|D) = \frac{h\&d}{d}$$

Looking at the examples above, it is a relatively simple matter to pick out the values d&h (30) and d (30 who have the disease plus the 120 who showed positive while not having the disease) and to complete the calculation 30/150. Relative, that is, to attempting to find the same data from the normalised frequencies, which requires the individual to not just identify the correct values involved, but also to then make the more complex calculations given above, returning to the base rate each time in order to arrive at the answer.

Mellers and McGraw (1999) conclude that natural frequencies facilitate Bayesian reasoning by making set structures clear, allowing for a clearer understanding, and easier manipulation,

of joint events such as 'has disease AND tests positive' 'does not have disease AND tests positive'. Similarly, Evans *et al.* (2000) had also concluded that the frequency formats only result in better reasoning if they are worded so as to encourage the reader to create a mental model of the sets involved. Yamagishi (2003) also stresses the importance of not only being aware of the nested sets, but in visualising them, with the aid of diagrams. This complements the findings of Cosmides and Tooby (1996) who also found increased accuracy when tasks are represented pictorially. Such studies enable us to examine what people might be capable of when given sufficient instruction, but do not necessarily represent how people reason on a day-to-day basis.

Gigerenzer and Hoffrage (1995; 1999) and Hoffrage *et al*. (2002) approach natural frequencies by using evolutionary theory, stating that the format leads to better reasoning due to the data being presented in the way that it would naturally be acquired. They suggest that often the reason frequency formats are found not to facilitate reasoning is that they have not been presented in this way, as natural frequencies, but have instead been normalised. Indeed, Gigerenzer and Hoffrage (1995) directly compare tasks written as normalised frequencies and as probability tasks and found no significant difference between them. They feel that those who describe the improved reasoning as being explained by 'nested sets' are misunderstanding the importance of natural frequencies' presentation. They feel that the suggestions of nested and subsets as explanations are 'nothing more than vague labels for the basic properties of natural frequencies' (p. 343, Hoffrage *et al.* 2002) and stress instead that 'natural frequencies result from natural sampling and thus carry information about the base rates' (p. 347, ibid).

For this study 2, it is anticipated that presenting tasks in a natural frequency format will again show a facilitating effect (when compared with tasks in a probability format) and will also elicit an interaction effect between age group and format, whereby older participants experience a greater facilitating effect of the frequency format. An association between the frequency format and attempts to manipulate the data contained within the task is also again anticipated.

Method

Participants

Fifty-one older individuals, with a mean age of 68.72 (5.53) years and a range from 60 to 79 were recruited through the U3A. There were 8 males and 42 females in this group, with one participant failing to give their gender. All participants in this group were given £10 in lieu of expenses for their participation.

Forty-one younger participants took part in this study, 11 male and 30 female, ranging from 18 to 25, with a mean age of 20.80 (2.56) years. As in study 1, all of this younger group were students of Liverpool John Moores University. They took part through the Psychology department's Research Participation Scheme, through which students earn Participation Points which enable them to recruit through the scheme in their own future research. Participants in the scheme are also offered the chance to pass their module through the completion of a reflective piece on their participation experience. For ethical considerations all students were free to choose an alternative activity to fulfil coursework requirements.

All participants were randomly assigned to either the probability or the natural frequency condition, and as far as could be ascertained, none of the participants had ever taken part in any similar research, at this or any other institution, and were naïve as to the issues being studied.

Materials

All participants completed two Bayesian reasoning tasks. As in the previous study, one was based on the 'disease X' problem (Evans *et al.* 2000) while the second was based on the non-causal version of the cab problem (Tversky & Kahneman, 1980). This new natural frequency format is modelled on that used by Giggerenzer and Hoffrage in their research (e.g. Giggerenzer & Hoffrage 1995, 2007) and will allow for greater comparison with their findings. The alteration from '1 out of 1000' to '10 out of 1000' for the base rate of the disease will also allow participants to work with integers, rather than fractions, decimals or percentages. Giggerenzer and Hoffrage (1995) suggest that this is one of the key benefits of the natural frequency format. The probability versions have been amended accordingly in order to better 'match' the new natural frequency tasks.

