

**ADULT AGE DIFFERENCES IN THINKING STYLES AND PROBABILISTIC
REASONING: THE EFFECT OF NATURAL FREQUENCIES**

ROSEMARY ANNE STOCK

**A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores
University for the degree of Doctor of Philosophy**

April 2012

Index

DECLARATION FORM.....	5
ACKNOWLEDGEMENTS	6
ABSTRACT.....	7
LIST OF TABLES	9
LIST OF FIGURES.....	11
LIST OF ABBREVIATIONS	12
CHAPTER 1: THESIS SUMMARY.....	13
CHAPTER 2 – PROBABILISTIC REASONING REVIEW.....	15
2.1 Probabilistic Reasoning	15
2.2 Conjunctions.....	16
2.3 Disjunctions	17
2.4 Bayesian Tasks	18
2.5 Explanations for Reasoning Errors.....	20
2.5.1 Linguistic misunderstanding	21
2.5.2 Signed summation.....	22
2.5.3 Representativeness	23
2.5.4 Potential surprise.....	25
2.5.5 Frequency interpretations.....	27
2.5.6 Applying the wrong probabilistic rule	29
2.5.7 Mental Models	30
2.5.8 Fast and Frugal.....	32
2.5.9 Averaging.....	33
2.5.10 Dual Process Theories.....	35
2.6 Conclusions	37
CHAPTER 3 - THINKING STYLES AND DUAL PROCESS THEORIES	40
3.1 Dual Process Theories of Reasoning.....	40
3.2 Evidence for Dual Process Theories of Reasoning.....	46
3.3 Conflict Between the Systems	52

3.4 Thinking styles.....	56
3.5 Links between thinking styles, dual process theories and problem format.....	60
3.6 Conclusions	63
CHAPTER 4 – COGNITIVE AGEING AND DUAL PROCESS THEORIES OF REASONING.....	66
4.1 Age related cognitive decline may affect reasoning	66
4.1.2 Information Processing Speed.....	71
4.2 Are older people reasoning differently? Links with dual process theory	73
4.3 Ageing and Thinking Styles	79
4.4 Ageing and Expertise.....	81
4.5 Conclusions	83
4.6 Summary of Literature Review and Introduction to Empirical Chapters.....	85
CHAPTER 5 – EFFECTS OF FORMAT ON PROBABILISTIC REASONING .	88
5.1 Introduction.....	88
5.1.1 Probability and Frequency Formats	89
5.1.2 Hypotheses	91
5.2 Method	92
5.2.1 Design	92
5.2.2 Participants.....	92
5.2.3 Materials.....	92
5.2.4 Procedure.....	96
5.3 Results	96
5.3.1 Conjunction Fallacy Results	96
5.3.2 Disjunction Fallacy Results	96
5.3.3 Error Results	97
5.4 Discussion.....	98
CHAPTER 6 – THINKING STYLES, TASK FORMAT AND PROBABILISTIC REASONING.....	102
6.1 Introduction.....	102
6.1.1 Hypotheses	104
6.2 Method	105
6.2.1 Design	105
6.2.2 Participants.....	106
6.2.3 Materials.....	107
6.2.4 Procedure.....	113

6.3 Results	114
6.3.1 Reasoning Performance	114
6.3.2 Analyses of Variance	115
6.3.3 Hierarchical Regression Analyses.....	120
6.4 Discussion.....	124
CHAPTER 7 – AGE, THINKING STYLES, PROBLEM FORMAT AND REASONING.....	128
7.1 Introduction.....	128
7.1.1 Age	128
7.1.2 Are the reasoning processes of older adults qualitatively different?	130
7.1.3 Hypotheses	131
7.2 Method	132
7.2.1 Design	132
7.2.2 Participants.....	133
7.2.3 Materials.....	133
7.2.4 Procedure.....	134
7.3 Results	135
7.3.1 Analyses of Variance	138
7.3.2 Regression analyses	144
7.4 Discussion.....	150
CHAPTER 8 – BAYESIAN REASONING	154
8.1 Introduction.....	154
8.1.1 Hypotheses	157
8.2 Method	157
8.2.1 Design and statistical analyses.....	157
8.2.2 Participants.....	158
8.2.3 Materials.....	158
8.2.4 Procedure.....	160
8.3 Results	160
8.4 Discussion.....	173
CHAPTER 9 –BAYESIAN TASKS WITH NATURAL FREQUENCIES	177
9.1 Introduction.....	177
9.1.1 Hypothesis.....	181
9.2 Method	181
9.2.1 Design and statistical analyses.....	181
9.2.2 Participants.....	182
9.2.3 Materials.....	182
9.2.4 Procedure.....	184

9.3 Results	184
9.4 Discussion.....	198
CHAPTER 10 – GENERAL DISCUSSION	202
10.1 Summary of Findings.....	202
10.1.1 Effect of Format	204
10.1.2 Lack of Age Effect and Measures of Individual Difference	207
10.1.3 Interaction Between Age Group and Task Format	209
10.1.4 Thinking Styles	210
10.1.5 Differences Between Exclusive and Inclusive Disjunctions.....	212
10.1.6 Dual Process Theory	214
10.2 Methodological Limitations	217
10.3 Conclusions	222
References	224
APPENDICES	240
Appendix 1 – Bob task (Chapter 5)	240
Appendix 2 – Venus task (Chapter 5)	241
Appendix 3 – Participant Instructions (Chapter 5)	242
Appendix 4 – Participant Information Sheet (Chapter 6).....	246
Appendix 5 – Conjunctive Tasks (Chapter 6)	247
Appendix 6 – Disjunctive tasks (Chapter 6)	253
Appendix 7 – Instructions to Participants (Chapter 6)	261
Appendix 8 – TDQ and REI as presented to participants.....	265
Appendix 9 – Cronbach’s alpha values for all thinking style measures	269
Appendix 10 – Descriptive statistics for all data from Chapter 6 tasks.....	270
Appendix 11 – Full beta weights for regression analysis (Chapter 6).....	272
Appendix 12 – Information Processing Speed Measure.....	278
Appendix 13 – Beta weights for regression (Chapter 7).....	284
Appendix 14 – Linear and quadratic curve estimations of thinking styles and Bayesian task responses.....	290
Appendix 15 – Correlations of AOT subscales.....	291

LIVERPOOL JOHN MOORES UNIVERSITY

Declaration Form

Name of candidate: Rosemary Stock

School: Natural Sciences and Psychology

Degree for which thesis is submitted: Doctor of Philosophy in Psychology

1. Statement of related studies undertaken in connection with the programme of research

Material presented below has been presented at conferences including, amongst others:

BPS national conference – 2009

Relationships between thinking styles, age, and the type and wording of task on probabilistic reasoning performance

Stock¹, R., Fisk², J. E., Brooks¹, P. & Montgomery¹, C.

BPS Cognitive section – 2008

Aging and the frequency effect

Stock¹, R., Fisk², J. E., Brooks¹, P. & Montgomery¹, C.

BPS national conference – 2008

Effects of aging and of problem format on conjunctive and disjunctive probabilistic reasoning

Stock¹, R., Fisk², J. E., Brooks¹, P. & Montgomery¹, C.

BPS Cognitive Section Conference – 2006

Reasoning errors made in tasks presented in frequency and probability formats

Stock, R., Fisk, J. E., Brooks, P. & Montgomery, C.

School of Psychology, Liverpool John Moores University

1. Liverpool John Moores University
2. University of Central Lancashire

2. Concurrent registration for two or more academic awards

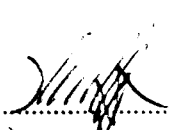
I declare that while registered for the University's research degree, I was, with the University's specific permission, an enrolled student for the following award:

Post Graduate Certificate in Teaching and Learning in Higher Education

3. Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award

Signed.....
(Candidate)



Date.....

12th April 2012

Acknowledgements

First I would like to thank Professor John Fisk, Dr Catharine Montgomery and Dr Phil Brooks for their supervision, wisdom, and incredible attention to detail. Also to everyone who has helped, advised and supported me within the School of Psychology at Liverpool John Moores University, notably Professor Andy Tattersall and Dr Steve Fairclough, technical and administrative staff, and all of my fellow postgraduate students (especially Val Todd, Alastair Gordon-Finlayson, Naomi Fischer and honorary post graduate Bethan Mead). The challenge of completing my PhD part time was made very much easier by having such good support within the department. Those who have helped me outside of my own department at the university include Suzy Hargreaves, of the Centre for Public Health, who offered a range of statistical, practical and motivational strategies. On a more personal level, I have to thank Suzy for being such a very good friend – similarly Sian Jones and Nicola Simcoe.

I have also received a lot of assistance from many researchers across the country at a number of different institutions. There are too many to name individually, but I am very grateful for the mix of practical and emotional support that they were able and willing to offer, both in person and online.

I also want to thank all of the U3A members who helped me to recruit participants, and/or took part themselves, as well as the undergraduate students of Liverpool John Moores University who took part. Many people were very generous with their time and energy, and I am very thankful for their help.

Finally, I want to thank my parents, Rodney and Mary Stock, my sisters, Emily and Jennifer, and Dr Steve Wright. Lastly I would like to dedicate this thesis to my brother, Michael Stock, who inspired me to commit to this programme of research, and to my nephews Callum, Jack and Lewis Wright, who have (unknowingly) motivated me to complete it.

Abstract

Probabilistic reasoning is a distinct type of reasoning which previous evidence has found to be particularly difficult for both naive and expert participants in laboratory research.

The current study looked at probabilistic reasoning performance in the light of dual-process theories of thinking and reasoning, using thinking style measures to investigate whether normative reasoning is indeed most often associated with a tendency to reason analytically, rather than heuristically.

The tasks were presented both as probabilities, requiring participants to think of the chance of a given event happening once and as frequencies, requiring participants to imagine a large number of times or people, and suggesting what number of these might involve the given event. The latter is believed to prime the analytical process of reasoning, particularly when natural, rather than normalised, frequencies are used.

Two age groups were used in order to examine the relationship between cognitive ageing and probabilistic reasoning, and to provide greater variability in a range of individual differences. Using samples of young participants (18-32 years) and older participants (>60 years) the studies reported in this thesis did find a consistent effect of format, whereby those in the frequency format showed both fewer fundamental reasoning fallacies on conjunctive and disjunctive tasks, and lower levels of error, as measured by absolute distance from the normatively correct answer. The format of the tasks – whether probability or frequency – was also an effective predictor of responses to two different Bayesian tasks. Many of the findings regarding the format of the tasks were consistent with dual process theories of reasoning.

There was no effect of age on reasoning performance, despite predictions that older individuals would show less analytical reasoning than the younger group. There was however an interaction effect between the format of the tasks and age group, whereby older participants' performance did not benefit from the frequency wording, indicating that they were either not primed to reason analytically, or that they were primed to do so but were unable to do so to the extent that they could obtain the normatively correct response. More surprisingly, there was no consistent relationship between thinking styles and reasoning performance.

Comparison between current results and previous literature continues to highlight the unique nature of probabilistic reasoning, and the above findings are considered as providing continued support for dual process theories of reasoning. Future research in this area may need to find more accurate ways of assessing an individual's preferred thinking styles, as well as further investigating the nature of the differences between the processes used in completing inclusive and exclusive disjunctive tasks. The measure of reasoning error developed in this current research would also benefit from greater application and further investigation of possible refinements in order to continue to increase our knowledge of how people reason with probabilities.

List of Tables

5.1 Percentage fallacies in disjunction tasks, by condition	96
5.2 Mean and standard deviations of error scores, by task format	97
6.1 Thinking style and Vocabulary Scale descriptive statistics	114
6.2 Table of means and standard deviations for number of fallacies in each group, by task	115
6.3 Table of means/SDs for mean error in each group, by each task	117
6.4 Model summaries for regression analyses of conjunction fallacy data	120
6.5 Model summaries for regression analyses of exclusive disjunction fallacy data	120
6.6 Model summaries for regression analyses of inclusive disjunction fallacy data	121
6.7 Model summaries for regression analyses of conjunction error data	121
6.8 Model summaries for regression analyses of exclusive disjunction error data	122
6.9 Model summaries for regression analyses of inclusive disjunction error data	122
7.1 Mean scores on background measures	134
7.2 Thinking style mean scores by age group	136
7.3 Mean and standard deviations for incidences of each type of fallacy by age group	137
7.4 Means and standard deviations for incidences of each type of error by age group	140
7.5 Model summaries for regression analyses of conjunction fallacy data	145
7.6 Model summaries for regression analyses of exclusive disjunction fallacy data	146
7.7 Model summaries for regression analyses of inclusive disjunction fallacy data	146
7.8 Model summaries for regression analyses of conjunction error data	147
7.9 Model summaries for regression analyses of exclusive disjunction error data	148

7.10 Model summaries for regression analyses of inclusive disjunction error data	148
8.1 Descriptive statistics for participants' judgments	160
8.2 λ and χ^2 values on disease task	164
8.3 Standardized canonical discriminant function coefficients for disease task	165
8.4 Correlation coefficients for disease task	166
8.5 Actual and predicted group memberships for disease tasks	168
8.6 Standardized canonical discriminant function coefficients for cab task	169
8.7 Correlation coefficients for cab task	170
8.8 Actual and predicted group memberships for cab task	172
9.1 Mean scores on background measures	184
9.2 Thinking style mean scores by age group	185
9.3 Descriptive statistics for participants' judgments	186
9.4 λ and χ^2 values on disease task	189
9.5 Standardised canonical discriminant function coefficients for disease task	190
9.6 Correlation coefficients for disease task	190
9.7 Actual and predicted group memberships for disease tasks	191
9.8 Means and standard deviations for response to disease task in each age group, by task condition	193
9.9 Percentage frequency of cab task response groups by task condition	195
9.10 Percentage frequency of cab task response groups by age group	196
9.11 Means and standard deviations for response to cab task in each age group, by task condition	196

List of Figures

6.1 Mean fallacies committed on each task	116
6.2 Mean error on each task	118
7.1 Mean fallacies committed in probability and frequency format	139
7.2 Mean error committed in probability and frequency formats	142
7.3 Task type by age group interaction	143
8.1 Distribution of responses to disease task	161
8.2 Distribution of responses to cab task	162
8.3 Percentage of participants completing the disease task in the frequency condition, by response group	167
8.4 Disease task combined groups plot	167
8.5 Percentage of participants completing the cab task in the frequency condition, by response group	171
8.6 Cab task combined groups plot	171
9.1 Distribution of responses to disease task	187
9.2 Disease task combined groups plot	190
9.3 percentage of participants completing the disease task in the frequency condition, by response group	192
9.4 Mean response on disease task, by task format	193
9.5 Distribution of responses to cab task	194

List of Abbreviations

A	Absolutism
AOT	Actively Open Minded Thinking
BI	Belief Identification
CEST	Cognitive Experiential Self-Theory
CT	Categorical Thinking
CSE	Certificate of Secondary Education
D	Dogmatism
DGC	Dynamic Graded Continuum
EPSE	Everyday Problem Solving Effectiveness
FI	Faith In Intuition
FOR	Feeling of Rightness
FT	Flexible Thinking
IPS	Information Processing Speed
GCSE	General Certificate of Secondary Education
IQ	Intelligence Quotient
MFFT	Matching Familiar Figures Test
MHVS	Mill Hill Vocabulary Scale
NFC	Need For Cognition
O-Level	Ordinary Level
PET	Positron Emission Tomography
QLI	Qualitative Likelihood Index
REI	Rational Experiential Inventory
SAT	Scholarly Aptitude Test
SD	Social Desirability
sfNFC	Short Form Need For Cognition
ST/LC	Superstitious Thinking/Luck Composite
TASS	The Autonomous Set of Systems
TDQ	Thinking Disposition Questionnaire
WAIS-R	Wechsler Adult Intelligence Scale – Revised
WISC-II	Wechsler Adult Intelligence Scale – Second Form

Chapter 1: Thesis Summary

The main body of the thesis is divided into the following 9 chapters.

The first three of these introduce and review the literature in three main areas. Chapter 2 introduces the probabilistic reasoning tasks to be used throughout the research, the ways in which people struggle with these tasks, and the main theories which attempt to explain why systematic errors are made. It looks specifically at the two formats in which tasks are frequently presented – as probability or frequency problems. Chapter 3 then looks at the existing dual process theories of thinking and reasoning and the way in which probabilistic reasoning tasks have been used to examine these theories. It also introduces measures of thinking styles which are believed to correspond to the two processes of reasoning. The final literature chapter, Chapter 4, looks at each of these things in the light of the ageing process, investigating the way in which ageing appears to affect reasoning performance in many contexts but not in the case of probabilistic reasoning. It is suggested that potential age-related deficits may be attenuated by age differences in thinking styles and other individual differences.

The following five chapters are empirical studies of probabilistic reasoning. The first, Chapter 5, looks at the effect of format on conjunctive and disjunctive reasoning and investigates a new way of assessing performance on the tasks by measuring a quantifiable error. No overall effect of problem format was observed in the study. However some limitations emerged with the error measure that was used. These were addressed in Chapter 6. This second empirical study uses improved reasoning tasks and an improved measure of reasoning error, as well as measuring individual differences in the form of thinking dispositions. A consistent facilitating effect of the frequency format was found, but little relationship was observed between reasoning and thinking style. As the ageing process is associated with decline in cognitive function, it was expected that an older group of participants would find the frequency effect particularly beneficial, due to its priming analytic processes which they may be less inclined to use unprompted. Chapter 7 investigated this possibility, and despite replicating the effect of format on all participants it found no interaction between format and age group. While thinking styles remain only occasionally related to performance, measures of information processing speed and verbal intelligence are also found to have mediating effects on reasoning performance. Due to evidence that age effects may only become apparent on the most cognitively demanding of tasks, the next chapter, Chapter 8, presented participants with the more challenging Bayesian reasoning tasks. In this case

the effect of format was present but less straight forward, with each format clearly eliciting particular types of responses, but neither being associated with greater levels of accuracy on the tasks.

The final study, presented in Chapter 9, presented participants with Bayesian problems framed either as probabilities or as natural frequencies, rather than the normalised frequencies used in the previous study. Again, there was a significant effect of format, whereby the natural frequencies did not appear to predict normative reasoning but did affect which task cues were attended to, and resulted in lower levels of overestimation of probabilities.

Chapter 10 discusses the results of each of the empirical studies in the light of previous research and specifically in relation to dual process theories of reasoning. The limitations of the methods used are also discussed in terms of advisable future directions in this area of research.

Chapter 2 – Probabilistic Reasoning Review

2.1 Probabilistic Reasoning

Probabilistic reasoning is a form of inductive extensional reasoning, with its origins being visible in Piaget's 'The Child's Conception of Number' (1952). Piaget presented five to six year old children with sets of beads, which were (for instance) all made from wood, with the majority of them coloured brown, and just two coloured white. Having established that the children understood that the beads were all made of wood, they were then asked 'are there more brown beads, or more wooden ones?' The children consistently answered that there were more brown beads, with Piaget concluding that:

"the difficulty found by children at the first stage in understanding the relationship between the part and the whole is due to the fact that they cannot see the whole as the result of an additive composition of the parts" (p. 172)

They could identify and indicate each of the sub sets of brown beads and white beads, but could not take the step to understanding these as being sets that were included in the global set of 'wooden beads'. This type of problem is still used today (for instance, by Fontaine & Pennequin, 2000) and such class inclusion tasks are very similar to those that are more often phrased as asking for probabilities. To ask the participants 'if we select a bead at random, are we more likely to find we have selected a brown bead or a wooden bead' is to ask them to use the same understanding of the class inclusion as Piaget was asking in his own study. This understanding of the extension principle (that the set of brown beads exists within the extension of the larger set of wooden beads) is thought to be necessary, although not sufficient, for the successful completion of such tasks (Yates & Carlson, 1986; Fisk & Slattery, 2005).

Reasoning is, of course, widely studied in cognitive psychology, and there is some debate as to how 'reasoning' could – or should – be subdivided into subsets or classes (see for instance Manktelow, 2004, addressing the duality of 'theoretical' and 'practical' reasoning). However, there is evidence to support the theory that deductive and inductive reasoning are quite distinct processes – closely related, yet fundamentally different. Osherson, Perani, Cappa, Schnur, Grazzi and Fazio (1998) and Parsons and Osherson (2001) used PET scans to confirm that deductive and probabilistic reasoning

tasks use different brain locations, with deductive reasoning showing more activation in the right hemisphere, and probabilistic reasoning showing more in the left. As such, the distinction between the two types of task appears to be a valid one, indicating that each type of reasoning will benefit from being studied in its own right.

Within the area of probabilistic reasoning, three main types of problem have been extensively investigated. These are conjunctions, whereby a participant is asked for the probability of A *and* B, disjunctions, the probability of A *or* B and Bayesian reasoning, the probability of A *given* some further evidence, B. All of these types of problems are frequently answered incorrectly by participants, who in doing so are disregarding some basic rules of logical inference.

2.2 Conjunctions

By far the most common type of task used in this area, the normative solution to a conjunctive task is found through the conjunction rule, where $P(A\&B)=P(A)\times P(B|A)$ (or $P(A\&B)=P(A)\times P(B)$ for 2 independent events). This rule is often broken by participants, in an error known as the ‘conjunction fallacy’. This fallacy occurs when participants erroneously estimate the conjunctive probability of two events to be greater than that of one or both of the components. This can be seen as an impossible ‘fallacy’ in probability, since probability is expressed as a number between 0 and 1, and the product of two such values will almost always result in a smaller value. Even if the value of either or both events is given as 1 (a certainty), the resulting conjunction can still not then exceed a value of 1. As such, the probability of the product of two components, the conjunction, can never exceed (and is usually less than) the value of either of the individual components alone. However, this fallacy is frequently found in the literature, with Fisk and Pidgeon (1996), Gavanski and Roskos-Ewoldsen (1991), Tversky and Kahneman (1983), Yates and Carlson (1986), Chiesi, Gronchi and Primi (2008) and West, Toplak and Stanovich (2008) all having found that 50-90% of participants are making such errors in conjunctive reasoning tasks, including those with high ecological validity such as tasks based on predicting outcomes of football matches (Nilsson & Andersson, 2010).

One of the most well known and most frequently used tasks in such studies is the Linda problem, first used in Tversky and Kahneman’s study in 1983.

The problem presents a brief vignette about ‘Linda’:

'Linda is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.'

The participant is then presented with (at least) three statements about Linda:

Linda is a bank teller.

Linda is active in the feminist movement.

Linda is active in the feminist movement and is a bank teller.

There have been variations to the study protocol, whereby the required response may be to rank the statements from most to least likely (as in Tversky & Kahneman's 1983 study), or participants may be asked to give an answer of 'how many chances in a hundred' (as in Fisk & Pidgeon, 1996) or they may be given the original information in terms of frequencies (Fiedler, 1988). In this last instance, after the above vignette the participants are asked to *'Imagine that we identified 100 individuals all closely resembling this description of Linda'*. They are then asked *'how many of the 100 would be bank tellers?'* and so on.

In each of the above versions in Fiedler's study (1988), the rates of fallacy were 85%, 70% and 22% respectively – this last being a quite noticeable reduction, but still indicating that a significant minority continue to make the error. Zizzo (2003) also found that there appears to be a 'lower bound' of 20%, with even the most detailed instructions failing to reduce the fallacy beyond this level.

2.3 Disjunctions

A second form of probabilistic reasoning used in this area is that of disjunctive reasoning. In this case, $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$, so that the disjunctive probability of the two components, $P(A)$ and $P(B)$, cannot logically be less than the individual probability of either one of them. That is, the probability that Linda is active in the feminist movement *or* is a bank teller cannot be less than the probability of either one of those things on their own.

This is specifically an *inclusive* disjunction, since the term P(AorB) includes the possibility that A occurs, the possibility that B occurs, and also allows for the possibility that the conjunction A&B occurs. An *exclusive* disjunction would be one where the term ‘or’ in P(AorB) was used to mean that either A or B occurred, *but not both*. The formula for this is P(AorB)=P(A)+P(B). So the probability that Linda was a full time bank teller or that she was unemployed could reasonably be seen as being an exclusive disjunction, since she cannot be both of those things at the same time (i.e. the value of A+B in this case would have to be zero, as an impossibility), while the probability that she was a full time bank teller or a member of any given political party would be an inclusive disjunction, as both occurring together would be quite possible. Inclusive and exclusive disjunctions will be discussed in more depth in later chapters, but it should be noted here that the ‘disjunction rule’, that the probability of the disjunctive event cannot be less than the probability of either single component, is applicable to both types of disjunction.

Again, this disjunction rule is often violated, as documented by Carlson and Yates (1989) who found an up to 80% incidence of the disjunction fallacy, using inclusive tasks, and by Fisk (2005) who found a lower, but still sizable incidence of up to just over 60% when using both exclusive and inclusive tasks.

However, there is less literature available in this area, with the bulk of the research conducted so far focusing on either conjunctive tasks or on Bayesian ones.

2.4 Bayesian Tasks

The third task addressed here will be the Bayesian task, an area that has been extensively investigated in the probabilistic reasoning literature, for instance by Phillips and Edwards (1966) and again from Kahneman and Tversky (1972 and onwards).

Bayes’ theorem refers to the probability of a certain event (represented below as E) *given* some relevant evidence (A):

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

In this equation, $P(E)$ is 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.

Fisk (2005) details how even if this normative process is not carried out exactly as shown, any individual who is attempting to solve the problem by utilising all the available evidence will still be manipulating a considerable amount of information as they do so (see also Stolarz-Fantino, Fantino & Van Borst, 2006). Studies using Bayesian tasks will often find that participants will largely ignore the base rate, and simply base their decision on the new evidence (Cobos, Almaraz & Garcia-Madruga, 2003; Tversky & Kahneman, 1983) while Birnbaum (2004) also suggests two other modal responses – that of apparently ignoring the new information, and responding with the base rate probability (also found by Philips & Edwards, 1966, who referred to this as conservatism) and that of an averaging technique, where the participants appear to multiply the base rate by the new evidence. As with disjunction and conjunction errors, research has suggested that rephrasing tasks as frequency problems can reduce the number of errors being made (Baratgin, 2002), for example, increasing the generation of the ‘correct’ response from 10% to 46% in an often cited sample of physicians tested by Gigerenzer (1996a).

Researchers in this area have put forward a wide range of reasons for the high incidence of each of these fallacies, the foremost of which are discussed in turn below. However before turning to these explanations, it should also be noted at this point that many of the researchers in this field can be identified (and indeed, would identify themselves) as being *either* ‘Bayesians’ or ‘frequentists’. Kahneman and Tversky (1996), as Bayesians, summarised the two labels as follows:

“Proponents of the Bayesian school interpret probability as a subjective measure of belief. They allow the assignment of probabilities to unique events . . . and require these assignments to obey the probability axioms. Frequentists, on the other hand, interpret probability as long-run relative frequency and refuse to assign probability to unique events ” (p. 582)

The paper in which this quote appears, is entitled ‘On the Reality of Cognitive Illusions’ and is a direct response to Gigerenzer’s previous criticism of Kahneman and Tversky’s approach and some of their writing in the area. Gigerenzer, a frequentist, feels that they

are restrictive in their definition of normative reasoning, and that the heuristics and biases research fails to give concrete and specific descriptions of the cognitive processes behind reasoning (Gigerenzer, 1996b). Whilst the debate between the above researchers is not purely based on the conflict between the frequentist and Bayesian approaches, it is clear that the different perspectives will colour not just researchers' approaches to data collection, but also their interpretation of their findings. For instance, frequentists feel that the term 'conjunction fallacy' is inappropriate, and may often be ignoring the context and the wording of the tasks involved (Gigerenzer, 1996b). It seems that the term *fallacy* is, to frequentists, somewhat reductionist in its presumption that the rules of mathematical logic should always apply, regardless of the context.

The frequentist approach also encompasses the view that people can and do reason intuitively, when materials are presented to them in a way that allows them to do so - that is, when they are not presented as individual events (Cosmides & Tooby, 1996), and are done so as natural frequencies, rather than as normalised ones (Hoffrage, Gigerenzer, Krauss & Martingnon, 2002). For an earlier review of the ways in which research has supported the concept that humans can be and often are intuitively using correct statistical rules in their reasoning, see Peterson and Beach (1967). Peterson and Beach do support the theory that individuals are intuitively capable of normative judgements, but state some reservations regarding, for instance, the tendency for people to consistently underestimate values. Cosmides and Tooby (1996) found that 92% of their participants could solve a Bayesian task when presented to them in what the authors deemed to be the most ecologically valid condition, the frequency condition, compared with a 12% accuracy rate in the non-frequency, Bayesian condition. They conclude that we can reason accurately when enabled to do so, and that when people appear to be reasoning incorrectly, and failing to arrive at the normative answer, there may still be some logic to their doing so – a suggestion which does take into account the fact that many errors in probabilistic reasoning do appear to be somewhat systematic, as illustrated below.

2.5 Explanations for Reasoning Errors

As mentioned above, the majority of research conducted so far has focused on conjunctive reasoning. As such, it is the conjunction fallacy that has been the main reasoning error examined and discussed, and the exceptions to this are clearly indicated.

2.5.1 Linguistic misunderstanding

For participants to be able to respond with the ‘correct’ answer as defined by the researcher, they must also be interpreting the questions as the researcher intends. Morrier and Borgida (1984) and Yates and Carlson (1986) have examined the suggestion that participants simply misinterpret the components as meaning, for instance, ‘Linda is a bank teller and *not* a feminist’, and they conclude that this is not enough of an explanation, through the use of what Morrier and Borgida refer to as ‘debiased’ problems. In these tasks, additional statements made it clear whether or not a component meant ‘this component and no other’ or ‘this component regardless of any other’. Such additional information did reduce the incidence of the fallacy in one task, but certainly did not remove it altogether.

Another possible misinterpretation of the tasks is that the term ‘and’ is given a causal interpretation, essentially being seen as ‘and then’, ‘and therefore’ or ‘and p is the cause of . . .’ Hertwig, Benz and Krauss (2008) report that 54% of participants identified one of these three interpretations as being the way they themselves responded to a conjunctive task, but this does then highlight the fact that the remaining 46% gave the simple ‘and’ as their interpretation, specifically rejecting the causal alternatives.

Tentori, Bonini and Osherson (2004) and Wedell and Moro (2008) have illustrated that even when participants understood the use of ‘and’ as intended, they still displayed high incidences of the fallacy, and similarly Sides, Osherson, Bonini and Viale (2002) found there to be no difference in performance between a group given the term ‘and’ and those given a range of other conjunctive terms. Sides *et al.* (2002) also felt that their work made clear to participants (who continued to make the fallacy) exactly what they were choosing. So by choosing that the occurrence of X was the most likely out of the three options presented to them (the occurrence of X, or Y, or X and Y) participants were aware that this did not mean that they were choosing ‘X *and only X*’. They were instead well aware that by choosing X as being mostly likely, they were not excluding the possibility that event Y might also occur. As such, while it is not possible to rule out the fact that some individuals may misinterpret tasks in this way, it seems unlikely that it is responsible for the levels of systematic error as described in the literature above.

Carlson and Yates (1989) also looked at this possibility in relation to disjunctive tasks, specifically addressing the possibility that individuals may be misinterpreting the term

'OR'. They suggest it is possible that participants are reading 'or' as meaning '*a or b but not both*' (an exclusive disjunction), rather than the logically accurate '*a or b or both*' (inclusive disjunction). They again found that addressing this problem did not reduce the incidence of the disjunctive fallacy, but as stated above the disjunction rule (that the probability of the disjunctive event cannot be less than the probability of either single component) applies whether the task is inclusive or exclusive. A more sensitive measure of error would consider that fact that the normative answer to an exclusive disjunction – $A \text{ or } B \ \& \ \text{not}(A\&B)$ – should be a smaller value than that for an inclusive disjunction – $A \text{ or } B \ \text{or } (A\&B)$. If participants are able to recognise and respond to this, levels of error should be the same for each task type.

2.5.2 Signed summation

Suggested by Yates and Carlson in 1986, signed summation is a heuristic whereby participants assign unlikely events/components with a negative value, likely ones with a positive value, and then add these two values together. This results in a conjunctive value that is either negative (unlikely) or positive (likely), with a large negative value for one component being partially offset by a smaller positive value for the other, and vice versa. Given the laws of arithmetic, the sum of a negative and a positive number must be greater than the negative number and less than the positive one. This means that the conjunctive probability would be assigned a number which exceeds the number assigned to the unlikely event, thus giving rise to the conjunction fallacy. This would also lead to the expectation that two positive or two negative components should result in a highly positive or highly negative conjunctive value, respectively. Thus the conjunction of two likely events might be expected to give rise to a double fallacy (a conjunctive value which is bigger than both of the components), while the conjunction of two unlikely events should not give rise to the fallacy. Fisk and Pidgeon (1996) show some support for this model, but there is little investigation of this theory in the literature in general. Gaynor, Wahio and Anderson (2007) have referred to this heuristic as algebraic summation, describing the conjunctive values obtained as being intermediate responses. However this blurs the line between signed summation and simple averaging, a model which is described in more depth below.

2.5.3 Representativeness

This heuristic has features in common with signed summation, and states that it is the most likely, or representative, component of a conjunction on which the value of the conjunction itself will be based, resulting in the fallacy occurring most frequently in instances where there is one very likely component, and one very unlikely component (Wells, 1985).

However, the signed summation heuristic is directly referring to the magnitude of the likelihood of each component, allowing either a negative or a positive value to have an impact upon the conjunction, whilst the representativeness heuristic would focus only on the component that was considered at all more likely, allowing that component to have the greatest impact on the conjunctive value. Nilsson, Juslin and Olsson (2008) suggest that the specific process involved is that the presented information (i.e. the conjunction) is compared with stored exemplars (of the components), and the probability of the presented information is then based on its similarity to the most frequently occurring exemplars (see also Juslin & Persson, 2002).

Bar-Hillel and Neter (1993) also provided support for the use of the representativeness heuristic in inclusive disjunctive tasks, although they also conclude that this approach cannot account for all occurrences of the fallacy. There is also some evidence that children aged ten and under are using a representativeness heuristic to answer conjunctive, although not (inclusive) disjunctive, tasks (Fisk, Bury & Holden, 2006).

The problem with both signed summation and the representativeness heuristic is that they suggest that a *very* likely (or representative) component should sway the conjunction to also be more likely/representative. That is, they should give it a higher probability. However, in Fisk's regression analyses, it is found that the likely component is usually not a statistically significant predictor of the conjunction (Fisk, 2002; Fisk & Pidgeon, 1996, 1997) Also, Fisk and Pidgeon (1998) and Thuring and Jungermann (1990) present further evidence that appears to contradict explanations based on representativeness.

Fisk (2002) also addressed the role of likely and unlikely components within inclusive disjunctions, and found that in this case the more likely/representative component seems to have more influence in determining the value of the disjunction in cases where both

components appear unlikely, or one is likely and one unlikely. In those cases where both were likely, to a greater or lesser degree, this influence was not apparent.

Gavanski and Roskos-Ewoldsen (1991) suggest that representativeness is being used only in estimating the value of each component, and not by judging the representativeness of the conjunction as a whole event. They claimed to have eliminated participants' reliance on the representativeness heuristic by using problems that were based around the imaginary situation of a planet, Kropiton, inhabited by the fictional Gronks. The participants were given the likelihood of the Gronks having blue hair, and then the likelihood of them having three eyes, and were asked for the conjunctive value of the likelihood of meeting a blue haired, three-eyed Gronk. By using imaginary creatures they theorised that 'subjects would have no basis for representativeness-mediated judgments' (page 184). However, it is not clear that this does truly eliminate the possibility of representativeness being used, as individuals may very quickly make their judgements about these new situations, and form a mental picture of a representative Gronk as they are reading the information. A person does not have to be an expert on any set of possible events, or even have heard of them before that day, to use the heuristic. More recently, Handley, Evans and Thompson (2006) found the conjunction fallacy occurring when participants were asked to endorse various outcomes in conditional reasoning. The fallacy was shown by participants being more likely to endorse the conjunction p and $not-q$ than they were to endorse either of the two components alone. It seems very unlikely that the fallacies in this type of conjunctive reasoning, with its abstract components, can be due to the representativeness heuristic. However, as with the 'Gronks' above, it is not entirely beyond the realms of possibility that participants may be rapidly forming their own impression of a representative card. See also Nilsson (2008) for similar attempts to manipulate the possible use of the heuristic.

Tversky and Kahneman (1983) also suggest that when a positive conditional relationship exists between two components, then the likelihood of the conjunction fallacy is *increased*. So, for instance, the likelihood that it will be snow and be below zero centigrade tomorrow is more likely to cause the fallacy than the likelihood that it will snow and be a Tuesday tomorrow, since the snow and the temperature would be in a conditional relationship, while the snow and the day of the week would not be. It might also be reasonable to suggest that a very cold day is more representative of a

snowy day than a Tuesday would be. They also assert that the strength of the conditional relationship is relevant in determining the extent to which the fallacy occurs. Fabre and Caverni (1995) also found that both the causal relationship and the strength of the relationship did have significant effects, while Fisk and Pidgeon (1998) support the suggestion that a positive conditional relationship makes it more likely that the fallacy will occur, but also found that the strength of that conditional relationship does not seem to affect the incidence of the fallacy.

Further evidence for participants basing their judgements, in some way, on how the conjunction 'fits' with their mental image or construct of the person or situation being described is provided by the findings of Stolarz-Fantino, Fantino, Zizzo and Wen (2003). They used a range of tasks, including the Linda problem, and found that if they did not provide the short vignette of information about the person in question, the incidence of the fallacy was significantly reduced (although still apparent, with the reduction being from 62% to 45%). Stolarz-Fantino *et al.* (2003) provided their participants with values for each component, asking them only to produce the conjunctive value. As such, it was not the case that those in the 'no vignette' condition produced component estimates, which may have led to more accurate calculations of the conjunction. This strongly suggests that those reading the vignette were swayed by the impression that they formed from the background information, while those with no vignette were reasoning more normatively due to less interference from background information.

2.5.4 Potential surprise

This is based upon Shackle's (1969) theory that a probability estimate of a component is based upon the potential surprise of that event actually occurring, and was developed as a theory by Christensen (1979) to explain the way conjunctions are estimated. This way of estimating the conjunction does not use any calculation, but instead suggests that the conjunction's surprise value (and therefore its estimated probability) is based upon the surprise value of only the *most* surprising/unlikely of the two components. Christensen uses the example of blood type, suggesting that as the chance of having blood group O is around 0.4, and the chance of having group Rhesus negative is around 0.15, then the conjunctive value of having O Rhesus negative blood must be 0.06. However, his interpretation of the theory of potential surprise is that since being blood group Rhesus negative is the most surprising thing, the conjunction's 'surprise value' is not actually

any higher (and therefore its probability is not any lower) than the value of that one component. As such, in conjunctive tasks a participant using surprise value as their guide would give the value of O Rhesus negative as simply being the same as that of being Rhesus negative, rather than doing the calculation to get the mathematically correct response. Christensen (1979) also stated that an event cannot have a negative surprise value – it is either surprising, with a surprise value of above zero, or not surprising, with a value of zero.

If this heuristic is being applied, then ‘the conjunction of two or more events is never more surprising than its most surprising constituent (and by implication not less likely than its least likely component)’ (Fisk, 2004) p.31. As such, the prevalence of this heuristic would mean that almost all conjunction tasks would result in the fallacy occurring – the conjunctive value given would always be equal to (if not greater than) that of one of its components. However, although occurrences of the fallacy are widespread they are not 100%. A further problem with this theory is that it fails to define exactly how the ‘surprise’ values are found.

This heuristic is in direct contrast to the representative heuristic discussed above, whereby the most representative (or least surprising) component has the most influence on the value given to the conjunction (Kahneman & Tversky, 1973). Evidence for Shackle’s theory is provided by Fisk (2002), Fisk and Pidgeon (1996, 1997 & 1998) and Fisk and Slattery (2005) as well as by Thuring and Jungermann, (1990), Kariyazono (1991) and Hertwig and Chase (1998).

Also relevant here is Gigerenzer and Goldstein’s (1996) one reason algorithm. They suggest that judgements are frequently made by focusing on only one cue, which is perhaps what is occurring when participants are relying on the one cue of ‘surprise’. This proposal would have some relevance ecologically, since in conjunctive judgements an alteration to the smaller component does have more impact on the conjunction itself – so in many real life situations, this would be a reasonable short cut to take.

Surprise theory is, however, somewhat under investigated and remains axiomatic in that while we can see that some of the data fits this model, the constructs underlying the theory remain unexplained in psychological terms and despite attempts by Fisk (2002),

and by Fisk and Pidgeon (1998), to further delineate the theory the underlying processes remain unspecified.

2.5.5 Frequency interpretations

The third version of the Linda problem in 2.2 above, whereby respondents were encouraged to '*Imagine that we identified 100 individuals all closely resembling this description of Linda*' and then asked '*how many of the 100 would be [bank tellers etc.]?*' greatly reduced the incidences of the fallacy (Tversky & Kahneman, 1983; Fiedler, 1988). This way of wording the problem is often known as a 'frequency version', and is thought to encourage participants to think in terms of numbers of actual people (or occurrences), rather than in terms of a percentage chance of a hypothetical person or event.

In the case of Bayesian reasoning tasks, Gigerenzer and Hoffrage (1995) found that up to 50% of participants would give the normative response to a problem phrased in the frequency format, whilst the probability formats they used showed a maximum of 28% correct responses.

Hoffrage *et al.* (2002) stress the importance of using natural frequencies in Bayesian tasks especially, as do Gigerenzer and Hoffrage (1995) and Mellers and McGraw (1999). They feel that these are presenting information in the same way as we naturally collect it during our lives, that is by using naturally occurring reference classes rather than with unnatural and normalised ones. That is, we may know that our post has a tendency to arrive before we leave for work on 2 out of 5 working days a week (5 working days being a naturally occurring reference class to most people living in the United Kingdom) and be more able to extrapolate from that information than if we are told to consider this information in terms of 40 out of 100 – the latter being a normalised reference class. This is discussed in more detail in Chapter 9.

Fantino and Stolarz-Fantino (2005) also claim strong support for the frequency effect with regards to Bayesian reasoning, but do so only by presenting the mean responses of their participants, with the mean for the frequency group being far closer to the normative value than for the group given a percentage version of the task. However, these means do not indicate how many participants in each case actually did give the

normative answer – it may be that the means represent an average of extreme responses in each direction, both over and underestimates.

The ecological validity of such versions is also questionable – in real life such decisions are far more often based on the percentage chance associated with one individual event or person (often ourselves), rather than our being required to think of a large number of people/events, and to decide upon the number of people/events that we think are likely to fit our current criteria.

Teigen, Brun and Frydenlund (1999) suggest that individuals do not always understand the importance of obtaining and using frequency information. For example Teigen (2005) found that when two types of frequency information are available, it was the least relevant which the participants used to make their estimate. Also, Evans, Handley, Perham, Over and Thompson (2000) found that such a format does not invariably improve reasoning performance. They suggest (and illustrate, by wording probability problems to get the same effect) that it is the wording used in the frequency format that enables participants to better understand the extensional relationship between conjunction and components. This raises the idea that the ‘frequency effect’ is something of a misnomer – such tasks do frequently lead to improved reasoning performance, but due to other differences in the wording used in the frequency style tasks, rather than due only to the removal of percentage probabilities. Sloman, Over, Slovak and Stibel (2003) also agree that framing the tasks as frequency problems can cause a significant reduction of reasoning errors, in both Bayesian and conjunctive tasks, but they feel that this is due to the framing revealing to participants that the tasks can be solved by using nested sets, an approach which has similarities to the mental models theory, as discussed below. As such, it may be possible for participants to reason just as well with problems that are not in the frequency format, so long as they are given similar framing details and instructions (see also Wedell & Moro, 2008, whose tasks did not elicit a ‘frequency effect’).

Both Kahneman and Frederick (2002) and Evans (2007a) have suggested that the frequency format is either cueing or enabling more analytic and less heuristic reasoning processes, by encouraging participants to see the nested sets in a task.

It does seem from the evidence above that the problems worded in what is described as the frequency format can often improve reasoning performance, and reduce the occurrence of the reasoning fallacies discussed earlier. It is worth considering that when the tasks are not presented in this format, but are instead phrased as asking for a percentage possibility, some individuals may still be approaching them as frequency tasks, and that this may account for the percentage of people who avoid committing the relevant fallacies.

2.5.6 Applying the wrong probabilistic rule

Wolford, Taylor and Beck (1990) feel that the Linda type problems may let participants think that they are being asked to look for which of the statements is actually *true* – so they are comparing the statements to each other, rather than using the conjunction rule. Respondents committing the fallacy in this case would effectively be saying ‘I think it *most likely that* Linda is a feminist and a bank teller,’ and *less likely* that she is only one of those two things. The participants are effectively estimating the Bayesian reverse probabilities, rather than the simple conjunctive probability. So they are considering the Linda problem to be asking them ‘Based on the information already provided about Linda, how likely is it that the person described is Linda given that they are a feminist and a bank teller?’ While $P(\text{feminist \& bank teller} | \text{Linda}) > P(\text{bank teller} | \text{Linda})$ violates the conjunction rule, $P(\text{Linda} | \text{Feminist \& bank teller}) > P(\text{Linda} | \text{bank teller})$ does not necessarily violate any probabilistic rule. As such, Wolford *et al.* (1990) would suggest, the individuals giving higher values for the conjunction than either component may in fact be reasoning normatively, albeit with the wrong rule.

However, Fisk (1996) found that clearer wording of the tasks, which addressed the possibility of participants misunderstanding its requirements as suggested by Wolford *et al.* (1990) still led to the rule being violated. As such, while the ‘wrong rule’ explanation does have some support, in that some participants will be misunderstanding what is required and applying the wrong probabilistic rule, it still fails to explain all incidences of the fallacy.