The probability version of the cab task was as follows (with bold to emphasise the difference between this and the natural frequency format):

Two cab companies, the Green and the Blue, operate in the city. There are more green cabs than blue so ***the probability of getting a blue cab is 15%, while the probability of***

getting a green cab is 85%. *A cab was involved in a hit-and-run accident at night. On the night of the accident a witness identified the cab as "blue."*

The court tested the reliability of the witness under the similar visibility conditions with Blue and Green cabs. When the cabs were really blue, ***there was an 80% probability that the witness would correctly identify the colour and a 20% probability that they would mistakenly report the colour as green****. When the cabs were really green,* ***there was also an 80% probability that the witness would correctly identify the colour and a 20% probability that they would mistakenly report the colour as blue.***

What is the probability that the cab involved in the hit and run was blue?

_____*%*

The natural frequency version of the cab task:

Two cab companies, the Green and the Blue, operate in the city. There are more green cabs than blue so ***of every 100 cabs in the city, 15 are blue and 85 are green****. A cab was involved in a hit-and-run accident at night. On the night of the accident, a witness identifies the cab as "blue".*

*The court tested the reliability of the witness under the similar visibility conditions with Blue and Green cabs. When the cabs were really blue****, the witness said they were blue in 12 out of 15 tests (while mistakenly identifying 3 as green).*** *When the cabs were really*

green, ***the witness mistakenly said they were blue in 17 out of 85 tests (while correctly identifying 68 as being green).***

What are the chances that the cab involved in the hit-and-run accident was blue?

_____*out of* _____

This wording provides the information in the form of natural frequencies, as opposed to the 'normalised' frequencies provided in the previous study (as discussed in the introduction to this study).

The probability version of the disease problem read as follows:

The probability that an individual from *the UK population* ***has*** *disease X* ***is 1%****. A test has been developed to detect when a person has disease X. In terms of accuracy,* ***the probability that a person who has*** *the disease will produce a positive test result* ***is 80%.*** *But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically,* ***the probability that a healthy person will produce a positive test result is 10%.***

Given the information above:

What is the probability that a person who tests positive for the disease actually has the disease?

_____ ***%***

The natural frequency format was as follows:

10 out of every 1000 people in *the UK population* ***have*** *disease X. A test has been developed to detect when a person has disease X. In terms of accuracy,* ***8 out of the 10 people who have*** *the disease will produce a positive test result. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically,* ***99 out of the 990 healthy people will also test positive for the disease.***

Given the information above:

On average, how many people who test positive for the disease will actually have the disease?

____ out of _____

All participants also completed the MHVS and IPS as detailed in Study 1.

Procedure

All participants completed the tasks individually or in small groups at Liverpool John Moores University. The IPS and MHVS were again completed before the reasoning tasks were presented. The presentation of the two Bayesian tasks was alternated to prevent any order effects.

Study 2 Results

The data will again be examined in two ways, first as values out of 100, and second as categories. For the disease task, the descriptive statistics are presented in Table 4.

<Table 4 around here>

Independent 2 x 2 ANOVA revealed that the natural frequency format again led to an effect of task format, $F(1, 81) = 7.76$. $p < .01$, partial $eta^2 = .087$ – an effect size over twice that found when using the normalised frequency format (where partial $eta^2 = .036$). Again, overall those in the probability condition produced larger estimates than those in the frequency condition.

The effect of age group was not significant, $F(1, 81) = .67$, $p > .05$, partial $eta^2 = .008$, but the interaction between the two factors was significant, $F(1,81) = 4.29$, $p < .05$, partial $eta^2 = .050$. This interaction is illustrated in Figure 3, where it can be seen that the effect of format is driven by the young group, while the old group are largely unaffected.

<Figure 3 around here>

<Table 5 around here>

Neither the MHVS nor the IPS scores (see descriptive statistics in Table 5) correlated strongly (or significantly) with the disease responses, both $r = -.15$ and $r = .15$ respectively, $ps > .05$.

When examined by task format, for the probability tasks the IPS coefficient was significant, at $r = .28$, $p < .05$, indicating that greater speed was again associated with greater over estimation. When examined by task format, and then by age group, no other correlations between estimate and MHVS or IPS were found, $rs < .24$, $ps > .05$

When added to the analysis of variance as a covariate, IPS was not significantly related to disease task score, $F < .01$, homogeneity of regression slopes being sound at $ps > .05$.