Sides *et al.* (2002) and Crupi, Fitelson and Tentori (2008) have suggested that the conjunction fallacy is actually a result of ‘confirmation relations’, with the apparently fallacious responses being the result of the given evidence (the vignette) increasing the credibility of a following component, leading to it appearing more credible, or probable.

than the conjunction. Crupi *et al.* (2008) do themselves conclude their research by acknowledging that this suggestion is currently purely theoretical, and has yet to have been supported empirically.

Birnbaum (2004) has also suggested that participants are in fact often reasoning ‘correctly’ on Bayesian problems, and that it is the researcher who has set an incorrect normative answer to the task. That is, the ‘wrong’ answers may in many cases appear more reasonable in a real life setting than those arrived at by the Bayesian equation. An example of this would be in the case of a task that requires participants to consider the following, taken from Evans *et al.* (2000):

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? ___ %

Those that give an estimate far greater than the normative response (which is 1.96%) may actually be giving a more realistic estimate based on the assumption that the person being tested has presented with some symptoms of the disease, or has some other indicator leading to the test being necessary. In this situation, it could seem reasonable to think that the prior probability of the person having the disease is in fact higher than the given ‘one out of 1000’ that applies to the rest of the (non-symptom showing) population, and this could justify the answer as being higher than the normative response that is deemed correct (by the researcher).

2.5.7 Mental Models

Already an influential and pioneering approach developed by Johnson-Laird with regards to other forms of reasoning (Johnson-Laird, 1983) this was also applied to probabilistic reasoning by Johnson-Laird, Legrenzi, Girotto, Legrenzi and Caverni in 1999. The theory suggests that when solving problems, we construct mental models of

the various possibilities, and the greater the number of models any given component fits into, the greater is its assigned probability.

In a review of mental models and their applications, Johnson-Laird further examined the way that models may be constructed in solving probabilistic problems (2005). He suggests that participants may follow a principle of equiprobability, whereby each model that they construct is given an equal probability. He uses the following inclusive disjunction task as an example (p197):

In the box, there is a green ball or a blue ball or both.

What is the probability that both the green and the blue ball are there?

Three mental models would be constructed – one which allows for only the green ball, one for only the blue ball, and one for both. This then leads to the conclusion that there is a one in three, or 33% chance of there being both balls in the box, where the normative answer would be 25%. In Johnson-Laird *et al.* (1999) he also provides illustrations of a similar pattern in disjunctive tasks, and again in the 2005 chapter mentioned above he further applies the theory to Bayesian reasoning. He uses the following example, (p.197):

The patient's PSA score is high. If he doesn't have prostate cancer, the chances of such a value is [sic] 1 in 1,000. Is he likely to have prostate cancer?

Johnson-Laird suggests that this leads immediately to two models. Firstly, an explicit one described within the task, where the patient does not have cancer, but has the high PSA score, which is given a value of 1. Secondly, an implicit one where all other options are combined and given a value of 999. To stop at these models would then lead to the incorrect reasoning, that the patient has 999/1000 chances of having the cancer, as he apparently only has a 1/1000 of *not* having it.

However Johnson-Laird suggests that the further two models of a) cancer but not high PSA (a value of zero, as an impossibility) and of b) cancer and high PSA (Johnson-Laird gives a value of 2 for this, the fact that even with high PSA prostate is relatively unlikely, without detailing how this is obtained) can be added, to create a Bayesian task with enough information to solve the problem. In his own notation:

¬prostate cancer	high PSA	1
¬prostate cancer	¬high PSA	999
Prostate cancer	high PSA	2
Prostate cancer	¬high PSA	0

Participants are aware that they are looking only for the subset of those patients who do have the high PSA level, and can see that out of the value of 3 assigned to the two conditions of high PSA, only 2 of them also have the cancer. The correct value then is that the patient has a 2/3 chance of having cancer. Failure to reason correctly, therefore, is due to a failure to produce enough models.

Most participants will make the error of overestimations as detailed above, and while this can be interpreted as support for the model theory, it is not clear how the mental model theory would account for the fact that some models (whether they are those that are more representative, as suggested by, e.g. Bar-Hillel and Neter, 1993, or those that are more surprising, as illustrated by Fisk, 2004), are apparently given more weight than others, above and beyond that suggested by the contents of the problems themselves.

Similarly, Over (2005) is supportive of mental models in many areas of reasoning, but expresses reservations about mental models in probabilistic reasoning. He describes one of the main flaws of the mental model account as failing to take into account the fact that individuals may be giving each model a different weighting of likelihood, rather than using ‘the principle of indifference’, in which each model is deemed equally likely.

Gigerenzer and Hoffrage (1999) suggest that one of the reasons that frequency formats can facilitate reasoning performance (as discussed above) is due to their encouraging the visualisation of the problems, and as such facilitate the use of mental models.

2.5.8 Fast and Frugal

Hertwig and Chase have suggested that the surface form of a problem determines the strategy used to solve it. Their 1998 paper looks at Gigerenzer’s fast and frugal heuristics theory, and specifically the idea of ‘one reason decision making’ – that is, the theory that participants are basing their judgements on just one cue selected from the available information (see Gigerenzer & Goldstein, 1996). In the case of the Linda task,

the cue could be any of the information presented to the participants, and in this way the cue that Linda ‘participated in anti-nuclear demonstrations’ could be the basis for a participant feeling that Linda was more likely to be a feminist *and* a bank teller, than simply a bank teller. They hypothesised that performance on the tasks would be unrelated to participants’ statistical expertise, and also that numerical problems would lead to rule based strategies (such as multiplying the values of the components, or the ceiling rule, whereby the conjunction is given the same value as the smaller/less probable component, as well as a form of ‘quasi-normative multiplication’), while ranking problems would lead to cue based strategies. This was supported, with participants in the numerical condition appearing to use the ceiling rule. They also found that naïve participants used the ceiling rule, while experts were more likely to be using multiplication.

Hertwig and Chase’s suggested method of estimating the cue value is essentially working out the reverse conditional probabilities (1998). That is, working out if the particular information in the vignette is true, given each of the components/conjunctions they are being asked to judge. This would concur with the work of Wolford *et al.* (1990), who suggested that the fallacy is a consequence of participants responding to conjunctions by applying the wrong probabilistic rule. Fisk (1996) argues against this proposition, demonstrating that even where participants are encouraged to distinguish between the two normative rules (i.e., the conjunction rule and Bayes’ theorem) many of them still produce non-normative results, suggesting that this particular account of the fallacy is again only a partial one.

This fast and frugal approach could be seen to be in accord with both representativeness and potential surprise, with the solver making their decision based only on the most salient component of the conjunctive, disjunctive or Bayesian task, whether that ‘saliency’ is due to its representativeness, or its surprise value. It would also be compatible with Fiske’s cognitive miser theory (1991), whereby individuals’ reasoning methods focus on being efficient, and taking the path of least cognitive effort, often at the cost of failing to be normatively correct.

2.5.9 Averaging

Averaging was one of the earlier approaches to this area, and continues to have an influence on the current literature. Wyer (1970) suggested that participants were simply

taking the average of the two components' values to make their conjunction judgement, but in fact it was found that the conjunctive judgements being made were actually more strongly correlated with the product of the two components and the normative response (that is, 'component \times component \times normative', as opposed to '(component + component)/2'). Zizzo (2003) also found that the models of averaging and product showed the best fit with their own data when compared with a simple summation rule. However, Juslin, Nilsson and Winman (2009) suggest that a better fit is a form of weighted additive integration, whereby the lower valued component is given greater weight, a suggestion similar to surprise theory, as discussed above.

Goldsmith (1978) found a range of arithmetic rules being used on a conjunction problem. These were the simple averaging of the two components to come to a conclusion about the conjunction, basing the conjunctive value on the value of the smallest (or most unlikely) of the two components, and lastly by multiplying together the two component values. Goldsmith's work did not suggest that the more probable/least surprising event plays a significant role in estimating the conjunctive value, but would instead support the view that it is the least likely, or most surprising component that has the most influence. Goldsmith also found little evidence that the conditional probability is being used in the process. Further support for the use of the averaging model is provided by Gavanski and Roskos-Ewoldsen (1991), who also suggested that participants often seemed to average the values and then adjust the resulting value down slightly – which would reduce the magnitude of the fallacy in each incidence.

An averaging model by Birnbaum, Anderson and Hynan (1990) looks at subjective probabilities of the component and conditional, and takes parameters and random error into account. They propose the following hypothetical equation:

$$q(A \& B) = s(A)^\beta s(B | A)^\alpha + e_{A\&B}$$

$s(A)$ being the estimate of the component event, and $s(B|A)$ is that of the conditional probabilities, while $e_{A\&B}$ is the random error.

If the model is correct, then when the parameters α and β equal 1, the model would give a normative response. In reality, their model fitted their own data far better when the parameters were actually considerably less than 1. As has been detailed above, the fallacy was again found to be occurring more often when the components had very different values (i.e. one high/probable/representative/likely, and one not). Birnbaum (2004) also details how averaging approaches can be seen in responses from Bayesian style reasoning tasks. Using log-log regressions, Fisk (2002) found that the parameters differed between the fallacy and non fallacy groups, with the non fallacy group's being nearer the normative value of 1, and the fallacy group's being less than 1. The influence of the likely component value in determining the conjunctive value was either not statistically significant (fallacy group) or substantially less in terms of the standardised beta weight (non fallacy group), again showing the stronger influence of the unlikely component.

This approach again fails to explain the psychological processes involved, or why some participants seem to give the conjunctive value as the same as the lower component, or why some participants appear to be giving the reverse probability estimates – that is, apparently responding to 'P(Linda|bank teller & feminist) rather than P(bank teller & feminist | Linda).

2.5.10 Dual Process Theories

Developed by Seymour Epstein (see for example Epstein, Lipson, Holstein and Huh, 1992), Cognitive Experiential Self-Theory (CEST) is a dual process theory which attempts to explain various judgements (and errors of judgement), including occurrences of the conjunction fallacy (Donovan & Epstein, 1996). They propose two systems – rational and experiential – and detail the characteristics of each. The latter utilises heuristic strategies, and is thought to be the 'default' system. As such, this theory need not be seen as in conflict with all those above, but instead would encompass them all as being part of, or perhaps a result of, the experiential system.

They set out two dimensions that exist within reasoning tasks – abstract/concrete, and 'natural/unnatural'. Abstract/concrete refers to whether abstract concepts, such as algebra, are being used, rather than concrete examples of real people, objects and/or events. Natural/unnatural refers to whether or not the problem elicits the appropriate strategy to obtain the correct solution, where natural tasks do so and unnatural tasks do

not. They illustrate that the concrete natural problems used in their study led to significantly fewer conjunction errors compared to the concrete unnatural problems, stating that the former triggers the rational system, the latter the experiential system. Thus it is the application of the experiential system which in many cases results in the fallacy being committed.

One problem with this approach (highlighted by Fisk 2004) is that the concrete natural problems used in the 1996 study by Donovan and Epstein involved the ranking of events, and so do not give the component values. Their problems also include two unlikely events, which, as shown by Yates and Carlson (1986), are far less likely to show the fallacy. As such, Fisk suggests that the findings of Donovan and Epstein may not be due to their abstract/concrete and natural/unnatural dimensions, but instead be dependent upon the size of the component probabilities involved.

Nonetheless CEST has the potential to offer useful insights into reasoning processes. The dual process perspective is capable of accounting for the normative reasoning which is evident in a minority of participants since these individuals may be utilising the rational system. The approach is also discussed by Stanovich and West (2000), who label the systems System 1 (the equivalent of Epstein's Experiential system) and System 2 (the Cognitive system). And further evidence for the existence of two discreet systems is provided by Sloman (2002) who defines them in rather different terms as the associative and the rule-based system. Similarly Zizzo (2003) characterises the dichotomy in terms of implicit and explicit learning processes which in turn are associated with discrete improvements in implicit and explicit reasoning.

One area for further investigation here is how these different systems are primed, and whether this could offer any explanation for the fact that phrasing problems in terms of frequencies (instead of probabilities) appears to reduce the occurrence of the fallacy (Fiedler 1988). While Donovan and Epstein (1996) name the heuristics based system 1 as the default system, Kahneman and Frederick (2002, 2005) suggest that system 2 supervises system 1. In the model that they propose, intuition – that is, system 1 – is only used when this has been approved by system 2, suggesting a fair amount of monitoring by this cognitive system. However, errors in simple reasoning tasks occur when the supervision of system 1 by system 2 is too superficial – they suggest that in

the case of the simplest tasks, system 2 can relinquish control too readily, leading to heuristics being inappropriately applied. This is discussed in greater detail in Chapter 2.

Another possibility is that the ageing process leads to a greater reliance on system 1 processes (Fisk, 2005; Yates & Patalano, 1999). Fisk (2005) found that older participants rely less on working memory – one of the cognitive functions most affected by age related decline – when solving probabilistic reasoning tasks, suggesting that they are adapting to their limited cognitive resources and moving away from analytic processes. However Fisk (2004, 2005) relies to a large extent on analysis of whether or not participants have completed a fallacy, which may not be a sensitive enough measure to identify more subtle differences in the *types* of response given by participants. A more sensitive way of assessing reasoning – the mean error – is examined in chapters 5 to 7, and in chapter 8 further attention is given to the shift in focus from ‘difference from normative’ to ‘difference from other participants’, using individual differences as potential predictors of these differences. The impact of age related cognitive decline on reasoning skills, and its value as a tool for examining probabilistic reasoning in both older and younger participants, is discussed in full in chapter 4.

2.6 Conclusions

It is clear that people do not reason normatively on the three main types of task used to assess probabilistic reasoning – conjunctions (Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Tversky & Kahneman, 1983; Yates & Carlson, 1986), disjunctions (Carlson & Yates, 1989; Fisk, 2005) and Bayesian tasks (Birnbau, 2004; Cobos, *et al.*, 2003; Tversky & Kahneman, 1983). However, it is also clear from the literature reviewed above that there are a wide range of contradictory explanations for this non-normative reasoning, each of which have a certain amount of empirical evidence in their favour.

Kahneman and Frederick (2005) suggest that all heuristics used in these tasks are essentially based on ‘attribute substitution’. This leads to errors when the participant is calculating ‘how many instances of x am I aware of?’ as opposed to answering the question ‘what is the chance of x?’ As such, various components may be given too much or too little weight in any conjunctive, disjunctive or Bayesian calculation.

Dual process theories may be useful in terms of ‘pulling together’ all the theories discussed above, none of which currently seem able to fully explain the reasoning processes, and all occurrences (and non-occurrences) of the various reasoning fallacies. However, this still leaves the issue of which heuristics the experiential system may be using, and also raises the question of how and when each process may be primed. These heuristics also fail to explain how a significant proportion of participants manage to avoid committing any of the fundamental reasoning fallacies. Individual differences such as experience and education, and personality factors such as the propensity to reason heuristically could be expected to play a large part. These two issues are addressed in the following chapter.

It may also be that the divide between the two processes in the dual process theory model is not always helpful. Some of our apparently analytical processes are actually immediate and preconscious, like times tables (example from Fisk, 2004), a process in which Kahneman and Frederick (2002, 2005) explain the skills as having moved over from system 2 to system 1, as they cease being conscious and deliberate, and become automatic and instinctive.

Chapter 5 of this thesis investigates probabilistic reasoning by looking at the possible facilitating effect of the frequency format, in both conjunctive and disjunctive tasks. It is also anticipated that by using a design that enables the participants’ errors to be quantified on a ratio scale (indicating how far from the normative answer participants’ responses are) rather than nominal or ranked data, a more detailed picture of reasoning performance may be obtained. Much of the data collected in this area so far (and therefore the conclusions reached about reasoning techniques and abilities) has been based on assessing simply whether or not a fundamental fallacy has been committed, frequently asking participants simply to rank options in order or likelihood. It is hoped that by obtaining ratio data it will be possible to see not just whether more or fewer fallacies occur in a frequency condition, but the extent to which participants are over or underestimating the likelihoods. It will also be able to take into account whether participants are avoiding fallacies and yet still reaching normatively incorrect responses. For instance, while a person may avoid committing the conjunction fallacy (which occurs when the combination is deemed more likely than either component) they may still none-the-less greatly underestimate the true likelihood of A and B, thereby still failing to reason normatively.

Chapter 3 - Thinking Styles and Dual Process Theories

3.1 Dual Process Theories of Reasoning

In the previous chapter, the types of reasoning tasks that are most commonly used to collect data on probabilistic reasoning skills were introduced, as well as the main theories used to explain apparently impaired – i.e. non-normative – performance in the tasks.

One of the theories discussed, the dual process theory of reasoning, has been developed extensively in recent years and is often used to integrate pre-existing theories of reasoning, and to explain why heuristics may be utilised in some situations and by some individuals, but not in others. This chapter will use the theory as a framework to examine different thinking styles, and how they might influence individuals' reasoning processes and performance.

Osman (2004) identifies three such conceptualisations of reasoning, these being Evans and Over's Dual-Process Theory (1996), Sloman's Dual-System Theory (2002) and Stanovich and West's Two-Systems Theory (2000), and as discussed in Chapter 2, Epstein also makes many contributions to this area with his Cognitive Experiential Self-Theory (Epstein *et al.* 1992; Epstein, Pacini, Denes-Raj & Heier, 1996).

Much of the research in this area has been produced by Evans and colleagues (notably Over), who, in 1996 published 'Rationality and Reasoning', a large part of which was devoted to examining the Dual-Process theory of thinking. They propose:

'a dual process theory of thinking in which tacit and parallel processes of thought combine with explicit and sequential processes in determining our actions' (page 143).

Their two systems are labelled System 1 (tacit and parallel) and System 2 (explicit and sequential). Crucially, System 1 is described as functioning domain specifically, while System 2 is more flexible, giving us the ability to reason about novel situations (Evans & Over, 1996).

Sloman's (2002) approach is known as a Dual-System Theory, and he focuses more on the computational distinction between the two different reasoning systems, rather than

on whether they are using implicit/tacit or explicit reasoning. Put simply, System 1, which corresponds to Evans and Over's System 1, is described as an intuitive processor, associative in nature and responding to its environment, while System 2 is a conscious rule interpreter, and is more rigidly rule based.

Stanovich and West's (2000) Two-Systems Theory is more focused on what causes individual differences in the way that people approach and respond to reasoning tasks. It identifies System 1 as being automatic, unconscious and context dependent, while System 2 is not context dependent, and is a controlled and analytical process.

A final interpretation of this perspective is the Cognitive-Experiential Self Theory (CEST), by Epstein (one of the earliest of Epstein's interpretations of his Self-Theory is found in Epstein, 1973, but for more recent and relevant developments see also Epstein *et al.*, 1992 and Epstein *et al.*, 1996). The CEST actually encompasses three systems, only two of which are of relevance here (for a full explanation of the entire CES Theory of Personality, see Epstein, 2003). The Experiential system can be seen to be equivalent to System 1, in that it is automatic, and:

“a relatively crude system that provides a quick and dirty way of assessing and responding” (Epstein *et al.*, 1992, p.328).

The Rational system can be seen to be equivalent to System 2, and is described in terms of being conscious, intentional, analytic, and relatively affect free (Epstein *et al.*, 1996). In their 1996 paper, Epstein *et al.* develop and assess a tool for measuring these constructs, and assessing the likelihood that an individual would (amongst other things) use heuristics over analytical processing. The outcome was a modified version of the Need for Cognition scale (see Cacioppo & Petty, 1982; Cacioppo, Petty & Kao, 1984; Cacioppo, Petty, Reinstein & Jarvis, 1996) and their own Faith in Intuition scale. These were found to be measuring distinct constructs that were both reliable and independent of each other, offering strong support for the existence of the two systems.

Clearly the systems are similarly defined in each of the above 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'. As briefly mentioned in the previous

chapter, it has been suggested that it is system 1 that is in use when any of the reasoning heuristics are being utilised, while system 2 is held to utilise, and be limited by, working memory resources. 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)

Despite these overarching similarities, there are some fundamental differences between the theories. For instance, Sloman disputes the suggestion that system 2 must always be conscious, instead believing that both systems can function unconsciously, while for each of the other three theories, the consciousness of system 2 is an important distinction. Sloman also stresses that the systems operate together, and may both be utilised on the same task (2002), while Evans (2007b) feels that this is an area still to be fully researched. In particular Evans points out that we do not yet know where conflict between the systems may be resolved, and suggests that much of the differences in *quantity* of reasoning undertaken may be dependent on dispositional and/or motivational factors, while much of the research produced so far has instead focused on the *quality* (in terms of accuracy) of reasoning. Stanovich and West's body of research in this area (Stanovich & West, 2000; Stanovich, West & Sá, 1999) also places particular emphasis on the importance of individual differences in terms of thinking dispositions, and also suggests that System 1 is in fact made up of a number of different systems, which he names The Autonomous Set of Systems (TASS, Stanovich 2004). As suggested by this name, Stanovich (2004) stresses the fact that these systems do act autonomously, with no conscious effort.

The CEST's definition of system 2 as being 'Rational' (e.g. Epstein *et al.*, 1996) is another matter of contention, with Evans, Over and Manktelow (1993) having suggested that rationality should be seen as a dichotomy of rationality₁ and rationality₂. It is the latter, rationality₂, that they use to refer to as exclusively using rules of logic, and which is most commonly assessed by laboratory tasks, while the former, rationality₁, is focused on achieving daily goals, which may not be facilitated by a laborious, logical process. Reyna and Brainerd (2007) also express reservations about the claim that the Need For Cognition Scale is a measure of objective rationality, or rationality₂ as Evans *et al.* (1993) describe it. This proposed dichotomy of rationalities contributes to what has been termed the 'rationality debate' (Stanovich & West, 2000), in which researchers have sought to address the question of whether humans are inherently rational. The

debate focuses on whether the high levels of non normative reasoning seen in laboratory tasks can be used to infer that humans cannot be considered rational beings. Rationality₁ is responsible for non normative reasoning that nevertheless leads to responses that can be deemed to be ‘correct’ in that they are beneficial for that individual. As such, it is possible to claim that humans may be ‘rational’ despite not following logical rules, so long as their decisions are based on prior experience and are ultimately beneficial to them. It is also suggested that what can be deemed rational must have some fluidity, as evolution depends on our ability to adapt our behaviours, and therefore thought processes, over time (Manktelow & Over, 1990). What was rational – that is, of benefit to us – before may not continue to be rational in the same sense today.

Fisk (2004) suggests that one of the strengths of such a dual process theory is that it allows for – although does not, in itself, explain - the fact that individuals may at some points reason normatively, and at other times fail to do so. Work by Over has also illustrated that the framework can be used to highlight limitations of theories in this area. For instance, in his work with mental models as an explanation of the probabilistic reasoning process, Over (2005) suggests that since mental models are used deductively they are a clear part of system 2 processes, and as such the theory does not encompass system 1 approaches. The only way that mental models could be seen as a full explanation of the processes in probabilistic reasoning is if we claim that no system 1 – heuristic and rapid – reasoning is occurring at all, which seems unlikely, given the evidence presented so far.

It should be noted at this point that when the expression ‘dual process theory’ appears throughout this thesis, it should not be taken to imply that the perspective adopted is exclusively, or even primarily, that of Evans and Over’s Dual-Process Theory. Indeed, the current research focuses on the use of thinking disposition measures to account for individual differences in reasoning performance, an approach that very much stems from Stanovich and West’s research programme as discussed below. Instead the phrase is being used in general terms to describe any such theory, with their commonalities, rather than their differences, being emphasised in what follows.

Osman (2004) offers a rare note of caution in this area, arguing that none of the evidence that she finds in the literature as evidence *for* the dual process account is actually entirely incompatible with a unitary system. She feels that a continuum from

implicit, to explicit and then to automatic, as suggested by Cleermans and Jimenez's dynamic graded continuum (DGC; Cleermans & Jimenez, 2002), gives a better framework than the attempt to bisect learning and reasoning into two discrete types. She suggests that one of the advantages of the DGC framework is that it allows for errors in analytical thinking. However, although much of the work on dual process theories does suggest that 'correct' answers are more often obtained through analytical processes, it should also be noted that errors can occur in either process, and that errors may also be due to analytical processes being inaccurately applied, frequently due to cognitive limitations. Evans (2007b) names this the quality hypothesis, whereby those with greater cognitive ability may perform better not because they are attempting more analytical reasoning, but because they are more adept at doing so. Equally, heuristic processes can yield judgements that are 'correct' in terms of being sufficiently accurate when taken in context. In both day to day and more exceptional 'real life' judgements the information available is often imperfect or incomplete, and precise analytical reasoning may not be possible. It can also be of more value to an individual to achieve a solution that is rapid but approximate, than one that is exact but takes longer to reach. A paramedic, for instance, may want to make a quick estimate of the costs and benefits of an immediate procedure, and want only to know if the latter outweigh the former. More precise information would not be likely to actually affect the decision made, and the time taken to reason it out may in itself be costly.

More recently, research in this area has begun to investigate the possibility of amendments to dual process theory, as in the case of Stanovich (2009) who suggests that a tri-process theory should be considered. In this model, the autonomous mind (that is, system 1) remains an autonomous set of systems, while the analytic (system 2) is subdivided into the algorithmic and the reflective minds. The ability of the algorithmic mind to reason effectively in any given situation is dependent on the individual's fluid intelligence, while the reflective mind is closely linked to individual differences in the disposition to think rationally. As such, the algorithmic mind can be assessed through measures of intelligence, while the reflective mind is better examined through measures of critical thinking, and the ability to avoid basing judgements on biases. The evidence that Stanovich (2009) presents in support of this is that thinking dispositions do frequently account for large amounts of variance in reasoning performance, once differences in intelligence have been accounted for (e.g Kokis Macpherson, Toplak, West and Stanovich, 2002, Stanovich & West 1997:1998:2000, Toplak & Stanovich,

2002. See section 3.4 for a review of this literature). In terms of rational (that is, rationality₁) thought, it is the reflective mind that triggers the cognitive processes that begin hypothetical reasoning about a given task. It is the algorithmic mind which then conducts this hypothetical thinking, and which conducts the decoupling required to prevent real world representations from affecting how we perceive such hypothetical thoughts (belief bias, for example, might be expected to indicate a failure in decoupling).

A similar proposal to the reflective system has been made by Thompson (2009), who suggests that systems 1 and 2 are linked by metacognitive processes. She supports the suggestion that heuristic judgements – system 1 – tend to be the default response to any problem. For the analytical system 2 to be engaged, there needs to be a metacognitive judgement regarding whether or not that greater cognitive engagement is necessary. Specifically, she proposes a Feeling of Rightness (FOR) which may be based on ‘the retrieval experience’, primarily the fluency with which the initial answer was achieved. Thompson (2009) directly refers to the similarity between her proposal and those of Stanovich (2009, as discussed above). She also reiterates Stanovich’s proposals that the likelihood that system 2 is engaged on any task is strongly linked to the person’s thinking dispositions (see also Stanovich and West 2007: 2008), in that those who are predisposed to think analytically may be more likely to override a strong FOR and engage system 2. This would especially be the case in experimental situations where participants may anticipate a certain level of deception – in such cases, analytical thinkers may not trust a strong FOR, thinking ‘it can’t be this easy’. Thompson’s theory (2009) does have one main difference from Stanovich’s (2009), which is that Stanovich appears to be suggesting that the reflective mind is a conscious/explicit system. Thompson’s own suggestion of a metacognitive monitoring system is conversely defined as being implicit, and therefore involving fewer intentional processes.

Similar to Stanovich’s (2009) assertion that system 1 is better understood as a collection of many systems, Evans (2008;2009;2010a;2010b) has come to prefer the terms ‘type 1 processes’ and ‘type 2 processes’. This terminology is used to make it clear that his interpretation of dual process theory is that each ‘system’ is in fact made up of many systems or processes, which can be clustered together into being fast, automatic, high processing capacity and low effort (type 1) or slow, controlled, limited capacity. high effort (type 2: Evans, 2009). Evans (2010a) describes type 1 systems as being of our

‘old mind’, in evolutionary terms, while type 2 systems are of the ‘new mind’. Furthermore, that the old mind is shared with many other animals is provided as evidence for the multiple type 1 systems, with many forms of implicit processing being necessary for processes such as vision and attention (Evans 2008).

Saunders and Over (2009) also prefer the ‘type of systems’ concept and terminology, explicitly stating their agreement with Stanovich’s proposal that system 1 is better identified as being The Autonomous Set of Systems (TASS). Saunders and Over present the example of an eating disordered individual who will *automatically* choose low calorie foods, despite this being in conflict with the evolutionarily powerful urge to choose high calorie foods. Each of these processes could be declared system 1, yet they are in conflict with each other, a state of affairs which can be best explained by the suggestion that there are a set of type 1 systems, rather than just one system.

These developments in the dual process theories of reasoning, moving away from a clear cut dichotomy, can imply that it might be reasonable to reassess the possibility of processing lying on a continuum, from truly intuitive to truly analytical, as suggested by Osman (2004, based on the Dynamic Graded Continuum of Cleeremans & Jimenez, 2002 as above). However, compelling evidence to continue to use the framework of intuition and analytical thought as being to some extent discreet processes is found in the work of Lieberman (2007). As a social cognitive neuroscientist, Lieberman presents neurological evidence to illustrate that use of the reflexive ‘X-system’, analogous to system 1, is revealed by activity in the ‘older’ parts of the brain (e.g. the amygdala, the basal ganglia), while the reflective ‘C-system’, or system 2, shows activity in quite different areas, but predominantly in the lateral prefrontal cortex. This clear division also supports Evans’(2010b) taxonomy of the old and new minds.

3.2 Evidence for Dual Process Theories of Reasoning

Epstein *et al.* found support for their dual process theory, CEST, in their work with the ‘if only’ effect in 1992. The if only effect is essentially what occurs when you (or a character in a vignette) aim for a goal which you then miss by a small amount, and are more upset by this than if you had missed out by a large amount. It describes the feeling of ‘if only’ you had left a few minutes earlier, or revised for just a few minutes longer. Epstein *et al.* (1992) asked participants who they thought felt the most foolish, the character who had been involved in an accident, lost out on a financial deal, missed

their flight or had their car damaged but had *very nearly avoided* their own particular misfortune, or the character who had experienced exactly the same misfortune, but with no suggestion that a small alteration to their behaviour could have allowed them to avoid it. The ‘if only’ effect is seen when those protagonists who had very nearly avoided misfortune are described as feeling more foolish than those who had no such near miss. They proposed four hypotheses:

1. That the ‘if only’ effect would be seen, and especially so when participants pictured themselves as the protagonists. Instructions in this case asked the participant to imagine how they ‘would probably react to the different versions of the situations if they happened to you’. Within CEST, system 1 is considered to be experiential, and closely associated with the emotions. As such, emotions elicited by the self-perspective condition (for instance, regret, embarrassment at making an apparently foolish choice) would engage system 1.
2. That the effect will be reduced when participants are specifically asked to be rational. They felt that this request would prime the rational system (equivalent of system 2) whereas the ‘if only’ effect is caused by the use of the experiential system (system 1). The rational system was primed with the request to ‘this time give a strictly logical response rather than one based on how you think people actually react . . . put your emotions aside’.
3. That the ‘if only’ effect will be stronger when the negative outcome (e.g. car accident) is more extreme (experiential system 1 being primed by stronger emotions), and that this will be eliminated when participants are asked to think more rationally.
4. That participants who are first reasoning experientially and then rationally will still be less rational in the rational condition than those who were never asked to reason experientially at all. This is a result of the experiential system having an influence over the ‘rational’ system which is sustained across the tasks.

Over the course of two studies, they found support for each of these hypotheses, and therefore for their proposed model. They report their findings as providing evidence for the idea that each of the two systems can be primed, with the experiential system providing a stronger ‘if only’ effect, due to its irrational nature and dependence on emotional states, and the rational system weakening the effect as participants see that ‘rationally’ a goal is either missed, or achieved. The rational system helps participants to see that the *degree* by which it is missed or achieved should not logically impact on a

person's feelings about their failure. In supporting their fourth hypothesis, that the experiential system has a lasting influence over the 'rational' system, Epstein *et al.*, (1992) suggests a relationship between the two systems which may be in contrast to the suggestion that system 2 monitors system 1, only intervening and over ruling the heuristic system 1 when certain conditions are met (see, for instance, Franssens & De Neys, 2009). The two are not necessarily mutually exclusive, however, and it is possible that system 2 both monitors system 1 while also being influenced by its activation.

In 2002, Kokis, *et al.* gave their participants an inductive reasoning task, a deductive task and probabilistic reasoning task, as well as their Thinking Dispositions Questionnaire (TDQ). The TDQ contains 53 items, and includes 7 subscales: Flexible Thinking, Belief Identification, Absolutism, Dogmatism, Categorical Thinking, Superstitious Thinking (ST) and Need for Cognition (NFC), as well as five items assessing social desirability response bias. The first five of these went into a composite scale, the Actively Open-minded Thinking scale (AOT), while the NFC and ST scales remained separate. The need for cognition subscale is a short form of that devised by Cacioppo *et al.* in 1996, and the majority of the others had previously been used in a similar form by Stanovich and West (1997). With many of the subscales containing items drawn from a wide range of sources, it is not appropriate to detail them all here. The scale is discussed in more detail in chapter 6, but to summarise Kokis *et al.*'s use of the scale (2002), it was intended to measure styles of epistemic regulation (the AOT) and of cognitive regulation (NFC). They found that cognitive ability (as assessed by the WISC-III block design and vocabulary subtests) was associated with analytic (as opposed to heuristic) reasoning on each of the reasoning tasks. That is to say, those achieving higher vocabulary and WISC-III scores were also more likely to avoid reasoning fallacies and errors. However, regression analyses also indicated that in each case, when entered after cognitive ability, one or more of the thinking dispositions accounted for significant amounts of the remaining variance.

Explaining these findings in terms of dual process theory, this gives further support for the idea of a separate, conscious system 2, which they call the analytic system and which is primed to work on these types of tasks in some individuals, while in others the pre-conscious system 1 – less reliant on cognitive ability, but more so on heuristics – is primed. They also suggest that the relationship between the two systems is such that system 2 deliberately overrides system 1 in such cases.

The AOT and ST have also been used by Macpherson and Stanovich in 2007, where they found that, along with the full version of the NFC scale (Cacioppo *et al.*, 1996) they were significant predictors of belief bias in a syllogistic reasoning task. Both the NFC and AOT were positively correlated with avoidance of the bias, while ST was negatively correlated. Kokis *et al.* (2002) had previously established that the AOT composite shows a strong positive relationship with analytical thinking, and Stanovich and West (2000) suggest that it measures the cognitive flexibility necessary for system 2 processes. The actively open minded person is someone who is not overly attached to or dependent on their current beliefs, and will challenge and change their beliefs in order to arrive at more accurate answers (Stanovich & West 1997). However, they did not predict ‘myside bias’, a bias which occurs when participants evaluate evidence and construct arguments for and against hypotheses on the basis of their own personal beliefs. In the case of belief bias, therefore, it seems that the analytical system 2 approach was effective in avoiding the bias, whereas in the case of myside bias the predisposition to think analytically was in some way overridden by system 1. Macpherson and Stanovich (2007) do suggest that participants may not have recognised the need for unbiased reasoning on the myside bias tasks, which were, for instance, asking participants to provide arguments around a given issue, although one condition did urge to participants to give reasons both for and against.

West *et al.* (2008) also used the AOT and NFC, combining them to create a composite scale. Cognitive ability was assessed by a self-reporting measure, whereby participants were asked their SAT scores – a measure that they indicate does correlate highly with actual SAT scores. They also gave participants a wide range of tasks designed to assess incidences of heuristics and biases, which were again combined to create a composite score, and syllogistic reasoning tasks, again looking at belief bias. While the thinking disposition composites were strongly positively correlated with many of the individual heuristics and biases tasks, including Bayesian tasks, and a form of exclusive disjunctive task, they did not correlate with the conjunction task. Cognitive ability showed a similar pattern. Regression analysis with syllogistic reasoning performance as the dependent variable revealed that when entered first, cognitive ability predicted 19.4% of the variance, but the thinking dispositions composite did predict a significant 1.9%. At 3.2% greater variance was accounted for by the thinking disposition composite in relation to the heuristics and biases aggregate outcome measure (after the 15.2%

accounted for by cognitive ability). Further experiments by Stanovich and West (2008) found that cognitive ability was not correlated with incidences of base-rate neglect (the fallacy often found in tasks requiring Bayesian reasoning) and this finding, coupled with those of West, *et al.* (2008), gives strong support to the supposition that the influence of thinking dispositions on probabilistic reasoning performance can be greater than that of cognitive ability.

With significant amounts of variance in reasoning performance being accounted for by individual differences as measured by the AOT and NFC, these findings do suggest that greater analytical thinking, as measured by these scales, does lead to lower reliance on heuristics, and more use of analytical system 2.

Ferreira, Garcia-Marques, Sherman and Sherman (2006) present a powerful argument that research in this area is very much ongoing, and that many of the previous findings are based on *assumptions* of the existence of the two systems, as opposed to being focused on finding empirical support for their actual existence. They refer to the two systems as being analytic or ‘rule based’ (corresponding to system 2) and ‘heuristic’ (system 1). Over the course of four experiments they concluded that the rule based system was primed by the use of rational instructions, rather than intuitive ones, and that this system was also affected by tasks that increased cognitive load, while the heuristic system was not affected by this increased load. They were also able to prime the heuristic system, by specifically instructing participants to ‘base their answers to the problems on their intuition and person sensitivity’, similar to the method used by Epstein *et al.* in 1992, in their work with the ‘if only’ effect, as discussed above.

One flaw with this over reliance on the description of the second system as being ‘rule based’ is that it may not discriminate clearly enough between that and the heuristic system, which is itself using rules, albeit less normatively logical ones. As discussed in Chapter 1, many heuristics rely on various ways of using the components in a probability task to arrive at a solution. These do rely on their own ‘rules’, which may be vague, apparently illogical and changeable, but do exist nonetheless. Saunders and Over (2009) also pick up on this point, suggesting that the emphasis should instead be placed on the difference between the implicit conformity to rules that occurs within system 1, and the explicit following of rules that occurs within system 2.

Verschueren, Schaeken and d'Ydewalle (2005) offered further support for the two processes, in their work with conditional reasoning, and specifically for the idea that the heuristic system 1 is the faster system, with analytic system 2 being more time consuming. Evans and Curtis-Holmes (2005) also examined the idea that the analytic system 2 requires more time to function. Using syllogistic reasoning tasks, they found that restricting the time that the participants had to complete the tasks did result in a greater reliance on the heuristic system, concluding therefore that the heuristic system is indeed the faster of the two. Within probabilistic reasoning, Oechssler, Roider and Schmitz (2009) found that participants found to be cognitively reflective – liable to spend some time on tasks – did commit fewer conjunction fallacies than those who were impulsive. Again, the indication is that the analytic system 2 is the more time consuming, and the faster system 1 is used when decisions are made more impulsively. Crisp and Feeney (2009) also looked at the conjunction fallacy within the context of dual process theory, and found that a memory load task condition led to greater incidences of the fallacy. They suggest that the memory load task led to greater reliance on the heuristic system 1, due to reducing the availability of cognitive resources.

In 2002, Kahneman and Frederick looked specifically at one of the possible heuristics used in reasoning, the representativeness heuristic, as being one of the tools used by the intuitive system 1 within a dual process system of reasoning. Using the Linda task previously discussed in Chapter 2.2, they found clear examples of the heuristic being utilised. They proposed that while their materials provided participants with all of the tools to arrive at a rational answer, the analytical system 2 processes were simply not prompted to do so. For whatever reason (and see 3.1 for details of a possible reflective process), the intuitive, representativeness based, answer was deemed adequate and was not challenged by system 2.

De Neys (2006b) also provides support for the existence of two separate processes, illustrating how a high load on executive resources affected syllogistic reasoning only when there was a belief/logic conflict. When the belief bias heuristic process and the analytic process both gave the same answer, no effect of load was found. However, when there was a conflict, with the heuristics (system 1) giving one answer, and the analytic process (system 2) giving another, there was an effect of load, whereby performance on the task decreased significantly. This supports the view that system 2 is putting a significantly greater load on executive function. De Neys also suggested that

all individuals can reason in both ways, and that the 2 processes are found within, and not between individuals (2006b).

3.3 Conflict Between the Systems

If we accept the existence of these two systems, either of which may sometimes be chosen over the other, this then leads to the examination of the processes by which either system is chosen. Evans (2007b) proposes three models to explain how this may occur – the pre-emptive conflict resolution model, the parallel-competitive, and the default interventionist.

The pre-emptive conflict resolution model would suggest that there is a very early decision to go with either one system or the other, and that this decision is based on some superficial aspect of the task involved. The parallel-competitive model instead involves the two processes working in parallel, and it is only when the two systems are not in agreement about a response that conflict occurs. Sloman (2002, and cited by Evans in 2007b) suggests that the heuristic system may often precede and ‘neutralize’ the rule based system 2, its speed giving it the advantage of obtaining the first response, with the slower system 2 then not being activated to produce its own solution. Lastly, the default interventionist model, for which Evans claims support from Kahneman and Frederick (2002) and Stanovich (1999), and now integrates into his own dual process theory (for example, Evans, 2006) describes system 1, the heuristic processes, as being a default system which on some occasions will be overruled by the analytic system 2. Moutier and Houdé (2003) also suggest specifically that some individuals are failing to ‘inhibit’ reasoning biases, implying again that system 1 is the primary system that will be used if no intervention from system 2 is somehow primed. This suggestion is also made by Franco (2009) who proposes that the fallacy frequently stems from an inability to inhibit first impressions – the response given by the rapid system 1 is incorrectly accepted. 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.

It appears that in each of the above possible models suggested by Evans, there is either an internal or an external prompt that primes either one of the systems to take over the reasoning task. Evans himself does not specify this (other than to suggest in the first

case there is an external prompt in the form of a superficial aspect of the task in question) but in each case there must ultimately be some factor influencing whether the heuristic or analytic response is the one chosen. This may be influenced externally (the task, and/or the circumstances in which it is completed) or internally (the individual's ability, for instance, or motivation).

In 2009, Evans further addresses this issue by suggesting the addition of type 3 processes, preconscious processes of which individuals may not be aware (see De Neys & Glumicic, 2008, below). In this model, type 1 and type 2 processes work in parallel, but the issue of the time consuming nature of type 2 processes can be addressed. Instead of every judgement having to wait until type 2 processes are complete, in order to then make a decision, type 3 processes can react to the rapid response given by type 1 processes, and quickly make the judgment as to whether this response is adequate. If so, the resulting behaviour occurs rapidly. If however, the type 3 processes detect a problem with the type 1 response, then the type 2 processes continue, as the additional time and cognitive demands of such processes have been deemed necessary. More time taken, however, may not always lead to the correct answer (De Neys & Glumicic, 2008) and it is reasonable to suggest two possibilities for this. First, that the type 3 processes choose incorrectly, and despite waiting for system 2 processes they decide that their response is incorrect and revert to type 1 process responses. Second, that the type 3 process has chosen the type 2 answer, but that the analytic processing has been done inaccurately, and an inaccurate answer is still the result. Thompson (2009) describes the Feeling of Rightness (FOR, as detailed above) and Evans (2010) also refers to the relevance of system 1 decisions *feeling* right in order to avoid any more in-depth and cognitively costly processing.

Evans (2008) also stresses that type 2 processes may sometimes be involved in confabulation – justifying the decisions made by system 1. This is also suggested by Thompson (2009), who uses the term 'rationalise' to describe the way in which participants may attempt to explain their own heuristic reasoning processes as containing more logical and/or analytical content than they actually do. Evans (2010a, 2010b) is also clear that probabilistic reasoning is a particularly noteworthy example of the way in which our intuitive reasoning can let us down, leading to both inaccurate responses and to post-hoc rationalisations by system 2 processes.

De Neys (2006a) adds some insight with a series of experiments which support the above assertions that the analytic system 2 is slower than system 1, and also requires greater executive resources. Using the Bill and Linda conjunction tasks, De Neys found that participants who were committing the fallacy were significantly quicker to arrive at their answers than those who answered correctly, and that a dual-task paradigm placing a greater cognitive load reduced cognitive performance, increasing the number of errors committed. This second finding suggests that in some instances the heuristic system will be used simply because the cognitive processes necessary for the analytic system are not readily available.

Further research by De Neys and Glumicic, in 2008, used verbal protocol analysis to examine whether or not participants were aware of the belief/logic conflict in base rate tasks. Explicit inclusion of the base rates in the participants' protocols was taken as evidence of their awareness of the conflict between the logic implied by the base rate, and their own belief about the further information presented. While these protocols did not reveal a great awareness of the base rates, a further task suggested that the base rate information was in fact being processed. When asked to recall the relevant base rates, participants could do so significantly more often for those tasks where a belief/logic conflict existed, suggesting that participants were aware of the conflict, and of the resulting importance of the base rate. They also found that tasks involving conflict took longer to complete, even if an incorrect answer was given. Again, this implies an awareness of the conflict, and support for the proposition that an attempt by the logical system 2 to intervene has occurred, but that a heuristic process has ultimately been used. Franssens and De Neys (2009) also found that participants completing Bayesian tasks were able to recall base rates effectively, and particularly so where a belief/logic conflict was present, despite showing no evidence of taking them into account in the reasoning process. Crucially, in examining the conflict between system 1 and system 2, Franssens and De Neys (2009) found that a condition of cognitive load, a visuo-spatial task, made no difference to participants' ability to identify belief/logic conflicts, as demonstrated by their better recall of base rates in these tasks. As such, the monitoring process was still triggered to prompt system 2 to solve the task, as evidenced by the fact that the extra cognitive load did then negatively affect reasoning performance on the conflict tasks. Non-conflict task performance was not affected by the cognitive load, reflecting the use of system 1, which is less dependent on such processes. That the cognitive load task did not affect the triggering of system 2 suggests that the conflict

detection process also is not cognitively demanding. De Neys and Glumicic (2008) also found that for tasks containing the conflict, the time taken by participants was longer, regardless of the accuracy of their final response. They suggest that this reflects that system 2 is indeed constantly monitoring system 1, and intervening in some way even if full analytic processes are not primed. They refer to this model of continuous but light monitoring by the analytic system as a shallow analytic monitoring process – something which Evans (2009) attributes to the preconscious type 3 processes.