The effect of task format remains almost unchanged, $F(1, 80^{1}) = 7.24$. $p < .01$, partial $eta^2 = .083$.

There is still no effect of age group, $F(1, 80) = .09$, $p > .05$, partial $eta^2 = .001$. The interaction is also now (marginally) non-significant, $F(1,80) = 3.71$, $p > .05$, partial $eta^2 = .055$. Estimated marginal means, presented in Table 6, indicate that when the difference in processing speeds is accounted for, the difference between the age groups' estimates do decrease.

<Table 6 around here>

The data were again categorised by response, as previously discussed. In this instance the proportion of 'unclassified' participants was far smaller, and was therefore included in the analysis.

1. 1 – the base rate (10 participants)
2. 7.48 – normative (3 participants)

3. 10 – value for 'positive test given the person is healthy' (11 participants)
4. 70 – indicative of some calculation (10 participants)
5. 80 – value for 'positive test given the person is healthy' (15 participants)
6. 90 – indicative of some calculation (12 participants)
7. Those responses which did not fit into any clearly defined group (24 participants)

As in Study 1, Chi square analysis of these groups found a significant association between task format and response category, χ^2 (6,85) = 14.33, p < .05, with Figure 4 indicating that those in the natural frequency group were far more likely to give responses that were unclassified, while far less likely to respond with the value of 90 that is thought to indicate that they have taken the chance of false positives from the total likelihood. The latter in particular is similar to the previous findings of Study 1.

Age group was not associated with response category, χ^2 (6,85) = 5.45, p > .05

<Figure 4 about here>

Cab task descriptive statistics are presented in Table 7.

<Table 7 about here>

ANOVA in this case found no significant effects (F < 1 for both the effect of age, and the interaction), although the effect of format approaches significance, F(1, 85) = 3.15, p=.08, partial eta^2 = .03, with those in the frequency condition responding with slightly lower estimates of likelihood than those in the probability condition group. As the normative value is 41.4 (12/29), it can be seen that the mean judgements in the frequency format are just slightly closer to being 'correct'.

Neither the Mill Hill nor the IPS scores correlated strongly (or significantly) with the cab task responses, r = .14 and r = .03 respectively, ps > .05, nor did they do so when the data were examined by format, or by age group, rs < .22, ps > .05 in all cases.

The cab task showed a greater number of clearly defined groups than in Study 1.

1. 15 – base rate (21 participants)
2. 41-42 – normative (2 participants)
3. 80 – value is given in the probability version as being the probability of the witness correctly identifying a blue cab (26 participants)
4. 20 – value is given in the probability version as being the probability of the witness stating 'blue' when it is a green cab (5 participants)
5. 12 - value is given in the frequency version as being the probability of the witness correctly identifying a blue cab (4 participants)
6. Those responses which did not fit into any clearly defined group (31 participants)

Chi square analysis of these groups found no significant association between task format and response group, χ^2 (5,89) = 8.20, p > .05, or between age group and response group, χ^2 (5,89) = 8.15, p > .05.

Study 2 Conclusions

Of the three hypotheses the first, that there would be a facilitating effect of the frequency format, has been partially supported, in the disease task. There was an interaction between age and format for the disease task only, offering partial support for that hypothesis, although the interaction ceased to be significant when processing speed was taken into account. Again

no clear indication that the frequency format was associated with responses indicative of attempts to manipulate more than one of the values present in the task. In both cases, a greater proportion of participants were able to be placed into groups for analysis, resulting in a reduction in the number of data 'lost' during this process.

When compared with Study 1, which used normalised frequencies, the use of natural frequencies did result in a larger effect of task format upon performance for the disease task, but there was no effect at all for the cab task. Giggerenzer and Hoffrage, (1999) have also suggested that there is something distinct about the cab problem which appears to suppress or counteract the beneficial effect of natural frequencies, therefore reducing the likelihood of finding an effect of format when using this task.

This second study also reveals an interaction between age and task format, on the disease task only, such that the effect of format is clearly driven by the young group, who are grossly overestimating the probability in the probability format, but showing closer to normative responses in the natural frequency format. The older participants show far less difference between the conditions – it could be said that they are not benefitting from the frequency format, but equally they are not suffering from the probability format.

Discussion

Across the two studies there is support for the effect of format (in both normal frequency tasks and in the disease natural frequency task), limited support for the interaction between age and format (in the disease natural frequency task) and no support at all for the association between task format and responses indicative of calculation.

Two studies of Bayesian reasoning have been presented here. The first, looking at normalised frequencies and the second at naturalised frequencies, with the expectation in each case that these frequency formats will lead to either more accurate reasoning, or at least more attempts to be more accurate, by using more of the information available. The ANOVA conducted in both studies demonstrated that for the disease task, there were consistently larger overestimates in the probability condition, with this effect being larger for natural frequencies than for normalised frequencies. For the cab task, however, the reverse was found – the difference was present when using normalised frequencies, but ceased to be significant when natural frequency tasks were used.

The second study found that older participants (that is, 65 and over) were neither advantaged by frequency format nor disadvantaged by probability format. Regardless of the format, they gave either estimates that were higher than the 'young frequency' group, or lower than the 'young probability' group. This is in accordance with findings by Mutter and Plumlee (2009), that not only did older participants find it more difficult to integrate a range of information in order to solve problems, they also did not appear to benefit from their being framed in meaningful contexts, something which significantly improved performance in a younger group. In the current study, it appears that the more transparent framing, and the use of the natural frequency format in particular, was of no benefit to older participants. However, rather than this being due to older participants performing poorly overall, group means suggest that this is a function of younger participants doing particularly poorly – making very large over estimates – in the probability condition.

In all cases, and in accordance with most previous research, very few participants correctly produced the normatively correct answer (see for instance Gigerenzer, 1996; Croskerry, 2009; Birnbaum, 2004; Gigerenzer *et al.*, 2008; Anderson *et al.*, 2014).

When the results were examined by response group, an association between format and response was found (in all but the cab task comparing natural frequencies with probabilities), but this was not in the form that was anticipated. It had been expected that the frequency formats would facilitate the use of System 2, presenting the information such that participants would make greater attempts to calculate the correct answer, even if they did not do so accurately. This may have been revealed by more 'correct' answers, but also by more answers given that showed some form of calculation, instead of simple 'base rate only' or 'base rate neglect' type responses that indicate participants have given as their answer one of the values shown in the wording of the task. No such association was consistently found, with all formats being associated with answers that indicated both calculation and lack of calculation in the majority of tasks. However for the cab task in Study 1, it was the case that the only response category in that data which was indicative of some attempt to integrate information in a multiplicative manner was only evident in the frequency format. Just as children can show intuitive judgements of likelihood when presented with complex but concrete examples (Téglás et al., 2011) it can be suggested that the frequency format does in some way facilitate this intuition but is not enough to initiate System 2.

Using the dual, and multi-system approach, one interpretation of the current findings is that System 1 (the rapid, autonomous system) is over ruling the external prompt to System 2 that the task format provides. Evans (2006) suggests this may occur when System 1 comes up with an answer so quickly that System 2 is simply not engaged, or that System 1, the is a

default system is only rarely over ruled by the by the analytic System 2. Stanovich (2004) also places particular emphasis on the autonomous nature of System 1 processes, as it may respond even as we are consciously aware that its response is incorrect and/or unnecessary. This suggests that we need a stronger external prompt to engage the more cognitively demanding System 2 to engage – that our intuitive processes are so strong, or ingrained, that this becomes a very difficult response to override..

A further possibility is that the frequency formats are providing an illusionary Feeling of Rightness (FOR: Thompson, 2009). The FOR is felt when an initial answer is given, by either system, and leads to the answer being accepted as true, rather than System 2 being (further) engaged to find another solution. If participants feel more confident working with the natural frequency values, this could result in less priming of System 2 due to the FOR. In a task asking about the probability of a positive diagnosis, for instance, participants may be primed to strongly associate a diagnosis with the occurrence of that condition, leading to a strong FOR when they reach an answer which suggests a high likelihood of the disease's presence.

Finally, it is always possible that System 2 is being primed, but that as it uses greater cognitive resources (Evans & Over, 1996; Sloman, 2002; Stanovich & West, 2000) participants either do not have the resources to use it correctly, or are being cognitive misers in taking the less effortful route of System 1. This might explain the large number of responses that did not appear to fit any recognisable pattern for all tasks used – such random data might be expected not from participants using erroneous rules, but from their attempting to use rules which they fail to execute accurately.