Kahneman and Frederick (2002) suggest the system used can be determined firstly by the task – for instance if it is time-limited then the faster system 1 is more likely to take over. Secondly, individual differences are likely to come into play, as some of those attempting the tasks may not have the skills needed to examine the problem analytically, and are therefore more reliant on system 1. Kahneman and Frederick also stress that when referring to systems 1 and 2, they are not meaning to describe two entirely discrete systems, but are instead talking about collections of processes that make up the two systems. They feel that often tasks that initially require complex cognitive operations do, with practice, move over from the rational system 2 to the intuitive system 1 as they become more automatic to the individual, so that system 2 may be used in the initial learning of a new skill, but once that skill has been well rehearsed it is likely to transfer to system 1 as it requires less time and cognitive effort. Evans 2010a is also clear that our ability to intuit effectively comes from our previous experiences making skills second nature.

This section of the literature review has been given the title ‘Conflict between the two systems’, and much of the discussion of how system 1 or system 2 processes and responses are chosen also use this term, for instance Franssens (2009), Evans, Barston and Pollard (1983), Evans (2007b), De Neys and Glumicic (2008). This terminology naturally implies a lack of agreement, but Saunders and Over (2009) suggest that the result of such ‘conflict’ does not always lead to one system’s response being accepted, and another rejected. Instead, outside of the psychology laboratory co-operation and integration can occur (something also suggested by Evans, 2010), as either the intuitive or the analytic may hold sway at any given time, maintaining a balance between them.

This current research aims to address the theories of Kahneman and Frederick (2002) and Stanovich and West (2000) by looking at both external factors and internal factors

that may prompt either system 1 or 2. The external factor is problem format: probability or frequency (see Chapter 2.5.5) while the internal factor will be thinking styles, or dispositions.

3.4 Thinking styles

The question of how and why system 1 is dominant at some times, and system 2 at others, will be examined in terms of individual differences, as discussed by Stanovich and West (2000) specifically in terms of individuals' thinking styles.

As discussed previously, both in this chapter and in the previous one, Epstein developed his own dual process theory, the Cognitive Experiential Self Theory (CEST) throughout the 1990s (for example, see Epstein *et al.*, 1992). In their 1996 paper Epstein *et al.* used their Rational-Experiential Inventory (REI), containing a version of the Need for Cognition Scale (Cacioppo & Petty, 1982) and their own Faith in Intuition scale (FI) to form a measure of these thinking styles – analytical-rational and intuitive-experiential respectively. In contrast to Kokis *et al.*'s definition of the NFC as 'a measure of epistemic regulation' (2002), Epstein *et al.* refer to it more specifically as being a measure of (tendency towards) analytical-rational thinking. They also define the FI scale as asking participants to respond to items concerning 'having confidence in one's feelings and immediate impressions as a basis for decisions and actions' (p. 394). This could be expected to directly measure an individual's propensity to use heuristics – given the limitation of self-reported measures, whereby participants may make such judgements more or less than they are consciously aware.

Again working with the same style of 'if only' task as discussed in relation to Epstein *et al.* (1992) above, participants in a study by Epstein *et al.* (1996) were asked in some cases to respond as they felt 'the average person' would do, in some to respond how they themselves would do, and lastly as a 'completely logical person' would do. Where participants made responses that suggested that a person would feel more foolish for missing out on something by a small margin or change of behaviour, Epstein *et al.* (1992) classified these responses as non rational, and based on heuristics rather than rational/analytical reasoning. Responding that a person would feel equally foolish in either situation (whether a goal was narrowly missed or by a large margin) was deemed rational, and thought to be based on analytical reasoning. Crucially, feeling more foolish

for having missed something by a small margin is 'irrational' because the protagonists in the vignettes have each experienced the same end result, and could have had no prior knowledge that their behaviour would lead to their failure to achieve their goal. As anticipated, they did find that participants gave non-rational answers when asked to think how they (the 'self-perspective' condition) or another 'average person' would react, but were less likely to be reasoning heuristically when asked to think how a logical person would react. That is, the materials did seem able to prime the participants to think either logically or heuristically. They were capable of thinking in both ways on very similar problems.

When Epstein *et al.* (1996) correlated the number of heuristic responses with the participants' scores on the NFC and FI scales, they did find some significant relationships. NFC correlated negatively, and significantly, with the levels of heuristic reasoning on the self-perspective version of the task in men ($r=-.5$, $n=55$) but less so (and non-significantly) in women ($r=-.23$ $n=129$). Conversely, FI positively and significantly correlated with heuristic reasoning in the same, self-perspective condition with women only ($r=.45$), while for men the correlation was as strong (at $r=.44$) but due to the smaller size did not reach significance. The men's data, however, did show a further positive and significant correlation between FI and the amount of heuristic thinking in the logical perspective ($r=.41$).

To clarify, within the self-perspective condition, the higher the males' NFC, the lower the chance of them reasoning heuristically, and the higher the women's FI, the higher the chance they would be reasoning heuristically. In each case, the direction of the relationship is as might be expected from the dual process theory, but it seems that for males it is the NFC that is important, while for females it is the FI. In the logical perspective condition, only FI for men was significant, with higher FI associated with higher levels of heuristic reasoning. Also, it is worth noting that in each of these cases the correlations were significant when male and female data was analysed as one set - i.e., as expected, high NFC predicts analytic thinking, high FI predicts heuristic thinking.

Dagnall, Parker and Munley (2007) also found an overall negative correlation with a measure of paranormal belief (a construct similar to Faith in Intuition) and probabilistic reasoning, indicating that those with high levels of paranormal belief were less likely to

respond normatively to problems involving probability. Four types of task were used, looking at perceptions of randomness, the use of base rates, conjunctive reasoning, and expected value problems. Regressions indicated that it was the ‘randomness’ tasks (for instance asking participants to judge the likelihood of various sequences of heads and tails when a coin is tossed a number of times) which accounted for the variance in paranormal belief, with none of the other types of tasks making a significant contribution. Of specific relevance to the current study is that neither conjunctive nor base rate task performance appeared to be related to paranormal belief. Rogers, Davis and Fisk (2009), however, found that those with a high belief in the paranormal did make significantly ($p < .001$) more conjunction errors than those who had a low belief, indicating that such a belief system is related to lower levels of (accurate) analytical thinking.

One issue that affects much laboratory based research is the issue of motivation – put simply, will participants who are more motivated to perform, complete the tasks more accurately? In the field of probabilistic reasoning, this has been examined by Stolarz-Fantino *et al.* (2003), who were testing the robustness of the conjunction effect – that is, the occurrence of the conjunction fallacy, as previously described. They found that increased motivation for a ‘good’ performance in the form of a possible financial incentive (those with correct answers would be entered into a draw for \$35) had no effect on performance, when compared with those who were given the same incentive for simply completing the task (being entered into the draw regardless of the accuracy of their answers). It may well be that the prize itself was insufficient as an incentive, and that if the participants had instead been offered a guaranteed \$35 for a correct answer, this may have provided a stronger incentive, thereby enhancing motivation and boosting performance. Furthermore the participants may well have been aware that this was a very large scale study, with 251 participants taking part, and so may have felt the chance of winning the money even with a correct answer was negligible.

Stanovich has conducted a number of studies examining the relationship between thinking dispositions and tasks which involve the use of heuristics and biases, for instance Stanovich and West in 1998 and Stanovich *et al.* in 1999. One of the studies by Stanovich and West (1998) used 546 participants who completed a range of reasoning tasks. These included syllogisms and ‘if only’ tasks similar to those used by Epstein *et al.* (1992, 1996). Stanovich and West (1998) found that performance on every one of

the nine heuristic reasoning tasks they used correlated significantly with both cognitive ability (as measured by SAT scores) and thinking dispositions. As was predicted, both cognitive ability and higher levels of the analytical thinking disposition were associated with avoidance of errors and biases on the tasks. Crucially, the thinking dispositions were also accounting for significant levels of unique variance, over and above that which was explained by cognitive ability. In 1999, Stanovich *et al.* refer to thinking dispositions as reflecting intentional-level processes. Put simply, the discrepancy between ability and performance can, in part, be explained by the participants' intentions.

Toplak and Stanovich also looked at a form of reasoning in order to investigate this area (2002). They gave their participants nine 'disjunctive' tasks, four nominated as 'decision-making' (for instance, prisoner's dilemma) and five as 'problem solving' (Wason selection task). These are clearly not the same simple disjunctive tasks as previously discussed (see section 1.3, as well as Carlson & Yates, 1989 and Fisk, 2005) but in each case are presenting participants with information and asking them to choose between two options – a form of exclusive disjunction, as they choose either/or, and not both. While the response in each case does require the estimation of an exact disjunctive probability, the authors suggest that in order to decide which of the two options is most likely to lead to a favourable outcome individuals must consider the relevant disjunctions. Toplak and Stanovich (2002) also measured cognitive ability, need for cognition (using the scale devised by Cacioppo *et al.* 1996), and gave their participants a reflexivity/impulsivity test. The measure from this task, the matching familiar figures test (MFFT, from Kagan, Rosman, Day, Albert & Phillips, 1964), was the number of errors made.¹

The MFFT involves picking out a target picture from a selection of six choices, with reflexivity being associated with longer times but fewer errors. Regression analysis revealed that participants' reflexivity (as indicated by their mean error score on the MFFT) and their NFC together accounted for 11.8% of the variance in reasoning performance, at 4.7% and 7.1% respectively, with those showing high NFC and/or high reflexivity being in general more likely to complete the tasks correctly. In comparison,

¹ Although they measured reaction times on this task, it was found that this variable did not account for any variance in any other measure over that accounted for by error – as such it was omitted from the analyses reported below.

Toplak and Stanovich (2002) found that cognitive ability only accounted for a further 5.2%, reflecting the importance of the two thinking styles.

Whether the tasks used are directly comparable with more traditional probabilistic reasoning tasks is debateable, as the Wason task in particular is far more often considered to be a deductive, rather than inductive, problem. With research by Osherson, Perani, Cappa, Schnur, Grassi and Fazio (1998) and Parsons and Osherson (2001) suggesting that quite different parts of the brain are activated by the processes involved in solving deductive versus inductive tasks, the different probabilistic reasoning tasks used in the present study raise the possibility that Toplak and Stanovich's results (2002) may not be replicated in the current research.

Another factor affecting whether or not system 1 or system 2 is chosen may well be age. A number of researchers (for instance Blanchard-Fields, 1996; Fisk, 2005; Klaczynski & Robinson, 2000; Yates & Patalano, 1999) have suggested that reasoning performance may not necessarily deteriorate with age, but that reasoning styles do change. For instance, Fisk (2005) found that in a young group of participants (<24 years), reasoning performance was closely related to working memory capacity, but in an older group (> 55) it was not, suggesting that some other processes are being relied upon. One possibility is therefore that the heuristic system 1 is being used more by the older group, and Evans (2008) does go so far as to suggest that to label system 1 as unconscious and 2 as conscious is a false distinction – instead, the main discriminatory factor should perhaps be that conscious use of working memory is involved in the latter, and not the former. If it is the case that analytical processing is heavily dependent on the conscious manipulation of information in working memory, this does have implications for older individuals, due to the well established link between ageing and deteriorating working memory (e.g. Salthouse, 1998; Salthouse & Babcock, 1991). Researchers such as Klaczynski and Robinson (2000) have also found that thinking dispositions as discussed above, and specifically the Need for Cognition, are affected by age. This is discussed in greater detail in the following chapter.

3.5 Links between thinking styles, dual process theories and problem format

Sprenger and Dougherty (2006) used the dual process framework to investigate differences in reasoning on probabilistic reasoning tasks that were framed in the two different formats: probability and frequency. They found that while frequency

judgements showed lower incidences of subadditivity, they did not show any difference from the probability versions in terms of relative accuracy – relative accuracy being measured by rank ordering the judgements. They cite two main further findings. First, they found that only probability judgements, and not frequency ones, were significantly related to working memory capacity. Second, that when asked for judgments of probability, participants' reaction times were significantly slower than when they were asked for judgements of frequency. These two findings do strongly suggest that each type of task is using different processing, with the probability format appearing to prime the slower, more conscious system 2, and the frequency format priming system 1, the rapid, more intuitive process less reliant on working memory capacity.

As discussed above, Kokis *et al.* (2002) have found, using their TDQ scale, that thinking style measures – including the Need For Cognition – could account for substantial amounts of variance in reasoning performance even after the greater variance attributed to cognitive ability had been controlled for. With Need for Cognition being an analytical disposition, it seems reasonable to expect a link between levels of this disposition and performance on the two problem formats. If Sprenger and Dougherty (2006) are correct that the frequency format primes system 1, for instance, then it may be that having a strong need for cognition can override this priming effect, and is the internal prompt needed to encourage the participant to reason analytically regardless of the wording of the task. This would lead to greater reasoning accuracy, both in absolute (occurrences of subadditivity/fallacy) and relative terms (actual distance from the normative answer), but could be expected, due to the nature of system 2 as conscious and deliberate, to be more time consuming.

However, there is another possible relationship between thinking styles and task format. Donovan and Epstein (1997) refer to different problems as being either concrete or abstract in their presentation. In their own example, a concrete task would refer to a specific 'house' that has been identified, while an abstract task would refer to 'house' as being a category. They suggest a further dimension of 'natural' and 'unnatural'², whereby a natural task is one that prompts the correct system to reach the correct solution. So if a task requires a rational reasoning process, the task would be 'unnatural'

² Tversky and Kahneman (1983) provide a different explanation of 'natural' in this context, instead suggesting that 'natural assessments' are those that we make in everyday life, which may well involve the use of heuristics such as availability and representativeness

if it actually elicited the experiential/heuristic system. In terms of probabilistic and frequency formats this leads to the supposition that frequency formats, in the case of probabilistic reasoning tasks, are what Donovan and Epstein would call natural tasks – they elicit the correct processes, the logical system 2 – to correctly solve them and reach the normative answer. Coupled with the lower demands on working memory suggested by Sprenger and Dougherty (2006, see above), the analytic system would also be able to work more efficiently, once primed, as it would no longer be so restricted by limitations in working memory capacity. The suggestion that the time consuming nature of the probability format is evidence that it primes system 2 is also not irrefutable, in as much as there is no set time limit after which it can be declared that either system is or is not responsible for processing. It may be that both formats are utilising system 2, but that the greater difficulty of the probability format (as implied by the greater numbers of incorrect answers) makes it more time consuming to complete. Just because the frequency format tasks are completed in less time, this does not necessarily tell us that they do not use system 2 at all. Indeed, it may be that the faster time is specifically due to the correct system 2 being primed initially, with no, or minimum levels of, conflict between the two systems occurring. If this were the case, and the frequency format does in fact prime the analytical system 2, then it would be reasonable to suggest that those with higher levels of NFC may benefit less from the format, being as they are already predisposed to think analytically. It may be those with lower levels of the thinking style that are able to most benefit from having their analytic thinking primed.

As discussed briefly in Chapter 2, Cosmides and Tooby (1996) believe that humans can and do often reason rationally. However, Evans (2007a) suggests that Cosmides and Tooby's claims about the frequency format representing innate reasoning fails to recognize the dual processing nature of reasoning. He feels that such tasks lead to lower incidences of reasoning fallacies due to their facilitating the analytic system 2, by making the nested-set structures in the tasks more apparent. Kahneman and Frederick (2002) also suggest the same – that tasks worded in terms of frequencies, by making the nested sets clear, enable a more analytical and less heuristic way of approaching the tasks. Sloman (2003) has also described the way in which frequency formats clarify the nested-set structures that exist in such tasks.

3.6 Conclusions

The four main dual process theories of reasoning discussed above (Evans & Over's Dual-Process Theory (1996), Sloman's Dual-System Theory (Sloman, 2002) Stanovich and West's Two-Systems Theory (Stanovich & West, 2000), and Epstein's Cognitive Experiential Self-Theory (Epstein *et al.*, 1992; Epstein *et al.*, 1996), all describe two, more or less discrete, core reasoning processes. To summarise, the four theories share certain features in common. System 1 is described as rapid, automatic, and, crucially, consisting of heuristic methods of reasoning. System 2 is slower, makes greater demands on cognitive processes, and is analytical in nature. As such, it is system 1 that is thought to be predominating when participants make reasoning errors that are caused by biases and heuristics, and system 2 that is felt to have been used when such errors are avoided. As discussed above, more recent research also suggests the existence of a third system, or set of processes, which serves a reflective, metacognitive role which may consciously (as implied by Stanovich, 2009) or unconsciously (Thompson's Feeling of Knowing, 2009) monitor the systems and resolve conflict between them (see also Evans (2009). Stanovich (2009), Saunders and Over (2009) and Evans (2008; 2009; 2010a; 2010b) also stress the importance of seeing each of the systems as being made up of a number of different processes, which can be generally classified as being either heuristic (traditionally labelled as being a part of system 1) or analytic (system 2). As stressed in 3.1, the 'dual process theory' referred to throughout this thesis is not meant to specifically represent any one of those discussed in this chapter, but is instead used as a general term to refer to any such theory, focusing primarily on the similarities between them.

Each of the theorists above provide their own research based evidence for the existence of the two (or more) systems, or sets of processes, providing evidence that it is not just cognitive ability that accounts for reasoning performance but that participants may have a disposition to reason either analytically or heuristically (Stanovich & West, 1997; 1998; 2000; Kokis *et al.*, 2002 Epstein *et al.* 1996). It is also possible to prime either the heuristic or the analytic system (Epstein *et al.*, 1992; Ferreira *et al.*, 2006) and it is established that analytical reasoning does frequently take more time than heuristics based processes (Verschueren *et al.*, 2005; Evans & Curtis-Holmes, 2005).

Further evidence consistent with this proposition is evident in the relationship between thinking styles and reasoning performance, whereby significant amounts of variance in

reasoning ability are accounted for by thinking style (Stanovich & West, 2000; Stanovich *et al.*, 1999), described by Stanovich as being an intentional measure, in that they assess a person's *intention* to reason in a certain way, and not their ability to do so (Stanovich *et al.*, 1999). Specifically, having a high NFC, an analytical disposition, leads to greater levels of analytical thinking, while high levels of FI frequently result in less accurate reasoning performance (Epstein 1992).

The contrast between the two systems can also be seen when looking at the effect of the frequency format on probabilistic reasoning, with Sprenger and Dougherty (2006) suggesting that the probability format is using working memory to compare hypotheses (analogous to using the slower, more cognitive ability dependent system 2) while the frequency format allows faster, less cognitively demanding reasoning (the heuristic system 1). This interpretation – an inference from Sprenger and Dougherty's work rather than their own stated conclusion – would be disputed by Evans (2007a) and Kahneman and Frederick (2002) who each suggest that the frequency wording of the tasks makes the nested set structure of the tasks clearer to participants, enabling system 2.

Following on from Sloman's (2002) assertion that the wording of frequency tasks allows participants to see the nested sets involved in any probability task, Barbey and Sloman (2007) also assert that reasoning (in this case, specifically Bayesian reasoning) is facilitated:

“when the set structure of the problem is made transparent, thereby promoting use of elementary set operations and inferences about the logical (i.e., extensional) properties they entail” (p. 245)

That is, the logical system 2 is being primed, or enabled, so that the participants are able to see the logic in the task, and to achieve the normatively correct answer through the mathematical, rather than heuristic, route.

Chapter 6 of this thesis examines the use of the TDQ, NFC and FI scales (as used by, for instance Epstein *et al.*, 1992, and Kokis *et al.*, 2002) to measure thinking styles, and will again use the two formats of task, known as probability and frequency. As such, one external prompt, the format of the task, and one internal prompt, thinking style or

disposition, will be investigated. It is expected that there will be an interaction effect between thinking style and format, which would indicate whether those with an analytical or a heuristic style find the exposure of the set structures to be more beneficial.

Teigen (2004) has suggested that whether or not a judgement will show any bias does depend on whether or not the intuitive response is the right response – as such, it is the use of tasks which cannot be correctly answered through intuition alone which can discover which system is being used (see also De Neys, 2006b, and Donovan & Epstein, 1997). As such, the use of the Linda style tasks, where intuition frequently leads to a reasoning fallacy, remains an appropriate method for assessing which style of reasoning has been used.

Chapter 4 – Cognitive Ageing and Dual Process Theories of Reasoning

4.1 Age related cognitive decline may affect reasoning

The remainder of this literature review will examine the research findings regarding ageing and reasoning skills, and will look at some of the underlying processes that may affect reasoning in old age. Evidence as to whether there is a change in thinking styles across the lifespan, and the resulting tendency to rely on one or other of the two systems discussed in the previous chapter, will be discussed. The primary rationale for using a cohort of older participants is to provide greater variability in the measures of thinking styles, and so lead to greater ability to discriminate between individuals on this measure. As discussed in sections 4.2 and 4.3, below, older individuals are thought to be more likely to utilise system 1, the heuristic system, when attempting to solve reasoning problems. This is thought to reflect their greater tendency to use an intuitive thinking style, but such age differences have not so far been satisfactorily addressed in the literature.

While reasoning skills are a part of everyday decision-making for everyone, the ability to make informed judgements about matters such as health care, or financial decisions, may actually be more important to older individuals, as such decisions become both more frequent and more important in later life (Chen & Sun, 2003; Peters, Finucane, MacGregor & Slovic, 2000). Peters *et al.* (2000) also suggest that more pressure is now being placed on older individuals to make such major decisions for themselves, without the support of their immediate and extended families, now that families tend to be more widely dispersed and less likely to live in the same home, or even the same town or country. It is also 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).

As such, it is reasonable to expect that reasoning performance will also show age related decline, and yet this is an area that has not yet been covered in any great depth (as noted by Peters *et al.*, 2000). The research that has been conducted has shown 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. However, EPSE is suggested by the authors to be quite different from what they call ‘traditional’ problem solving, of the sort addressed in this thesis, whereby the problem and its solution are often more stylised, and designed by the researcher to examine particular phenomena.

One example of the way that ageing can negatively affect reasoning comes from Chasseigne, Mullet and Steward (1997), who gave participants a multiple cue probability learning task, whereby the cues involved could have a direct or inverse relationship to the criterion outcome. Participants were asked to estimate the temperature of a boiler from the cues that they were given, the cues being firstly a card with a green vertical bar on it, and secondly a picture of a boiler, with a temperature knob coloured green so as to match the colour of the bar on the card. The bar could be from 1 to 9 centimetres high, and the number of centimetres was the cue’s value. This then indicated the temperature of the boiler, in either a direct relationship (a tall bar meaning a high temperature) or an inverse one (tall bar meaning a low temperature). As well as these single cue conditions, multiple cue conditions occurred whereby participants had two cards, each with a different colour vertical bar on, to go with two colour coded temperature knobs. These could be two direct relationships, two inverse ones, or one of each, and the participant would have to combine these cues to estimate the correct temperature of the boiler.

Chasseigne *et al.* (1997) found that older participants could perform as well as younger ones when all of the cues were direct ones, but an age affect emerged when one of the cues given was inverse. Chasseigne, Ligneau, Grau, Le Gall, Roque and Mullet (2004) also found similar results, in that older individuals struggled more (in comparison to a younger cohort) as the tasks involved became more complex. Suspecting that the older group’s performance was being affected by the burden on working memory that the inverse cue produced, Chasseigne *et al.* (1997) designed an experiment to reduce this burden, by providing more written task information, and found that this did improve the performance of the 65 to 75 year old participants, but did not do so for those over 76. Similarly, Saczynski, Willis and Schaie (2002) had found that while all their participants, aged 64 to 95, did benefit from instruction on an inductive task, those aged 64 to 70 showed significantly greater improvement on strategy use than did the older participants. As such, perhaps research looking at the effects of younger old age – that is, those in their 60s – needs to be using measures that do include some quite complex

tasks, to avoid a ceiling effect where all participants, young and old, achieve the same high standard.

Further to the findings of Chasseigne *et al.* (1997), Mutter and Williams (2004) gave young (with a mean age of 20 years) or old (mean age of 73 years) participants a contingency task, asking them to determine whether pressing the spacebar had any effect on the display on the computer screen – a triangle, which either flashed or remained constant. Similar to the findings of Chasseigne *et al.* (1997), that older people were less able to detect inverse relationships, the age-related differences in Mutter and Williams' study were greater when the contingencies were negative (that is, pressing the space bar prevented the on screen response) than when they were positive. They also found that older people were particularly poor at giving an actual numerical estimate for the relationship. Further support for the theory that older people find the negative relationships more difficult was also found by Mutter, Haggblom, Plumlee and Schirmer (2006), where participants were required to identify whether a target was 'correct' or not. In the positive condition, the target was 'correct' if it did contain a preordained symbol, while in the negative condition it was 'correct' if that symbol was absent. While all participants found the negative rule more difficult to learn, this was particularly the case for older adults. Mutter and Plumlee (2009) also found that not only did older participants find it more difficult to integrate a range of information in order to solve problems, they also failed to benefit from their being framed in meaningful contexts, something which significantly improved performance in a younger group.

Chasseigne *et al.*'s 1997 study does then support the supposition that the age related decline in working memory can result in impaired reasoning performance. Also, while in some cases older individuals will not show higher levels of illusory correlation (e.g., overestimating the prevalence of undesirable traits in a minority group), an age effect was found when participants had been given a distraction task to perform whilst the target information was presented (Mutter, 2000). Again, the suggestion is that the greater cognitive load is more difficult for the older participants to deal with, and as such their judgments become negatively affected far more than those of the younger group.

Gilinsky and Judd (1994) found age related decline in syllogistic reasoning, in both the ability to construct and to evaluate conclusions. Older people were also more susceptible to belief bias, whereby they were more likely to accept a conclusion as valid if it appeared to be 'believable', or likely to occur in the real world, regardless of its logical validity. Similarly, they were also more likely to reject unbelievable conclusions, even if they were in fact valid. There was also an effect of the number of mental models required on the task. The more complex syllogisms, requiring greater use of working memory (Johnson-Laird, 1983), were found to be even more difficult by older participants. Again, this would suggest that any age related decline is due, at least in part, to the older participants' poorer working memory.

Indeed, De Neys, (2006b), states that 'erroneous reasoning in the case of a belief/logic conflict is not only associated with, but also directly caused by, limitations in executive resources' (p432). He concluded that when the heuristic system 1 and the analytic system 2 both came to the same answer on a given task, cognitive loading had no effect, but when there was a conflict between the two systems, an effect was found.

Peters *et al.* (2000) also suggest that older people may be more likely to show a reasoning bias, perhaps due to an overreliance on the representativeness heuristic, which could lead to base rate neglect in real life situations. They offer the example of choosing a hospital on the basis of the food and service you experience on a visit (individual case information), rather than the basis of the effectiveness of the treatment it provides for their condition (the base rate information). In this case, the food and service are 'representative' of having a good hospital experience, despite being entirely independent of the more important factor of the likelihood of a successful treatment outcome.

While the literature discussed above illustrates how older people can struggle with a range of reasoning tasks (and especially more complex ones), one area where older people have been found to not only equal, but to actually out perform their younger counterparts regards making 'irregular decisions' about preferred options. Tentori, Osherson, Hasher and May (2001) asked their participants which type of hypothetical supermarket discount card they preferred – Card A, which gave a 15% discount and had a minimum spend of \$20, or Card B which gave a discount of 25% but a minimum spend of \$45. In a between subjects design, other participants were also shown cards A

and B, but were also offered Card C, which has a discount of 25%, but a minimum spend of \$100. An irregular decision occurs when participants in the AB group prefer card A, but the ABC group go for Card B. It appears that the existence of card C has irrationally influenced the participant's attitudes to cards A and B, and it was found that younger participants were consistently making this kind of irregular decision more often than the older group, even when the researchers took into account the possibility of the older participants having some sort of market expertise or greater familiarity with the concepts involved in the materials. This task may be relatively simple, however, and it may well be that with a more complex task, and an increased cognitive load, age-related deficits might have been apparent.

With regards to probabilistic reasoning, the limited amount of research that has looked at the effects of ageing includes a series of studies by Fisk (2005), who found that while all participants made errors in Bayesian inference, and on conjunctive and both types of disjunctive problems, this was equally evident in both younger and older persons. However, Fisk did not measure the actual magnitude of the conjunctive and disjunctive errors, instead simply categorising responses according to whether or not an error had been made. It may be that quantifying the errors made will reveal that while there is no age difference in the actual number of errors, there may be an age difference in the magnitude and perhaps the type of error that is made. In his research on the role of schemas in memory and task performance, the results obtained by Hess in 1990 could offer one explanation for the lack of an age effect with this form of reasoning. He found that when given some traits about a person previously described to them, both old and young people were more likely to remember the traits that were inconsistent with the primed stereotypes (or schemas) about that person. In the case of the probability tasks, for example, like the conjunction fallacy paradigm, it would seem that both old and young participants would be more likely to pay greater attention to the unlikely elements of the vignette and statement, with no age-related effect therefore apparent on that aspect of the tasks.

Mutter and Goedert (1997) presented a young and an older group of participants with a series of words, informing them that some of these words would be repeated, and that they would later be tested on their memory of the words. Half of each age group were also given a distracting addition task. When they were later asked about the frequency of a given word's presentation, it was found that older people's estimates of its

frequency deviated from the correct answer significantly more than the younger group's did. However, controlling for higher backward digit span eliminated the age-related variance. Similarly, when participants were asked instead to discriminate between two words, saying which had been presented more frequently, the older group did not perform significantly worse. However, the difference again became significant once the age-related variance associated with performance on the WAIS-R had been controlled for. Again, this indicates that the effects of age on judgement making may be quite subtle, but also closely linked to a range of cognitive functions.

The literature examined above suggests that older people do show some decline in a range of reasoning skills, although this is often only evident in more complex tasks. It has been suggested that this is in large part due to the decline of working memory with age (Chasseigne *et al.* 1997; Gilinsky & Judd, 1994).

4.1.2 Information Processing Speed

One of the cognitive processes of importance in both cognitive ageing and reasoning performance is that of information processing speed. Salthouse (1993) and Salthouse and Babcock (1991) have looked at the associations between the age-related cognitive processes, and conclude that much of the decline in cognitive function is directly related to that in information processing speed. Fisk and Sharp (2002) also found that the age affect on syllogistic reasoning performance disappeared once information processing speed had been accounted for, and further support for the impact of information processing speed can be found in Fisk and Warr (1996). Clark, Gardner and Brown (1990) also looked at information processing speed. Their participants were given analogical reasoning tasks, in the form of 'x is to y as a is to . . .' and they again found a significant effect of age group, in that the older participants found this reasoning task more difficult, but there was also a significant effect on response time, with older participants (those over 50 years) responding more slowly than their younger counterparts. In exploring the basis of age-related differences in information processing speed, Salthouse (1996) also investigated the 'limited time mechanism' and 'simultaneity mechanism' The former suggests that (age-related) slower processing speed impairs performance as participants spend a disproportionate amount of time on the initial stages of a task, and run out of time to complete remaining stages effectively. As such, the limited time mechanism is most likely to be involved when the task is such that participants have time limits placed upon them. The simultaneity mechanism

suggests that the slower speed of processing may lead to the results of early stages of processing (on a given task) being forgotten or corrupted as the later stages of processing are completed – a mechanism that would affect time unlimited tasks as much, if not more so than those that are time limited.

Salthouse (2000) further investigated this area, looking at a wide range of data on a variety of cognitive tasks, and found that the influence of information processing speed is not simply attributable to participants failing to reach the later items in timed tests – that is, the influence cannot be due to the limited time mechanism alone. While older persons' slower completion of time-limited tasks such as Raven's Progressive Matrices did result in them being unable to complete later test items, Salthouse found that performance on those tasks that older persons had completed was still adversely affected by the increasing difficulty relative to younger persons. Further to Salthouse (1996) above, Salthouse (2000) also discusses the suggestion that speed of reading and responding cannot have too great an impact on reasoning ability when participants are given enough time to complete the task. However, Salthouse (2000) does suggest that the 'speed of internal mental operations may be a factor' (p.594), suggesting that it may be at the processing stage, rather than at input or output, that the effect may lie. Wareing, Fisk, Montgomery, Murphy and Chandler (2007) also suggest a third mechanism, whereby processing is aborted too soon due to impulsivity. While impulsivity may be present in both age groups, the implications of this would be that it is more evident among older participants, with less processing having been completed at the time the process was aborted.

The evidence presented above suggests that the older participants' slower information processing speed is affecting task performance for some process based reason, over and above simple time limitations, although it also does not rule out the possibility that older participants may be aware of the possibility of running out of time, and become increasingly anxious about this as the task progresses. As such, the current study will impose no time restriction on completion of any of the reasoning tasks.

It should also be noted here that older participants may experience some decline in their ability to inhibit unnecessary information. Viskontas, Morrison, Holyoak, Hummel and Knowlton (2004) found that older people were finding it more difficult to inhibit irrelevant information in an analogy task, and especially so when the tasks became more

complex. This is an indication that inhibition is one of the processes that deteriorates with age, and therefore one of the processes that cause the age related decline in many of the more complex reasoning tasks. However, in a review of the literature on inhibition deficits in older people, Burke and Osborne (2007) conclude that inhibition is not the most evident age-related deficit in laboratory-based tasks, and is in any case difficult to measure as a discrete process.

4.2 Are older people reasoning differently? Links with dual process theory

There is a wealth of research to support the theory that there is an age related decline in various cognitive functions. In his 1998 paper, Salthouse examined a range of variables which he summarised under seven headings – reasoning, knowledge, quantitative, short term memory, perceptual speed, closure (making sense of incomplete words, pictures and sounds) and associative memory. Each of these declined to some extent with age, in a group of participants ranging from childhood to 94 years.

One of the most consistent findings in the ageing literature is that memory, specifically working memory, declines with age, (see, for example, Salthouse & Babcock, 1991; Craik, 1994; Hultsch, Hertzog, Small & Dixon, 1999). More recently, Mutter *et al.* (2006) found that reduced working memory capacity, whether this was age related or due to the experimental design creating a concurrent load on working memory, did lead to a decline in their participants' ability to learn predictive relationships. Similarly, work by De Beni and Palladino in 2004 found that older participants had particular problems with updating information in their working memory, and specifically in excluding information that was relevant previously, but was now no longer needed. This could be particularly important to participants in a laboratory setting where they may be asked to perform several tasks in succession, and in the case of this current research, where they are being asked to complete a series of probabilistic reasoning tasks that may have some similarities.

In a review of the research, Salthouse (2005) also concludes that working memory clearly has a part to play in accounting for the effect of ageing on reasoning performance although the exact nature of the relationship remains unclear. Beyond working memory influences there is evidence of differing strategy use, although this evidence is somewhat inconsistent (Salthouse, 2005).

Fisk (2005) found that many of the cognitive measures that were correlated with reasoning ability in a younger group showed no such correlations in an older group, suggesting that older people may be using different methods, and therefore different cognitive processes. One example of such a method is Gigerenzer and Goldstein's (1996) theory of 'fast and frugal' reasoning, whereby the process of reasoning is conducted using an algorithm which negates the need to use and integrate all of the available material (see 2.5.8). Instead the participants may 'Take the Best' cue available to them, with 'the best' being defined by whether or not that cue is recognised, over and above other cues. If this cue discriminates effectively between various solutions, then the task is considered solved, with no further cues needing examination. If, however, the cue is found wanting, then further cues are examined, in turn, and in order of recognition. Gigerenzer and Goldstein tested their Take the Best algorithm and found that it produced the correct answer in almost 90% of cases, despite not necessarily taking account of all of the available evidence (1996). However, the examples that they used may of course have been ones that particularly lent themselves to the use of the algorithm – that is, they chose tasks where the most representative cue was frequently the correct one. Having said that, this may not reduce ecological validity if the representative cue is also most frequently the correct one in everyday life. If, as suggested by Fisk (2005) older people are not using the same cognitive processes as younger people, then perhaps older people are more likely to be using such algorithms and in that way working around their limitations, whether consciously or unconsciously. Further evidence for this tentative hypothesis comes from Blanchard-Fields (1996) who suggests that as we age our reasoning skills mature to cope with more complex, 'real life' problems, as opposed to the more structured and simple problems that young people encounter in the academic environment. Again, in dealing with more complex problems, it may be that older people are relying on different processes to solve them effectively and yet frugally.

As summarised above, older people often do not begin to show any decline in reasoning performance until tasks become particularly complex. In their work on older individuals' strategy use, Chen and Sun (2003) gave two groups of participants – young and old – a yard sale task, whereby they had to try and sell their (hypothetical) products for the highest of three bids. However, the bids were presented sequentially, with no indication as to which of them may be the highest, and as such the participants needed

some kind of strategy to decide when to accept a bid, and when to reject it in the expectation that the next bid will be higher. Thus participants were making judgements of risk, for instance, if bid A was rejected, while B was even lower, should they then accept B as the bid may drop further, or risk rejecting B in the hope that C may be higher. The results revealed that individuals performed consistently better than chance. Furthermore, the older group performed as well as their younger counterparts, although they were more likely to choose one consistent strategy which did not load heavily on working memory, whereas the younger participants were more likely to switch strategies, creating a heavier load on their working memory processes. Fisk (2005) has also suggested that younger adults may be using different strategies from their older counterparts when making judgements on reasoning problems. This leads to the possibility that older individuals may be less inclined to rely on the highly cognitively demanding system 2, and be more likely to utilise the more heuristic system 1 (as defined by Evans & Over, 1996, Sloman, 2002, and Stanovich & West, 2000, and discussed in Chapter 3).

Chen and Sun (1997) suggest that their findings reflect Baltes' theory of selective optimization (Baltes, 1997), whereby as we age we select strategies and approaches that we can continue to use to full effect, using practice to optimize our performance and to compensate for the shortfalls caused by cognitive decline. This may, of course, be an entirely unconscious shift.

Yates and Patalano (1999) have also suggested that as we age we move towards a more automatic style of reasoning. They also point out that an older person has, almost by definition, made more decisions across the course of their life compared to a younger person, and as well as resulting in a more automatic process, this form of practice could also account for the lack of age-related deficits found in so many studies in this area. It could be that in many cases the automatic, system 1 processes are actually achieving the correct answer, meaning that it is fallacious to assume that an accurate answer must represent analytical processes. If older individuals are more likely to be reliant on the heuristic system, this could offer one explanation of why only the most complex of tasks are leading to any measurable age-related deficits in performance. It is possible that for the less demanding tasks, either system will reach the correct answer, leading to old and young groups showing no difference in ability, while the more complex tasks can really only be solved through analytical reasoning, which then disadvantages older.

heuristic dependent, individuals. However, Fisk (2005) did not find any such age effect when using probabilistic tasks, despite their well documented difficulty. It may well be that such tasks are utilising quite different processes not only from deductive problems, but also from other inferential tasks.

Similarly, Mutter and Pliske (1994) also concluded that older people may rely more on heuristics and rules which bypass the limitations of their memory. In a study in which participants were asked to judge whether fictional patients were likely to be exhibiting certain clinical behaviours, Mutter and Pliske found that older people were less able to utilise the information that they were given. That is, while the younger participants' performance on the task improved as the information made available became more salient, older people showed significantly less improvement when the same information was provided. However, when the information given to the participants was not particularly salient, there was no effect of age upon judgement making ability, suggesting that the younger participants were able to appreciate the importance of the more salient information, while the older group were not. Johnson's (1993) work also suggested that younger people might be making more use of the information available to them, as they appeared to be rechecking this more frequently than their older counterparts. In a task asking them to choose an apartment from a list of five with various attributes there was no significant difference between the two age groups' choices – the older participants' apparent lack of rechecking did not seem to affect their overall reasoning. Again, this reinforces the idea that both system 1 and system 2 processes may lead to similarly 'right' answers, and be equally valid strategies in many cases.

While many of the researchers cited above suggest that older participants are able to effectively compensate for cognitive decline Park, Willis, Morrow, Diehl and Gaines (1994) found this not to be the case. They reviewed a range of research looking at instructions on medication (both in laboratory and field studies), and found that older individuals frequently showed problems comprehending these instructions, especially when they involved making some kind of inference. For instance, realising how much longer their pills would last them if taking them three times a day, and when they would need to get a new supply. Although Park *et al.* (2004) were not directly comparing an older group with a young cohort, Finucane, Slovic, Hibbard, Peters, Mertz and MacGregor (2002) did find an age effect when participants were asked to judge which

of a range of health plans was best for them. Their older group were again showing significantly greater comprehension errors, as well as a higher level of inconsistency in their choices.

These two examples, by Park *et al.* (1994) and Finucane, *et al.* (2002) do offer further support for the suggestion by Chasseigne *et al.* (2004), Chasseigne *et al.* (1997) and Mutter and Pliske (1994) that older people are particularly affected by the increase in cognitive load that some tasks require. With regards to probabilistic reasoning, and specifically to the two formats (probability and frequency) addressed in Chapter 2, it could be that expressing the tasks in the frequency format reduces some of this cognitive load by reducing the inferences made. For instance, for individuals not using percentages in day to day life, knowing that 10% can also mean 1 in 10, or even that it means 10 in 100, is an inference in itself, and not necessarily an easily accessible concept. As such, this could lead to older individuals benefiting to a greater degree from the more explicit '10 out of 100 people' phrasing which characterises the frequency format. However, Park *et al.* (1994) has noted that when asked for their preferences in the way medication information was arranged, they did not show any differences in their choices than a younger group – the two age groups were equally aware of the benefits of the different formats (in this case the information was as a list, or a paragraph, the latter showing greater memory load). If older participants do benefit from more explicit information, they may not be aware that this is the case.

Evidence for this proposal that participants may be unaware of their limitations comes from Copeland and Radvansky's (2004) syllogistic reasoning study. These researchers found that on those occasions when participants with lower memory spans were thought to be constructing fewer mental models, they did not report lower levels of confidence in their abilities. In other words, perhaps they construct fewer models because they are unaware that additional models need to be constructed, rather than because they are aware of their limited memory span and are attempting to work around it. This would suggest that any shift towards an automatic, heuristic based system may well not be a conscious choice driven by an awareness of reduced cognitive ability. Older people would increasingly benefit from the frequency format if, as suggested in Chapter 3.5, it primes the analytic system 2 while also allowing it to work more effectively by reducing the load on working memory.

Regarding the lack of age effect on confidence levels, regardless of the lower ability (Copeland & Radvansky, 2004), it may be the case that such a lack of awareness of cognitive limitations might lead to more frequent instances of a Feeling of Rightness (FOR). In her development of the dual process theory, Thompson (2009) suggests that it is this FOR which leads to participants failing to engage the analytic system 2 when it is appropriate to do so. That is, if a participant has a strong FOR from the answer that is rapidly obtained through system 1, then they accept that answer as being correct with no further analysis. If older people are experiencing higher numbers or levels of inaccurate Feelings of Rightness (due to a lack of awareness that there are further steps that could be taken to reach the correct answer), then they will be more likely to accept the non-normative responses obtained by system 1.

Further evidence for a shift in strategy use in older adults comes from Fontaine and Pennequin (2000), who examined the effect of ageing on inferential reasoning about class inclusion. The results revealed that from 60 years of age onwards, adults showed signs of a deterioration in their ability to make correct inferences. For instance, in one of the measures used, participants were shown pictures of tulips and daisies, and were then (having previously been presented with questions about ‘the white flowers’ and ‘the red tulips’) asked whether there were more flowers or more daisies. Older participants were shown to be finding it harder to recognise that the subclass of daisies should be included in the super ordinate class of flowers. Crucially, those over 60 were also giving different justifications for their choice of answer, with older adults giving what the authors called ‘empirical’ justifications (for instance, ‘there are more flowers than daisies because there are 8 flowers and only 5 daisies’), while younger adults gave more ‘logical’ justifications (‘there are more flowers because daisies are flowers’). Also, the oldest adults often gave justifications that should lead to the correct response, and yet were still ultimately giving an incorrect answer. Contrary to previous research, which suggested that older participants used different strategies from their younger counterparts but could still be as successful in achieving correct answers (e.g. Johnson, 1993), in this case it seems that the older participants’ different strategies were ultimately ineffective. Fontaine and Pennequin (2000) felt that the elderly adults were struggling in two main ways – first, they often failed to integrate the relevant information, and second they were actually less able to extract the relevant information and had more difficulty in inhibiting what was irrelevant. With older participants finding this task particularly difficult, it seems that the set structures remain unclear to

them. In the current study, the frequency format is expected to facilitate reasoning by making the set structures involved more explicit (as discussed by Sloman, *et al.* 2003), and it may therefore be that this is particularly helpful to older participants.

Saunders and Over (2009) note that while system 1, or The Autonomous Set of Systems (TASS) includes many processes that are evolutionarily advantageous, some of the processes are those that have actually been learnt (see also Fisk, 2004, and Kahneman & Frederick, 2002; 2005). This could lead to an advantage for the older participants if they have successfully learned and made automatic – transferred from system 2 to system 1 – the correct, or normative, way of calculating the answer. However, see section 4.4 for a full discussion of the unlikelihood of an age advantage due to expertise.

To conclude this section, it is anticipated that older participants will find the frequency format to be particularly beneficial to them in terms of avoiding the reasoning fallacies (conjunctive and disjunctive) and in making estimates that are closer in magnitude to the normative answers than are found when using the probability format. This expectation is supported by the research cited above which suggests that older participants are less able to manipulate complex information (e.g. Chasseigne, *et al.* 2004, Chasseigne, *et al.* 1997, and Mutter & Pliske, 1994), and that the frequency format clarifies important elements of the task, such as set structure (Sloman, *et al.* 2003), which older participants may be expected to find particularly complex (Fontaine & Pennequin, 2000).