A key strength of the current studies is the focus on examining not whether responses are normatively correct – what has been termed 'rational' – but at how close they are to the correct answer, as a way of assessing intuitive reasoning of comparative likelihood. In such Bayesian tasks, and in the disease task in particular, the errors made are usually that the likelihood of the disease is vastly over estimated. As such, the fact that the frequency format led to less overestimation is an indication that it aids reasoning. Evans (2014) states that the research in this field should move away from categorising responses, and people, as being either 'rational' or 'irrational', and Mandel (2014) also suggests that there needs to be "better discussion of whether being non-Bayesian is necessarily irrational" (p3). Moving the focus towards degrees of accuracy, rather than absolutes, is following this trend in suggesting that the aim should be in moving towards reasoning that is functional, and advantageous to the individual.

Stupple, Ball, Evans and Kamal-Smith (2011) found that participants who take longer over syllogistic tasks are more logical, in terms of getting more correct answers and being less susceptible to belief bias. They suggest that this supports the dual processing theories. If we couch the disease task as being susceptible to the belief bias that 'positive test = having a disease' then we can again conclude that to fail to get the 'normative' entirely 'correct' answer is not indicative of a reasoning failure. It may serve us well to have greater awareness of the possibility of false positives, but our general rule that a positive test is highly associated with having the disease will more usually be to our advantage.

Future directions

As demonstrated by the very low number of participants getting the 'correct answers' on both the current and previous studies (i.e. Gigerenzer, 1996; Croskerry, 2009; Birnbaum, 2004; Gigerenzer et al., 2008; Anderson et al. 2014). Such bayesian tasks are very difficult, and when considered in terms of 'getting them right' there is a strong floor effect. The current study minimises the impact of this in terms of our ability to analyse the results by looking at categories of responses, and degrees of accuracy, it is also possible that the difficulty of the tasks leads to a lack of a 'feeling of solvability' (see Thompson 2009) which may be disheartening and lead to low motivation and lack of engagement of System 2 at all. We would propose a number of ways to address this. One would be to use diagrams, as have already been found to facilitate reasoning (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Yamagishi, 2003), or to create more concrete tasks, and/or tasks with a more limited range of answers. such as those used by Gonzales and Girotto (2010). This may have low ecological validity in terms of diagnostic and financial decision making, but does have great value as a tool to understanding the processes.

Other possibilities would be to look in more detail at measures of cognitive ability as mediators/predictors of reasoning performance (cf Fisk 2005, also Stanovich & West). This study found very little influence of processing speed, and none of vocabulary – aspects that did differ between the groups despite attempts in recruitment to match them on these aspects. One aspect that has been omitted (due to time constraints and concerns of fatigue in participants) is mathematical ability. This would influence participants' ability to use the analytic system effectively, and a lack of correlation between performance and cognitive ability would reflect a greater likelihood of the response having been created by the intuitive System 1 . Stanovich has frequently used SATs results as a measure of cognitive ability, (for instance Stanovich & West, 1998). Nonetheless exam results across age cohorts are not likely

to be comparable, so future research may be required to administer such measures within a test battery. However, fatigue is a concern here, as is motivation (as discussed above) and is likely to affect older participants disproportionately, so it is an important consideration when designing studies in this area.

One of the challenges for this area, as in so many attempts to understand cognitive processes, is to identify what the process has actually been when confabulation is an issue (Evans 2008; Thompson 2009). The current study has attempted to get to the root of this by looking at whether any calculation has occurred and categorising responses accordingly, but this remains a post hoc interpretation. Another method is in testing the participant's memory for each cue (Franssens & De Neys, 2009), on the premise that if a person has used the value in their calculation at all, they will be more likely to be able to recall it. This may be true even in cases where they have previously stated that they did not use or notice the particular cue when completing the task.