4.3 Ageing and Thinking Styles

If, as much of the literature suggests, older people are less likely than younger persons to utilise system 2 processes then this should be evident in indicators of their thinking style. Specifically, are these lower levels of analytic processing reflected in lower levels of the analytical thinking dispositions? There is very little research in this area that directly compares ‘old’ and ‘young’ participants in terms of thinking styles, but one study that has done so was conducted by Klaczynski and Robinson (2000). Similar to findings reported above (for instance, Chasseigne *et al.* 1997;1994 – see Chapter 4.1), Klaczynski and Robinson suggest that age-related deficits may only be found in certain, manipulated situations. They found an age difference in relation to biases in reasoning, whereby older people did show more reasoning bias – for example in failing to make

judgements that follow the law of large numbers. Klaczynski and Robinson's (2000) older participants were also more likely to accept evidence consistent with their views with little evaluation, while by contrast rejecting evidence that contradicted them, using apparently logical/scientific methods. On a measure of 'justification' of their judgements, the older participants would often accept invalid conclusions by extrapolating from their own personal experience, using quasi-logical statements such as "I'm a Catholic, and so are all my friends, and their faith in God is just as strong as mine", which would reinforce the validity of the syllogistic conclusion that 'all Catholics have such strong faith'. Crucially, they also found that 'middle aged' adults (mean age 47.5) and older adults (69.7) had significantly *higher* levels of NFC than a younger group (mean age 21.5). This would suggest that possessing a higher NFC does not necessarily lead to a more accurate performance, perhaps due to the cognitive limitations of older participants preventing them from making effective use of the analytical processes that characterise system 2 (see also Chapter 4.5, below). The inference made from older participants' greater utilisation of system 1 (heuristic) processes, discussed in 4.2, is that they are more likely to show a tendency to think intuitively, which was not directly addressed by Klaczynski and Robinson (2000). They instead measured only analytical thinking. As these measures are not orthogonal, and it is possible for an individual to show high or low tendencies on both, the finding that the older participants showed greater levels of NFC cannot be used to infer that they therefore showed lower levels of intuitive thinking. In the current research, a wider range of thinking styles will be addressed, using both Epstein *et al.*'s Rational Experiential Inventory (1996) and Kokis *et al.*'s Thinking Dispositions Questionnaire (2002). These measures are discussed at length in the materials section of Chapter 6, but to summarise they assess a range of thinking styles, which can be broadly categorised into being either 'intuitive' or 'rational'. It is anticipated that older participants will show a greater tendency to intuition, and will show greater variability in thinking styles overall, allowing the current study to discriminate more accurately between those showing high and low levels of each style or disposition.

However, Klaczynski and Robinson (2000) do suggest that everyday reasoning may not be particularly affected. Thus while reasoning biases may become more pronounced with ageing, it does not necessarily follow that older persons' reasoning performance, and their ability to arrive at the best answer, will be seriously impaired in all situations. The application of heuristic processes may reach useful and/or accurate conclusions in

many cases. The implication here is that tasks may need to be specifically designed to highlight any conflict between heuristic and analytical processes. As discussed by Donovan and Epstein (1997) it is possible for tasks to be ‘natural’ or ‘unnatural’, with natural tasks being those that allow the correct answer to be reached through heuristic thinking alone. In order to use inaccurate reasoning as an indicator of non-analytical thinking, this current research uses probabilistic reasoning tasks which can be considered to be ‘unnatural’, in that analytical reasoning is believed to be needed if reasoning fallacies are to be avoided.

4.4 Ageing and Expertise

The literature discussed so far focuses on ageing and its associated cognitive decline, and the ways in which this will either lead to a detriment in reasoning performance (e.g. Chasseigne, *et al.*, 2004; Mutter *et al.*, 2006) or, at best, may be ‘accommodated’, or ‘compensated for’ to enable an equal performance (Baltes, 1997; Yates & Patalano, 1999; Fisk, 2005). A further possibility is that older individuals may actually outperform younger ones on some tasks due to increased knowledge, expertise, and/or ‘wisdom’.

Wisdom as a concept is somewhat different from crystallised intelligence (Stuart-Hamilton, 2006), but similar to crystallised intelligence it is often taken to be a form of intelligence, or an indicator of ability, which increases across the lifespan. Staudinger (1999) investigates wisdom from the ‘theoretical conceptualisation of wisdom as expert-level knowledge and judgement in the fundamental pragmatics of life’ (p.643), which suggests that wisdom is in itself a specific domain of expertise. Expertise in such ‘real life’ judgements (Staudinger’s materials involve making decisions about reacting to emotionally charged social situations, 1999) may have little impact on the type of task used in the field of probabilistic reasoning, where participants are far more likely to achieve the correct answer through considering the statistical data only, and may be more likely to fall prey to reasoning errors if previous experience is called into account. Staudinger (1999) finds however that wisdom does not increase markedly between the ages of 20 and 75, or at least if there is any increase, it is then involved in again *compensating* for cognitive decline, leading to a levelling out of performance over all (see also Blanchard-Fields, 1996, and Krampe & Charness, 2006). Mickler and

Staudinger (2008) later investigated personal wisdom – a field of wisdom relating specifically to one’s own life – and found it to be negatively related to age.

In terms of expertise, the field of social cognition has again suggested that age is associated with *greater* sensitivity to certain cues, such as using behaviours to determine personality traits and make social judgements about others (Hess, Osowski & Leclerc, 2005). Hess *et al.* (2005) also found that greater sensitivity to such cues was associated with social experience, and that social experience did moderate the age effect such that it seems the age related increase in performance was in part due to older participants’ greater levels of experience in making such judgements. As discussed in 4.1 above, Tentori *et al.* (2001) found that older participants made ‘better’ – that is, more regular – decisions in a task involving choosing between supermarket discount cards with varying properties. They suggest that their findings show that the older group were demonstrating ‘judgmental wisdom’, and that they may have made more sensible decisions based on realistic estimates of their own spending.

Whether or not people have expertise in any domain is dependent on their knowledge and experience of the subject (Rybash, Royer & Hoobin, 1986), with expertise being something that is very much domain specific. Just as the participants involved in the study by Tentori *et al.* (2001) would have had greater expertise in shopping expenditure and money saving decisions, and Hess *et al.* (2005) used participants with greater social experience, for older participants to show greater expertise in probabilistic reasoning tasks they must have a greater level of experience and knowledge of solving such tasks. And as found by Hess *et al.* (2005) above, this must then be to the extent that their expertise is able to not only compensate for cognitive decline, but to produce a performance that overrides this to the extent that they outperform younger participants. There are two ways in which such greater expertise may occur. The first would be for there to be a cohort effect, caused by changes in the content of maths curricula over the past few decades. School pupils in England and Wales have since the late 1980’s taken mathematics General Certificate in Secondary Education (GCSE) which contain a module on ‘handling data’, a small part of which is concerned with probabilities, with students encouraged to use tree diagrams and data in the forms of fractions in their calculations. While not all school pupils will pass, or even take, mathematics at GCSE, all young participants in the current research will have done so to grade C or above, as it is a part of their degree course requirements. The vast majority of older individuals will

have taken either the CSE (Certificate in Secondary Education) or, if deemed more academically able, the O-Level (Ordinary-Level). Neither of these directly assessed the pupils' ability to reason with probabilities, and as such any cohort effect would be more likely to exist in the opposite direction, with the younger participants having the greater, and the more recent, level of experience. However, it should be noted that previous research suggests that expertise in statistical problems may link only to the method used to tackle the tasks, and not to the accuracy of the responses given (Hertwig & Chase, 1998).

The second way in which older participants as a group may have greater experience and therefore expertise in probabilistic reasoning is by their simply having made (or witnessed) more such judgements. Just as Hess *et al.* (2005) found that older participants made better social judgements in part because of their greater social experience, it may be that greater experience in situations requiring reasoning with probabilities would also lead to better judgements being made on the problems used in the current research. While it seems reasonable to assume that older participants will have made many such judgments across their lifetime, on every decision from the daily and inconsequential to the unusual and life changing, it seems unlikely that such 'real life' decisions will lead to a true level of expertise, which is usually created by intensive and purposeful practice in the field. If older participants are to be advantaged by greater levels of experience in making judgements of probability, it seems likely that the use of natural frequencies will be particularly advantageous to them, representing, as it is suggested, the way in which such data is collected and acted on in real life (Gigerenzer & Hoffrage, 1995), as discussed above in 4.2.

4.5 Conclusions

To summarise the literature discussed above, it is clear that a range of functions (notably working memory and information processing speed) do decline as we age (Salthouse & Babcock, 1991; Salthouse 1993, 1998, and 2000). However, despite the documented deterioration in many cognitive processes, there are mixed findings with regards to the way in which the ageing process may be related to the evolution of reasoning skills. As demonstrated by Chasseigne *et al.* (1997), Chasseigne *et al.* (2004) and Mutter (2006), it is frequently the case that age-related declines in reasoning skills are only found when the tasks utilised become particularly complex, and involve less

straightforward cause and effect relationships. However, in the limited amount of research that is available on probabilistic reasoning and ageing, no clear age-related deterioration has been found (Fisk, 2005). This then leads to the suggestion that perhaps the tasks involved in probabilistic research are too simple to show up the age effect. However, as addressed in the previous chapter, between 50 to 90% of adults of any age will make errors in conjunctive tasks (Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Tversky & Kahneman, 1983; Yates & Carlson, 1986), while Carlson and Yates (1989) and Fisk (2005) found that the incidence of errors in both inclusive and exclusive disjunctive tasks ranged from just over 60% to 80%, and Birnbaum (2004) suggests that over 80% of student participants give incorrect answers on Bayesian tasks. Clearly then, the majority of participants are not finding these tasks to be simple, and the implication is that they are cognitively demanding to the point where so many errors are made, either because the tasks are too difficult to complete correctly, or because participants *perceive* them to be difficult or unimportant, and do not apply themselves.

If it is the case that most people find these kinds of tasks difficult, but they do not show the kind of age related decline associated with other complex reasoning tasks, there must be some factor – or process – allowing older participants to equal their younger counterparts' performance. Chen and Sun (2003) suggest that older people may be reasoning differently by using a greater number of heuristics to reach their answers, and if this were the case then in many instances it is possible that these heuristics might be as effective in reaching the correct, or normative, answer as any other method. Fisk (2005) and Yates and Patalano (1999) also support the proposition that we may move towards a greater reliance on system 1 as we age. Using the framework of a dual process system of reasoning, as discussed in the previous chapter, it seems that older people may be showing a tendency to use the heuristic system 1 (as opposed to the analytic system 2) more often than their younger counterparts.

A final explanation for the unexpected findings regarding ageing and probabilistic reasoning may well be due to the role played by individual differences. In his work examining the process of cognitive ageing, Salthouse (1998, 2005) proposes that individual differences may account for much of the age-related variance found in cognitive functions, both within and between age groups. Individual differences in terms of cognitive ability and processing speed may be relevant, and in addition to these

the current research will also look at thinking dispositions. With previous research having shown a clear relationship between analytical thinking and reasoning performance on a range of tasks (for instance Stanovich & West, 1998, and Stanovich *et al.* 1999) it may be that older people's greater reliance on heuristics can be accounted for not only as a consequence of their reduced cognitive capacity, but also by an increased tendency to show a heuristic thinking style in general during the course of reasoning.

As such, this current research will examine older participants' performance on tasks (presented in both probability and frequency formats) in the light of their individual differences in intelligence, processing speed and thinking style, with the intention of comparing such participants' performance on these measures with that of younger persons.

4.6 Summary of Literature Review and Introduction to Empirical Chapters

The current chapter and the preceding two have summarised a number of overlapping bodies of literature. They address probabilistic reasoning, dual process theories and their relevance to such reasoning, individual differences in the form of thinking styles, and finally possible findings and theories that may explain the lack of effect of age upon probabilistic reasoning performance. The current literature leads to a number of research questions, which will be addressed in five empirical chapters.

Chapter 5 examines the association between format (frequency or probability) of reasoning tasks and fallacies committed on conjunctive and disjunctive tasks. The latter type of task has so far been largely absent from studies addressing the so called 'frequency effect', and it is anticipated that this facilitating effect will be seen in disjunctions, both inclusive and exclusive, just as has previously been found in a number of studies using conjunctive tasks (e.g. Fiedler, 1988). This study will also introduce a measure of error, whereby the magnitude of the difference between the normative response to a problem and the participants' own response can be assessed. Again, it is anticipated that this will reveal the frequency format to produce more accurate reasoning on both conjunctive and disjunctive tasks. It is also anticipated that this measure will aid the investigation of whether participants are able to discriminate between, and thereby respond appropriately to, inclusive and exclusive disjunctions.

Chapter 6 then refines the materials used, both in terms of this measure of error and the reasoning tasks themselves. This chapter also measures the participants' thinking styles and verbal intelligence in order to assess the impact of these individual differences upon reasoning performance. It is anticipated that verbal intelligence and thinking style will to some extent mediate the effect of the frequency format, with those showing higher levels of intelligence benefitting less from that format. Due primarily to Sloman's (2002) suggestion that the frequency format will prime analytic thinking in those not normally disposed to do so, it is anticipated that those high in analytic thinking will not benefit so greatly from the frequency effect. In this case, it is the investigation of thinking styles and their relation to the frequency effect which is novel to this study.

With only limited support for a link between problem format and thinking style having been found in Chapter 6, Chapter 7 then addresses this further by using a cohort of older participants who are believed to reason primarily using system 1, heuristic, processes (see Chapter 4.2, and particularly Fisk, 2005). An effect of format upon both measures of fallacy and error was again anticipated, and that thinking styles would also account for significant levels of variance within reasoning performance, and would attenuate any task format related variance. Similarly, it was expected that information processing speed and verbal intelligence would also attenuate any such variance, with information processing speed in particular being known to account for much of the age related detriment to many cognitive tasks (Salthouse, 2005). It was also expected that there would be an interaction effect between age (old and young) and task format (probability and frequency) whereby older participants would find the frequency format particularly beneficial. This is due to their having a greater tendency to reason intuitively, while the frequency format primes more analytical processes.

The fourth empirical chapter will again examine the frequency effect and age related variance in thinking styles, but uses the more cognitively demanding Bayesian tasks discussed in Chapter 2.4. Just as the use of older participants was intended to provide greater variability in the measures of individual differences, the use of more complex tasks was expected to illustrate the way in which older participants show poorer performance on such tasks only when the cognitive demands involved are particularly high (e.g. Chasseigne *et al.* 1997; Mutter & Williams, 2004). Such tasks also tend to show a pattern of responding which reveals the extent to which participants have

attended to the information presented to them (Birnbaum 2004), which it is anticipated will reveal an age effect in terms of the method used in completing the tasks. This would produce more directly observable evidence of the two age groups' differing methods than has previously been presented.

With the previous chapter having found no observable age differences in performance on Bayesian tasks presented in probability and *normalised* frequency formats, the final chapter will again present results from Bayesian reasoning tasks, but in this case they are presented as both probabilities and *natural* frequencies. The latter presentation is believed to be particularly advantageous to older participants as they facilitate analytical reasoning (Hoffrage *et al.*, 2002).

Chapter 5 – Effects of Format on Probabilistic Reasoning

5.1 Introduction

As discussed in Chapter 2, probabilistic reasoning has been addressed in the literature for decades, notably by Kahneman and Tversky (1972 onwards) and more recently by Gigerenzer and Hoffrage (1995) and Fisk (2002, 2005). Traditionally, probabilistic reasoning tasks provide participants with some background knowledge about a hypothetical person or situation, and will then ask them to use that information to estimate the probability of some other personal characteristic, or event.

Perhaps the best known task used in this area of research is the conjunction problem in which individuals are asked to estimate the probability of two events occurring together, i.e., the conjunctive probability of A *AND* B. In most tasks, one of the components will be likely (or representative), and one unlikely. In this context most people erroneously assign a higher probability to the conjunctive event than to either one or both of the components, typically larger than the unlikely event. This behaviour breaks a fundamental rule of probability, and has become known as the conjunction fallacy. Less frequently used are disjunction tasks, asking for the probability of A *OR* B. A disjunction fallacy is committed when participants produce a value for the disjunctive probability that is less than one or other (or both) of the two component event probabilities. Disjunctions can be either inclusive, with the desired answer being ‘A or B or (A&B)’, or they can be ‘A or B & not(A&B), and it is the wording of the tasks themselves which makes (or fails to make) this distinction clear to participants (see, for instance, Carlson & Yates, 1989).

Errors in the above reasoning tasks are incredibly common, with 50-90% of participants committing the fallacy in conjunctive reasoning tasks (Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Tversky & Kahneman, 1983; Yates & Carlson, 1986) and similar numbers doing so in the disjunction fallacy (Carlson & Yates, 1989, found an up to 80% incidence and Fisk, 2005, found just over 60%). Reasons why these errors are made so consistently were discussed in depth in Chapter 2. To briefly summarise, these include the representativeness heuristic (Kahneman & Tversky, 1972, 1973; Moutier & Houdé, 2003) and potential surprise (Shackle, 1969; Fisk, 2002), each of which suggest that the conjunctive or disjunctive value is disproportionately affected by one or other component, the one that is most representative or most surprising, respectively. A

further theory is that people use a ‘fast and frugal’ heuristic – that is, one that used the least possible amount of time and knowledge to arrive at an answer. This would not rule out the possibility that, as part of this frugality, it is just one component (representative or surprising) that is being given more prominence. Yates and Carlson (1986) suggest a further six procedures that they found, by verbal protocol analysis, being used by their participants to solve conjunctive problems. These include overly simplified and inaccurate extensional reasoning, and attempts at calculation, as well as misinterpretation of the question they are being asked. Again, further details and evidence for and against the value of these explanations are presented in Chapter 2. The focus here, however, is the so-called ‘frequency effect’, whereby presenting tasks as natural frequencies has been found to have a facilitating effect on conjunctive and Bayesian reasoning ability (Fiedler, 1988; Gigerenzer & Hoffrage, 1995). In particular, the present study addresses the current lack of evidence regarding the effect of presenting disjunctive reasoning tasks in frequency formats.

A further aim of the current research is to examine the magnitude of the error made by participants in reasoning tasks, in order to move this field away from its current dependence on the dichotomous ‘fallacy committed’/‘no fallacy committed’ style measure. One of the limitations of the fallacy measure is that ‘no fallacy’ can inaccurately be taken to imply normative reasoning, whereas in reality participants may still be making large amounts of error, but not in the expected direction. For instance, in the disjunctive tasks a fallacy is committed if a participant estimates that the disjunction is *less* likely than either of its components. If they greatly overestimate the disjunction – for example, stating that the probability of a disjunction of two values of 0.2 is greater than .9 – then this is labelled as ‘no fallacy’, obscuring the fact that the participant has in fact made a very large reasoning error. The fallacy/no fallacy approach may also be obscuring differences in ability to discriminate between inclusive and exclusive disjunction fallacies. If participants are (as suggested by Noveck *et al.* 2002) misreading exclusive disjunction tasks – $A \text{ or } B \ \& \ \text{not}(A \ \& \ B)$ – as being inclusive this may be indicated by a greater underestimation of the disjunction, illustrating that they are thinking (implicitly or explicitly) of the formula for ‘ $A \text{ or } B \text{ or } (A\&B)$ ’.

5.1.1 Probability and Frequency Formats

Probabilistic reasoning tasks will often require participants to respond by giving their probability estimates in percentage terms. The alternative to this ‘probability format’ is

the ‘frequency format’, where participants are instead asked to think of a certain number of individuals (or events) and to give their answer as a discreet number. So instead of being given the vignette about (for instance) Linda and then asked for the percentage chance of her being a bank teller, participants would be given the vignette and asked to imagine 100 people who fit Linda’s description, then asked to estimate how many of those people would be bank tellers. Tversky and Kahneman (1983) found that the occurrence of the conjunction fallacy was reduced when the wording of the problem utilised an absolute frequency that is asking participants to imagine a number of real events or people, rather than produce a percentage probability. In this case the effect was a reduction in the incidence of the fallacy from 65% of responses to only 25%.

Fiedler (1988) found similar results across a series of studies, and concluded that while many participants in the frequency condition will continue to commit the fallacy (although to a lesser extent), this does not appear to occur as systematically as when participants are given the probability version. Those committing the fallacy in the frequency condition were very much the exception, while those committing it in the probability condition could be considered the norm. The strength of this effect was further confirmed by Fiedler’s findings that priming the participants (teaching them how to approach the problems) did not seem to reduce the effect of format.

To date, very little research has directly addressed the effect of the frequency format on disjunctive reasoning, with Costello (2009) providing one notable exception. Using inclusive disjunction tasks, Costello found that participants given the task in the probability format did commit the disjunction fallacy significantly more often than those given the same problem in a frequency format.

The purpose of this Chapter is to identify whether the fallacy does occur less often when problems are presented in a ‘frequency’ format. Both disjunctive and conjunctive tasks are presented, with the use of disjunctions being particularly noteworthy, as these are currently underrepresented in the literature. This is investigated in each case by comparing the occurrence of the fallacy. Aside from this it had been intended to examine the actual magnitude of the conjunctive and disjunctive errors. However a degree of ambiguity in the working of the problems used in the present study made this possible only in the case of exclusive disjunctions. This was done by examining the extent to which participants deviate from the normative answer, the ‘normative’ being

calculated from their own judgements of each component. An error value was calculated, by subtracting the normative disjunctive value from the participants' disjunctive estimate, with a greater error value indicating poorer reasoning. This error value may be positive, indicating an overestimate, or negative, indicating an underestimate.

For example, a participant may respond that component A has a likelihood of 0.4 and component B has a likelihood of .3. Following normative reasoning, the value of the exclusive disjunction of these two should be $A + B = .7$. If the actual response given by the participant as the exclusive disjunctive value is *less* than either of the two component values they have given, they are noted to have committed a fallacy. However, stating that they have committed the fallacy does not differentiate between a small margin of error – suggesting a disjunctive value of .35, for instance, and resulting in an error value of -.35 – and a larger one – suggesting a disjunctive value of .05, an error value of -.65. In this way the error measure being introduced here can address this previous limitation. Similarly, there may be times when the participant does not make the disjunction fallacy and does so by getting the normative answer, and therefore an error value of 0, but it is also possible that participants will deviate from the normative by over estimating the disjunctive value, for instance by suggesting a probability of .95 – an error value of .25. Again, the new measure will be able to take this into account, and instead of categorising all normative and overestimated responses as 'no fallacy' will be able to provide a subtler measure of the level of reasoning accuracy being recorded.

A between subjects design was used, in order to avoid participants seeing both formats and being primed to think in terms of probability on a frequency task, or vice versa.

5.1.2 Hypotheses

1. There will be an association between format and performance on the conjunctive and disjunctive tasks, whereby participants in the frequency condition will commit the respective fallacies less often.
2. There will also be a significant reduction in the difference between the normative and given values (on exclusive disjunctive tasks) in the frequency format.

5.2 Method

5.2.1 Design

The study was between participants, with each individual completing problems in either the frequency or the probability format. The ‘problems’ refers to probability reasoning tasks which were either conjunctive or disjunctive. The independent variable was the format of these problems, either frequency or probability. There were two dependent variables. The first was whether or not a fundamental reasoning fallacy had occurred, and the second was the magnitude of the difference between the normative and actual probability judgements for the conjunctive and disjunctive statements. However, due to possible ambiguities in the way some of the statements were framed, the magnitude of difference could not be calculated for the conjunction tasks or the inclusive disjunctions (see 5.2.3.1).

5.2.2 Participants

78 participants, all undergraduate students of Liverpool John Moores University, took part in the study. 37 of these were in the probability condition, 4 males and 33 females, with a mean age of 20.75 years (SD 2.83, minimum 19, maximum 32). Of the 41 participants in the frequency condition, 9 were male and 32 female, and the mean age was 20.53 (SD 2.57, minimum 19, maximum 28). In each condition, one participant did not report their age. All of the students were taking part in a second year research methods module and were naïve as to the issues being investigated.

5.2.3 Materials

All of the tasks within this study were completed with pen and paper.

5.2.3.1 Conjunctive problems

The participants in each condition were given two conjunctive reasoning problems. One of these was based on the Linda problem (Fisk, 2005; Tversky & Kahneman, 1983), while the second was based on the Bob problem (Fisk, 2005). For the Linda problem, those in the *probability condition* received the following vignette and statements (bold added here to emphasise the different wording):

***Linda** is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and*

social justice, and also participated in anti-nuclear demonstrations. How likely is it that:

Linda is a bank teller (cashier) and is active in the feminist movement

Linda is a bank teller (cashier)

Linda is active in the feminist movement

Participants were asked ‘how many **chances** in 100?’, responding with a number between 0 and 100.

Those in the *frequency condition* received the following:

Imagine a woman who is 31 years old, single, outspoken and very bright. At university she studied philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Now imagine that there are exactly 100 women fitting this description. Out of these 100, how many do you think:

Are bank tellers (cashiers) and are active in the feminist movement

Are bank tellers (cashiers)

Are active in the feminist movement

They were asked ‘how many in 100?’, and responded with a number between 0 and 100.

The Bob problem was similarly worded in the two formats, and is available in Appendix 1.

The probability version also included the conditional statement ‘Linda is active in the feminist movement given that she is a bank teller (cashier)’. It was intended to include an equivalent conditional statement in the frequency version, that is participants were asked to indicate ‘of those that are feminists, how many are also bank tellers (cashiers)’. However it was subsequently realised that the latter statement was deficient in two ways. First the reference class was feminists (with participants asked to estimate the number of these who were bank tellers). This was in contrast to the conditional statement in the probability version where the reference class was bank tellers with participants asked to estimate the likelihood of a teller being a feminist. Second, this latter statement, used in the frequency condition, was effectively asking the participants the same as the conjunction they had already seen – ‘are bank tellers (cashiers) and are active in the feminist movement’ since the actual numeric answer would be the same in

both cases. It was clear that future studies needed to include equivalent unambiguous conditional probability statements in both the probability and frequency formats so as to fully calculate the participants' normative value for the conjunctions, and thereby compute an error score.

For the current study, only one of the disjunctions, the exclusive disjunction, could be used to calculate the normative and error scores, as the conditional is not necessary. In this case the calculation, $P(A \text{ or } B) = P(A) + P(B)$, requires only the two components to which the participants had responded.

5.2.3.2 Disjunctive Problems

Two disjunctive problems were also used, and these were again based on ones used by Fisk (2005). These were the James problem, and the Venus Williams problem. There follows the wording for the James problem in each condition, while details for the Venus Williams problem are again available in Appendix 2.

For the probability version, participants were given:

*James is in secondary school and taking GCSE examinations. He is to sit papers for eight different subjects. His best subjects are biology and chemistry and he has consistently achieved excellent results in these. His worse subjects are mathematics and physics. He did not want to study these but his parents had insisted. While he has put a lot of work into these two subjects and his performance has improved somewhat he is not optimistic concerning his prospects in the exam. **How likely is it that James:***

Achieves a grade A in mathematics or a grade A in biology

Achieves a grade A in mathematics

Achieves a grade A in biology

Fails mathematics

Achieves a grade A in biology given that he achieves a grade A in mathematics.

Achieves a grade A in mathematics or fails mathematics

They were then asked 'how many **chances** in 100?', as in the conjunction problems.

Those in the frequency condition were given:

Imagine a boy who is in secondary school and taking GCSE examinations. He is to sit papers for eight different subjects. His best subjects are biology and chemistry and he

has consistently achieved excellent results in these. His worse subjects are mathematics and physics. He did not want to study these but his parents had insisted. While he has put a lot of work into these two subjects and his performance has improved somewhat he is not optimistic concerning his prospects in the exam. Now imagine that there are exactly 100 students fitting this description. Out of these 100, how many do you think:

Will achieve a grade A in mathematics or a grade A in biology

Will achieve a grade A in mathematics

Will achieve a grade A in biology

Will fail mathematics

Will achieve a grade A in mathematics or fail mathematics

Of those that achieve a grade A in biology, how many will achieve a grade A in mathematics?

And were then asked ‘how many in 100?’

It can be seen that there is an inclusive disjunction (it is possible for James to get an A in maths and a grade A in biology), and one that is clearly intended to be exclusive (it is not possible to get a grade A in mathematics and to also fail mathematics). The Venus Williams problem also contained both an inclusive and an exclusive statement, as shown in Table 5.1.

Each set of problems (whether probability or frequency version) were also prefaced by a front page of instructions, with participants in both conditions receiving the same wording. These instructions were followed by a completed example of the kind of problem they would be shown which was worded appropriately to each format, either probability or frequency. The instructions and examples can be found in Appendix 3.

Counterbalancing was achieved by presenting the tasks in quasi-random order, in order to avoid any proactive interference that may have been caused by presenting the different types of task within the same session. They were also asked not to return to any problems once they had been completed, and were also asked both before and after completing the booklet to ensure that they had completed every one of the tasks presented to them.

5.2.4 Procedure

The measures were administered to the students within their usual classroom environment. The students were all aware that they could choose not to participate, that such a decision would not adversely affect their chosen course at the university, and that their confidentiality was assured. They were randomly allocated to either the frequency or the probability condition as the measures were handed out, and they were not fully aware of the specific aims of the study, or that were being put into different groups.

The research conformed to the British Psychological Society's ethical principles and guidelines, with participants giving informed verbal consent, and being made aware of their rights to choose not to participate. The research was also approved by Liverpool John Moores University Research Ethics Committee.

5.3 Results

5.3.1 Conjunction Fallacy Results

Two conjunctive tasks were completed – the Linda problem, and the Bob problem (as described above). In the Linda problem, the frequency condition found a 56% occurrence of the conjunction fallacy, while the probability format did find a greater occurrence of the fallacy, at 74%. This is a more moderate effect than in previous research. Fiedler (1998), for example, found that participants given the 'probability' version were making the fallacy around 75% of the time, while those in the frequency condition showed an error rate of only 25% .

In the Bob problem, the frequency format also found a surprisingly high fallacy rate of 56%, and the probability condition produced only a slightly higher rate of 62%.

Neither of these differences proved to be statistically significant. For the Linda Scenario, χ^2 (df = 1, N=74) = 2.59, $p > .05$ and for the Bob scenario χ^2 (df = 1, N=78) = 0.30, $p > .05$.

5.3.2 Disjunction Fallacy Results

Within each of the disjunctive problems given to participants there were two disjunctive values required (one inclusive and one exclusive), giving four disjunctive judgements in total. These are set out in Table 5.1, and it can be seen that only in one case was there a

significant difference, with the frequency format actually showing significantly more fallacies than the probability format in the case of ‘is Venus Williams more likely to lose the match or break her racquet’.

Table 5.1: Percentage fallacies in disjunction tasks, by condition.

Task Type		Probability	Frequency	χ^2 (df, n)
Inclusive disjunction	James ‘A in maths or A in biology’	67.6	51.2	2.15(1, 78)
	Venus ‘lose or break racquet ’	48.6	34.1	1.69(1, 78)
Exclusive disjunction	James ‘A in maths or F in maths’	55.6	58.5	0.07(1, 77)
	Venus ‘break racquet or one racquet for whole match’	54.1	85.0	8.79(1, 77)**

* p<.05; ** p<.01; *** p<.001

5.3.3 Error Results

As detailed above, it was only possible to accurately calculate error scores in the case of the exclusive disjunctions. A MANOVA was conducted, with the between participants factor of problem format, and the two dependent variables of error on the James problem and the Venus problem. The means and standard deviations of the error scores on each are displayed in Table 5.2. The greatest level of error was in the frequency condition, with the Venus task showing error levels of around 53, greater than for the same task in the probability format, at 28.

Table 5.2: Mean and standard deviations of error scores, by task format

	Probability Format		Frequency Format		Total	
	Mean	SD	Mean	SD	Mean	SD
James	-27.44	27.05	-21.98	25.24	-24.57	26.08
Venus	-28.14	28.22	-52.80	33.09	-41.12	33.09

The MANOVA confirmed that there was a significant effect of format, at $F(2,73)=7.19$, $p=.001$, Wilks' $\lambda = 0.84$ with an effect size (partial η^2) of .17, while univariate analyses also confirmed that this was due to the difference in mean error scores in the Venus task only, $F(1, 74)=12.08$, $p<.001$, partial $\eta^2=.14$ with no effect on the James task, $F<1$, partial $\eta^2=.01$. Levene's test indicated equality of error variance in each case.

As such, neither the fallacy nor the error data suggested any facilitating effect of the frequency format.

5.4 Discussion

This study found little association between format and number of fallacies committed. That is, wording the problems as 'frequency' tasks did not reduce the occurrence of fundamental reasoning fallacies in either the conjunction or the disjunction tasks, and in fact increased the number of fallacies in one of the exclusive disjunctions. As such, there was no support for Hypothesis 1. This is in conflict with much of the research in this area which in most cases not only shows an effect of format, but also illustrates that this effect is robust despite various manipulations in the wording of the tasks and instructions (Tversky & Kahneman, 1983; Fiedler, 1988; Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Costello, 2009). Hypothesis 2, that the frequency format would lead to less error on the exclusive disjunction tasks, must also be rejected, as there was in fact *greater* error found in the frequency condition. This is a direct contradiction of findings by Sprenger and Dougherty (2006) who found that absolute accuracy was increased (i.e. error was decreased) by the frequency format.

While the majority of previous research does indicate that the frequency format would have a facilitating effect on reasoning performance, there is tentative support for the results of the present study in the literature. Epstein *et al.*, (1996) and Evans *et al.* (2000), for instance, provide support for the fact that while the frequency format can support improved reasoning, this is not always the case.

Previous studies, such as Fisk and Pidgeon (1996), Gavanski and Roskos-Ewoldsen (1991), Tversky and Kahneman (1983), Yates and Carlson (1986), Chiesi *et al.* (2008) and West *et al.* (2008), had found that 50 to 90% of participants would commit the conjunction fallacy on a traditionally presented conjunction task. In the probability format, the percentage of fallacy found here on conjunctive tasks does correspond with that, at 62% and 75%. Conversely, Fiedler (1988) found that problems worded as the frequency tasks in the current study could reduce fallacy rates to as low as 22%, and while on one task participants here made only 25% fallacies, on the other task this was as high as 56%. This suggests that the lack of effect found may be due to participants not benefitting from the frequency format to the extent that had been expected. The lack of difference is due to unexpectedly high levels of fallacy on the frequency version, and not unexpectedly low levels on the probability version. Findings are similar in terms of the disjunction fallacy, with the mean percentage of 56 in the current study being somewhat lower than would be suggested by the literature (Carlson & Yates, 1989; Fisk, 2005). The only such literature available regarding frequency versions of disjunctive tasks, Costello (2009) indicates a level of 39% for such problems, a far lower value than the 57% obtained here. Again, the participants were not unusually good at the probability version, but appeared to simply not be benefitting from the frequency format (and indeed, in one case, the latter format was associated with greater levels of both fallacy and error).

The majority of the frequency versions of the tasks in this study (and in previous studies, such as Fisk 2005) did not allow for the calculation of the normative values of conjunction in each case. As such, the responses were coded as either ‘fallacy’ or ‘no fallacy’. A richer form of data can be collected, and indeed was done so in the case of the probability format on all tasks, and the frequency format for exclusive disjunctions. This can be achieved in the conjunctive tasks by including a question requiring the participants to estimate the conditional probability (i.e., ‘Linda is active in the feminist movement *given* that she is a bank teller (cashier)’) which would allow for the calculation of the implicit normative probability that Linda is active in the feminist movement and is a bank teller. The implicit normative probability being calculated from the participant’s own judgements of how likely the relevant component and the conditional are. This then allows for a measure of the extent to which the participant’s actual estimate of the conjunctive probability deviates from the normative value.

Quantifying the participants' responses in this way would enable analysis of the amount by which they deviate from the normative, and in so doing could lead to a closer analysis of any existing differences between the way they are responding in the two conditions of probability and frequency formats. While the frequency format did not facilitate participants in avoiding fallacies in the present study, this does not examine the *extent* to which participants are over or underestimating the conjunctive/disjunctive values. By measuring the magnitude of the error, more subtle differences in performance can be examined than is made possible by the categorical fallacy measure. While it had been intended to do this in the present chapter, ambiguous or inappropriate wording of the conditional statement in the frequency format meant that this was not possible. Further limitations with the tasks used in this chapter are discussed in full in Chapter 9.

Given that the participants in this study showed similar performance to published fallacy rates on tasks in the probability format, but made greater errors in the frequency format, it is likely that the current sample were not able – or inclined – to take advantage of the presented frequency information. In order to examine the conditions in which the frequency format does lead to lower levels of reasoning fallacy, and error, it is necessary to consider the different propensities to utilise heuristic versus analytical strategies. As detailed in Chapter 3, there is evidence to support the suggestion that individuals with either thinking style – analytic or intuitive – may be likely to particularly benefit from the frequency format (see for instance Sprenger & Dougherty, 2006 and discussion of Donovan & Epstein's 1997 findings). It is possible that the thinking styles of those in the present sample were such that they were less able to take advantage of the frequency format.

Kokis *et al.*, (2002) looked at various types of reasoning in children, including probability tasks, and concluded that in this age group of 10 to 13 year olds, measures of thinking styles were important indicators, often accounting for much of the variance in their scores. However, this is under-investigated in relation to the effects of the probability and frequency formats. Consequently, research should assess whether the different cognitive styles of participants may be accounting for the differences *within* the two groups (completing either probability or frequency versions of the tasks), over and above the differences occurring *between* the groups.

Like Kokis *et al.*(2002), Epstein *et al.* (1996) had looked at individual differences in reasoning tasks, examining the processes involved in terms of a dual process theory of reasoning, whereby System 1 is a natural process reliant on heuristics, while System 2 is instead more explicit, rational and analytical (Epstein, 1992: Tversky & Kahneman, 1983). The measurement of thinking styles could enable the identification of between subjects differences in these thinking styles, and any relationship such differences may have with performance on each version (probability and frequency) of the probability tasks.

Chapter 6 – Thinking Styles, Task Format and Probabilistic

Reasoning

6.1 Introduction

In the previous chapter, the effect of problem format on probabilistic reasoning performance was investigated in a group of young (aged 19 to 32) participants, and no overall significant effect was found.

The problems used in the previous study inadvertently did not allow for the normative values of the conjunctions to be calculated and compared between the two formats (due to the ambiguous wording of many of the frequency problems) so that the data in the case of these problems was limited to being either a ‘fallacy’ or ‘not a fallacy’. In this second study, the appropriate data will be collected to allow for normative values to be calculated in every case, as detailed in Chapter 5 (discussion) and below. This will then enable the measurement of the distance between participants’ actual responses from the normative values, giving a richer data source which may reveal more detail of the processes involved.

The main focus of this second study, however, will be the consideration of the possible mediating role of thinking styles. Chapter 3 details the ‘two process’ theory of reasoning, and its relevance to the effect of format on performance in probabilistic reasoning tasks. As noted in the literature review, many theorists in this field support the view that cognitive reasoning processes can be conceptualised under two systems (such as Evans & Over, 1996; Sloman, 2002; Stanovich & West, 2000). What each of these theories have in common is that system 1 is considered a more automatic, heuristic and rapid system, while system 2 is conscious, analytic and requires both more time and greater cognitive resources. Although these two systems exist within each individual, there is also evidence that differences exist between individuals, whereby some are more likely to reason heuristically (system 1) and others to reason analytically (system 2). Evans (2007b) proposed three models that would explain how either system may be primed into processing any given problem, and in each case it seems that there is either an internal or external prompt which leads to one or other system being used. That is, either an influence from the individual or from the task is crucial in this process.

With regards to the internal prompt, Stanovich and West (2000) also stress the importance of individual differences in this area. Furthermore, measures have been developed in order to measure the individual's propensity to reason with either system (Epstein *et al.* 1996). One of the most notable is the Rational Experiential Inventory (REI – Epstein *et al.* 1996). The REI contains two scales, Cacioppo and Petty's (1982) Need for Cognition Scale, and Epstein *et al.*'s own Faith in Intuition scale (FI). The former measures a person's propensity to reason analytically – using system 2 – while the latter assesses their predisposition to use heuristics in their reasoning – system 1. Epstein *et al.*'s own results support the notion that the scale measures two reliable and discrete constructs (1996).

A second measurement tool is the Thinking Dispositions Questionnaire (TDQ) devised by Kokis *et al.* (2002). Kokis *et al.* constructed the TDQ from several subscales: Flexible Thinking, Belief Identification, Absolutism, Dogmatism, Categorical Thinking, Superstitious Thinking, Social Desirability, and a nine-item version of the Need for Cognition scale. They found that the NFC was significantly positively correlated with probabilistic reasoning ability, suggesting that it was related to the type of analytic thinking that would avoid the occurrence of the reasoning fallacies. Equally, Superstitious Thinking correlated significantly but negatively with probabilistic reasoning – again, lending support to the idea that this type of thinking style is related to the heuristic reasoning processes that lead to poor performance on probabilistic reasoning tasks. It also correlated negatively with cognitive ability, lending further support to the suggestion that heuristic reasoning may be associated with lower ability to reason analytically.

Both of these scales were used in the current research, with the aim of investigating what role the different thinking styles may play in the type of probabilistic reasoning that is being used. It is anticipated that by measuring participants' thinking styles it may be possible to identify any mediating effect that they may have in performance on the tasks in each of the two different formats. Kokis *et al.*'s findings (2002) do seem to suggest that a person's thinking style may account for much of the variance in reasoning performance, and if individuals with the heuristic style (which often does not lead to normative reasoning) do benefit greatly from either format of task, then this would indicate that the format in question may be priming analytical reasoning more

effectively than the other. Thus perhaps those with particular thinking styles might benefit more from the frequency format effect.

Where Evans (2007b) refers to a ‘superficial aspect of the problem that can cue [either system to respond]’ (p. 326), he gives the example of syllogistic tasks that have believable conclusions cueing heuristic processing. In the present research this superficial cue will be the two formats of reasoning task, the probability and frequency formats. As noted in Chapter 2, in the probability format participants will be asked to produce actual probability estimates while in the frequency format they will be asked to produce absolute frequencies. It will then be possible to establish which response format is more likely to prime the normative response, perhaps implicating analytical system 2 processes, as opposed to system 1 heuristic processes, which have been found to lead to the biases and errors as described in Chapter 1. If the richer form of data being used here does show that the frequency format facilitates normative reasoning, or even that it qualitatively changes the responses being given, then this may indicate that the frequency format is either priming the analytic system, or facilitating its function, so that the heuristic system is not being relied upon. This will be further supported if it is found that, as expected, thinking styles do mediate this particular source of variance in probabilistic reasoning performance.

This study will also measure participants’ verbal intelligence, using the Mill Hill Vocabulary Scale (MHVS) to control for underlying differences in intelligence, which might co-vary with reasoning performance, and thus mediate some of the beneficial effects of the frequency format. As it has been suggested that the wording of tasks in the frequency format makes the underlying structure of the task particularly clear (e.g. Evans, 2007a), those with poorer verbal intelligence may be less able to benefit from this function.

6.1.1 Hypotheses

It is expected that, for each type of reasoning task:

1. Those in the frequency condition will have a lower mean number of fallacies than those in the probability condition.
2. Those in the frequency condition will have a lower mean error than those in the probability condition.

3. The individual differences of thinking style and verbal intelligence will mediate the facilitating effect of the frequency condition.

6.2 Method

6.2.1 Design

There were again three types of probabilistic reasoning task – conjunctions, exclusive disjunctions and inclusive disjunctions. Each type of task yielded two scores – the number of fallacies committed (each participant having a total out of five for each type of task), and the level of absolute error. Absolute error was obtained by calculating the normative value of the conjunction or disjunction suggested by the participant's responses to each component, and then finding the difference between that 'normative' and the participant's actual response (as an absolute value, so as to avoid over and under estimates across different individual tasks from cancelling each other out). Where participants' given values for the components led to a normative value of greater than 100 (subadditivity) these were scaled down to avoid the impossibility of a normative probability equalling more than 100.

To illustrate using a case of an exclusive disjunction, component values of .6 and .7 would lead to a normative disjunctive value of 1.3 for that participant. To calculate error without the issue of subadditivity the 'normative' value for the disjunction should be subtracted from the 'given' value for the disjunction. However, as participants were guided to give values that were logically viable as probabilities this would mean that they would never be able to give an answer that was 'normative' by this definition – they would always end up with a negative level of error, indicating that they had underestimated the disjunction. To account for this, the following formula was used:
(given – normative) * (100/normative)

Supposing the participant in the current example gave a disjunctive value of .9 (not committing the fallacy) the calculation for their level of error would be:

$$(.9 - 1.3) * (1/1.3) = -4 * .77 = -.31$$

This is converted to an absolute value of 31. It is acknowledged that some variance in the data is lost through this process, and this is discussed in depth in Chapter 10.2.

The error score in each case is converted to an absolute value, in order to create a mean error score for each participant, on each task. This was necessary as one participant may over estimate one conjunctive value but underestimate the next, and the mean from such a combination of positive and negative scores would greatly under represent the actual value of error involved.

Despite this new measure, the fallacy measure was still retained for three main reasons. The first is that it allows for comparisons with previous research. The second is that the limitations of the error measure – regarding the adjustment for subadditivity and the loss of positive and negative values to allow for the creation of a meaningful mean value – are acknowledged. Finally, the continued use of the fallacy measure allows for a direct comparison of the two measures (fallacy and error) and a further discussion of their relative merits.

Subadditivity was only an issue in the case of the two types of disjunctive task, as the calculations involved in the conjunction ($P(A) \times P(B)$, or $P(A) \times P(B|A)$) for conditionally related events, mean that the normative based on the participants' own components could never exceed 1.

Each participant had a mean error score out of 100 for each type of task. As such, there were two dependent variables, each being a measure of reasoning performance.

6.2.2 Participants

There were 81 participants in total. Of these, 33 were in the probability condition, 3 males and 30 females, with a mean age of 18.88 (.89) and a range from 18 to 22. The frequency group contained 39 participants, 4 male and 35 female, with a mean age of 18.97 (1.22) and a range from 18 to 24. As in the previous study (see Chapter 4) all participants were students of Liverpool John Moores University and the experimental tasks were integrated into their coursework. For ethical considerations students were free to choose an alternative activity to fulfil coursework requirements, although none chose to do so.

As far as could be ascertained, none of the participants had ever taken part in any similar research, at this or any other institution. They were also naïve as to the issues

being studied, receiving only the information provided on their participant information sheets (Appendix 4).

6.2.3 Materials

As well as the probabilistic reasoning tasks, participants also completed the Rational Experiential Inventory and the Thinking Disposition Questionnaire (detailed below).

6.2.3.1 Mill Hill Vocabulary Scale (MHVS)

The participants completed the Multiple Choice section of the MHVS, using the 1988 revision of the Form 1 Senior version. This multiple choice section of the scale presents the participant with a target word (their own example being ‘malaria’) and then gives 6 possibilities (basement, theatre, ocean, fever, fruit, tune) out of which the participant then indicates which they feel to be the best synonym for the target word. The scale progresses in difficulty through its 33 items. The open-ended section of the scale was omitted due to the time constraints that were present in relation to testing the participants. However, for the group of participants involved (i.e. literate adults) they could be expected to gain similar results on either of the two sections of the test (Raven, Raven & Court, 1998). Raven *et al.* also report good reliability for the scales, both test-retest and split half, for a range of groups, with the vast majority being over 0.9. This included a group of “under 30s” at .97. In terms of the scale’s validity, the authors report correlations with other vocabulary scales ranging from .51 to .87.

6.2.3.2 Probabilistic reasoning tasks

The participants were presented with five disjunctive problems (each written problem including one inclusive, and one exclusive disjunction), and five conjunctive problems. In this between participants design, each individual was again seeing problems that were either all worded as probability tasks or all as frequency tasks. They were also presented within the tasks in quasi-random order, in order to address fatigue and practice effects.

The Linda task was omitted from this current study, due to problems identified in previous research, e.g. Donovan and Epstein (1997). They detail how, over and above other such tasks, it provides a particularly strong prime for the heuristic system 1 mode of reasoning, making a normatively correct answer very hard for participants to achieve.

Conjunctive problems

Five conjunctive problems were used, one of which was a variation on the Bob problem used in the previous study, and originally developed by Fisk (2005). The probability version of this problem remained unchanged, but the wording of the frequency version was altered as follows, with the bold again added to indicate the changes:

*Imagine a man who works five days a week. Most days he drives to work but occasionally he takes the bus when his wife uses the family car for shopping. He usually buys a sandwich at the local delicatessen (deli) but sometimes he eats at the local public house (pub) with his work mates. **Now imagine that that we assemble a very large number of men fitting this description.***

If we select 100 of these men at random on any given day, how many do you think take the bus to work and buy a sandwich at the deli?

If we select 100 of these men at random on any given day, how many do you think take the bus to work?

If we select 100 of these men at random on any given day, how many do you think buy a sandwich at the deli?

From this group, if we select 100 men who have all taken the bus to work on a certain day, how many of this number do you think buy a sandwich at the deli?

This new wording was intended to achieve two things. Firstly, it better framed the task so as to encourage the participants to think of a quantity (that is, a frequency) of actual people, from which a random sample can be taken and examined. Secondly, the last statement was reworded so that it was now asking for a conditional estimate that is the equivalent to the conditional statement used in the probability condition, which reads ‘Bob buys a sandwich at the deli given that he takes the bus to work.’

The remaining four problems were the Helen problem (likes nightclubs, works in a library, likes nightclubs and works in a library), the Morris problem (collects stamps, plays team sports, collects stamps and plays team sports), the Jim problem (has a maths

lesson, has games, has a maths lesson and games) and a marble problem, all of which are shown in their entirety in Appendix 5, as is the Bob problem discussed above.

Disjunctive problems

Five disjunctive problems were used in this study, one of them being a version of the James problem, as discussed in the previous chapter. As with the conjunctions, alterations were made to the frequency version of this task, as follows (again, the probability version remained unaltered, and in the task below bold is added to highlight the changes):

*Imagine a boy who is in secondary school and taking GCSE examinations. He is to sit papers for eight different subjects. His best subjects are biology and chemistry and he has consistently achieved excellent results in these. His worse subjects are mathematics and physics. He did not want to study these but his parents had insisted. While he has put a lot of work into these two subjects and his performance has improved somewhat he is not optimistic concerning his prospects in the exam. **Now imagine that we assemble a very large number of students fitting this description.***

If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics or a grade A in biology?

If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics?

If we select 100 of these students at random, how many do you think will achieve a grade A in biology?

If we select 100 of these students at random, how many do you think will fail mathematics?

From this group, if we select 100 students, all of whom achieve a grade A in mathematics, how many of this number will achieve a grade A in biology?

If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics or fail mathematics?

Again, this wording was changed to better frame the task to encourage participants to think in terms of frequencies, and also in order to provide a conditional statement equivalent to that in the probability version. The four other disjunctive problems were the Smiths (took dog to shops or played bridge), Mr F (is short sighted or has had a heart attack), a marbles problem and the Bill problem from Fisk (2005; Bill plays jazz or is an accountant). These are presented in full in Appendix 6.

As illustrated by the example above, each disjunctive task now required participants to provide judgements for three components (e.g., grade A in maths, grade A in biology, fail mathematics), one conditional event (for a sample of those achieving a grade A in maths, how many will achieve a grade A in biology) and two disjunctives. One of these can be defined as exclusive ‘grade A in mathematics or fail mathematics’ in that they cannot possibly do both, while the other was the inclusive ‘grade A in maths or grade A in biology’ as it would be quite possible to achieve both. Since each disjunctive task solicited both an inclusive and an exclusive disjunctive judgement, the five problems resulted in ten disjunctive judgments in total, and no prediction was made at this stage with regards to performance on these two types of disjunction. Due to the suggestion by Markus and Zajonc (1985) that conjunction errors are often due to participants interpreting the component of A as being ‘A but not B’, Carlson and Yates (1989) investigated the possibility that participants were making similar mistakes in disjunction tasks. They found that, contrary to the suggestion that participants failed to understand ‘A or B’ as being an inclusive disjunction, participants responded no differently when given the more explicit ‘A or B, or both’ form of a task. Noveck, Chierchia, Chevaux, Guelminger and Sylvestre (2002) found that people are actually more likely to assume a disjunction is inclusive, unless the language and/or context makes it clear that the disjunction should be considered exclusive. As such, the wording of the current tasks is designed to make it implicitly clear to participants which disjunctions may be inclusive or exclusive. As Roberge found that participants find exclusive disjunctions easier to work with than inclusive ones, it is to be expected that the same might be found here, although Roberge was working with slightly different materials, and was in fact looking at propositional, rather than probabilistic reasoning (Roberge, 1976).

The participants were also given a front page of instructions as detailed in the previous study, explaining how to fill in their answers, and that they should be sure to complete

all of the problems, without returning to any of them once they had been completed. They were also shown similar examples to those used in Chapter 4, but the frequency version was again altered to be in keeping with the new frequency versions being used in this study, so that the participants were, at the end of the vignette, asked to ‘imagine that a very large number of men fitting this description will be running a mile race this weekend’. Each statement was then prefaced by the guidance ‘If we select 100 of these men at random, how many do you think. . .’ (see Appendix 7).

6.2.3.3 Thinking Styles

Rational Experiential Inventory

The REI is a composite of two scales – the Need for Cognition Scale (Cacioppo & Petty, 1982) and the Faith In Intuition Scale (Epstein, *et al.*, 1996). The REI as a whole was devised and evaluated by Epstein *et al.* (1996), and uses a short form of the NFC, consisting of 19 of the scale’s original 45 items. They require responses on a five point scale from ‘1 = extremely uncharacteristic’ to ‘5 = extremely characteristic’, with 14 items being reverse scored.

Sample items from the NFC include ‘I would prefer complex to simple problems’ and ‘I don’t like to have the responsibility of handling a situation that requires a lot of thinking’, the latter being a reverse scored item. The minimum possible score on this scale is 19 – 19 responses scoring 1 each – while the maximum is 95 – 19 responses scoring 5. As such, a high score indicates a high need for cognition, while a lower score indicates a relatively lower need for cognition.

The FI scale contains 12 items, including statements such as ‘I believe in trusting my hunches’ and ‘I can typically sense right away when a person is lying’. None of the items within this subscale were reverse scored, and they were all scored on the five point scale as detailed above, leading to a minimum score of 12, and a maximum of 60.

Items from the two scales were combined (a total of 31 statements) and randomised before being presented to the participants. Epstein *et al.* (1996) report that the NFC scale (in this form) has an internal reliability of .87, while the FI’s is only slightly lower at .77. Their factor analysis also suggests that each scale is a truly independent construct, with all items loading more highly on their respective factors. The current

study's internal consistency values for these and the Thinking Disposition Questionnaire subscales, and the scales as presented to the participants, are available in Appendices 8 and 9 respectively.

Thinking Disposition Questionnaire (Kokis *et al.* in 2002)

The TDQ is a 53 item scale, consisting of a number of subscales. Each item is scored on a four-point scale from 1, 'strongly agree' to 4, 'strongly disagree'. Each of the subscales is summarised below, and in each case the lowest possible score would be a value of 1 for each question, while the highest possible would be 4. Also for each scale in this questionnaire, a *low* score indicates a *strong* presence of that disposition. In previous studies (Kokis *et al.*, 2002; Stanovich & West 2007) five of the following subscales – Flexible Thinking, Belief Identification, Absolutism, Dogmatism and Categorical Thinking – were combined to create an Actively Open-minded Thinking scale (AOT). In the current study, it was felt that using the separate items held greater utility.

Flexible thinking scale (FT) (Based on a scale by Stanovich & West 1997).

Containing 10 items (5 reverse scored) this is the longest of the subscales. It contains items such as 'A person should always consider new possibilities' and 'It's OK to be undecided about some things'.

Belief Identification (BI) (Originally developed by Sá, West & Stanovich, 1999).

This component contains 6 items (2 reversed) including 'It's fantastic when someone famous believes in the same things as me' and 'I never change what I believe in – even when someone shows me that my beliefs are wrong'.

Absolutism (A) (Erwin, 1983)

Five items (1 reversed) define this aspect including 'Right and wrong never change'.

Dogmatism (D) (Troidahl & Powell, 1965).

Six items reflect the concept of dogmatism, none of which are reversed. Items include 'I really hate some people because of the things they stand for.'

Categorical Thinking (CT) (Sa *et al.*, 1999).

Three items assess this aspect with a low score indicating a strong tendency towards a 'black and white' view of the world, with no reversed items. For instance, 'There are basically two kinds of people in this world, good and bad'.

Superstitious Thinking/Luck Composite (ST) (Based on a scale by Stanovich & West, 1997).

Containing 8 items, with one reverse scored, this subscale contains items that refer directly to superstitions 'the number 13 is unlucky', and items that refer to luck in general 'I don't believe in luck'.

Need for Cognition Scale (sfNFC) (A short form of the scale by Cacioppo *et al.*, 1996)

This shortened NFC scale contains just 9 items, four of which are reverse scored.

Social Desirability Response Rate Bias (SD) (Paulhus, 1991).

There were 5 items assessing social desirability, 3 reverse scored, including statements such as 'I sometimes tell lies if I have to' (a reversed item). Taken from the Balanced Inventory of Desirable Responding (Paulhus, 1991).

6.2.4 Procedure

Materials were presented to the participants in two separate sessions. In the first, they completed the thinking style questionnaires and MHVS, and in the second they completed the probabilistic reasoning tasks.

As detailed in the previous chapter, all data collection took place within Liverpool John Moores' School of Psychology, during timetabled lesson times. Again, all of the British Psychological Society's ethical principles and guidelines were adhered to and the study was approved by Liverpool John Moores University's Research Ethics Committee.

6.2.4.1 Analyses of Variance

Two separate analyses of variance were conducted. Each involved the within participants factor of task type, which had three levels, these being conjunction, exclusive disjunction and inclusive disjunction. The between participants factor was problem format with two levels – probability or frequency. The inclusion of the within participants factor made it possible to establish whether or not the magnitude of the

error was similar across the different problem types and whether the frequency format would benefit reasoning performance in one problem type relative to another.

NFC and FI scales as well as the MHVS scores were then added as covariates. One of these two analyses used fallacy as the dependent variable, and one used error.

6.2.4.2 Regression analyses

With the above analyses of variance establishing that task format did have a significant effect on the number of fallacies as well as the mean error, regression analyses were designed to investigate whether individual differences, as measured by the MHVS and TDQ, might account for significant levels of variance in reasoning performance.

Outcome variables in each case were one of the six measures of performance, these being the level of fallacy on each of the three tasks and the level of error on each of the three. For each measure, three analyses were conducted. In the first, the scales of the TDQ were entered as step 1, with format as step 2. In the second, the MHVS score was step 1, with format again as step 2. In the third, both the TDQ scales and the MHVS score were entered as step 1, in order to ascertain whether the combined measures accounted for significant variance before the addition of format, as step 2. The number of potential independent variables comprising the TDQ favoured a regression approach as opposed to the ANCOVA design that was utilised for the NFC and FI scales.

6.3 Results

6.3.1 Reasoning Performance

While a high score on the REI subscales indicates a high FI or NFC, the TDQ subscales are scored in the opposite direction, so that a low score indicates a strong level of that disposition. Scores are displayed in Table 6.1, on the following page, and the weak levels of Social Desirability indicate that such a response bias was not an issue in this sample.

Table 6.1: Thinking style and vocabulary scale descriptive statistics

Scale	Subscale	Mean (SD)	Min/max possible scores
REI – Rational Experiential Inventory	FI – Faith in Intuition	42.83 (6.14)	12/60
	NFC – Need for Cognition	61.46 (11.41)	19/95
TDQ – Thinking Dispositions Questionnaire	FT – Flexible thinking	18.46 (2.96)	10/40
	A - Absolutism	14.91 (1.66)	5/20
	D - Dogmatism	14.34 (2.26)	6/24
	CT – Categorical thinking	9.23 (1.51)	3/12
	ST/LC– Superstitious Thinking/Luck Composite	22.04 (4.65)	8/32
	sfNFC – Short form Need for Cognition	20.45 (3.72)	9/36
	SD – Social Desirability	14.66 (2.25)	5/20
	BI – Belief Identification	16.05 (2.03)	6/24
MHVS		15.51 (2.61)	0/33

6.3.2 Analyses of Variance

In each of the three tasks, the mean number of fallacies made was greater in the probability condition, and Table 6.2 also illustrates that it was the exclusive disjunction task which led to the least fallacies being made in either condition. In each case, the maximum number of fallacies that participants could have made is five.

Table 6.2: Table of means and standard deviations for number of fallacies in each group, by task

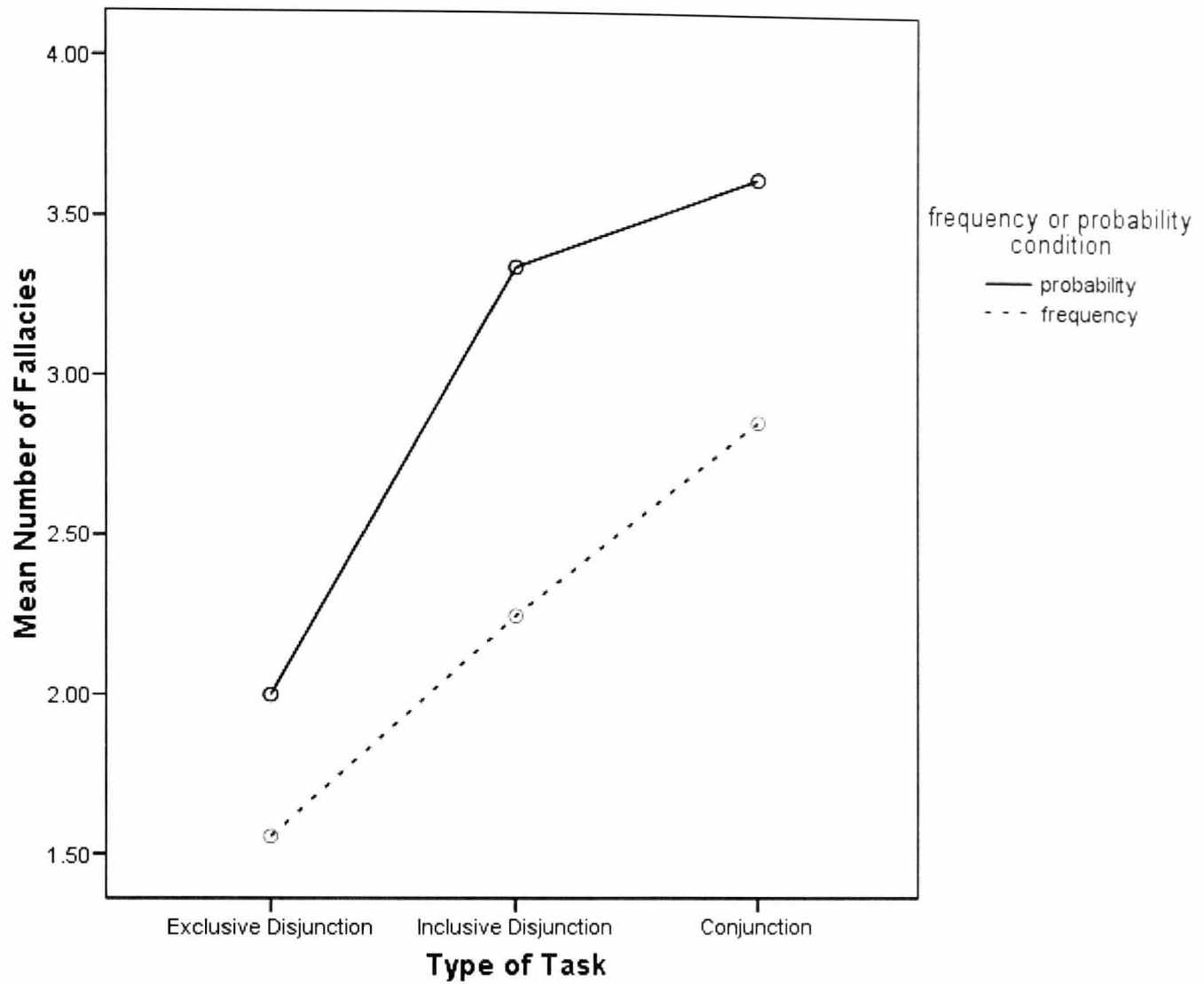
	Probability format	Frequency format	All Data
	Mean (SD)	Mean (SD)	Mean (SD)
Conjunction fallacies	3.63 (1.21)	2.86 (1.44)	3.22 (1.38)
Exclusive disjunction fallacies	2.00 (0.95)	1.56 (1.68)	1.76 (1.39)
Inclusive disjunction fallacies	3.34 (1.18)	2.25 (1.42)	2.76 (1.42)

An ANOVA confirmed that there was both an effect of task type, $F(2,65)=22.21$, $p<.001$, with an effect size (partial η^2) of .41 and an effect of format, $F(1,66)=11.92$, $p<.01$, partial $\eta^2= .15$.

Bonferroni pairwise comparisons revealed that while the exclusive disjunction task differed significantly from both the inclusive disjunctive and conjunctive tasks ($p<.001$), the inclusive disjunctions and conjunctions did not differ significantly from each other, in terms of the number of fallacies occurring ($p>.05$).

There was no interaction effect between the two factors ($F(2,65)=1.56$, $p>.05$, partial $\eta^2= .05$) indicating that the effect of format was no larger or smaller for any one type of task. The above findings are further illustrated by Figure 6.1.

Figure 6.1 Mean fallacies committed on each task



When Need For Cognition, Faith in Intuition and the MHVS scores were entered as covariates, the effect of format was barely attenuated, remaining significant at $F(1,62)=10.94, p<.01, \text{partial } \eta^2= .15$. Each of the covariates themselves were non-significant, $F<1$ in each case, and none violated the assumption of homogeneity of regression. ($p>.05$ for the group by covariate interaction in all cases).

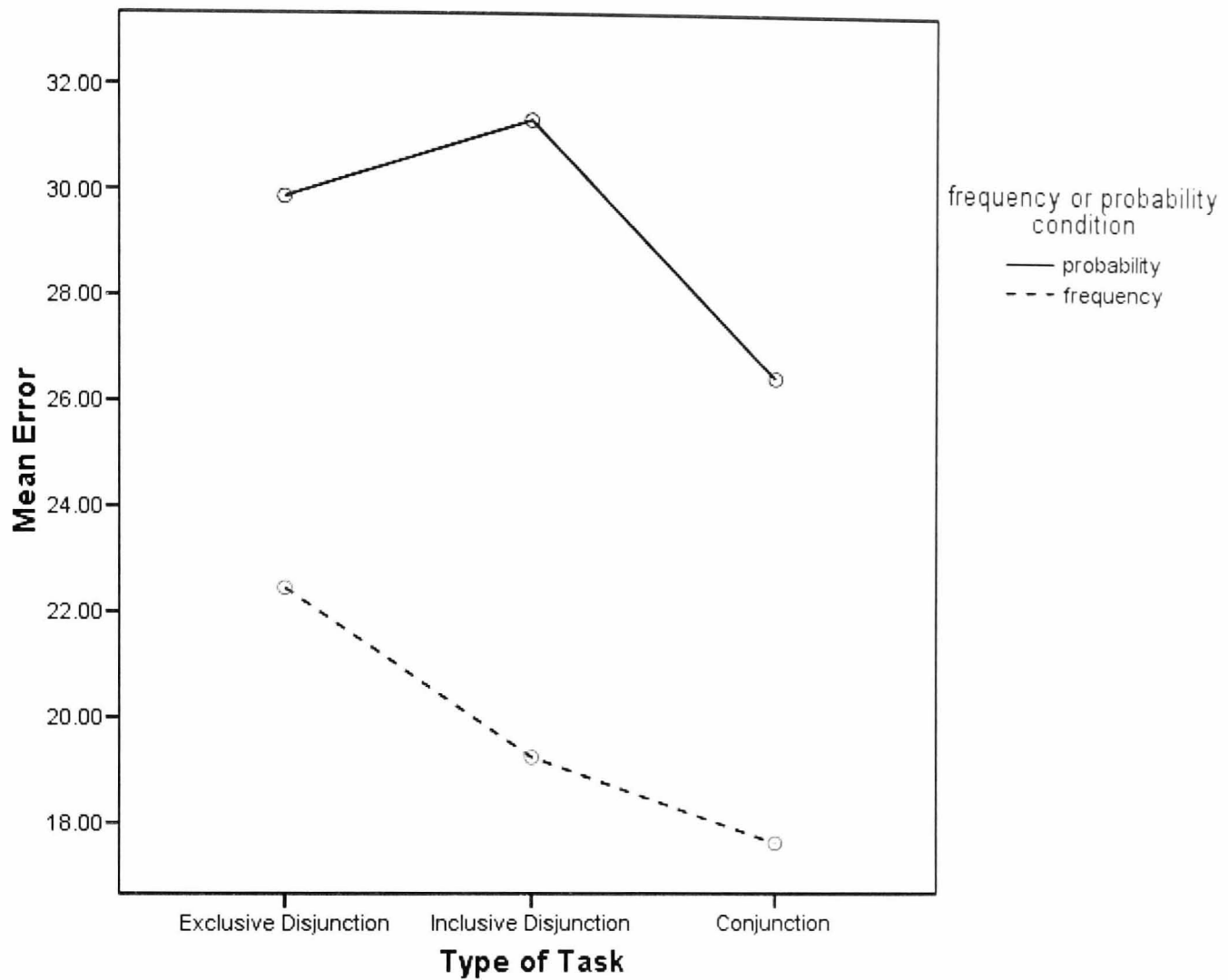
Table 6.3: Table of means/SDs for mean error in each group, by each task

	Probability format	Frequency format	All data
	Mean (SD)	Mean (SD)	Mean (SD)
Conjunction error	26.47 (6.25)	18.05 (9.00)	22.01 (8.85)
Exclusive disjunction error	29.85 (10.06)	22.86 (13.99)	26.15 (12.70)
Inclusive disjunction error	31.34 (9.62)	19.80 (12.32)	25.23 (12.48)

As in the fallacy data, levels of error were consistently higher in the probability condition. However, in contrast with the number of fallacies, where fewest fallacies were found in the exclusive disjunction task, the conjunctive task showed the least error, while the exclusive disjunction task actually showed the most.

ANOVA again indicated that although the effect of task type was smaller in this case, it was still significant, at $F(2,65)=4.16$, $p<.05$, partial $\eta^2= .11$, albeit in a different direction from the analyses above. The effect of format was larger ($F(1,66)=21.64$, $p<.001$, partial $\eta^2= .25$) while there continued to be no significant interaction ($F(2,65)=1.39$, $p>.05$, partial $\eta^2= .04$). Bonferroni pairwise comparisons in this case indicated that the conjunctive tasks differed significantly from each of the disjunctive tasks, at $p<.05$, but the disjunctive tasks did not differ from each other.

Figure 6.2 Mean error on each task



When Need For Cognition, Faith in Intuition and the MHVS scores were entered as covariates, the effect of format was again only slightly attenuated, and remained highly significant at ($F(1,62)=20.81$ $p<.001$, partial $\eta^2= .25$). None of these covariates were significant, with no F value exceeding 1.5, and none violated the assumption of homogeneity of regression.

The values presented above are all absolute error values, as detailed in the design of this study (section 6.2.2). It should be noted at this point that the individual error means for each conjunctive task were positive values, indicating that the general trend was to overestimate the conjunctive value, while in both types of disjunctive task the individual error means were negative, indicating an underestimate of each kind of disjunction. The

implications of this with regards to the disjunctive tasks are discussed in 6.4, and the individual means for this and the following study are presented in Appendix 10.

6.3.3 Hierarchical Regression Analyses

Having established an effect of format in the data, the purpose of the regression analyses was to investigate whether any further variance in performance could be accounted for by the variance in the measures of thinking disposition (TDQ), as well as by the variance in the Mill Hill Vocabulary Scale scores, with the latter expected to show a positive relationship with reasoning ability. Most important, the extent to which the effect of format was mediated by the thinking dispositions measures could be established through examining the incremental variance accounted for by format when entered last in the model after the TDQ subscales.

To achieve this three analyses were conducted for each of the six outcome variables (conjunction fallacy, inclusive disjunction fallacy, exclusive disjunction fallacy and the corresponding three error measures) as detailed below.

Fallacies:

- Step 1 MHVS, step 2 format
- Step 1 TDQ subscales, step 2 format
- Step 1 TDQ and MHVS, step 2 format

6.3.3.1 Conjunction Fallacy Regression

In the case of the conjunction tasks, neither the MHVS, nor the TDQ subscales nor the two in combination accounted for significant levels of variance in the number of fallacies committed, with R squared change at less than .1 in each case ($p > .05$). The task format accounted for significant levels of variance in each of the three analyses (although the whole model ceases to be significant with the inclusion of the TDQ subscales), and these results are summarised in Table 6.4, with full beta weights for each model in the current and each of the following analyses all being presented in Appendix 11.

Table 6.4: Model summaries for regression analyses of conjunction fallacy data

Predictors	R Squared	R squared change	F
MHVS	.00	.00	
MHVS, Format	.10	.10**	(2, 65)=3.63*
TDQ subscales	.10	.10	
TDQ subscales, Format	.20	.10**	(9,58)=1.58
MHVS, TDQ subscales	.10	.10	
MHVS, TDQ subscales, Format	.20	.10**	(10, 57)=1.42

* p<.05; ** p<.01; *** p<.001

6.3.3.2 Exclusive Disjunction Fallacy Regression

The exclusive disjunction fallacy data showed that while the MHVS and TDQ again failed to account for significant variance in performance, Format also failed to show significance in these current analyses, accounting for negligible levels of variance when entered after either or both of the two other predictors. This may reflect the smaller difference between the means in this data, as illustrated in Figure 6.1.

Table 6.5: Model summaries for regression analyses of exclusive disjunction fallacy data

Predictors	R Squared	R squared change	F
MHVS	.01	.01	
MHVS, Format	.03	.02	(2, 65)=1.93
TDQ subscales	.19	.19	
TDQ subscales, Format	.19	.00	(9, 58)=2.85
MHVS, TDQ subscales	.19	.19	
MHVS, TDQ subscales, Format	.20	.00	(10, 57)=1.34

* p<.05; ** p<.01; *** p<.001

6.3.3.3 Inclusive Disjunction Fallacy Regression

In this final fallacy data analysis, the effect of format was highly significant, accounting for 18% of variance after the MHVS only was entered. Neither the MHVS nor the TDQ subscales were significant in any of the three analyses.

Table 6.6: Model summaries for regression analyses of inclusive disjunction fallacy data

Predictors	R Squared	R squared change	F
MHVS	.00	.00	
MHVS, Format	.18	.18***	(2, 65)=7.21**
TDQ subscales	.21	.21	
TDQ subscales, Format	.30	.09**	(9, 58)=2.71*
MHVS, TDQ subscales	.21	.21	
MHVS, TDQ subscales, Format	.30	.09**	(10, 57)=2.40*

* p<.05; ** p<.01; *** p<.001

6.3.3.4 Conjunction Error Regression

The error data from the tasks was analysed in the same way as the fallacy data, again using hierarchical regression analyses to investigate the amount of variance that was accounted for by the MHVS, the TDQ subscales (separately and together) and the further unique amounts accounted for by task format.

When compared with the analysis of the conjunction fallacy data, task format accounts for far greater variance in each mode, leading to higher significance in each case. The MHVS and TDQ remain non significant as predictors.

Table 6.7: Model summaries for regression analyses of conjunction error data

Predictors	R Squared	R squared change	F
MHVS	.06	.06	
MHVS, Format	.32	.26***	(2,65)=15.02***
TDQ subscales	.16	.16	
TDQ subscales, Format	.41	.25***	(9, 58)=4.52***
MHVS, TDQ subscales	.18	.18	
MHVS, TDQ subscales, Format	.43	.26***	(10, 57)=4.38***

* p<.05; ** p<.01; *** p<.001

6.3.3.5 Exclusive Disjunction Error Regression

The exclusive disjunction error data showed some difference from the corresponding fallacy data. In the error data, the format of task accounted for a small but significant level of variance when added after the MHVS, at 7%, $p < .05$, while in the fallacy data this was a non-significant 2%. When entered first, the TDQ subscales accounted for a significant 23% of the variance in scores, with the format failing to account for more than an additional 1%. Beta weights (available in full in Appendix 11) revealed that the most relevant subscale of the TDQ was the Absolutism scale, at $\beta = -.24$, $t(58) = -2.03$, $p < .05$. This negative relationship indicated that greater levels of error were associated with stronger levels of Absolutism.

Table 6.8: Model summaries for regression analyses of exclusive disjunction error data

Predictors	R Squared	R squared change	F
MHVS	.02	.02	
MHVS, Format	.08	.07*	(2, 65)=2.97
TDQ subscales	.23	.23*	
TDQ subscales, Format	.25	.01	(9, 58)=2.11*
MHVS, TDQ subscales	.24	.24	
MHVS, TDQ subscales, Format	.25	.01	(10, 57)=1.94

* $p < .05$; ** $p < .01$; *** $p < .001$

6.3.3.6 Inclusive Disjunction Error Regression

Again, when compared with the fallacy measure on these tasks, the error data here showed that a greater level of variance was accounted for by task format in each model, while the TDQ and MHVS scales remained non-significant as predictors.

Table 6.9: Model summaries for regression analyses of inclusive disjunction error data

Predictors	R Squared	R squared change	F
MHVS	.00	.00	
MHVS, Format	.25	.25***	(2, 65)=10.73***
TDQ subscales	.18	.18	
TDQ subscales, Format	.33	.15***	(9, 58)=3.18**
MHVS, TDQ subscales	.18	.18	
MHVS, TDQ subscales, Format	.33	.16***	(10, 57)=.286**

* $p < .05$; ** $p < .01$; *** $p < .001$

6.4 Discussion

Unlike the findings of the previous chapter, where no consistent effect of format was found, analyses of variance on the current data showed a consistently significant difference between the two formats of the task, with the frequency format appearing to facilitate the avoidance of the fallacy in all three tasks. The same was found with the error data – significantly lower levels of error were made on the frequency versions of the tasks. They also indicated an effect of task type, whereby the conjunction led to the most fallacies but the least error, and the exclusive disjunction led to the *least* fallacies but the *greatest* error, with the inclusive disjunction being between the two in each case. No variance was accounted for by either of the thinking styles of the REI – the NFC and the FI – or the Mill Hill Vocabulary Scale on either fallacies or mean errors.

Regression analyses also showed the significant relationship between task format and reasoning performance, but showed only limited support for the theory that thinking dispositions would account for variance in reasoning performance, with such a relationship being found only in the case of the exclusive disjunctive errors. No variance in performance was accounted for by verbal intelligence, as measured by the MHVS, in any of the tasks. In relation to the hypotheses, both 1, that those in the frequency condition will have a lower mean error than those in the probability condition, and 2, that those in the frequency condition will commit significantly fewer fallacies, have been strongly supported by the analyses of variance, as well as by the regression analyses.

There are some interesting findings regarding the effect of task type. While both exclusive and inclusive disjunctions are consistently underestimated (each task having a negative mean error value) the inclusive tasks are done so more frequently, as indicated by greater numbers of fallacies. However, although the exclusive tasks did show greater levels of error (representing the magnitude of underestimate), this was not a significant difference.

This fallacy data would concur with the findings of Roberge (1976) who suggested that exclusive tasks were more easily solved than inclusive ones, on the basis of nominal data similar to the ‘fallacy’/‘no fallacy’ approach. Roberge gave participants tasks that gave disjunctive premises (‘either there is a P or there is a Q (or both)’ or ‘either there is a P or there is a Q (but not both’)), then a further premise of ‘there is a Q’ or ‘there is a

P'. The task was then to assess a third premise, 'there is a Q', based on the preceding two. Participants could answer yes (that the third premise is true), no (the premise is false) or maybe (indicating the participant's indecision). Participants were more likely to correctly accept or reject the third premise in the exclusive (P or Q *but not both*) tasks.

Noveck *et al.* (2002) found that their participants were more likely to assume that a disjunction was inclusive (included the chance of A occurring, of B occurring, and also the chance of A and B occurring) if exclusivity was not made explicit, in that a greater amount of error was found on the exclusive than the inclusive tasks. In terms of fallacy, the current study shows the opposite, as more mistakes are being made on the exclusive tasks. The error data here also reveals an interesting picture. To reiterate, for inclusive tasks the value of the disjunction is $pA + pB - p(A+B)$, while for the exclusive tasks the disjunction is calculated simply by $pA + pB$.

With more inclusive disjunctive fallacies, but no greater magnitude of inclusive disjunctive error, this is consistent with the proposition that each inclusive fallacy is represented by a smaller amount of error. It could also be the case that the two types of fallacy (inclusive and exclusive) are associated with similar levels of error, but that any non-fallacy responses (which are either normative or overestimates by definition) are associated with *larger* levels of error for the exclusive disjunctions.

It could be that on the exclusive tasks – showing a larger amount of error for each fallacy – those committing the fallacy are doing so by failing to (implicitly or explicitly) appreciate that in calculating their exclusive disjunctive estimate, the conjunctive value of the components is zero, as an impossibility, and should not be deducted from the disjunctive value. This would suggest that those who are reasoning incorrectly on these exclusive tasks, and committing the fallacy, are incorrectly scaling down their disjunctive values, as if they were inclusive tasks. That participants are doing this consciously and deliberately seems very unlikely – it would be a gross misunderstanding to believe that, for instance, the protagonist in a scenario could get both an A and fail on a single exam. In either case, the participants appear to be failing to adequately appreciate the differing demands of the two types of disjunction.

Subadditivity was present in each of the disjunctive tasks, but far more so in the exclusive tasks. Where subadditivity occurs, the transformation used sets the normative value of the disjunction to 1, and scales down the participant's actual disjunction estimate accordingly (see 6.2.2). In the case of the inclusive tasks, where subadditivity rarely occurs, the normative value is almost invariably less than 1, and the participant's actual estimate is not scaled down.

As anticipated, the error measure has shed greater light on the accuracy of participants' responses. Not just in terms of the gross underestimation of the exclusive disjunctive values, but also in highlighting a clear effect of format that is sometimes not found when using only the fallacy measure (e.g. in the exclusive disjunctive tasks).

Hypothesis 3, however, that thinking style and verbal intelligence would mediate the facilitating effect of the frequency condition has received only very limited support, with only Absolutism showing such a relationship, and only in the case of the exclusive disjunction error.

With regard to the two process theories of reasoning (Evans & Over, 1996; Stanovich & West, 2000; Sloman 2002), the lack of relationship between thinking style and reasoning performance may in some part be due to the narrow range of scores on these measures. Table 6.1 illustrates that the standard deviations of the scores were all quite low, suggesting little variation in thinking style across the sample. A population which varies more greatly in this regard may be a more effective way of assessing whether thinking styles do relate to reasoning performance.

In conclusion, the clear effect of problem format found in this study may be argued to be priming the analytical system within participants, but the lack of clear evidence from the thinking style data does not indicate that individuals with a tendency towards analytic thinking in general are likely to perform significantly better on the tasks presented than those with no such tendency, or even a predisposition to reason intuitively. As mentioned above, the sample used in this study (as in the previous one) was somewhat limited in age range, as well as with regard to thinking style scores, which were somewhat clustered towards the upper end of the scale. This leads the research to look to a different population, one which may be better suited to illustrate any existing relationship between thinking style and reasoning ability. As discussed in

Chapter 4, an older population may show a different pattern of thinking dispositions and be more inclined to reason heuristically. For example Fisk (2005) has suggested that older persons may be more likely to deploy heuristic strategies as a consequence of their limited working memory capacity. Thus it might be easier to detect the effects of thinking styles and their possible mediation of the format effect in a more age-heterogeneous population. Furthermore the extension of the present analytical approach to the cognitive ageing context would have utility in its own right in so far as it would extend and further develop Fisk's (2005) results and in doing so potentially shed more light on the relative contributions of heuristic and analytical strategies to probabilistic reasoning more generally. This was the basis of the next chapter.

Chapter 7 – Age, Thinking Styles, Problem Format and Reasoning

7.1 Introduction

In the preceding chapter, the effects of problem format (frequency versus probability) and thinking style on reasoning performance were examined for different types of problems (conjunctive, exclusive disjunctive and inclusive disjunctive). There was a consistent effect of problem format, with both fewer fallacies and lower mean error being found when tasks were worded as frequencies, illustrating the occurrence of the frequency effect, and supporting the findings of Fiedler (1988) and Costello (2009). There was also an effect of task type as measured by both fallacy (exclusive disjunctions leading to significantly fewer than either inclusive disjunctions or conclusions) and by error (conjunctive tasks leading to less error than either type of disjunction). The hypothesis that thinking style and verbal intelligence would mediate the effect of format upon reasoning performance was not supported.

The sample used to collect the data so far has been somewhat homogenous, with very little variance evident in participants' thinking styles and verbal intelligence. This current study will compare data from two age groups – young (<30 years) and old (>60 years). As discussed in Chapter 4, and summarised below, it is to be expected that older participants may reason differently from their younger counterparts, due in part to cognitive limitations associated with the ageing process which may lead to greater reliance on the heuristic system 1, rather than the analytical system 2.

7.1.1 Age

It is well established that a certain amount of cognitive decline is a usual part of the human ageing process, with particular declines in working memory performance and information processing speed, both of which may have progressively detrimental effect on higher order factors and processes (Salthouse 1998; Salthouse & Babcock, 1991). Salthouse (2005) concludes that, while an effect of age has been found in many reasoning tasks, 'there is still no convincing explanation of the causes of age-related effects on reasoning' (p. 604). While working memory decline may account for some of the reasoning deficits in older individuals, Salthouse suggests that information processing speed is also an important factor. As discussed in Chapter 4.1.2, Salthouse (1996; 2000) hypothesised that slower processing speeds affect older participants' performance on time constrained tasks as they simply run out of time (the limited time

mechanism). On those tasks without time constraints performance is still affected as slower processing means that the earlier stages of processing are forgotten or corrupted as the individual moves onto late stages (the simultaneity mechanism). With age related slowing of information processing speed mediating the detrimental effect of age on analogical and syllogistic reasoning (Clark *et al.*, 1990 and Fisk & Sharp, 2002, respectively) it is reasonable to expect a similar effect on probabilistic reasoning, and this current study will measure information processing speed in order to control for the expected age related decline.

As discussed in Chapter 3, Chasseigne *et al.* (1997, 2004), Mutter (2000), Mutter, *et al.* (2006) and Mutter and Williams (2004) have all found that age related reasoning deficits often only become apparent when the tasks are particularly complex, for instance those that involve learning inverse cue relationships, or distracting tasks, with no age effect found on more simple tasks. Gilinsky and Judd (1994) and Peters *et al.* (2000) have also found that older people are more likely to be susceptible to biases, such as 'belief bias' in tasks such as syllogisms. Again, this is an indication that age effects may only be shown when tasks are more cognitively demanding.

With regards to probabilistic reasoning, Fisk (2005) found no significant effect of age in conjunctive, disjunctive and Bayesian tasks. However, he did find that many measures of cognitive function correlated with reasoning ability in a younger group, while there were no such relationships in the older group. This would indicate that the older participants may be using different processes (ones which are less dependent on cognitive abilities), while the lack of age effect suggests that these different processes are just as effective as those being used by the younger group. As previously discussed, Fisk (2005) was looking primarily at incidences of the reasoning fallacies associated with each task, and did not measure how far participants' responses were from the 'correct', or normative response. As seen in the previous chapter, participants were making significantly more inclusive disjunction fallacies than exclusive ones, but were making slightly larger amounts of error on the latter. Again, this current study will collect both types of data in order to be able to see any otherwise obscured effects of age on performance.

7.1.2 Are the reasoning processes of older adults qualitatively different?

Consistent with Fisk's (2005) findings that older people's ability to reason does not appear to be related to many of the cognitive measures that relate to younger people's reasoning performance, differential use of working memory capacity by younger and older adults was demonstrated by Chen and Sun (2003) in their yard sale decision making vignette. In this task, younger participants chose strategies that did load on their working memory, while older participants did not. However, as with Fisk's probability reasoning tasks, the old group were performing just as well on the task as their younger counterparts. Chen and Sun (2003) suggest that this could be due to what Baltes (1997) terms Selective Optimization with Compensation, whereby, throughout the ageing process, we perform to our full potential by compensating for our limits, as well as by benefiting from prior experience. In terms of dual process theory, this would suggest that older people move away from the analytical reasoning that characterises system 2, and towards the more intuitive and heuristic system 1. This may or may not be a conscious decision. This movement from analytical to heuristic reasoning as part of the ageing process would also have theoretical support from the idea that the analytic system is more time consuming, while the heuristic system is more rapid. As older participants show longer information processing speeds they would find the analytic processes increasingly time consuming, subject to the forgetting and/or corruption of information proposed by the simultaneity mechanism (Salthouse, 1996, 2000). Johnson (1993) has also found that younger people may be making more use of all of the information presented in a task, as they appeared to be rechecking the information provided more frequently than their older counterparts. Again, this difference in strategy was not connected to any significant difference between the two age groups' reasoning ability.

Lastly, in a discussion of the importance of emotion in problem solving, Blanchard-Fields (1996) identifies more qualitative differences in the way that older people approach reasoning tasks, with their reasoning on concrete tasks being more often guided by social schemas than their younger counterparts. Blanchard-Fields proposes a sort of maturation of the reasoning process, which again suggests that older people become more reliant on experiences, and less so on analysis, but that their reasoning performance will not necessarily suffer as a result of this.

Fontaine and Pennequin (2000) suggest that some of the apparently contradictory findings in the literature in this area may be due to a lack of measuring and controlling for levels of education, and Saczynski *et al.* (2002) also suggest that better educated participants may be more able to learn and utilise reasoning strategies than those with a lower educational level. As such, a self reported measure of education duration was used in this study to enable any such differences to be controlled for. Furthermore, the measure of mean error, assessing the magnitude of reasoning error, rather than the dichotomous ‘fallacy/no fallacy’ will again allow for a more sensitive measure of reasoning ability. With previous research having found no ageing affect when measured by fallacy (Fisk, 2005) it is anticipated that the effects will be revealed by the mean error data, with older participants showing greater over and under estimates of conjunctive and disjunctive values respectively, when compared with a younger cohort.

7.1.3 Hypotheses

Five hypotheses were tested in this current study.

1. It was predicted that the effect of problem format would again be found in this sample, with the frequency task continuing to be associated with fewer fallacies and lower levels of error.
2. Older participants will show significantly greater levels of error than their younger counterparts.
3. Older people will show significantly lower levels of analytical thinking than the younger group.
4. An interaction effect was therefore predicted, whereby older participants would be differentially affected by the frequency format. It is anticipated that the older group will find the format more beneficial than the younger group due to Sloman’s assertion (2002) that the format enables analytical thinking in those who may not normally use system 2 to solve such tasks.
5. Individual differences, as measured by the MHVS, IPS, REI and TDQ, will attenuate the age and format related variance.

7.2 Method

7.2.1 Design

7.2.1.1 Analyses of Variance

This study used a mixed design, with age (young or old) and format (frequency or probability) as between participants independent variables, and problem type (conjunction, exclusive disjunction or inclusive disjunction) as the within participants independent variable. As in the previous study, both dependent variables were based on the participants' probability judgements. The first dependent variable was the error score as detailed in the previous Chapter, while the second was whether the conjunctive or disjunctive fallacy had been committed in each case. The subscales of the REI (Need For Cognition and Faith in Intuition) and the Mill Hill Vocabulary Scale were again used as covariates, as was a measure of information processing speed, as described below.