In conclusion, the two studies presented here demonstrate that frequencies are easier to work with than single event probabilities, with both normalised and natural frequencies eliciting more accurate (or perhaps 'less inaccurate') responses on the 'disease' task. While the reasoning shown even in the natural frequency group is by no means normative, such reduction in overestimates is beneficial in situations such as clinical settings, when a more realistic understanding of risks and benefits can facilitate discussion and decision making. It may be as simple as the presentation of the reference class having lead to the magnitude of the subgroup in question – its smallness – being made salient to the participants, and they are then anchoring their responses more appropriately. However, the interaction of age by format in the natural frequency versus probability versions of the disease task reveals that for the

younger group a poor performance in the probability formatted tasks is particularly reduced by the natural frequency format, suggesting that they are using the information more effectively than their older counterparts..

When examined by task format, there was no evidence of greater use of System 2 – no evidence that people in the frequency groups were attempting to get the correct answer with some kind of statistical/rational method. However, in both studies large amounts of apparently random variation was found. This latter issue is a function of the fact that participants are required to generate their own answers to complex tasks, rather than being invited to select from a range of possible solutions, and future research may look for ways of inviting participants to select possibilities while still engaging System 2 to devise an answer, rather than relying on a FOR (Thompson, 2009) when presented with various options.

The studies presented here demonstrate once again the difficulty of eliciting and identifying the use of System 2 process on complex reasoning tasks, but that the frequency formats used here both lead to more accuracy from participants in the 'Disease task' in particular continues to illustrate their usefulness. While 'correct' Bayesian reasoning may continue to be beyond most participants in this study and in day to day life, a focus on improving accuracy by considering its utility should be at the centre of future research in this field.

References

Anderson, B. L., Gigerenzer, G., Parker, S., & Schulkin, J. (2014). Statistical Literacy in Ostetricians and Gynecologists. *Journal for Healthcare Quality, 36*, 5-17. doi: 10.1111/j.1945-1474.2011.00194.x

Birnbaum, M. H. (2004). Base Rates in Bayesian Inference. In R. F. Pohl (Ed.), *Cognitive Illusions: a handbook on fallacies and biases in thinking judgement and memory* (43-60). Hove, UK: Psychology Press.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin and Review, 15*, 248-289. doi: 10.3758/PBR.15.5.284

Chasseigne, G., Ligneau, C., Grau, S., Le Gall, A., Roque, M., & Mullet, E. (2004). Aging and probabilistic learning in single- and multiple-cue tasks. *Experimental Aging Research, 30*, 23-45. doi: 10.1080/03610730490251469

Chasseigne, G., Mullet, E., & Steward, T. R. (1997). Aging and multiple cue probability learning: the case of inverse relationships. *Acta Psychologica, 97*, 235-252. doi: 10.1016/S0001-6918(97)00034-6

Chen, Y., & Sun, Y. (2003). Age differences in financial decision-making: using simple heuristics. *Educational Gerontology, 29*, 627-635. doi: 10.1080/03601270390218152

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1 - 73. doi: 10.1016/0010-0277(95)00664-8

Croskerry, P. (2009) Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in Health Sciences Education, 14*, 27-35. doi: 10.1007/s10459-009-9182-2

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking.*Cognition, 106,* 1248-1299. doi: 10.1016/j.cognition.2007.06.002

Evans, J. S. B. T. (2008). Dual -processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278. doi: 10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition, 77,* 197-213. Retrieved from http://faculty.arts.ubc.ca/pbartha/p520w07/evans_freq.pdf

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. New York, Oxford University Press.

Evans, J. S. B. T. (2014). Rationality and the Illusion of Choice. *Frontiers in Psychology, 5,* 1-4. doi: 10.3389/fpsyg.2014.00104

Fisk, J. E. (2005). Age and Probabilistic Reasoning: Biases in Conjunctive, Disjunctive and Bayesian Judgements in Early and Late Adulthood. *Journal of Behavioural Decision Making, 18,* 1-28. doi: 10.1002/bdm.488

Fisk, J. E., & Sharp, C. (2002). Syllogistic reasoning and cognitive ageing. The Quarterly *Journal of Experimental Psychology, 55A*, 1273-1293. doi: 10.1080/02724980244000107

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

Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making, 16,* 273-280. doi: 10.1177/0272989X9601600312

Gigerenzer, G. (2011). What are natural frequencies. *British Medical Journal,* 343. doi: 10.1136/bmj.d6386

Gigerenzer, G., & Galesic, M. (2012). Why do single event probabilities confuse patients? *British Medical Journal, 344*. doi.org/10.1136/bmj.e245