7.2.1.2 Regression Analyses

The current data, and previous literature (e.g. Gigerenzer & Hoffrage, 1995) support the proposal that the frequency format has a facilitating effect on reasoning performance, reducing levels of fallacy (as in the literature) and magnitude of error (as in the current data). The current data also suggests that, while there is no straight forward effect of age upon probabilistic reasoning, the old group are less affected (albeit not significantly) by the format of the task – see 7.3.1.

However, there is also a strong possibility that information processing speed may account for large amounts of variance in performance on a range of reasoning tasks, especially age related variance (Fisk & Sharp, 2002; Salthouse and Babcock, 1991). Equally, evidence has also been found for the suggestion that thinking dispositions account for large amounts of variance in reasoning performance (Stanovich & West, 1998), as do measures of cognitive ability (Toplak & Stanovich, 2002).

As such, four hierarchical regressions were conducted for each dependent variable, enabling examination of the remaining variance once that which is accounted for by a) format, b) the MHVS and IPS measures, c) the TDQ subscales or d) age group has been accounted for. In each case the variable of primary interest was entered first, identifying the total variance accounted for by that that predictor, and all remaining variables were

entered as a second step. This allows for a comparison between the variance associated with the first predictor and the incremental effect of all of the other predictors, indicating the amount of *unique* variance accounted for by each.

The regression then tells us about any additional variance that our second predictor (or group of predictors) can account for, above that which is accounted for by the first predictor.

7.2.2 Participants

The older group consisted of 77 individuals, 24 male and 53 female, with a mean age of 70.53 (7.12), and a range from 60 to 88. They were contacted through local groups of the national organisation of the University of the Third Age (U3A), a group for older people which provides a range of classes aimed to provide education on a wide range of topics. Hultsch *et al.* (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 would be matched to the younger group of students, in the way 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 contacted the researcher and 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.

Eighty young participants took part in this study, 18 male and 62 female, ranging from 18 to 29, with a mean age of 19.66 years (2.49).

As in the previous study (see Chapter 5) these participants were all students at Liverpool John Moores University.

7.2.3 Materials

Participants in this study were asked how many years they had spent in full time education. This education measure was used in this current study, as the elder cohort may have spent significantly less time in education than the university students in the young cohort.

Participants also completed the multiple choice MHVS (found by Raven *et al.*, 1998, to have good reliability for adults over 50, at $\alpha=.9$) and Information Processing Speed 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. They were given a full page of instructions, as well as five practice items. Participants were given 30 seconds to complete the items at each level of difficulty. These materials are available in Appendix 12.

7.2.3.1 Probabilistic Reasoning tasks

The participants were presented with the same 5 disjunctive problems, and 5 conjunctive problems as in the previous study (Chapter 5). The participants either received all 10 problems in the probability format or all 10 in the frequency format. The individual problems were again presented within the booklets in quasi-random order, in order to address fatigue and practice effects.

7.2.3.2 Thinking Styles Measures

All participants also completed the Rational Experiential Inventory, consisting of the Need for Cognition and Faith in Intuition scales (Cacioppo & Petty, 1982; Epstein *et al.*, 1996). The Thinking Disposition scale (Kokis *et al.*, 2002) was also completed, and full details of each of these measures can be found in the preceding chapter.

7.2.4 Procedure

The measures were administered in two separate sessions, with the vast majority of participants completing them exactly a week apart. In the first session, the participants completed the health and education questionnaire and the MHVS, and in the second they completed the information processing speed task and the probabilistic reasoning

tasks. Within the booklet of reasoning tasks the current participants also completed the Bayesian reasoning tasks for the following study (Chapter 8), and all were presented in a quasi-random order.

As with the previous study, the younger participants completed the measures during two of their teaching sessions, in their usual classrooms. The older participants came to the university at times that were convenient to them and completed the tasks either individually, or in groups of up to six participants. The British Psychological Society's ethical principles and guidelines were adhered to.

7.3 Results

Scores on the background variables are presented below, in Table 7.1. There was a significant age difference in each case, with young participants exhibiting significantly faster information processing speed and lower Mill Hill scores. Older people showed less than a year's difference in education, with a mean of 13 years to the younger group's 14, a difference that can be explained by changes in statutory requirements of school attendance.

Table 7.1: Mean scores on background measures

	Young Mean (SD)	Old Mean (SD)	(df)t	All Data Mean (SD)
Years of Education	14.32 (1.58)	13.43 (2.93)	(153) 2.67*	13.88 (2.38)
MHVS	15.03 (3.15)	24.52 (3.87)	(151) -16.63***	19.80 (5.92)
Information Processing Speed	103.19 (15.70)	72.27 (16.09)	(142) 11.63***	86.66 (22.16)

* $p < .05$; ** $p < .01$; *** $p < .001$

There were differences also between the age groups with regards to the thinking style measures (Table 7.2). A series of t tests revealed that the younger group actually showed *higher* levels of Faith in Intuition than the older group, $t(151)=3.13$, $p < .01$, as

well as higher levels of NFC, $t(151)=2.36$, $p<.05$. With regards to the TDQ (where lower scores indicated stronger levels of that trait), the young group showed significantly stronger Dogmatism, $t(151)=-3.61$, $p<.001$, and Superstitious Thinking/Luck, $t(151)=-6.16$, $p<.001$, but less Absolutism, $t(149.87)=2.84$, $p<.01$, and Social Desirability bias, $t(151)=3.49$, $p<.001$. As such, it is not the case that older participants were reporting consistently lower (or higher) levels of analytic thinking. Although the older group showed a significantly higher level of Social Desirability bias, it was still indicating a low level in absolute terms, and as such it is felt to be acceptable. The table (7.2, overleaf) also presents the standard deviations in each case, and contrary to expectations, these do not indicate that using an older cohort of participants led to greater variability in thinking styles across the sample.

Table 7.2: Thinking style mean scores by age group

Scale	Subscale	Young Mean (SD)	Old Mean (SD)	Min/max Possible score	t value (df=151)	All Data Mean (SD)
REI – Rational Experiential Inventory	FI – Faith in Intuition NFC – Need for Cognition	43.22 (6.32)	40.17 (5.74)	12/60	3.13**	41.69 (6.21)
		64.56 (10.41)	60.47 (11.05)	19/95	2.36*	65.50 (10.89)
TDQ – Thinking Dispositions Questionnaire	FT – Flexible thinking A - Absolutism D - Dogmatism CT – Categorical thinking ST/LC – Superstitious Thinking/Luck Composite sfNFC – Short form need for cognition SD – social desirability BI – belief identification	18.88 (3.24)	18.39 (2.94)	10/40	.97	18.64 (3.09)
		14.49 (2.13)	13.58 (1.81)	5/20	2.84** (df=149.87)	14.04 (2.02)
		13.62 (1.99)	14.84 (2.20)	6/24	-3.61***	14.24 (2.18)
		8.83 (1.54)	8.97 (1.41)	3/12	-.61	8.90 (1.48)
		22.86 (4.69)	27.06 (3.71)	8/32	-6.16***	24.97 (4.71)
		19.55 (3.02)	19.19 (3.83)	9/36	.65	19.37 (3.44)
		14.76 (2.21)	13.64 (1.73)	5/20	3.49***	14.19 (2.05)
		16.40 (2.14)	16.62 (1.81)	6/24	-.70	16.51 (1.98)

* p<.05; ** p<.01; *** p<.001

7.3.1 Analyses of Variance

Fallacy Data

Table 7.3: Mean and standard deviations for incidences of each type of fallacy by age group

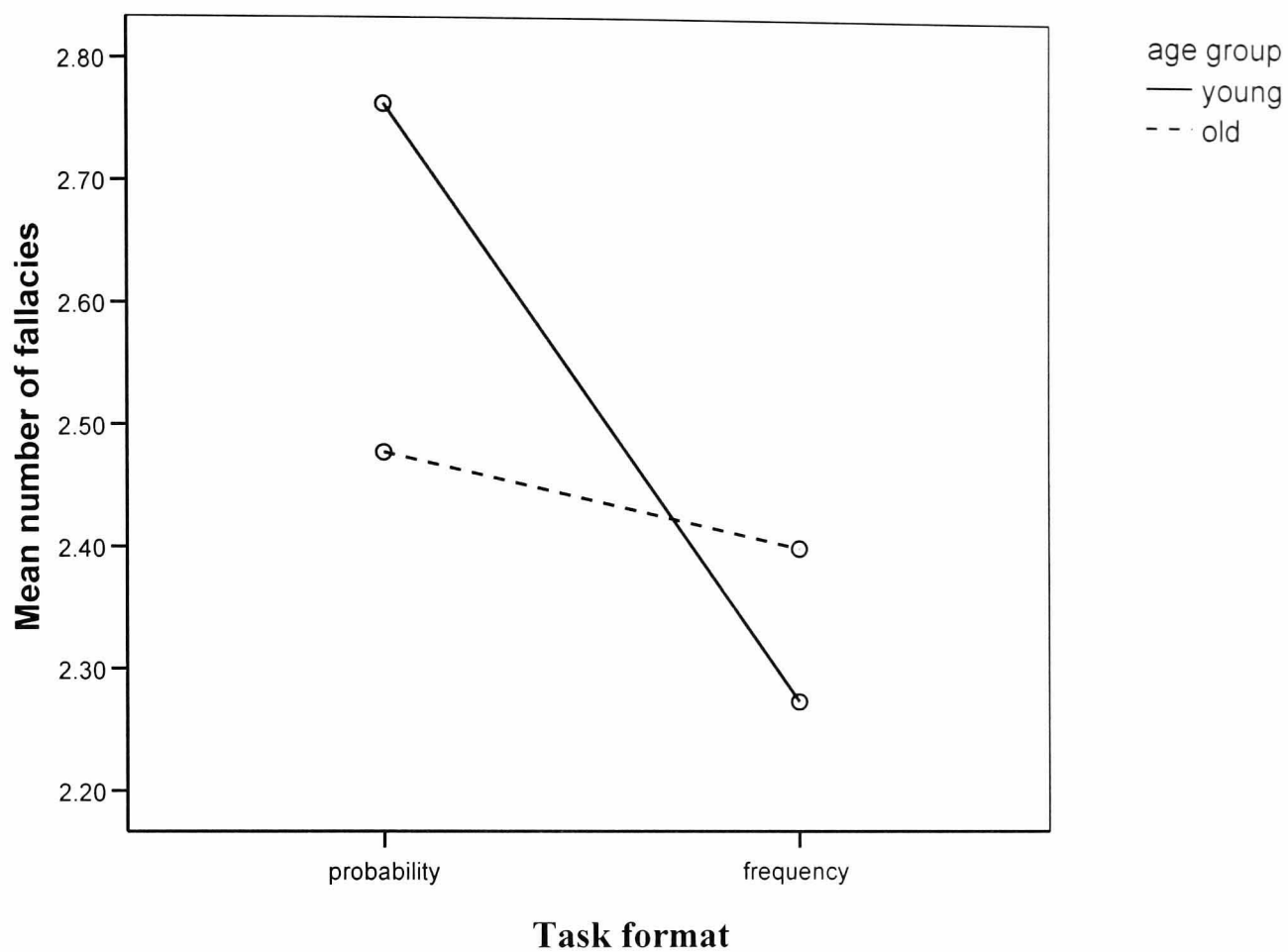
	Age Group	Probability	Frequency	All data
		Mean (SD)	Mean (SD)	Mean (SD)
Conjunction	Young	3.42 (1.03)	2.89 (1.40)	3.20 (1.22)
	Old	3.17 (1.53)	3.24 (1.30)	3.20 (1.42)
	Combined	3.31 (1.27)	3.06 (1.35)	3.20 (1.31)
Exclusive Disjunction	Young	2.00 (1.12)	1.43 (1.36)	1.76 (1.15)
	Old	1.83 (1.29)	1.80 (1.35)	1.82 (1.31)
	Combined	1.93 (1.19)	1.60 (1.25)	1.79 (1.22)
Inclusive Disjunction	Young	2.87 (1.04)	2.50 (1.53)	2.71 (1.27)
	Old	2.43 (1.45)	2.16 (1.14)	2.31 (1.32)
	Combined	2.68 (1.25)	2.34 (1.36)	2.53 (1.30)

The mean number of fallacies in each condition indicate that there was no clear difference between the age groups, with the older group making slightly fewer fallacies in the inclusive disjunctions and slightly more in the exclusive disjunctions, and no noticeable age difference at all on the conjunction tasks. Regarding the format of the task, the probability format led to a greater number of fallacies in each case.

Figure 7.1 overleaf illustrates the contrasting effects of problem format. Inspection of the Figure reveals that the younger group still appeared to be benefiting from the ‘frequency effect’, making fewer fallacies when problems are framed in terms of frequencies. By way of contrast, the older group performed similarly irrespective of problem format. In fact, Table 7.3 shows that in the conjunction tasks they actually performed slightly better in the probability condition. However, this apparent interaction proved not to achieve significance, as detailed below.

Within participants comparisons showed that there was a significant effect of task type, $F(1.90, 222.61)=49.67, p<.001$, with an effect size (partial η^2) of .30 (Greenhouse-Geisser adjusted degrees of freedom are shown here, as in this one case the data violated the assumption of sphericity). This is illustrated by the combined values in Table 7.3, with conjunctions leading to the greatest number of fallacies with a mean of 3.20 (SD=1.31), and exclusive disjunctions led to the least with a mean of 1.79 (1.22). Bonferroni pairwise comparisons indicate that each type of task differed significantly from each other ($p<.001$). There were no other significant main effects or interactions. For the effect of problem format $F(1, 117)=2.87, p>.05$, partial $\eta^2 = .02$ while for all interactions $p>.05$.

Figure 7.1 Mean Fallacies Committed in Probability and Frequency Format



To investigate the relationship between NFC, FI, MHVS and the effect of format, these were added as covariates. It was found that neither thinking style nor the MHVS was significant, $p > .05$, indicating that controlling for these variables had no clear impact on the main effect.

In this case, the participants' information speed was also added as a covariate, and this measure was significant $F(1,109)=6.19$, $p < .05$, effect size of partial $\eta^2 = .05$. After controlling for information processing speed, the between participants factor of problem format became marginally significant, at $F(1,109)=4.21$, $p < .05$, partial $\eta^2 = .04$. Before adding the covariates, estimated marginal means had been 2.62 and 2.33 fallacies for the probability and frequency conditions respectively. After controlling for IPS, the number in the probability condition rose to 2.63, and the fallacies in the frequency condition fell to 2.29. As such, it is apparent that when the older participants' slower information processing speed was controlled for, the probability format again led to a significantly greater number of fallacies, but this result should be observed with some caution, as the

changes in the means (as stated above) were very slight. The age group x format interaction remained non-significant. Homogeneity of regression was established for every covariate in the above analysis, with $>.05$ for the factor x covariate interaction.

Error Data

A second analysis of variance again looked at the within participants factor of task type, and between participants factors of problem format and age group, but using the error data as the dependent variable. Mean scores are displayed in Table 7.4 below.

Table 7.4: Means and standard deviations for incidences of each type of error by age group

	Age Group	Probability Mean (SD)	Frequency Mean (SD)	All Data Mean (SD)
Conjunction	Young	24.38 (9.42)	19.28 (11.98)	22.23 (10.79)
	Old	23.91 (13.83)	21.36 (11.93)	22.75 (12.95)
	Combined	24.17 (11.51)	20.28 (11.89)	22.47 (11.79)
Exclusive Disjunction	Young	29.05 (12.23)	19.81 (11.90)	25.15 (12.85)
	Old	30.25 (16.03)	23.85 (15.44)	27.34 (15.95)
	Combined	29.59 (13.96)	21.75 (13.73)	26.16 (14.34)
Inclusive Disjunction	Young	29.66 (8.74)	18.08 (10.10)	24.77 (10.91)
	Old	23.13 (12.57)	19.41 (9.28)	21.44 (11.25)
	Combined	26.74 (11.04)	18.72 (9.64)	23.23 (11.15)

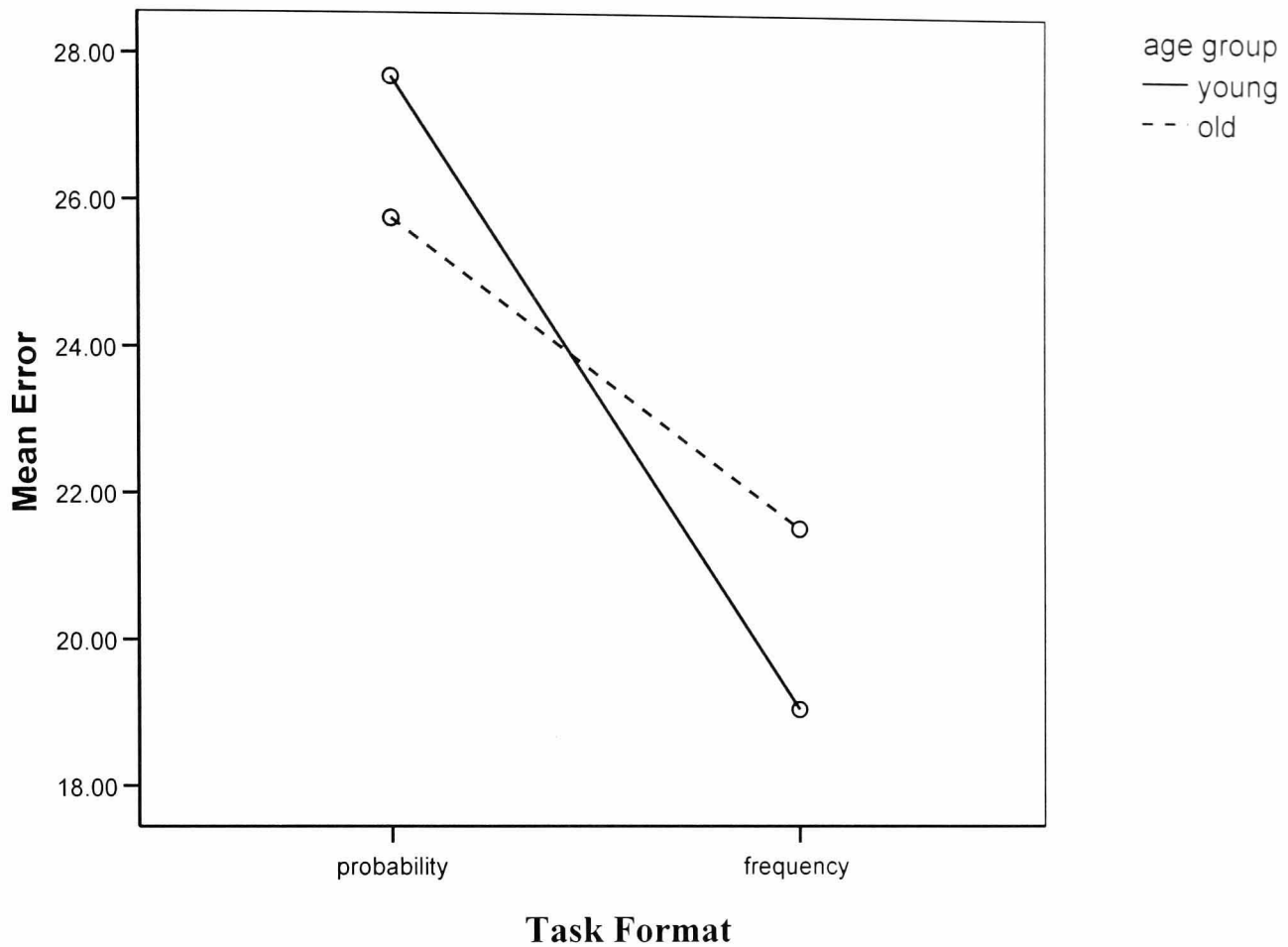
By this measure, performance in the probability condition was consistently worse, as demonstrated by higher levels of error in this condition for each type of task. The data presented here showed heterogeneity of covariance, therefore Greenhouse-Geisser adjusted degrees of freedom are reported as appropriate.

With regard to problem format, the effect size was greater than in the analysis of the fallacy data, and there was a significant main effect, $F(1,115)=15.23$, $p<.001$, partial $\eta^2 = .12$. This can be seen clearly in Figure 7.2.

There was a significant effect of task type on errors ($F(1.87, 215.21) = 4.38$, $p<.05$, partial $\eta^2 = .04$). As in the previous chapter, exclusive disjunctions showed the largest error, followed by inclusive disjunctions and then conjunctions. Bonferroni pairwise comparisons reveal that the exclusive disjunctions generated larger errors than both the inclusive disjunctions and the conjunctions, $p<.01$ and $p<.05$ respectively, while these last two did not differ significantly from each other (in the previous study the conjunction errors had been significantly higher from both disjunctions, which did not differ from each other). These values are presented in Table 7.4, above.

Again, there was no main effect of age ($F<1$) and there were no significant interactions – each age group is similarly affected by the effects of format ($F(1,115)=1.79$, $p>.05$, partial $\eta^2 = .02$) and task type ($F(1.87, 215.21)=2.05$, $p>.05$, partial $\eta^2 = .02$), and there was no significant interaction between the two ($F(1.87, 215.21)=1.50$, $p>.05$, partial $\eta^2 = .01$).

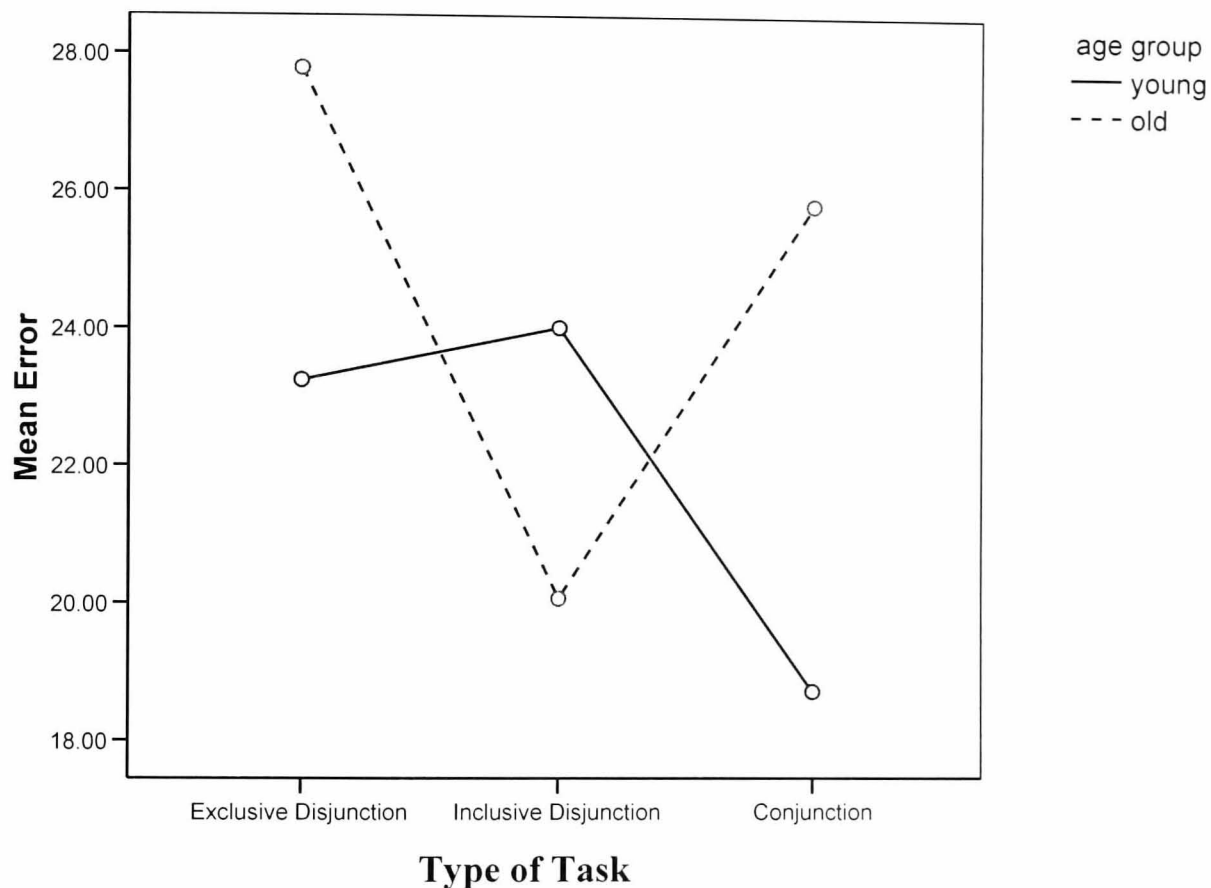
Figure 7.2 Mean Error Committed in Probability and Frequency Formats



When included as covariates, neither the Mill Hill, the IPS nor the scales of the REI achieved significance. However, the effect of all three covariates combined was to lead to the task type by age group interaction achieving significance, at $F(2, 106)=3.04$, $p=.05$. The old group performed worse than the younger group on both the conjunctive and the exclusive disjunctive tasks, but better on the inclusive disjunctions, as indicated in Table 7.4 and Figure 7.3, but it was only when individual differences were accounted for that this became statistically significant. The older participants' mean scores remained exactly the same when adjusted, at 22.75, 27.34 and 21.44 for conjunctions, exclusive disjunctions and inclusive disjunctions respectively. The difference in performance was reduced in the case of conjunctions, with the younger participants' adjusted means showing an increase of error from 22.23 to 22.39 (i.e. older group is 'worse' by a smaller difference), and in the inclusive disjunctions, where the young group's error was reduced from 24.77 to 24.18 (older group became 'better' by a smaller difference). The difference increased in the exclusive disjunctions, however, with younger participants' level of error having been reduced from 25.15 to 25.02,

leading to the older group doing worse by a greater margin after controlling for individual differences, indicating that they are even less able to discriminate between the inclusive and exclusive tasks. Again, assumptions of homogeneity of regression were not violated for any of the analyses of covariance.

Figure 7.3 Task type by age group interaction



7.3.2 Regression analyses

As in the previous study, six hierarchical regression analyses were conducted, one for fallacy on each type of task, and one each for the error data. With effect of format apparent in each analysis of variance above (after controlling for processing speed in the case of the fallacy data) this pattern was expected to be repeated here, but as detailed below, the effect was less consistent across all measures of probabilistic reasoning. Age, MHVS, IPS and the TDQ scales were also examined as predictors of variance in each of the dependent measures used, in order to examine the amount of variance that could be attributed to each and the extent to which they attenuated the age and format related variance.

For each outcome variable, four sequences were conducted. In the first sequence, problem format was entered as the first step, with all remaining predictors (the TDQ subscales, Mill Hill score, information processing speed (IPS) and age group) as the second, in order to examine whether any effect of format remained significant once these measures of individual difference had been included. For the second sequence, Mill Hill score and IPS were entered first, as measures of cognitive ability, with the remaining predictors second. For the third the TDQ subscales were the first step, followed by the remaining measures, and finally age was entered as a first step, again followed by all other predictors in the second, to examine whether any effect remained significant after task format and individual differences had been accounted for.

As indicated by the results reported below, age was not significant as a predictor of variance in any of the dependent variables (a difference in error only had been hypothesised), and as such the question of whether individual differences would attenuate an age effect has now become redundant for this study. It should also be noted that format was a significant predictor of reasoning performance in only two cases, exclusive disjunction error and inclusive disjunction error, and it is therefore only in these two measures that the question of attenuation of this effect can be addressed.

Conjunction Fallacy Regression

Conjunction fallacy was the first outcome variable examined, with the results summarised in Table 7.6. No predictor accounted for significant levels of variance when entered as the first step, neither did the remaining predictors account for significant further variance. Of the predictors, the TDQ scales showed the greatest R squared change in each case, at 10%, with no single disposition showing significance (full beta weights for all regression analyses presented here are in Appendix 13).

Table 7.5: Model summaries for regression analyses of conjunction fallacy data

Predictors	R Squared	R squared change
Format	.00	.00
All remaining	.11	.11
MHVS and IPS	.02	.02
All remaining	.11	.09
TDQ subscales	.10	.10
All remaining	.11	.01
Age	.00	.00
All remaining	.11	.11

Significance of the regression for the full model: $F(12, 115)=1.15, p>.05$

* $p<.05$; ** $p<.01$; *** $p<.001$

Exclusive Disjunction Fallacy Regression

With exclusive disjunction fallacies as the outcome variable, it was again the case that no individual step nor the overall model accounted for significant levels of variance, although beta weights indicate that IPS was significant when entered in the first step ($\beta=-.22, p<.05$), and remains so in the full model ($\beta=-.27, p<.05$), with a faster processing speed being associated with fewer fallacies. Full beta weights are again available in Appendix 13. These standardised beta weight indicate how much, in standard deviations, the outcome variable will alter as a result of a change of one standard deviation change in the predictor. So in this case as IPS increases by its SD of 15.90, the negative beta weight indicates that the number of exclusive disjunction fallacies would (if the model were perfect) decrease by $-.22$ of its own SD (1.22). $-.22*1.22 = -.27$, so one standard deviation change in IPS leads to the number of fallacies dropping by a little under a third of a fallacy. Please see Tables 7.1 and 7.2 for the SD values involved.

Table 7.6: Model summaries for regression analyses of exclusive disjunction fallacy data

Predictors	R Squared	R squared change
Format	.01	.01
All remaining	.08	.07
MHVS and IPS	.04	.04
All remaining	.08	.05
TDQ subscales	.03	.03
All remaining	.08	.05
Age	.00	.00
All remaining	.08	.08

Significance of the regression for the full model:F(12, 114)=.88, p>.05

* p<.05; ** p<.01; *** p<.001

Inclusive Disjunction Fallacy Regression

Finally, looking at the inclusive disjunction fallacy, it is apparent that the TDQ accounts for significant amounts of variance in performance, at 17% when entered as the first step. It can be seen that the subscale leading this is Social Desirability, at $\beta = -.20$, $p < .05$. In the full model, Social Desirability remains a significant predictor at $\beta = -.24$, $p < .05$ and Superstitious Thinking/Luck is also significant ($\beta = -.22$, $p < .05$). In each case a weak level of the disposition was associated with fewer fallacies. IPS was also a significant predictor in the full model, at $\beta = -.32$, $p < .01$, showing that a faster processing speed was associated with fewer fallacies.

Table 7.7: Model summaries for regression analyses of inclusive disjunction fallacy data

Predictors	R Squared	R squared change
Format	.03	.03
All remaining	.24	.21**
MHVS and IPS	.04	.04
All remaining	.24	.19**
TDQ subscales	.17	.17**
All remaining	.24	.07
Age	.01	.01
All remaining	.24	.22**

Significance of the regression for the full model:F(12, 112)=2.90, p<.01

* p<.05; ** p<.01; *** p<.001

Conjunction Error Regressions

The conjunction error data revealed that when entered as a first step, both verbal intelligence and IPS were significant predictors of the conjunction error, ($\beta=-.28, p<.01$. and $\beta=-.24, p<.05$ respectively) indicating that higher verbal intelligence and faster processing speeds were associated with lower levels of error. In the full model, only verbal intelligence remains a significant predictor, at $\beta=-.39, p<.01$.

Table 7.8: Model summaries for regression analyses of conjunction error data

Predictors	R Squared	R squared change
Format	.02	.02
All remaining	.18	.16*
MHVS and IPS	.07	.07*
All remaining	.18	.11
TDQ subscales	.11	.11
All remaining	.18	.07
Age	.00	.00
All remaining	.18	.18*

Significance of the regression for the full model:F(12, 113)=2.01, $p<.05$

* $p<.05$; ** $p<.01$; *** $p<.001$

Exclusive Disjunction Error Regression

In this case, the format of the tasks became significant, accounting for 8% of the variance in performance, with a beta weight of $-.28, p<.001$ ($\beta=-.30, p<.01$ for the full model). As in the analyses of variance, this reflected the fact that those in the frequency condition showed a lower level of error on the exclusive disjunction tasks. It is also noteworthy that format remained statistically significant in the full model after controlling for the effects of the TDQ and the MHVS and IPS.

Table 7.9: Model summaries for regression analyses of exclusive disjunction error data

Predictors	R Squared	R squared change
Format	.08	.08**
All remaining	.15	.07
MHVS and IPS	.03	.03
All remaining	.15	.12
TDQ subscales	.03	.03
All remaining	.15	.11**
Age	.01	.01
All remaining	.15	.14
Significance of the regression for the full model:F(12, 114)=1.63, p>.05		

* p<.05; ** p<.01; *** p<.001

Inclusive Disjunction Error Regression

Format was again a significant predictor, accounting for 16% of variance in performance when entered as the first step. First step and full model beta weights for this predictor were $\beta=-.40$ and $\beta=-.34$ respectively, both at $p<.001$, indicating that the ‘frequency effect’ was present. Again, as in the previous set of analyses, format remained statistically significant in the full model after controlling for the effects of the TDQ, the MHVS and IPS.

Table 7.10: Model summaries for regression analyses of inclusive disjunction error data

Predictors	R Squared	R squared change
Format	.16	.16***
All remaining	.23	.07
MHVS and IPS	.01	.01
All remaining	.23	.22***
TDQ subscales	.01	.01
All remaining	.23	.13**
Age	.02	.02
All remaining	.23	.21**
Significance of the regression for the full model:F(12, 112)=2.78, p<.01		

* p<.05; ** p<.01; *** p<.001

7.4 Discussion

There were five hypotheses made for the current study, and as in the previous study the first hypothesis – the facilitating effect of the frequency format – was strongly supported by the data. This was particularly marked in the error data, whereas the effect was only apparent in the number of fallacies once the covariate of information processing speed had been controlled for. Again, a clear indication of the effect of format is apparent when the dependent variable is the mean error measure i.e., the magnitude of the deviation from the normative value. The effect was not evident in the analysis of the categorical ‘fallacy/not fallacy’ data.

The hypothesised effect of age upon mean error was not found, indicating that the second hypothesis must be rejected. There was no main effect of age on either measure, and neither did age account for significant levels of variance in any of the regression analyses. The third hypothesis was that the older group would show significantly lower levels of analytical thinking (Baltes, 1997; Chen & Sun 2003), with the expectation that this increased variance in the thinking style data would lead to more detailed examination of the relationship between thinking styles and reasoning performance. This hypothesis was also rejected, as although older participants showed significantly weaker levels of Dogmatism (as measured by the TDQ) they also showed stronger levels of Absolutism, and weaker levels of Superstitious Thinking/Luck, a measure which indicates a lower tendency to think analytically. Similarly, the REI showed that the older group reported lower levels of Faith in Intuition, but also lower levels of Need For Cognition.

There was also no interaction between age and format – the older group did not find the format to be particularly beneficial, which fails to support the fourth hypothesis. The older participants actually appeared to benefit *less* from the frequency format, but due to their performance across the two being very similar, the interaction was not significant in either the error or the fallacy data, and the fourth hypothesis must therefore be rejected. The lack of age effect is similar to the findings of Fisk (2005), who was measuring reasoning performance by incidences of fallacies, and the use of error data further confirms the lack of difference between the groups. It could be that the older cohort are performing as well on the tasks as the younger group because the tasks are not complex enough (see discussion of Chasseigne *et al.*, 2004, and Mutter & Pliske, 1994, in Chapter 4), but the high incidences of fallacy (out of a possible five. the

conjunction, exclusive disjunction and inclusive disjunction show 3.31, 1.93 and 2.68 respectively) may suggest otherwise.

The error data in the current study does again show some interesting differences between performances on the different types of task (as in Chapter 6). In this case the effect of disjunction type is even more pronounced, as the exclusive disjunctions again showed significantly lower numbers of fallacies, but also significantly higher amounts of error, with the older group being particularly susceptible to this effect. This represents the fact that while each disjunctive task shows a mean underestimate of the disjunction value, it is the exclusive tasks that show a greater level of underestimate, and in particular a greater level of underestimate *per fallacy*. This is due, at least in part, to the fact that when calculating the normative response to the inclusive task, the value of the conjunctive event, $A+B$, is deducted – leading to a lower implicit normative value and one which is perhaps likely to be closer to the participant's own estimate of the disjunctive probability. No such deduction is made in calculating the normative value of the exclusive disjunctions and the disparity between the actual estimate and the normative one is likely to be even greater where there is subadditivity. This does not rule out the possibility that the participants are incorrectly scaling down their exclusive estimates, as if they were inclusive tasks requiring the deduction of the conjunctive value, nor does it rule out them failing to deduct the conjunctive value in the inclusive tasks, when it would be correct to do so. The latter example could occur in situations where the participant mistakenly believes there to be a negative conditional relationship between events (the occurrence of one making the occurrence of the other less likely) on a task where the compound probabilities are inconsistent with such a possibility, leading to superadditivity (see also section 6.4).

Regarding hypothesis five, it had been anticipated firstly that thinking style would attenuate the age related variance. However, as age did not have an effect upon reasoning, and did not account for large amounts of variance, this meant that this part of the hypothesis could not be addressed.

The TDQ subscales did account for significant amounts of variance in the inclusive disjunction fallacy data. In this one regression model it was the Social Desirability subscale that appeared to account for the most variance, closely followed by Superstitious Thinking/Luck. In each case, a weak level of the disposition was

associated with a lower incidence of the fallacy. In the case of Superstitious Thinking/Luck, this would correspond to the theory that a strong tendency to superstition may be associated with lower levels of analytical thinking, as proposed by Kokis *et al.* (2002). In the case of SD, however, it is a little more surprising that the desire to appear socially acceptable should be associated with greater errors. At least, it seems to suggest that a desire to please is not associated with greater accuracy in these tasks, even if it may be associated with greater effort.

In examining the effect of format, the MHVS and the IPS were also controlled for. Controlling for Information Processing Speed did lead to the effect of task format achieving significance, with the adjusted means showing fewer fallacies in the frequency condition, and increased fallacies in the probability condition. It should be noted that, due to missing data, there were fewer participants in the analysis of covariance than in the original ANOVA. The participants missing from the second analysis, however, were in the younger group, and it is the younger group that were the most affected by the tasks' format. As such, the effect of controlling for IPS appears to be despite the missing data, and not because of it. In the regression analyses, the only outcome variable predicted by IPS was conjunction error, with faster participants showing lower levels of error.

As discussed in Chapter 4, age differences in reasoning tasks are often only apparent when the tasks become highly cognitively demanding. As such, the following study uses a more challenging probabilistic reasoning task, in the form of two Bayesian problems. While concerns are expressed about causing a ceiling effect in performance, it is felt that if an older cohort does indeed find these tasks significantly more challenging it is highly unlikely that both groups will do so badly that no difference can be discerned. The focus will not be on whether participants respond with the normative answer, but on any differences between the groups in the *types* of answer they give. As such, the following study will not measure reasoning fallacy or error, but will examine any pattern in the responses. It is anticipated that the pattern of responses will be able to reveal effects of age and format upon the processes being used by participants. By being more likely to reveal an effect of age upon reasoning, it is also anticipated that the Bayesian problems will provide a better context within which to explore any mediating role of individual differences, specifically thinking styles as measured by the REI and

TDQ. As noted above, the attenuation of variance by such measures can only be addressed when such variance – in this case, an effect of age – is present in the data.

This final study will also address the lack of attenuation by thinking styles on the effect of format. Again, it is anticipated that the pattern of responses given by participants will allow for the examination of the ways in which participants are *attempting* to address the tasks.

Chapter 8 – Bayesian Reasoning

8.1 Introduction

The studies reported so far have looked at two main types of reasoning task – conjunctions and disjunctions. As discussed in Chapter 2, there is a third form of probability problem that has been used extensively to investigate reasoning performance – the Bayesian 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 E *given* the evidence, A .

Utilising this formula 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 above, an individual who is attempting to solve a Bayesian task 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 facilitating frequency format was used. Despite evidence that participants have processed the relevant base rate information, the most common response to such tasks is to respond only by giving the value of the further evidence, a tendency known as base rate neglect (De Neys & Glumicic, 2007; Johansen, Fouquet & Shanks, 2007; Franssens & De Neys, 2008)

Gigerenzer and Hoffrage (1995) demonstrated that the frequency effect does appear to be present with Bayesian problems. Similarly, Cosmides and Tooby (1996) found that phrasing the Bayesian ‘disease X’ problem in a frequency format would consistently elicit more correct responses from participants than the probability format. This effect was so robust that to even phrase the actual question in this way, with the previous information remaining in a probability format, would 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 Bayesianly.

Interestingly, Cosmides and Tooby (1996) refer to the condition that led to the 92% accuracy level as being the condition with the highest ecological validity. This condition involved the participants being given diagrams to help them reach a conclusion. The diagram consisted of 100 squares each representing one person. Participants were then instructed to circle the number of people (squares) who would have the disease, then to colour in the number of diseased people who would also test positive. This then does not seem to have ecological validity in terms of a person actually being given such a diagnosis of a disease, as they are actually being given far more help than would be available to them in such a real situation.

Research does suggest that Bayesian reasoning can be greatly improved when tasks are worded as frequency problems (Fantino & Stolarz-Fantino, 2005, 2006; Sloman, *et al.*, 2003, see also Evans, 2003, for a brief review), with Sloman *et al.* (2003) suggesting that the facilitating effect comes from the wording of such tasks making the nested sets in the problem more apparent to the participants, as described in Chapter 2. In discussion of Dual Process Theory, Evans (2003) also concludes that the frequency format is facilitating the use of the analytical system, and therefore the correct use of the base rate information. Croskerry (2009) extols the use of a dual process account in explaining (and facilitating the reduction of) errors in diagnostic procedures, presenting a model for the application of dual process theory to Bayesian reasoning in diagnostics.

With regards to the age differences being examined, 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. With the current research also finding no clear ageing effect on reasoning performance so far, the main aim of the current chapter is to investigate any influence of age and task format on the way in which participants respond to the more challenging Bayesian tasks, and also to assess

the interrelationships, if any, between thinking styles, age, and format and the implications for performance on these tasks. While the relationship between analytic and heuristic thinking style has been examined as mentioned above (see also Franssens & De Neys, 2009), this has to date come from the inference that analytical thinking leads to normative responses, while heuristic processes lead to ‘incorrect’ responses on such tasks. The literature does not directly address the question of whether those with rational or experiential thinking styles do show a tendency to respond to such tasks in the way that dual process theory would predict.

Chasseigne *et al.*, (2004), Chasseigne *et al.* (1997), Mutter (2000), Mutter, *et al.* (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. 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 primarily at the magnitude of error made by participants, but also at the apparent underlying processes.

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%, will give 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 former condition is known as base rate neglect, with participants apparently failing to take into account the relationship between base rate number of cabs and the witness’s accuracy. The latter has been called the reverse base rate fallacy (Teigen & Keren, 2007), 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. 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%. As suggested above, it is the *type* of answer that is of interest in the current study. It is anticipated that those higher in analytic thinking – and specifically with a high Need for Cognition – will be more likely to attempt the complex Bayesian reasoning process.

By looking at the participants' distribution of responses in these terms, it is felt that any age differences may be better revealed than by using previous chapters' approach of focusing on a mean response value, or mean error. In the case of Bayesian tasks, identifying which of the available cuing material the participants are using, and how, may help to illustrate how they are approaching the problems, and whether either age or task format has any effect on the distribution of responses in these terms.

8.1.1 Hypotheses

1. That the way in which participants respond – whether with the correct answer or one of the common errors described by Birnbaum, above (2004) – will vary according to thinking style. It is anticipated that those with an analytical thinking style will be more likely to give a response that indicates an awareness of each of the relevant pieces of information, and/or an attempt to manipulate these to reach the normative answer.
2. Older participants will be less likely to give such responses, and more likely to show clear base rate neglect or base rate only responses.
3. Consistent with previous research (e.g., Gigerenzer & Hoffrage, 1995; Sedlmeier & Gigerenzer, 2001), those receiving problems posed in the frequency format will be more likely to achieve the correct result.

8.2 Method

8.2.1 Design and statistical analyses

The primary interest in collecting the current data was not whether or not participants give the normatively correct answer (although this is of course one of the aims), so much as investigating the *types* of answer they give and what may lead to an individual reasoning in any given way. As such, instead of a measure of how far participants were

from the normative answer the current study looks at whether each response conforms to any of the groups as identified by Birnbaum (2004) and discussed above and/or any other apparent groups or clusters. Responses to each task were allocated to groups (for full details of how this was done see 8.3), and discriminant function analysis was used to establish whether participants' responses to each of the reasoning tasks could be predicted by the variables of gender, age, format, and thinking styles. This analysis allows us to examine the dimensions (or 'functions') along which the response groups may differ, by illustrating what factors are most important in predicting the responses. The predictors in this analysis were therefore gender, age, format and each of the thinking style scales, while the outcome or 'group' variable was the type of response given to the task.

8.2.2 Participants

The participants in this current study were the same as those used in the previous one. 77 older individuals, with a mean age of 70.53 (7.12) years and a range from 60 to 88 were recruited through the U3A. There were 24 males and 53 females in this group. 80 younger participants took part, 18 male and 62 female, ranging from 18 to 29, with a mean age of 19.66 (2.49) years.

8.2.3 Materials

8.2.3.1 Thinking Styles

All participants also completed the thinking style scales described in detail in Chapter 5. These were the Rational Experiential Inventory, consisting of the Need for Cognition and Faith in Intuition scales (Cacioppo & Petty, 1982; Epstein *et al.*, 1996), and the Thinking Disposition scale (Kokis *et al.*, 2002).

8.2.3.2 Bayesian problems

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 *et al.* (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? _____

8.2.4 Procedure

As described in previous studies, the measures were presented to each age group in two separate sessions, with the thinking style measures being completed in the first, while the reasoning tasks themselves – the Bayesian tasks of interest here, and the conjunctive and disjunctive tasks discussed in Chapter 7 – were completed in the second session. 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.

8.3 Results

In the case of the cab problem, the participants' responses in each condition were already presented as values out of 100, while in the frequency version of the disease problem the participant chose themselves what quantifier to use, so their score was then converted so that it was directly comparable to the percentage given in the probability format of the task.

8.3.1 Descriptive Statistics

Table 8.1: Descriptive statistics for participants' judgments

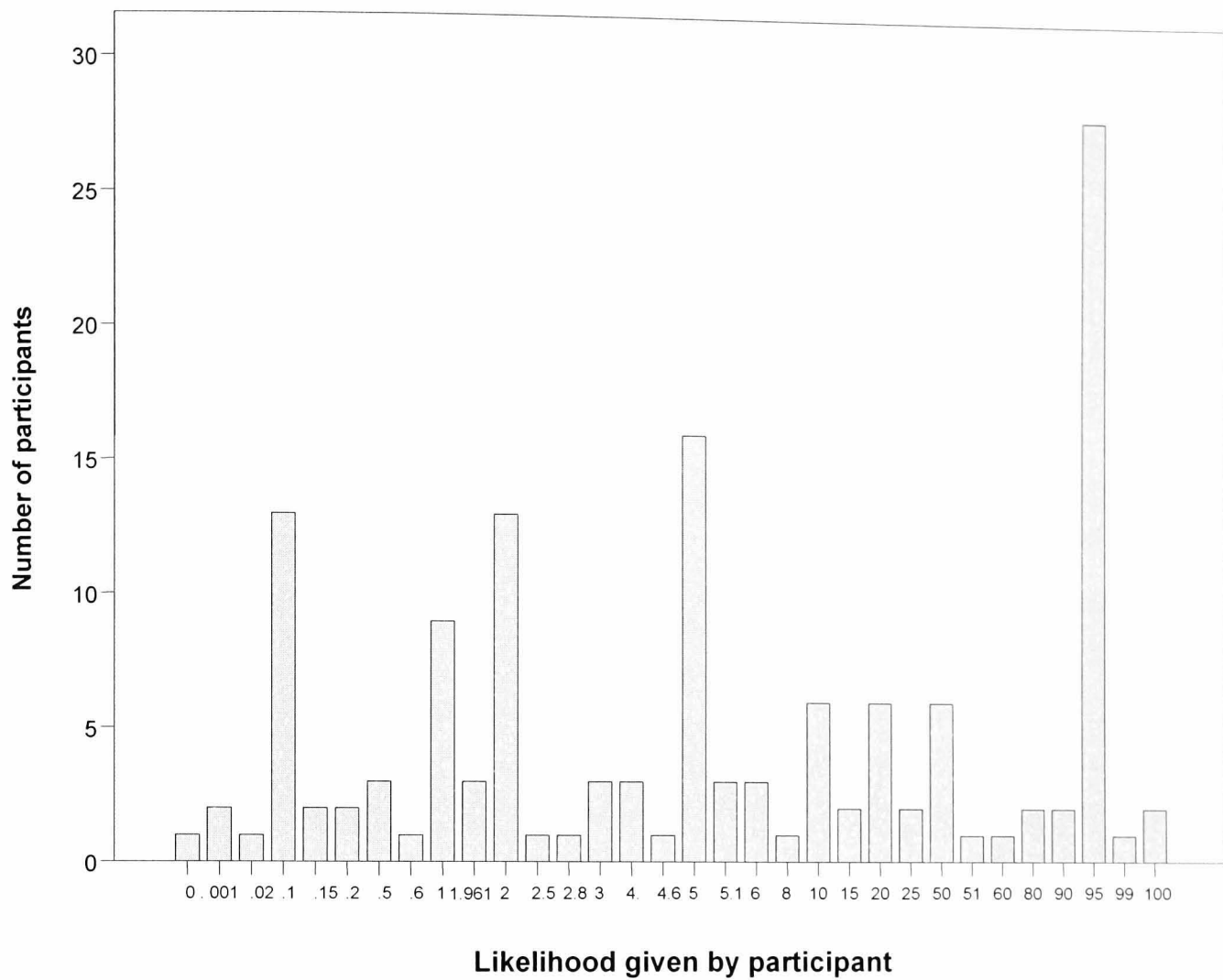
	Mean	Median	Mode	Standard Deviation	N
Disease problem	29.55	5.00	95.00	39.20	141
Cab Problem	47.98	60.00	80.00	31.68	143

Descriptive statistics for both tasks are shown in Table 8.1. The mean, median and mode are all presented in order to illustrate the nature of the data's distribution. It can be seen that while the mean score on the disease task is 30, the median was only 5 – fairly close to the normative value of 1.96 – and the most common response was in fact 95. This was the accuracy of the test described in the task, indicating that the largest group of participants was those who simply took this value and did not attempt any manipulation of the values involved. In other words, this indicates base rate neglect.

The cab task shows a similar pattern, as in this case the most common response was 80, the witness's level of accuracy. The normative answer was 41, and it may be tempting to see the mean of 47 as being indicative of participants responding with something approximating the correct value; however, as the measures of central tendency suggest, each set of scores shows large amounts of variance and the mean is not expected to be a good indicator of the scores' true distribution.

The distributions are illustrated in figures 8.1 and 8.2, overleaf.

Figure 8.1 Distribution of responses to disease task



In the case of the disease task, Figure 8.1 indicates that the modal response of 95 was given by 28 participants (out of 141 who completed the task), while just 3 gave the exact normative answer of 1.96. Allowing for ‘rounding up’, the 13 who suggested a value of 2 out of 100 can also be considered to be successfully manipulating the data to achieve the correct answer, giving a total of 16 participants. This is still many less than those who gave the 95 value, and exactly the same amount as those who gave an answer of 5, which was the false positive value of the diagnostic test described. A further peak is seen at 0.1, the base rate value of people in the general population who have the disease, and also at 1, a value which can perhaps be interpreted as showing some understanding of the calculations involved.

Figure 8.2 Distribution of responses to cab task

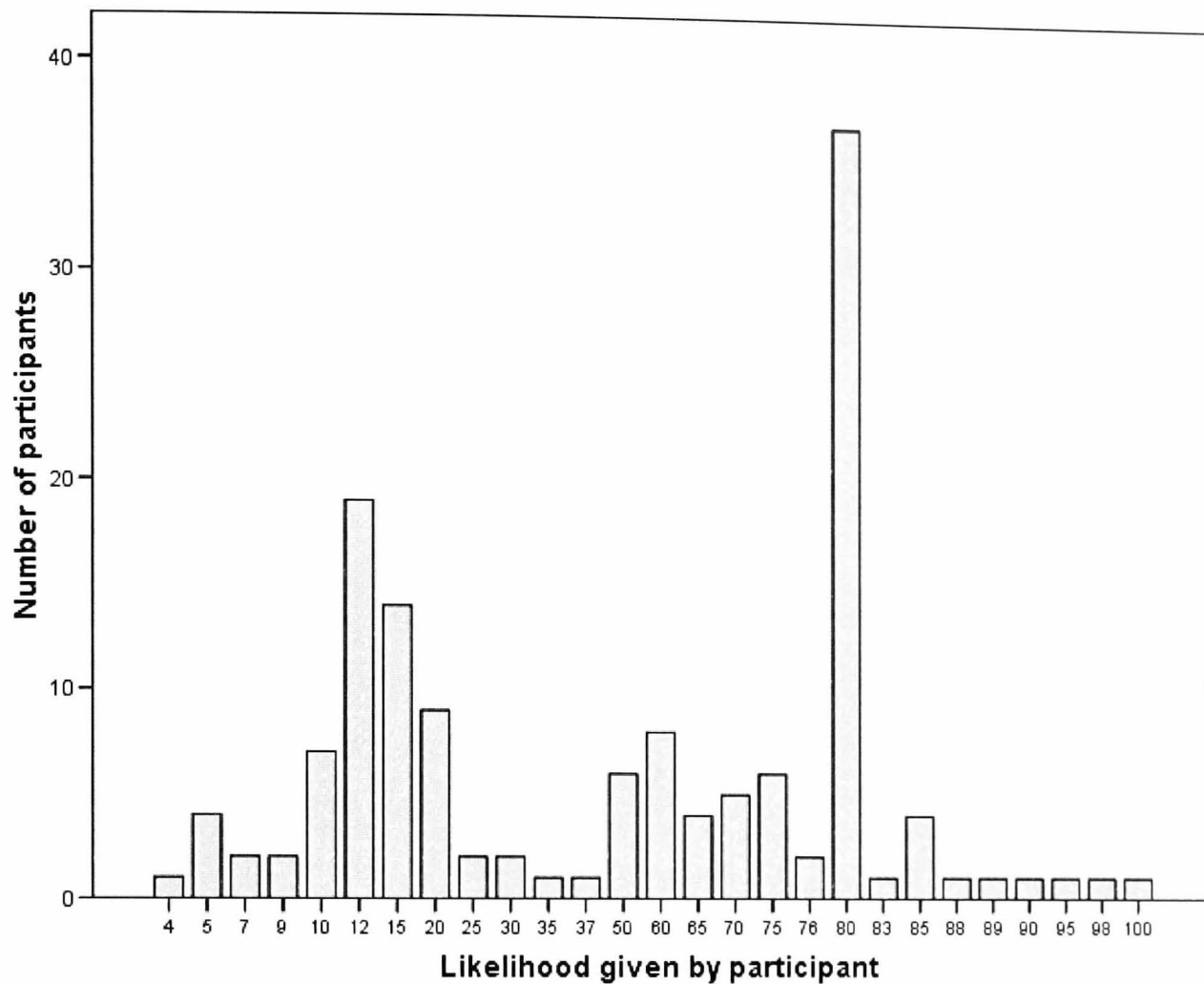


Figure 8.2 illustrates the scores given for the cab task, and in this case 37 (out of 143) participants gave the modal response of 80 (the witness accuracy rate, indicating base rate neglect), while none achieved the normatively correct answer (41). Two other noticeable peaks are at 15 (the base rate number of blue cabs) and 12. The latter is likely to be indicative of participants multiplying the base rate by the witness accuracy rate (.15 by .8).

Discriminant function analysis was conducted on each set of data, in order to examine whether or not the factors of age, task format and thinking styles could predict the response that a participant would give to each task.

The primary focus in choosing the groups to use for this analysis was the intention to identify those who were showing base rate neglect, using the base rate only, successfully achieving the normatively correct answer, or showing some evidence of attempting to manipulate the information given but failing to give the correct answer.

In both data sets the data was very 'noisy' and it was apparent that substantial amounts of data would be lost to this process. However, it was felt that widening the groups – including response groups that were less clearly indicative of any of the categories above – would reduce the benefit of this method of analysis. Furthermore the inclusion of additional responses that were endorsed by only one or two participants adversely impairs the overall fit of the model, and in situations where there are large numbers of predictors can this can produce singularity in the variance covariance matrix preventing a solution from emerging (Tabachnick & Fidell, 2007).

8.3.2 Disease task discriminant analysis

In the case of the disease task, five outcomes were selected. These were:

1. 95, the modal response. This may be a result of taking the false positive rate (5%) away from 100, in the belief that if the test falsely identifies as positive 5% of people without the disease, it is therefore also inaccurate 5% of the time when testing people with the disease. This would be despite having been explicitly informed in the instructions that the test is 100% accurate when testing those with the disease. Nevertheless, a calculation of $P(\text{pos test}|\text{Disease}) - P(\text{pos test}|\text{Not Disease})$ seems to be the most likely explanation. (28 participants fell into this group)
2. 5, the false positive rate (16 participants)
3. .1, the base rate of people within the population who have the disease (13)
4. 1, believed to be indicating some manipulation of both base rate and evidence. It is also possible that these participants found the value of 1 in the phrase 'one out of every 1000 people has the disease' to be particularly salient and mistakenly gave this answer as if they were giving the base rate –that is, saying '1 out of 100' in error, when they had intended '1 out of 1000'. (9)
5. 1.96 through to 2, as being the normatively correct answer (16)

A total of 82 participants were therefore included in this analysis, 58% of those who attempted the task.

Four discriminant functions were revealed (that is, dimensions along which the response groups may differ). The first explained 56.1% of the variance (canonical correlation of .62) while the remaining three accounted for 25.6% (.47), 15.3% (.38) and 3.0% (.18)

respectively. When taken in combination, these four functions were able to significantly differentiate the participants' judgements, (see Table 8.2) Wilk's Lambda = .40, $\chi^2(48) = 65.32$, $p < .05$, but when the first and thereafter any remaining functions were removed it is clear that the remaining functions did not discriminate significantly between the groups.

Table 8.2: λ and χ^2 values on disease task

Function(s)	λ	Df	χ^2
1 through 4	.40	48	65.32*
2 through 4	.65	33	31.05
3 through 4	.83	20	13.40
4	.97	9	2.32

* $p < .05$; ** $p < .01$; *** $p < .001$

The individual predictors which have the strongest relationship with the outcome can be examined by looking at Tables 8.3 and 8.4, containing all predictors' standardised canonical discriminant function coefficients and correlation coefficients respectively. The standardised canonical discriminant function coefficients indicate the size and direction of the relationship between each variable and each function. So in Table 8.3 it can be seen that at 1.11, the coefficient for CT and Function 1 shows the greatest relationship, indicating that an increase of one standard deviation in CT would lead to an increase in the factor score by 1.11. For each individual, the total factor score is the sum of all of these effects (i.e., the sum of the individual's standardised score on each predictor multiplied by the standardised canonical discriminant function coefficient for that predictor). The individual's predicted group is then that which has been assigned the value closest to this summed value. A strong relationship can also be seen between Function 1 and FT and Format, and with A and D but in the opposite direction to the first two.

While CT showed the greatest standardised canonical discrimination function coefficient, at 1.11 (for Function 1), it can be seen that task format had the most consistently large relationship across each of the four functions, and particularly Function 2. Figure 8.4 illustrates the way in which the responses are distributed along the two first functions, or dimensions, which (as detailed above) account for a total of 81.7 of the variance on the outcome/grouping variable. It can be seen that Function 1 (and therefore potentially CT) discriminated between those responding with the base

rate (centroid 3, with each individual case being illustrated by a Δ) and those giving a judgement of 1 (centroid 4, each case illustrated by \diamond). Function 2 appeared to discriminate between those giving an answer of 5, and those giving an answer of 95, but this was not to a significant extent. The figure also shows that there was a large amount of overlap between the groups, although this is often the case in such analysis.

Table 8.3: Standardised canonical discriminant function coefficients for disease task

	Function			
	1	2	3	4
Age group	-.20	.23	-.47	.28
Frequency or Probability format	.63	.63	.31	.38
FI – Faith In Intuition	-.13	.05	-.01	.25
NFC – Need For Cognition	-.35	.17	.28	.68
FT – Flexible Thinking	.63	-.02	-.39	.29
A – Absolutism	-.60	.38	-.17	.32
D – Dogmatism	-.73	.29	.17	-.07
CT – Categorical Thinking	1.11	-.33	-.06	-.15
ST/LC – Superstitious Thinking/Luck Composite	.09	-.33	.50	.71
sfNFC – short form Need For Cognition	-.45	-.08	.09	.74
SD – Social Desirability Subscale	-.35	.02	-.57	.38
BI – Belief Identification	.44	-.54	.34	.07

Table 8.4: Correlation coefficients for disease task

Predictor	Function			
	1	2	3	4
CT – Categorical thinking	.41	-.31	-.03	-.08
age group	.18	-.05	.07	.11
frequency or probability format	.43	.75	.25	.26
BI – belief identification	.09	-.44	.41	.15
FT – Flexible thinking	.12	.13	-.57	.07
sfNFC – Short form need for cognition	.00	-.23	-.48	.31
ST/LC – Superstitious Thinking/Luck composite	.07	-.36	.45	.36
NFC – Need for Cognition	-.24	.15	.43	.16
SD – social desirability	-.00	-.16	-.42	.39
FI – Faith in Intuition	-.15	.19	-.25	.04
D – Dogmatism	-.00	.06	.29	-.33
A – Absolutism	-.13	.05	.01	.17

A further stepwise analysis was conducted, examining the effect of each predictor on the groups. The only significant predictor was the task format, $F(4,76)=5.33$, $p<.001$. As illustrated in Figure 8.3, the group with the highest proportion of those in the frequency format was group 3, those responding with the base rate value only. The group with the second highest proportion of participants in the frequency condition was group 2, those giving the false positive rate (at 68.75% of participants). Of the two groups showing the lowest proportion of participants in the frequency format, group 4 (a probability value of 1, suggestive of some manipulation) had the least, at just 11.11%, while group 1, those giving a value of 5 (possibly also suggesting data manipulation, as described above) contained 25.93 percent frequency format versions. Perhaps most strikingly, those getting the correct answer, group 5, were almost equally likely to have been presented with the frequency or the probability format, with 56.25% of them having completed the frequency version.

Figure 8.3 Percentage of participants completing the disease task in the frequency condition, by response group

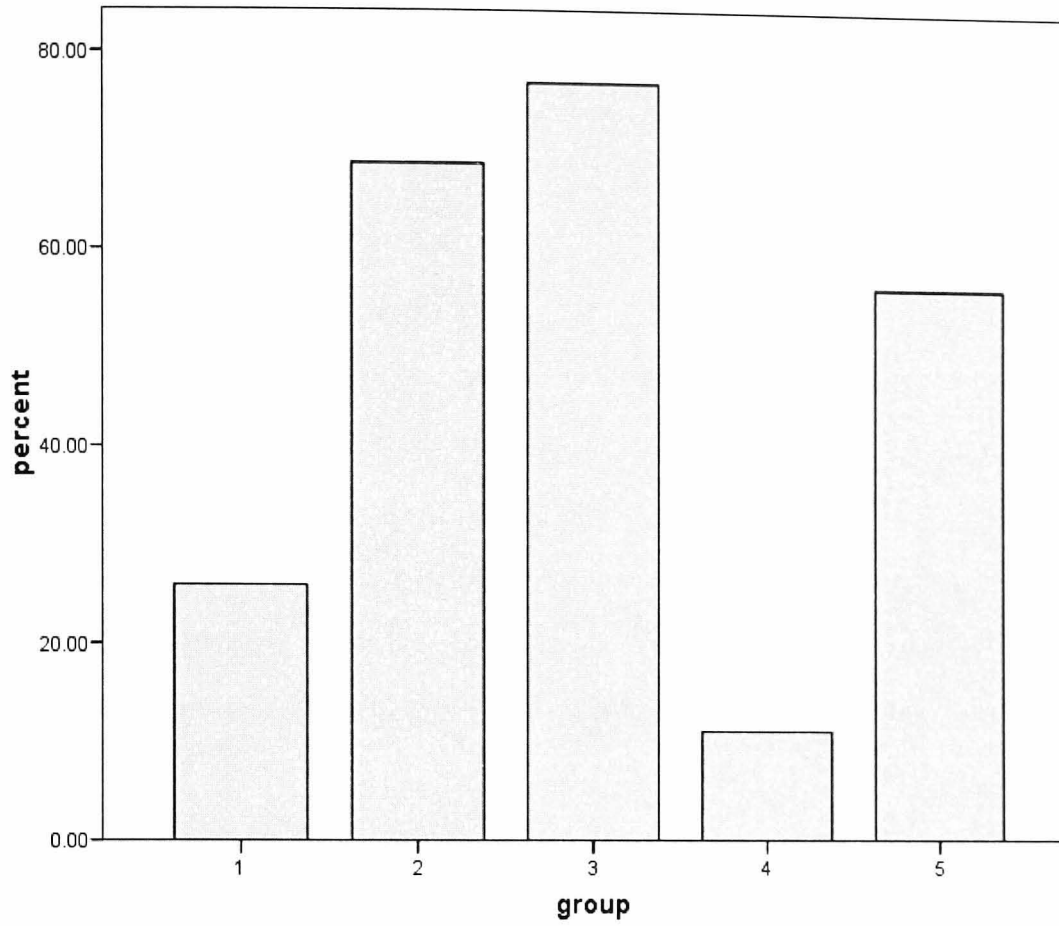


Figure 8.4 Disease task combined groups plot

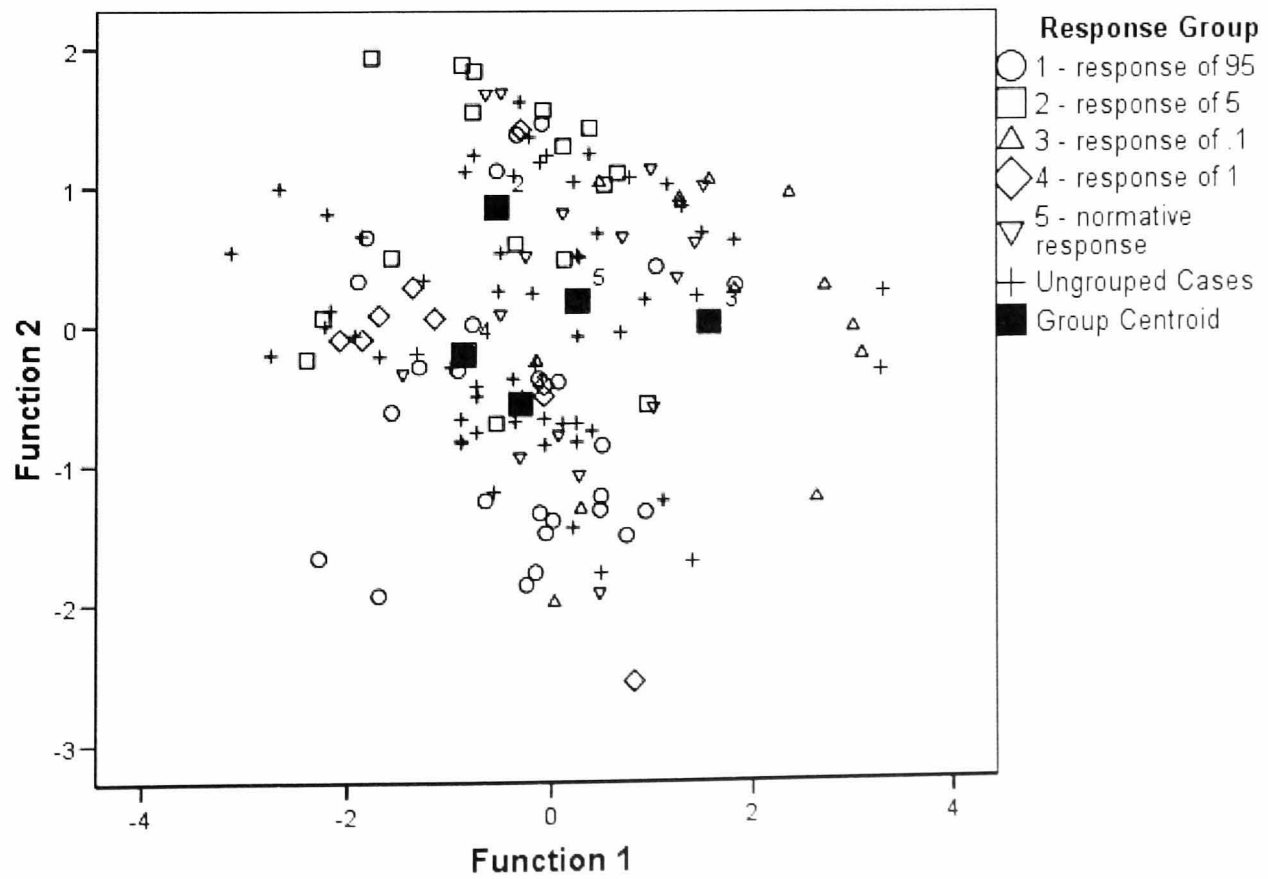


Table 8.5: Actual and predicted group memberships for disease task

		Predicted Group Membership %					Total N
		1	2	3	4	5	
Actual Group	1	63.0	18.5	.0	11.1	7.4	27
	2	18.8	68.8	.0	6.3	6.3	16
	3	15.4	7.7	61.5	7.7	7.7	13
	4	22.2	.0	.0	66.7	11.1	9
	5	37.5	25.0	12.5	.0	25.0	16

A total of 56.8% of the participants in this analysis were correctly classified as giving one of the five responses. Four of the five groups were correctly classified over 60% of the time, the highest being those who gave a value of 5, at 68.8%, while the remaining group, those who actually gave the normative answer, were the least well predicted by this model, with only 25% being correctly classified. 37.5% of the 16 people who gave the correct response were incorrectly predicted to give a value of 95 (group 1).

8.3.3 Cab task discriminant analysis

For the cab task, only three outcomes were used (see Figure 8.2). These were:

1. 80, the modal response and the accuracy of the witness (37 participants fell into this group)
2. 15, the base rate of blue cabs (13 participants)
3. 12, the product of .80 and .15 (17)

A total of 67 participants fitted into these four categories, 43% of those who completed the task.

The normatively correct answer was 41, and no participant gave this response. A value of 40 would also have been accepted as indicative of normative reasoning, but again no one gave this value.

Two discriminant functions were revealed, the first accounting for a vast majority of the variance, at 77.5%, with a canonical correlation of .63. The second function's 22.5% had a canonical correlation of .40.

The two functions together were able to significantly differentiate the participants' judgements, Wilk's Lambda .51, $\chi^2(24) = 39.96$, $p < .05$, while the second alone was not significant ($\lambda = .84$, $\chi^2(11) = 10.26$, $p > .05$).

Again, the individual predictors which had the strongest relationship with the outcome can be examined by looking at Tables 8.6 and 8.7, containing all predictors' standardised canonical discriminant function coefficients and correlation coefficients respectively. The strongest predictor here is Flexible Thinking, with the discriminant function coefficients indicating that it discriminated equally well on Function 1 and 2, at .54. Figure 8.6 shows us that Function 1 discriminated between groups 1 (a response of 80) and 3 (the product of .80 and .15), and between groups 3 and 2 (response of 15) to almost the same extent, while function 2 discriminated most clearly, although not significantly, between 1 and 2 (base rate only) but may also have been able to discriminate between groups 1 and 3. As such, if FT was a significant predictor, it would be discriminating between all 3 of the groups used in the analysis. Format was again a strong predictor, on Function 1 only. This is illustrated in Figure 8.6, below, which indicates the clear difference between the group centroids of groups 1 and 3, but again also indicates a large amount of overlap between the groups.

Table 8.6: Standardized canonical discriminant function coefficients for cab task

	Function	
	1	2
age group	-.02	.26
frequency or probability format	1.06	-.10
FI – Faith In Intuition	-.09	.47
NFC – Need For Cognition	.15	-.21
FT – Flexible Thinking	.54	.54
A – Absolutism	.01	-.33
D – Dogmatism	-.11	-.15
CT – Categorical Thinking	.38	.11
ST/LC – Superstitious Thinking/Luck Composite	.21	.63
sfNFC – short form Need For Cognition	-.10	.25
SD – Social Desirability Subscale	-.25	-.12
BI – Belief Identification	-.27	.43

Table 8.7: Correlation coefficients for cab task

	Function	
	1	2
frequency or probability format	.83	-.12
SD – social desirability	-.18	-.14
BI – belief identification	-.16	.05
CT – Categorical thinking	-.05	-.04
age group	.16	.53
sfNFC – Short form need for cognition	.02	.46
A - Absolutism	-.04	-.43
NFC – Need for Cognition	-.13	-.42
FT – Flexible thinking	.19	.33
ST/LC – Superstitious Thinking/Luck Composite	-.05	.32
FI – Faith in Intuition	-.02	.17
D - Dogmatism	.11	-.11

Stepwise analysis again showed that format of task was the only significant predictor, $F(2,64)=14.48$, $p<.001$, and Figure 8.5 indicates this is in part because 100% of those in group 3 (the product of base rate and witness accuracy rate) completed the frequency format. A slight minority of those in groups 1 and 2 were in the frequency condition, at 35.14% and 38.46% respectively. This is different from the results of the disease task, in which a majority of those in the base rate neglect/base rate only groups were in the frequency format.

Figure 8.5 Percentage of participants completing the cab task in the frequency condition, by response group

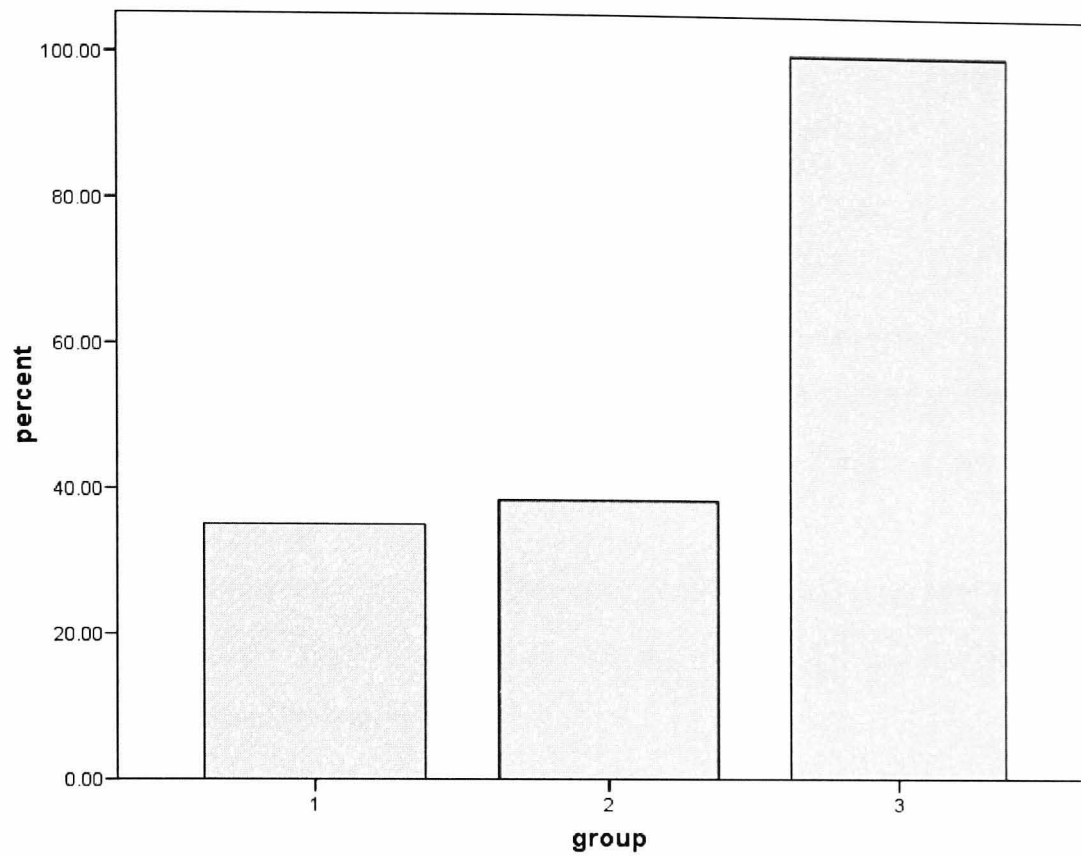


Figure 8.6 Cab task combined groups plot

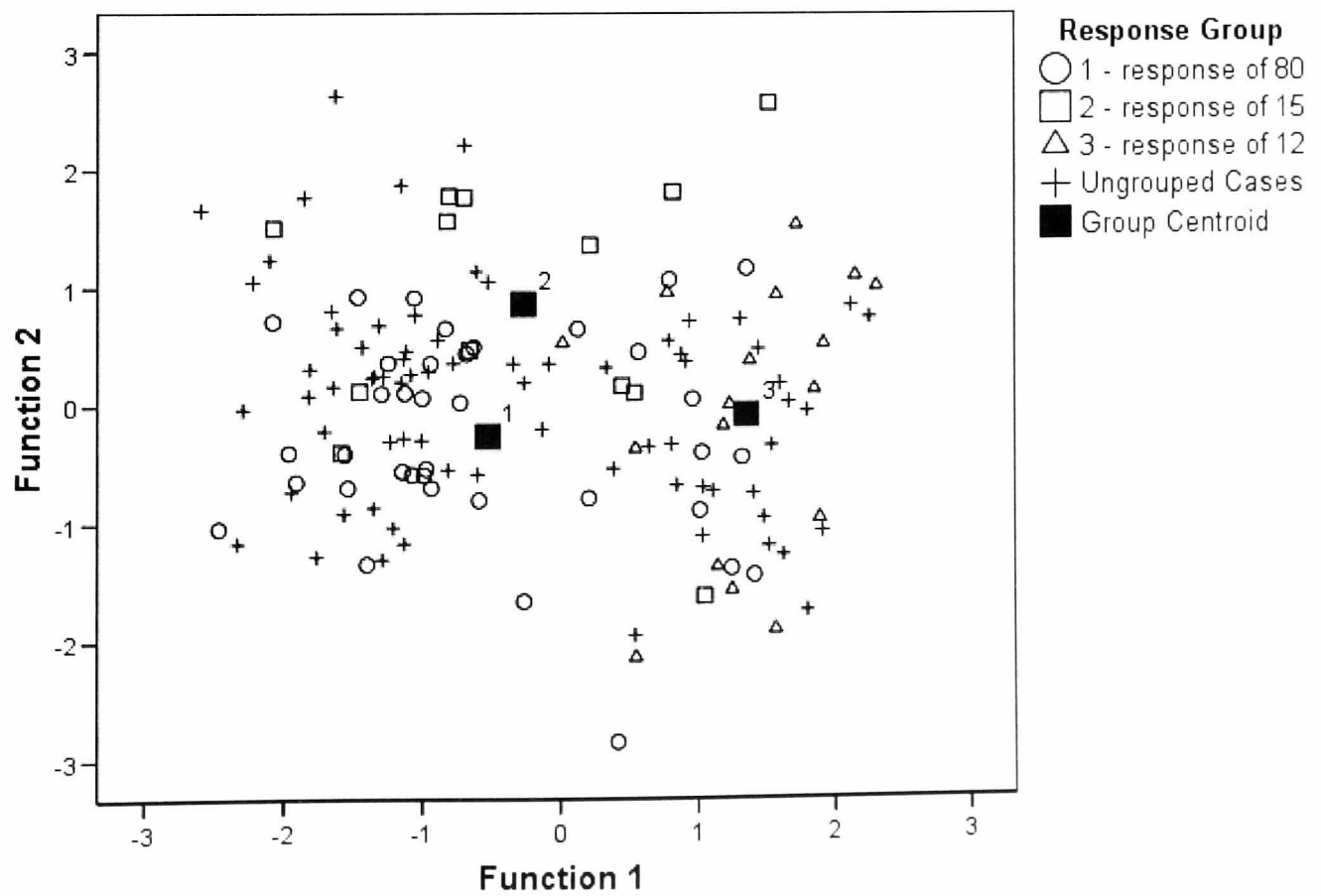


Table 8.8: Actual and predicted group memberships for cab task

		Predicted Group Membership %			Total N
		1.00	2.00	3.00	
Actual Group	1.00	75.7	.0	24.3	37
	2.00	46.2	38.5	15.4	13
	3.00	11.8	.0	88.2	17

A total of 71.6% of cases were correctly classified in this case, a greater number than in the case of the disease task. The model is particularly accurate in the case of those who gave 12 as their response, correctly predicting 88.2% of these cases.

Finally, Chi squared analysis indicates that in neither task was age group significantly associated with task response, at $\chi^2(4) = 1.66$ $p > .05$ and $\chi^2(2) = 2.80$ $p > .05$ for the disease and cab task respectively.

8.4 Discussion

As anticipated and suggested by previous research, participants did not appear to be responding to the tasks by using Bayesian reasoning, as only a minority achieved the normative answer in the disease task, and no participant did so in the cab task. Response rates were quite different to those found by Birnbaum (2004) and Hinsz *et al.* (2008), as the current participants showed lower incidences of each of the expected fallacies (base neglect and base rate only) and greater levels of ‘noise’ in the data – on each task the answers were spread across the entire possible response range. For many of these answers, it was impossible to infer any particular reasoning method.

The predictors of format, gender and thinking styles were able to predict a total of 57% of outcome groups in the disease X problem, and 72% in the taxi cab problem (however, see the following paragraph). In each case, format was the strongest predictor of the type of response, but not in terms of either format actually corresponding to greater incidences of the normative answer. In the disease problem, the frequency format was associated with the false positive rate and the base rate only answers, while the probability format was associated with base rate neglect and a response of 1. This latter value, although not the correct answer, is deemed to indicate some manipulation of the data (although it is also possible that it indicates a gross misunderstanding of the

base rate, as suggested in 8.3.2), and an appreciation of how unlikely the disease is, given the available information. In the cab problem, the frequency format was associated with an averaging model (base rate x evidence) while the probability condition was associated with both base rate only and base rate neglect as responses. These findings directly contradict those reported by Evans *et al.* (2000), who found, using versions of the ‘disease X’ task, that the frequency condition was associated with greater incidences of participants responding with the base rate only.

The values given above of 57% of responses in the disease task, and 72% in the cab task, should be considered in the context of the data that was ‘missing’ from the analysis. The nature of DFA is that the outcome variable must be a nominal variable, and as such the data collected here was categorised into groups as explained in detail in the results section. This resulted in a large proportion of the data (57% and 42% respectively) being omitted from the analysis as it did not fit into any of the groups created. Therefore the true number of correctly categorised responses could be considered to be 32.9% for the disease task and 30.89% for the cab task, approximately a third of the original data in each case.

One suggestion for the influence of format would be that the different formats could be highlighting different cues within the problem as being particularly salient. Gigerenzer and Goldstein’s ‘fast and frugal’ heuristic (1996) is such that only this one most salient cue would then be used to make a judgement, rather than going on to integrate any other cues within the problem. If this is the case, then there does not appear to be a pattern across the two tasks. It is not that the frequency format is consistently highlighting either base rate or the individuating evidence, for example, and neither does it appear to be consistently priming participants to attempt greater manipulations of the data. In the case of the disease task, those attempting any form of manipulation (the value of 5, of 1, or the correct answer of approximately 2) were either no more likely, or actually less likely to be in the frequency format, although in the cab task those attempting manipulation and calculating the product of base rate and evidence (giving a value of 12) were exclusively within the frequency group. It may be, however, that the frequency format is priming manipulations of which the participants are not in fact capable, leading to a wide range of non-systematic wrong answers, which were not covered in the current analysis.

In parallel with previous findings throughout this research, age group was not a strong predictor in either case. As well as neither age group performing 'better', in terms of being more likely to give the normative answer, they were actually not performing differently at all, as indicated by the χ^2 analysis.

Regarding thinking styles, there was a lack of consistency across the two tasks, with Categorical Thinking emerging as an important predictor of group outcome in the disease task, while only Flexible Thinking emerged as a predictor in the cab task. As with task format, the thinking styles did not differentiate those who got the normative answer from those who did not, but instead seemed to discriminate between the different types of 'wrong' answer. In the former task, CT did potentially discriminate between those giving the base rate only and those giving a judgement of one, while in the latter FT appears to discriminate between those giving either the base rate or the witness reliability weight, and those giving an answer that was the product of the two. As such, there is some evidence that thinking styles may contribute to whether participants give only a value that was present in the vignette, or attempt some form of manipulation. However, neither of these two thinking styles was found to be a significant predictor.

Within the disease problem, format did load onto the same function as many of the thinking styles – Categorical Thinking, Dogmatism, Absolutism and Flexible Thinking. This would suggest that they may all map onto an inter-related factor, presenting a link between task format and thinking styles which has so far been elusive in this research. This link was not apparent in the cab task.

Although there were large numbers of responses to each task that could not be categorised, there were nevertheless clear clusters of responses that were as expected from previous literature – i.e. base rate neglect, positive/false positive rates and some form of calculation (Birnbaum, 2004). There is also evidence that participants appreciated the importance of the false positive, but were unable to use it to achieve a normative answer, instead giving that value as their response. Krynski and Tenenbaum (2007) also suggest that participants are aware of the relevance of this value, but use it inappropriately, resulting in base rate neglect. However, the predictions made for this study, that an analytic thinking style, young age, and the frequency format would each be associated with either the normative response, or one that suggested some

manipulation of the various cues, were not supported. While the models presented here do accurately predict group membership in some cases, there is not sufficient evidence to support any of the presented hypotheses.

The accuracy in the cab task was exceptionally poor. In this task, the response value that indicated manipulation of both base rate and case evidence was correctly categorised by the discriminant function analysis in 88% of cases, but with no normative responses given it was, of course, not possible to include them in the model. In the previous chapter, it was suggested that such tasks might just be too difficult for participants, and while this may well have been the case here the use of discriminant function analysis has allowed us to look more specifically at the different types of responses, rather than being reliant only on whether responses are normative or not.

The current study had used normalised frequencies, which may not make set structures as clear as those presented with natural reference classes (Hoffrage *et al.* 2002). For the reasons set out in the introduction to the chapter that follows, this emerged as being a potentially serious limitation. Thus before a more definitive statement can be made regarding the effect of problem format, it is necessary to consider whether presenting the information in terms of natural frequencies would lead to substantially different results. Therefore the following – and final – empirical chapter will attempt to elicit greater levels of analytical reasoning by presenting problems in natural frequency formats.

Chapter 9 – Bayesian Tasks With Natural Frequencies

9.1 Introduction

In the previous study, participants from two age groups (a young group with a mean age of 20 years, an older group with a mean age of 71) were given Bayesian reasoning tasks in one of two formats: probability or frequency. It was anticipated that these two factors, along with thinking styles and age group would be able to predict the type of responses given to the task. As anticipated, many participants did respond by either showing ‘base rate neglect’ (see Birnbaum 2004, but also De Neys & Glumicic, 2007; Johansen, Fouquet & Shanks, 2007; Franssens & De Neys, 2008), base rate only (ignoring the additional evidence, *ibid*) or some other response suggesting an attempt to integrate both values (such as a straight forward multiplication). However, the data collected was very ‘noisy’, with many participants giving responses that appeared to be quite random, rather than falling into the groups predicted by, for instance Birnbaum (2004) and Chapman and Liu (2009). As a result, only around half of the data collected was actually used in the analysis, leading to greatly reduced power. The rate of correct responses was also very low when compared with previous studies (e.g. Birnbaum, 2004; Gigerzner & Hoffrage, 1995). While the number of people getting the task correct does not affect the predictive power of the analyses, it may indicate that our tasks were simply too difficult for our group, which may have led to lower motivation and less engagement on the part of our participants.

One of the main reasons why the frequency format did not have a facilitating effect, in terms of participants in that condition being more likely to achieve the correct answer, is thought to be the use of normalised, rather than natural frequencies. To illustrate this distinction, Hoffrage *et al.* (2002) present the following examples:

“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 *et al.*, 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 25%.

Both the standard probability format, and the normalised frequency format (as used in the previous chapter) 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:

$$P(H|D) = \frac{P(H) \times P(D|H)}{P(H) \times P(D|H) + P(\text{not } H) \times P(D|\text{not } H)}$$

Gigerenzer and Hoffrage (1995) and Mellers and McGraw (1999) illustrate how the natural frequency format no longer requires that you use this full calculation, as the values presented already contain information about the base rate – they give you the information *given* the base rate. As such, the formula becomes (with lower case letters indicating ‘data and hypothesis’):

$$P(H|D) = \frac{d\&h}{d\&h + d\¬h}$$

In other words, the number showing both a positive test result and actually having the disease, 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, but also the 120 who showed positive while not having the disease) and to use the calculation $30/150$. Relative, that is, to attempting to find the same data from the normalised frequencies, which requires you 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.

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, as in the previous chapter. 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.*).

This being the case, natural frequencies may be particularly advantageous to older participants who have greater experience than their younger counterparts of (probably unconsciously) assimilating data regarding probabilities in day to day life. Previous research has suggested that age related detriments in reasoning are often elicited by the addition of cognitive load tasks (Mutter, 2000). With natural sampling giving participants clearly stated values that already contain base rate information, this may reduce the cognitive load on older participants, enabling them to use system 2 processes (which are thought to be primed by the frequency format, Kahneman & Frederick, 2002; Sloman, 2003) more effectively.

Peters *et al.*, (2000) suggest that base rate neglect found to be more common in 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 lead to answers that indicate greater attempts to reason, 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 two conditions, as they would not benefit from the natural frequency information in the same way as the younger participants.

Younger participants are also expected to benefit from natural frequencies, as per previous research (for instance Brase, 2008; Gigerenzer & Hoffrage, 1999), but it is anticipated that this benefit will be less than for the older participants, as they are already able to deal with the cognitive load.

It is therefore anticipated that by using natural frequencies instead of normalised ones the tasks will be more approachable, whether by making set structure clear and/or making the base rate information easier to deal with. As such, the hypotheses for the current study are the same as those that were previously unsupported in chapter 8, with only the small addition italicised in hypothesis 3.

9.1.1 Hypothesis

1. That the way in which participants respond – whether with the correct answer or one of the common errors described by Birnbaum, above (2004) – will vary according to thinking style. It is anticipated that those with an analytical thinking style will be more likely to give a response that indicates an awareness of each of the relevant pieces of information, and/or an attempt to manipulate these to reach the normative answer.
2. Older participants will be less likely to give such responses, their inability to integrate information making them more likely to show clear base rate neglect or base rate only responses. While it is acknowledged that their greater verbal intelligence may be an advantage, it is anticipated that this will be outweighed by their slower processing speed.
3. Those receiving problems posed in the natural frequency format will be more likely to achieve the correct result *and/or to respond with a value indicative of having some understanding of the set structures involved.*

9.2 Method

9.2.1 Design and statistical analyses

As in the previous study, there were two dichotomous between participant factors of age (old/young) and format (frequency/probability). Further independent variables were thinking style, verbal intelligence and information processing speed, and the two dependent variables were the participants' responses to each of the two reasoning tasks. Again, the dispersion of scores was unique to each task and as such they were analysed separately. Discriminant function analysis was used, with age, format and relevant thinking styles as predictors and the participants' raw scores coded into groups according to which cue(s) was apparently utilised. In the case of the cab task, no measure of individual difference was related to response, and as such a chi square

analysis was used. Analyses of variance were also used, to examine whether the between participant factors of age group and task format affected responses given to each task. For full details see the results section of this study as well as the details of the discriminant analytical method within sections 8.2 and 8.3.