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M. & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8,* 53-96. doi: 10.1111/j.1539-6053.2008.00033

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704. Retrieved from http://www.cogsci.ucsd.edu/~coulson/203/GG_How_1995.pdf

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review, 106,* 425-430. doi: 10.1037/0033-295X.106.2.417

Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behavioral and Brain Sciences, 30,* 264-267. doi: 10.1017/S0140525X07001756

Gonzales, M. & Girotto, V. (2010). Combinatorics and probability: Six- to ten-year-olds reliably predict whether a relation will occur. *Cognition, 120,* 327-379. doi: 10.1016/j.cognition.2010.10.006

Hinsz, V. B., Tindale, R. S., & Nagao, D. H. (2008). Accentuation of information processes and biases in group judgments integrating base-rate and case-specific information. *Journal of Experimental Social Psychology, 44,* 116-126. doi: 10.1016/j.jesp.2007.02.013

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition, 84,* 343-352. doi: 10.1016/S0010-0277(02)00050-1

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: engaged lifestyle as a buffer of cognitive decline in aging? *Psychology and Aging, 14*, 245-263. doi: 10.1037/0882-7974.14.2.245

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007). Paradoxical effects of base rates and representations in category learning. *Memory & Cognition, 35,* 1365-1379. doi: 10.3758/BF03193608

Johnson, M. M. S. (1993). Thinking about strategies during, before and after making a decision. *Psychology and Aging, 8*, 231-241.

Kahneman, D. (2011). *Thinking, Fast and Slow.* London, Penguin.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* New York, NY, US: Cambridge University Press.

Mellers, B. & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review,* 106, 417-424. doi: 10.1037/0033-295X.106.2.417

Mutter, S. A. (2000). Illusory correlation and group impression formation in young and older adults. *Journal of Gerontology, 55B,* 224-237. doi: 10.1093/geronb/55.4.P224

Mutter, S. A. & Plumlee, L. F. (2009). Aging and integration of contingency evidence in causal judgment. *Psychology and Aging, 24,* 916-926. doi: 10.1037/a0017547

Mutter, S. A., Haggbloom, S. J., Plumlee, L. F., & Schirmer (2006). Aging, working memory, and discrimination learning. *The Quarterly Journal of Experimental Psychology, 59,* 1556-1566. doi: 10.1080/17470210500343546

Mutter, S. A., & Pliske, R. M. (1994). Aging and illusory correlation in judgments of co-occurrence. *Psychology and Aging, 9,* 53-63. doi: 10.1037/0882-7974.9.1.53

Mutter, S. A., & Williams, T. W. (2004). Aging and the detection of contingency in causal learning. *Psychology and Aging, 19,* 13-26. doi: 10.1037/0882-7974.19.1.13

Peters, E., & Bruine de Bruin, W. (2012). *Aging and decision skills Judgment and Decision Making as a Skill: Learning, Development, and Evolution*. New York: Cambridge University Press.

Raven, S., Raven, J. C., & Court, J. H. (1998). *Manual for Ravens Progressive Matrices and vocabulary scales, Section 5, Mill Hill vocabulary scale*. Oxford, UK: Oxford Psychologists Press.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27,* 763-776. doi: 10.1037/0012-1649.27.5.763

Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment. New York: Cambridge University Press.

Stanovich, K. E. (2004).*The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press

Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology, 23,* 931-941. doi: 10.1080/20445911.2011.589381

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Process and Beyond*. New York: Oxford University Press.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011) Pure reasoning in 12-month-old infants as probabilistic inference. Science, 332, 1054-1059. doi: 10.1126/science.1196404

Teigen, K. H., & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition, 103*, 337-357. doi: 10.1016/j.cognition.2006.03.007

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty In M. Fishbein (Ed.), *Progress in Social Psychology*. New Jersey: Erlbaum.

Yam, A., Gross, A. L., Prindle, J. J., & Marsiske, M. (2014). Ten-Year longitudinal trajectories of older adults' basic and everyday cognitive abilities. *Neuropsychology, 28,* 819-828. doi: 10.1037/neu0000096.

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology, 50,* 97-106. doi: 10.1026//1618-3169.50.2.97