9.2.2 Participants

51 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.

41 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.

9.2.3 Materials

9.2.3.1 Measures of individual differences

All participants completed the Rational Experiential Inventory, consisting of the Need for Cognition and Faith in Intuition scales (Cacioppo & Petty, 1982; Epstein *et al.*, 1996), the Thinking Disposition scale (Kokis *et al.*, 2002), the Mill Hill Vocabulary Scale and a measure of Information Processing Speed. These are all described in detail in Chapters 6 and 7.

9.2.3.2 Bayesian problems

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).

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 _____

This new natural frequency format is modelled on that used by Gigerenzer and Hoffrage in their research (e.g. Gigerenzer & 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. Gigerenzer and Hoffrage (1995) suggest that this is one of the key benefits of the natural frequency format.

Again, bold is used to emphasise the differences in the tasks.

9.2.4 Procedure

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

9.3 Results

The descriptive statistics for the background variables are presented below in Table 9.1. As in the previous sample, young participants do show significantly greater processing speed, as well as significantly lower scores on the Mill Hill Vocabulary task. The difference between the years of education are again indicative of the differences in statutory school attendance requirements.

Table 9.1: Mean scores on background measures

	Young Mean (SD)	Old Mean (SD)	(df)t
Years of Education	15.43 (2.33)	13.76 (3.46)	(84.17) 2.71*
MHVS	16.20 (3.64)	22.80 (5.34)	(87.85) -7.04***
Information Processing Speed	115.29 (21.48)	70.41 (15.33)	(90) 11.68***

* $p < .05$; ** $p < .01$; *** $p < .001$

Note: Levene's test for equality of variances was significant for Years of Education and MHVS, so the values given here are for unequal variances.

With regards to thinking styles, as in the previous age comparison (see Table 7.2) older participants showed significantly lower Need For Cognition, $t(90)=2.67$, $p < .01$, but stronger levels of Absolutism, $t(87.01)=4.02$, $p < .001$, and Social Desirability, $t(90)=3.42$, $p < .001$ (remembering that for the TDQ subscales high values = weak levels, and vice versa). In this sample the old group also showed significantly stronger levels of Categorical Thinking than the young group, $t(90)=3.36$, $p < .01$. In the previous study there had been no difference. The other differences found in the previous sample – with younger participants showing higher Faith in Intuition, stronger Dogmatism and stronger Superstitious thinking, are not present in the current study.

The only thinking style to correlate significantly with either reasoning task in either linear or quadratic curve estimations was Superstitious Thinking, which showed a positive relationship of $R^2 = .09$ in each case (therefore predicting just 9% of the variance) with responses to the disease tasks. See Appendix 14 for all correlations.

Table 9.2: Thinking style mean scores by age group

Scale	Subscale	Young Mean (SD)	Old Mean (SD)	Min/max Possible score	(df) t value	All Data Mean (SD)
REI – Rational Experiential Inventory	FI – Faith in Intuition NFC – Need for Cognition	39.92 (7.42)	42.97 (7.50)	12/60 19/95	(90) -1.944 (90) 2.67**	41.61 (7.58) 65.52 (11.44)
TDQ – Thinking Dispositions Questionnaire	FT – Flexible thinking A - Absolutism D - Dogmatism CT – Categorical thinking ST/LC – Superstitious Thinking/Luck Composite sfNFC – Short form need for cognition SD – social desirability BI – belief identification	17.58 (3.03) 15.18 (1.69) 14.78 (2.21) 9.88 (2.14) 24.93 (4.52) 18.17 (19.25) 14.07 (1.77) 16.73 (2.19)	18.22 (3.56) 13.40 (2.55) 14.36 (2.57) 8.45 (1.93) 26.52 (3.65) 19.25 (3.63) 12.75 (1.90) 16.33 (2.76)	10/40 5/20 6/24 3/12 8/32 5/20 6/24	(80) -.922 (87.01) 4.02*** (90) .84 (90) 3.36** (90) -1.87 (90) 3.42*** (89.99) .77	17.93 (3.33) 14.19 (2.37) 14.55 (2.41) 9.09 (2.14) 25.81 (4.12) 18.77 (3.89) 13.34 (1.95) 16.51 (2.52)

* p<.05; ** p<.01; *** p<.001

Note: Levene's test for equality of variances was significant for Absolutism and Belief Identification, so the values given here are for unequal variances.

9.3.1 Descriptive Statistics for Bayesian Tasks

Table 9.3: Descriptive statistics for participants' judgments

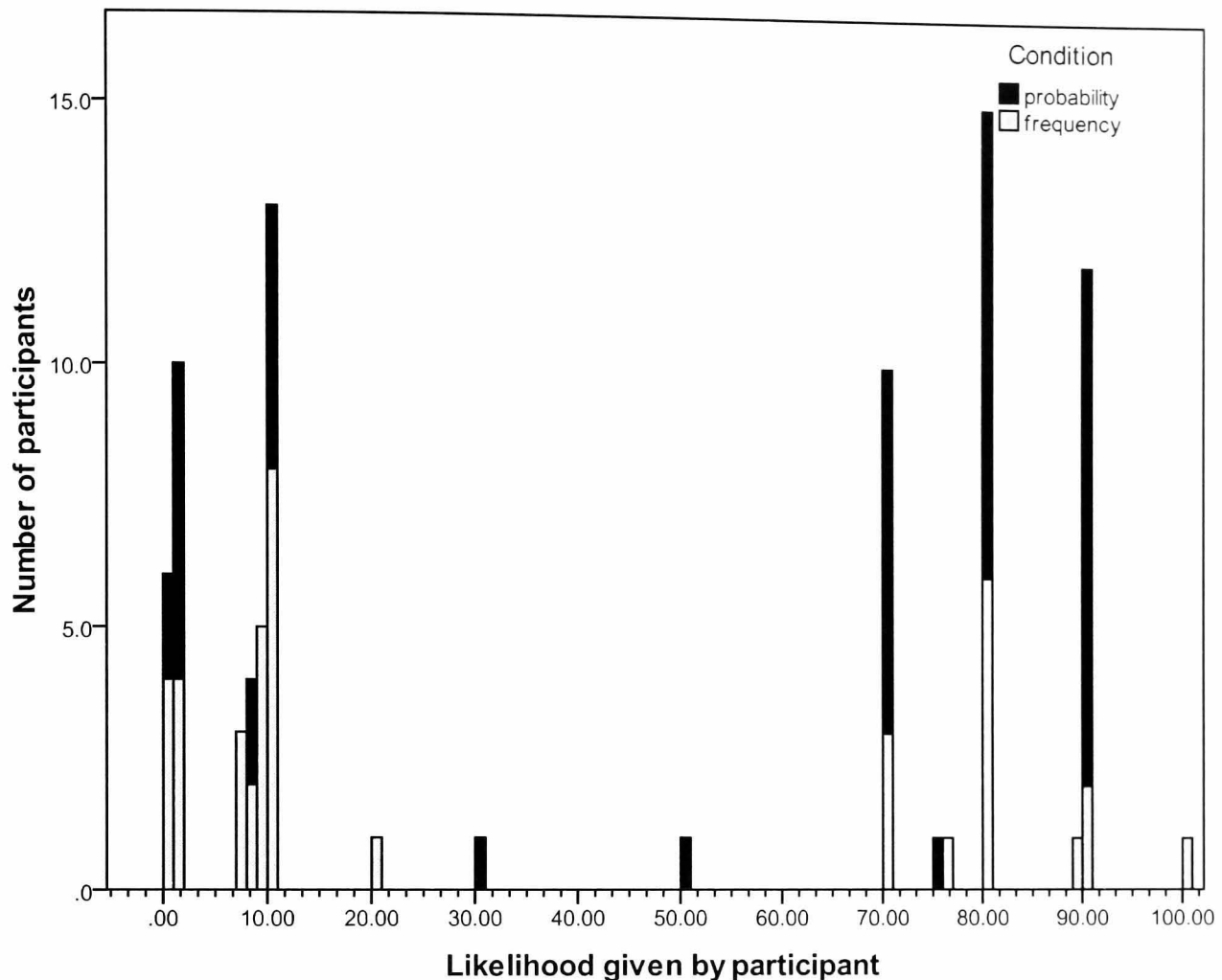
	Mean	Median	Mode	Standard Deviation	N
Disease problem	43.14	30.00	80.00	37.69	85
Cab Problem	43.04	29.00	80.00	30.37	89

Descriptive statistics for both tasks are shown in Table 9.3. It can be seen that the values are incredibly similar across both tasks, despite the counterbalancing used, and participants being asked (and observed) not to return to the previous task once they had turned the page and seen the next one. Regarding the identical mode values of 80, the value itself does appear in the probability versions of both tasks.

Again, the measures of central tendency and of dispersion indicate a skewed data set, and despite the cab problem's mean of 43.04 being very close to the normative, 41.38, this does not suggest that a large number of people actually gave answers of this magnitude. The distributions of both response sets are illustrated in figures 9.1 and 9.2, overleaf.

9.3.2 Full analysis of disease task

Figure 9.1 Distribution of responses to disease task



In the disease task data presented in Figure 9.1, it can be seen that there are two distinct clusters – those who thought that the likelihood of having the disease was 15% or less and those who thought it was 70% or more. The majority of the latter were in the probability condition, while the majority of the former were in the frequency condition. Fifteen participants gave the modal response of 80, the probability of a person who has the disease also having a positive test.

Three participants got exactly the right answer, 7.48% (expressed as 8/107 in the frequency version) while 4 gave a value of 8, which will not be accepted as being indicative of ‘rounding’ as the value of 8 was provided in frequency version the task itself – it is the number of individuals who have the disease and also test positive.

There were four further common responses. The base rate (1%) and the false positive rate of the test (10%) are relatively self explanatory, and to be expected from previous

research. Another fairly common response was 90%, which is likely to be a result of concluding that if the test has a 10% false positive rate, then it must also be 90% accurate – therefore, mistakenly concluding that the answer must be 90%. The last group is a little more surprising, a value of 70%. This is perhaps from the value of Positive|disease, minus the value of positive|healthy.

To summarise, the groups identified are as follows:

1. 1 – the base rate (10 cases)
2. 7.48 – normative (3 cases)
3. 10 – positive|healthy (11)
4. 70 – indicative of some calculation (10)
5. 80 – positive|disease (15)
6. 90 – indicative of some calculation (12)

This would include a total of 61 participants, 72% of those who attempted the task. Clearly, this is already an improvement on the previous study, when only 58% of participants were included in the analysis. To further improve the analysis, a final 7th group was created, for all responses that could not confidently be classified (this group is labelled as 0).

A further refinement was made to the analysis, in that only the thinking style measure that predicted any variance in reasoning performance, Superstitious Thinking, was included as a predictor.

Three discriminant functions were revealed, with the first accounting for 61.7% of the variance (canonical correlation of .50). The remaining two functions accounted for 30.1% (.37) and just 8.2% (.21) respectively. In combination, the three functions did significantly differentiate the responses to the task (see Table 9.4), Wilk's Lamda = .62, $\chi^2(18) = 37.89$, $p < .01$. This does show greater discriminative power than the previous 'disease tasks', (Wilk's Lamda = .40, $\chi^2(48) = 65.32$, $p < .05$, see Chapter 8 for details). When the first function is removed, the remaining two do not discriminate significantly between the response groups (either singly or together).

Table 9.4: λ and χ^2 values on disease task

Function(s)	λ	Df	χ^2
1 through 3	.62	18	37.89**
2 through 3	.82	10	15.25
3	.96	4	3.41

* $p < .05$; ** $p < .01$; *** $p < .001$

To find the individual variable with the best ability to predict the outcome of disease task response, Tables 9.3 and 9.4 present the standardised canonical discriminant function coefficients and correlation coefficients respectively. The correlation coefficients show that for Function 1, the only significant function, it is age group which shows the highest loading, at $-.71$. Figure 9.2 shows the distribution of the responses across the first 2 functions, and it can be seen that Function 1 discriminates between those responding with the base rate only, represented by a small square, and those giving the response of 90, represented by a rectangle, indicative of some attempt at calculation. The negative relationship between age group and Function 1 suggests that as the loading on the function gets higher, the response is more likely to come from someone in the younger group, suggesting that it is the older participants who are more likely to be attempting the level of calculation suggested by the value of 90 as discussed above.

Function 2, for which task format showed the greatest loading at $.84$, is clearly discriminating between those who give the base rate only and those who give the correct answer (represented by Δ), with the positive value indicating that those in the frequency condition would be more likely to get the correct answer, and less likely to give a base rate only response. However, this second function was not significant and therefore this discrimination cannot be taken to have any great implications.

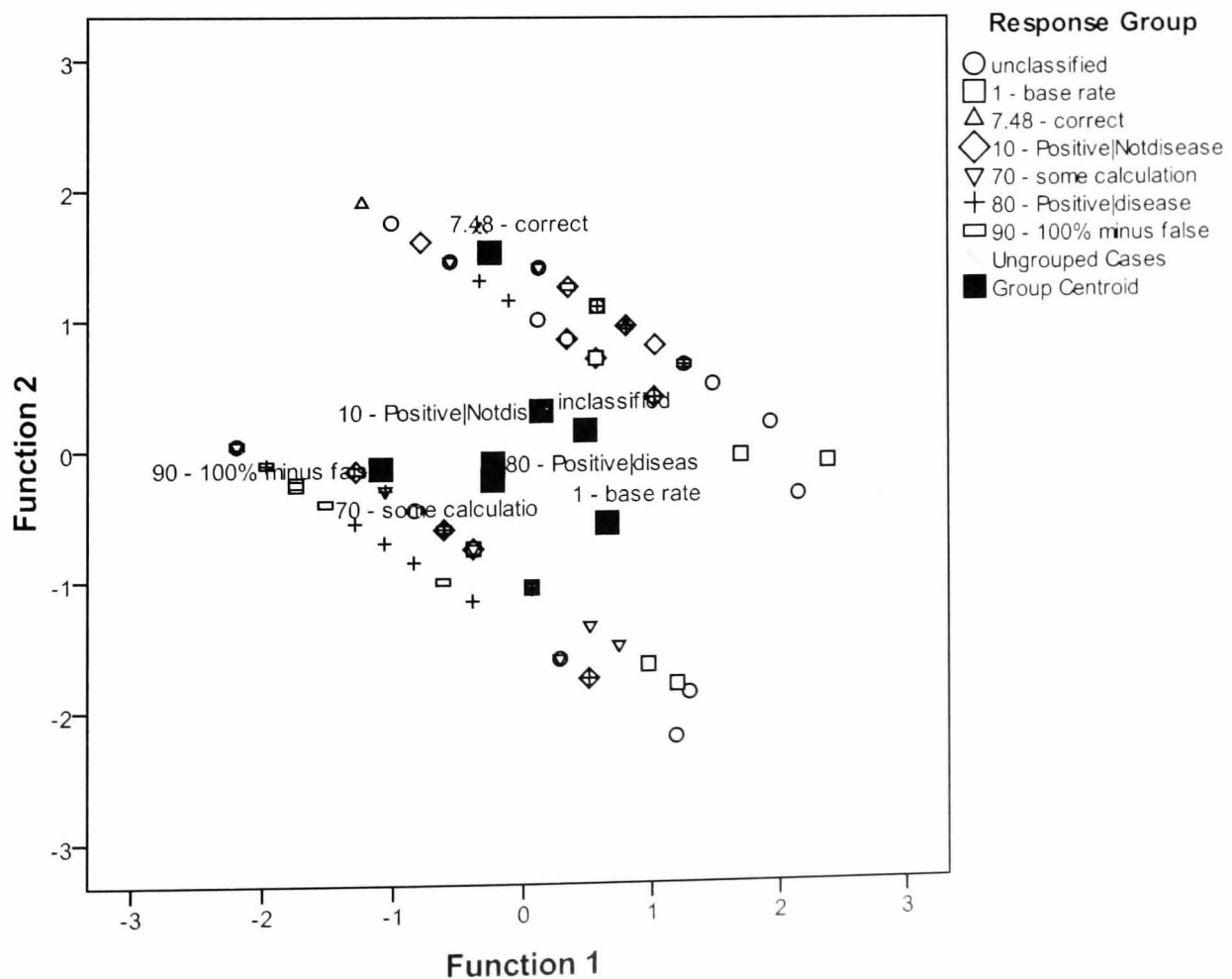
Table 9.5: Standardised canonical discriminant function coefficients for disease task

	Function		
	1	2	3
Age group	.57	-.18	.87
Frequency or Probability format	.45	.89	-.12
ST/LC – Superstitious Thinking/Luck Composite	-.85	.57	.27

Table 9.6: Correlation coefficients for disease task

Predictor	Function		
	1	2	3
age group	-.71	.44	.55
frequency or probability format	.51	.84	-.16
ST/LC – Superstitious Thinking/Luck Composite	.29	-.02	.96

Figure 9.2: Disease task combined groups plot



Stepwise analysis examined the effect of each predictor on the groups. Two of the predictors were significant at $p < .05$, with ST and condition showing very similar effect sizes, at $F(6,78) = 2.74$, and $F(6,78) = 2.64$ respectively.

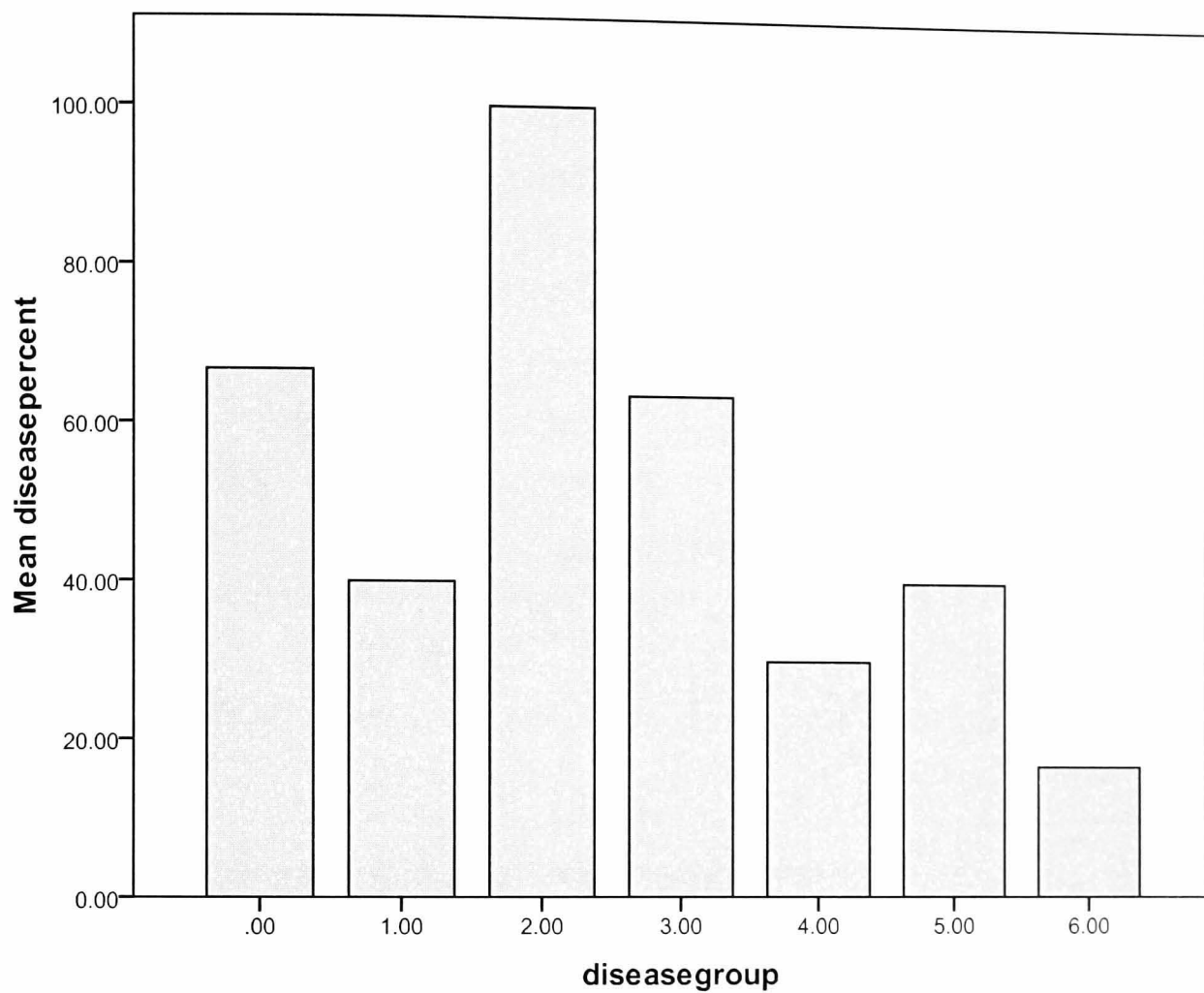
A total of 30.6% of the responses were correctly classified by the analysis (see Table 9.7, in which each response group is numbered as detailed on p. 188), with the normative response, and the response of 90, indicating some calculation, showing the highest levels of correct classification, at 66.7% and 58.3% respectively. It is also of interest that the unclassified responses were relatively well predicted, at 33.3%, although many of the false positive rate responses were also mistakenly grouped into this category (36.4%).

Table 9.7: Actual and predicted group memberships for disease task

		Predicted Group Membership %							Total N
		0	1	2	3	4	5	6	
Actual Group	0					16.7	0.0		
	1	33.3	16.7	20.8	8.3			4.2	24
	2	20.0	40.0	10.0	0.0	20.0	0.0	10.0	10
	3	0.0	0.0	66.7	33.3	0.0	0.0	0.0	3
	4	36.4	9.1	18.2	9.1	27.3	0.0	0.0	11
	5	0.0	30.0	30.0	0.0	30.0	0.0	10.0	10
	6	13.3	6.7	13.3	13.3	20.0	6.7	26.7	15
	6	8.3	0.0	8.3	0	16.7	8.3	58.3	12

Figure 9.3 below illustrates the effect of condition, with 100% of all normative responses being in the frequency condition. This condition also contained the majority of unclassified responses (66.7%) and those giving the value of the false positive rate (63.6%). The lowest rate of response in this condition was when participants indicated some level of calculation by giving a response of 90 (16.7%).

Figure 9.3 Percentage of participants completing the disease task in the frequency condition, by response group

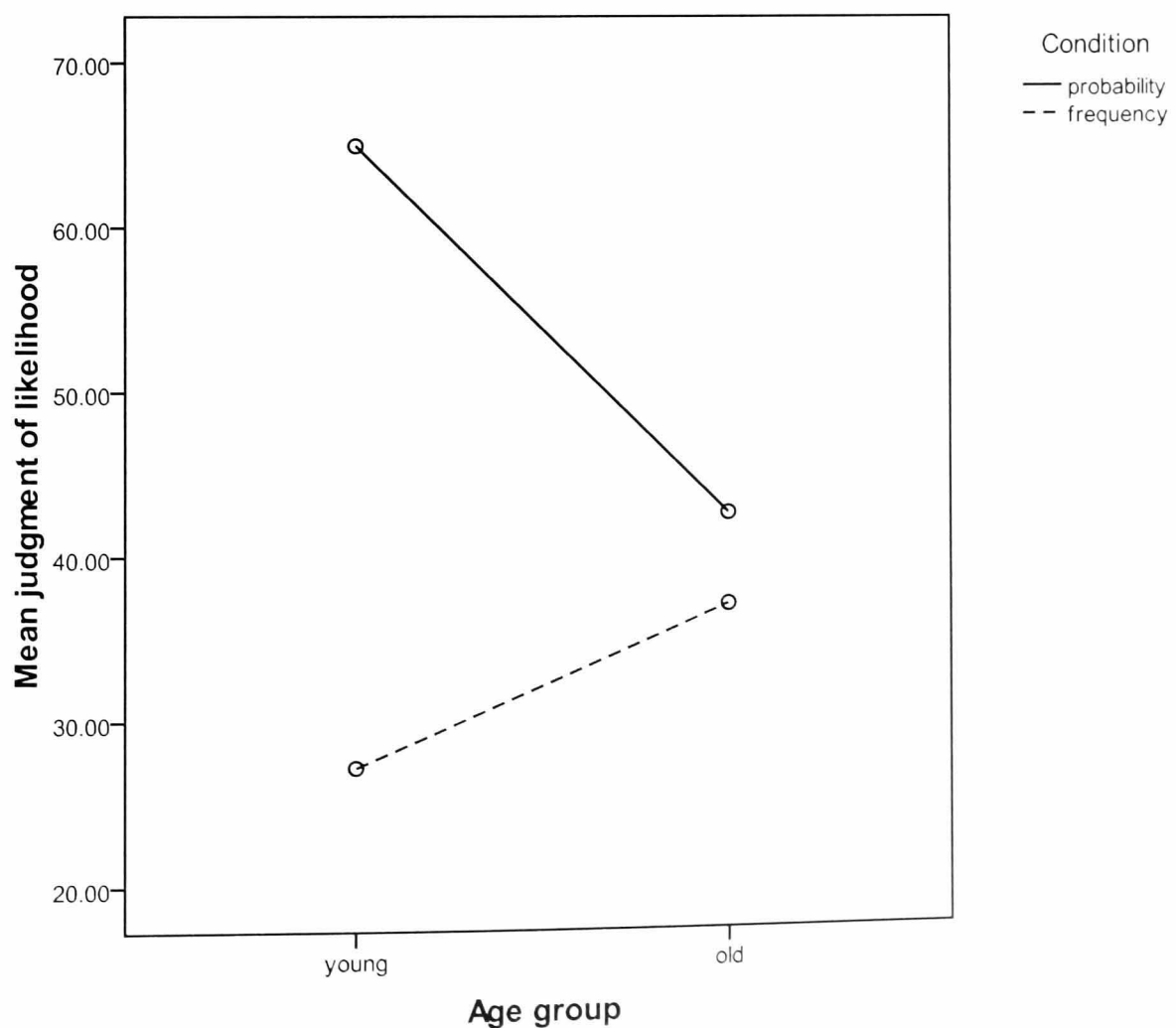


An analysis of variance using the between participants factors of age and condition, and the dependent variable being the raw responses to the disease test confirmed the effect of condition $F(1, 81)=7.76, p<.01$, partial $\eta^2= .087$, and also that there was no effect of age group ($F<1$). An interaction effect was revealed, $F(1,81)=4.29, p<.05$, partial $\eta^2= .050$. This interaction is illustrated by Figure 9.4 and Table 9.8, below. This shows that while the younger participants' mean response was affected by whether the task was presented in the natural frequency version (mean = 27.14, sd = 34.22), or the probability version (64.92, 35.14), the older group were far less affected, showing only a slight reduction in size of response from the probability (42.43, 35.64) to the frequency (36.89, 37.75) conditions.

Table 9.8: Means and standard deviations for response to disease task in each age group, by task condition

	Probability format	Frequency format	All Data
	Mean (SD)	Mean (SD)	Mean (SD)
Young group	64.92 (35.14)	27.15 (34.22)	46.98 (39.23)
Old group	42.43 (35.64)	36.89 (37.75)	39.72 (36.37)
All Data	53.17 (36.79)	32.37 (36.04)	43.14 (37.69)

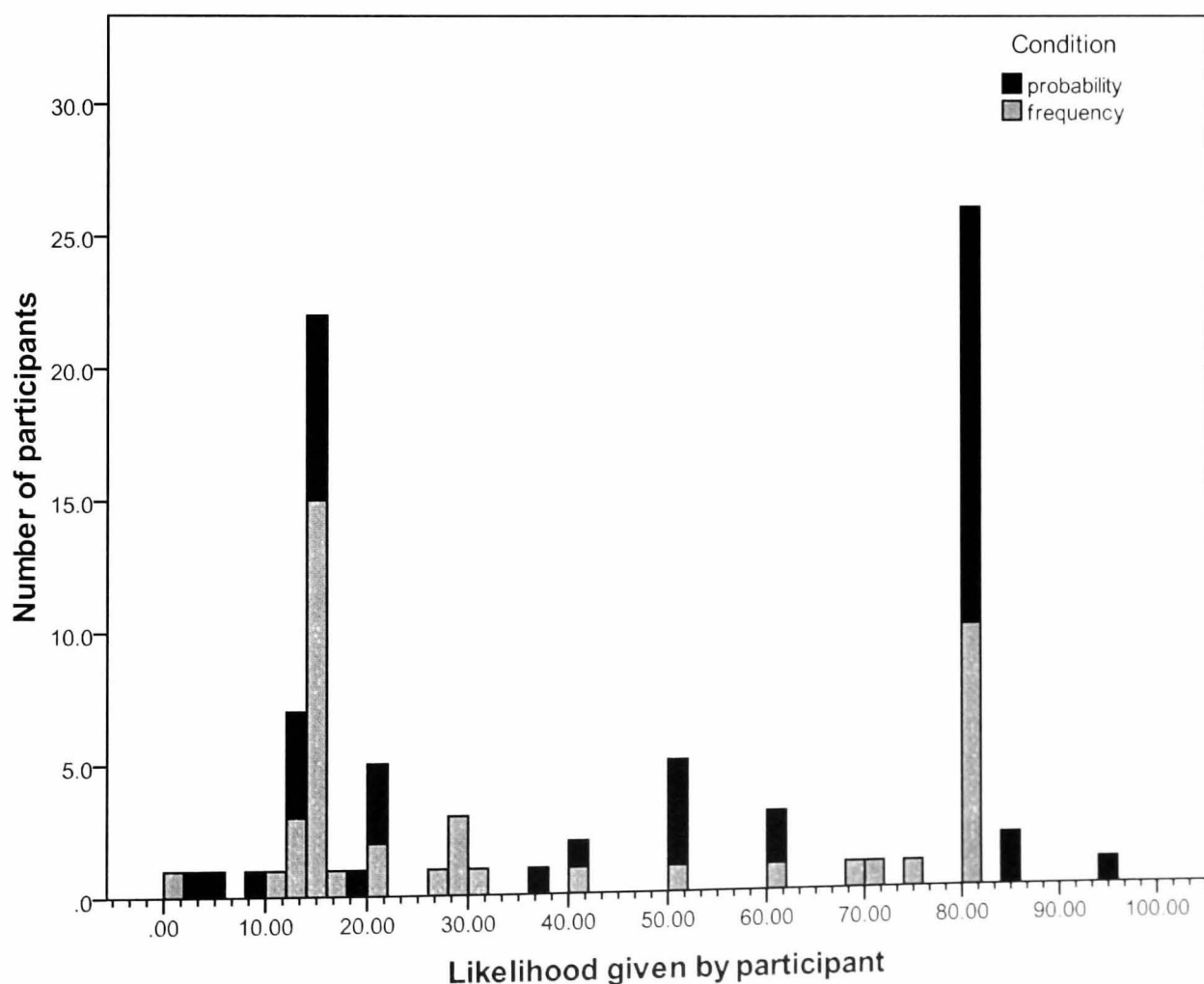
Figure 9.4 Mean response to disease task, by task format



9.3.3 Full analysis of cab task

The cab task data is presented below, in Figure 9.5. Again, the mode of 80 is clearly illustrated, and in this case this value is the witness accuracy rate in the probability version of the task – the probability of their having said they saw a blue cab, given that it was a blue cab. The second most common answer was 15, the base rate of blue cabs in the city. Only one person got the task exactly right, with the answer of 41.38, while one other said 41 which was also accepted as correct on the basis of rounding. There are two other groups of interest – those giving a response of 12 (in the frequency version $P(\text{witnessBlue}|\text{bluecab})$) and those saying 20 (in the probability condition, $P(\text{witnessBlue}|\text{greencab})$) The latter also is $P(\text{witness Green}|\text{blue cab})$ (in the probability version), which seems unlikely. As in the previous task, the higher estimates are disproportionately given by those in the probability condition, with the frequency task being more likely to lead to the lower estimates.

Figure 9.5 Distribution of responses to cab task, by task format



The groups are therefore as follows:

1. 15 – base rate (21 cases)
2. 41-42 – normative (2)
3. 80 – probability version witness'blue'|bluecab (26)
4. 20 – probability version witness'blue'|greencab (5)
5. 12 – frequency version witnessBlue|bluecab (4)

The group of participants who responded with 50 were not included as it was not possible to see where this came from – it may be indicative of participants feeling that as they did not know the answer, they should just guess entirely.

This includes a total of 58 participants, 65% of those who attempted the task. This is again an improvement when compared with the previous version, in Chapter 8, where only 43% were able to be classified. As with the disease data, a final group, labelled 0, was created for all those whose responses could not confidently be attributed to any cue or combination of cues within the tasks. In this case, none of the thinking styles or measures of individual differences correlated with the cab data (see Appendix 14), and therefore the only predictors involved in the analysis were age and condition. As such, DFA was not appropriate in this case and two chi square analyses were conducted, one to look at the association between cab task response group and age group, and one between cab task group and task condition.

Regarding task condition, the 2 by 6 chi square analysis was not significant, $\chi^2(5, 89) = 8.15, p > .05$. although Cramer's V shows a medium effect size, .30 (it should be noted that in this case some of the expected frequencies were less than 5).

Table 9.9: Percentage frequency of cab task response groups by task condition

Response Group	Probability	Frequency
Unclassified	45.2	54.8
15 base rate	33.3	66.7
41-42 normative	50.0	50.0
80 P(witnessblue bluecab)	61.5	38.5
20 – P(witnessblue greencab)	60.0	40.0
12 – P(witnessblue bluecab)	100.0	0.0

The 2 by 6 chi square looking at the association between age group and response group was also non significant, $\chi^2(5, 89) = 8.20, p > .05$, and Cramer's V again suggests a medium effect size of .30.

Table 9.10: Percentage frequency of cab task response groups by age group

Response Group	Young	Old
Unclassified	41.9	58.1
15 base rate	47.6	52.4
41-42 normative	0.0	100
80 P(witnessblue bluecab)	50.0	50.0
20 – P(witnessblue greencab)	20.0	80.0
12 – P(witnessblue bluecab)	100.0	0.0

This was followed with an analysis of variance to examine any effect of age and condition upon the raw scores on the cab task. This again showed that, despite the probability format showing a greater mean error, this was not a significant difference, $F(1, 85) = 3.15, p > .05$, partial $\eta^2 = .036$ and $F(1, 85) = .03, p > .05$ partial $\eta^2 = 0$, and no interaction effect ($F < 1$). As such, the cab task showed none of the effects found in the disease task.

Table 9.11: Means and standard deviations for response to cab task in each age group, by task condition

	Probability format	Frequency format	All Data
	Mean (SD)	Mean (SD)	Mean (SD)
Young group	49.77 (35.56)	37.41 (30.380)	43.74 (33.31)
Old group	47.71 (27.93)	37.06 (27.58)	42.43 (27.97)
All Data	48.67 (31.36)	37.27 (28.54)	43.04 (30.37)

9.4 Discussion

This second study looking at Bayesian tasks again found that the vast majority of participants did not appear to be using normative Bayesian reasoning in their responses. In the case of the disease task, three participants gave the normative response, and in the case of the cab task just two participants did so.

The current data set was closer to the pattern found by Birnbaum (2004). In his study using a version of the cab problem, 60% of participants gave the witness reliability rate, while 20% gave the base rate only. A similar proportion of the current study gave the base rate (24%), but only 29% responded with the witness reliability rate. There were also two other small groups, where participants had responded either with a value of 12, or with 20. The latter is likely to be from the probability of the participant responding with the value of $P(\text{witness blue}|\text{greencab})$, or possibly $P(\text{witness green}|\text{bluecab})$, both of which indicate a clear lack of understanding of the requirements of the task.

A relatively large number of participants also chose the base rate in the disease task, at 12% of those who completed the task, but this was the same as the number who chose a value of 70 (not directly in the written tasks, and indicative of some calculation), and a little less than those who chose 10, the value of $P(\text{positive}|\text{healthy})$, at 13%, and those who gave a value of 90 (again suggesting some attempt at calculation), at 14% of the responses. The value of 80, $P(\text{positive}|\text{disease})$ was chosen by just 18% of participants, notably less than Birnbaum's 60% choosing this additional evidence and showing base rate neglect (2004).

Overall, as in the previous data set, the current participants showed lower incidences of each of the predicted fallacies (base rate neglect and base rate only) than had been anticipated based on Birnbaum (2004) and Hinsz *et al.*, 2008. However, the 'noise' in the current data set was less, with greater numbers falling into one of the clear groups of responses. However, there was still up to 35% of responses that could not be readily attributed to any cue within the task, and remained unclassified. This applied to both the probability and the natural frequency condition, and despite the set structure being clearly set out in each version.

One of the main purposes of this study was to see if using natural frequencies produce a facilitating effect where normalised frequencies fail to do so. Such an effect was not consistently found. Although Discriminant Function Analysis indicated that condition loaded heavily onto Function 2, which in turn discriminated between the normative response (all of which were in the frequency condition) and the base rate only response. this second Function was not able to significantly discriminate between the groups. This may in part have been due to the low numbers of normative responses given. The analysis of variance did find a significant effect of format upon response to the disease task (although not to the cab task), which again suggests that the format does affect the way participants approach the task, with the frequency format reducing the level of overestimates being made. This analysis also revealed an interaction between age group and task format, whereby the older group seemed to respond to both conditions in the same way, instead of making wild overestimations (perhaps pegged to the base rate) as shown by the young participants in the probability condition. While this interaction had also been suggested in Chapter 7, using the measure of conjunctive and disjunctive reasoning fallacies, it had not in that case achieved significance.

No such effect was found in the case of the cab task, and chi square analyses also found no association between either age or task format and response group. The medium effect size does again suggest, however, that larger group sizes may lead to a significant effect becoming apparent. Gigerenzer 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.

One of the reasons why participants may still be struggling with the tasks is suggested by Brase (2008), who stresses that the important factor is not the literal representation of the task by the experimenter per se, but instead whether or not the participant themselves forms a mental representation of the task as being in natural frequencies. Brase demonstrated that some participants would still interpret a frequency format as if it were a single event, and in this way would impair their own ability to respond normatively. This could illuminate the processes behind the age/task interaction, if the older participants are in fact already primed by previous experience to consider such tasks as probabilities. It may also be the case that the university laboratory environment prompts all participants to think in such terms, although the old group would be less

familiar with the university setting and therefore perhaps more susceptible to such effects.

To summarise, Hypotheses 2 and 3 anticipated an effect of age and of format whereby older participants, and those in the probability condition, would be less likely to achieve the normative response or to respond in a way that indicated some level of calculation. These must both be rejected, as despite the effect of condition, it was not clear that the frequency format had a *facilitating* effect, with no evidence presented suggesting that it was in fact priming system 2 and allowing for more accurate analytical processing to take place.

Older participants did respond differently to the younger group, however, in that they appeared unaffected by task format. There was also some evidence to support this in Chapter 7 of this thesis, where a significant effect of format on conjunction and disjunction fallacies was found, but did not appear to apply to the older group to the same extent as to the younger group (although no significant interaction between format and age was found). This supports suggestions by Mutter and Pliske (1994) and Johnson (1993) that older participants do not make good use of the relevant and salient information presented in tasks. As Brase (2008) stresses that participants may still interpret natural frequency information in terms of one event probabilities, this may be more likely to occur in the older group, who may have established patterns of behaviour in problem solving which over ride further details given within the tasks.

Hypothesis 1 also did not find strong support, as only one thinking style, Superstitious Thinking, showed any relationship with either Bayesian task – weak levels of superstitious thinking were associated with higher response values on the disease task. The Discriminant Function analysis revealed ST as a significant predictor, loading most heavily onto Function 1, discriminating between responses of 90, indicative of some calculation, and 1, the base rate only response. Again, the direction of the relationship indicates that weak levels of Superstitious Thinking are indicative of some attempts to manipulate the data, and to understand that the task requires such information processing, rather than replying only with one of the values already presented.

To conclude, using the natural frequencies rather than normalised frequencies does appear to have resulted in more participants responding with answers that would be

expected from the literature – i.e. the base rate only and base rate neglect groups – but did not show any facilitating effect in the form of greater normative responses. It seems that either system 2 was not primed, or it was primed but participants were still not able to accurately process the various cues involved. The lack of normative responses makes data analysis more difficult, with such a small group size, but to obtain a larger group size could necessitate the sample itself growing exponentially, as to get just 2 or 3 normative responses required almost 90 participants in each case. As such, future directions could include more visual representations of the tasks (see, for instance, Cosmides & Tooby, 1996; Yamagishi, 2003) although this leads such lab based tasks further away from ‘real life’ examples, where such visual information is unlikely to be available.

Chapter 10 – General Discussion

10.1 Summary of Findings

This present research focused on the factors that may affect and/or explain adults' ability to solve probabilistic reasoning tasks. The first study conducted, and detailed in Chapter 5, looked at the effect of format on reasoning ability. In this case it was found that contrary to the majority of previous research in this area (e.g. Fantino & Stolarz-Fantino, 2005; Fiedler, 1988; Tversky & Kahneman, 1983), there was no significant effect of format on reasoning ability as measured by the number of fallacies. When measured in terms of error, participants actually performed more poorly on tasks presented in the frequency format than on those in the probability format. This was found with one task in particular, which asked about the probability that Venus Williams might (for instance) lose the first set or break her racquet in a tennis match. As this related to a subject which some participants may well have known little about, this may have been of concern to them (i.e., although it does not matter to a person's ability to complete the task, participants may have been confused by this), and therefore this type of task was avoided in the remainder of the research. It is of course possible that the participants in this study were particularly unusual, in terms of the factors that may affect the impact of the frequency format, specifically thinking style. This was addressed in the following studies.

The second study used an amended and improved set of materials to assess participants' performance on conjunctive, exclusive disjunctive and inclusive disjunctive tasks. In this case a significant facilitating effect of the frequency format was found, when performance was measured by both fallacies and by error. Two further hypotheses had been proposed, predicting that measures of thinking style and verbal intelligence would mediate the facilitating effect of the frequency condition. These were deemed to be important factors in the context of using the dual process theory of reasoning to elucidate the processes involved, as those with faster processing speeds may be more able to manipulate the information needed for analytic processing, while those with higher verbal intelligence could be expected to be more able to benefit from the frequency format's wording. Neither of these hypotheses was supported.

The third study (Chapter 7) introduced the between participant factor of age. While no age differences were anticipated in the older cohorts' responses to the task, as measured

by fallacy, it was predicted that the greater variation in thinking styles and in the other variables of interest in a more age heterogeneous sample would better reflect differences in the underlying processes, as measured by error. There was again a strong effect of task format, and no effect of age with regard to normative reasoning and there was also very little evidence to suggest that the older cohort were reasoning differently from the younger group. There was partial support, however, for the influence of verbal intelligence and information processing speed on reasoning performance.

The fourth study required participants to complete the more cognitively demanding Bayesian tasks. Discriminant function analysis revealed that while format significantly affected the pattern of responses that were observed, in fact neither format led to normatively ‘correct’ reasoning in either of the tasks presented. Age was also not a statistically significant predictor, but in this case thinking styles did predict performance in terms of some of the types of response given, although they did not actually predict *correct* reasoning. A further finding from this analysis was that, in the disease task only, format and a number of the analytic thinking styles loaded on to the same function, suggesting that they may both be linked to one underlying factor.

The final study again looked at Bayesian reasoning, but in this case used tasks phrased as natural frequencies (and compared them with probability versions), rather than the normalised frequencies used in the fourth study. It was anticipated that this would facilitate Bayesian reasoning for both age groups, and while this was not found to be the case in terms of predicting normative reasoning, it was found that the natural frequency format did have an effect on the size of estimate, leading participants to overestimate the likelihood in the disease task to a far lesser degree than in the probability condition.

To summarise the above, this research has generated six main conclusions:

- The frequency format does show an effect on reasoning performance, and in particular a significant facilitating effect in the case of conjunctive and disjunctive tasks (Chapters 6, 7).
- Age does not show any significant effect on performance in these tasks (Chapter 7).
- Older participants are less affected by problem format, when compared with a younger group (Chapter 9, see also Chapter 7)

- Of the individual differences examined in this research, verbal intelligence and information processing speed show the most consistent relationship with reasoning performance (Chapter 7)
- Thinking styles show no consistent relationship with reasoning performance (Chapters 6,7,8)
- There is a difference between responses to exclusive and inclusive disjunctions, in terms of the numbers of fallacies and amounts of error, indicating that participants do not appreciate the differing requirements of these two types of task (Chapters 6, 7).

This discussion will now focus on each of these conclusions in turn.

10.1.1 Effect of Format

The effect of format was consistent in Chapters 6 and 7, and format was also found to be a significant predictor of responses in the two Bayesian reasoning tasks in Chapter 8 and also showed significance in the disease task in Chapter 9.

The format clearly facilitated normative reasoning on conjunctive, inclusive disjunctive and exclusive disjunctive tasks, leading to fewer fallacies and to lower levels of error. In the case of exclusive disjunction tasks, this had not been directly reported in the literature. Costello (2009) looked only at inclusive disjunctions of weather events.

This facilitation effect was not seen so clearly in Chapters 8 and 9, using Bayesian tasks. This may have been due to a floor effect in the data – in the first study using the disease problem, only 16 out of 143 participants (11%) responded with the normative answer, with even fewer doing so in Chapter 9, and in the cab problem, no participant obtained the normative answer in the first version of the tasks (Chapter 8) and only 2 out of 89 participants did so for the second version (Chapter 9). In each case, this reduces the ability of the analysis to predict membership of the ‘normatively correct’ group, and in the first version of the cab task it clearly becomes an impossibility, with no such group created. When compared with those reported by Evans *et al.* (2000), the rates of normative responses obtained here are within the same range, but at the lower end. Evans *et al.*’s most difficult version of the disease task elicited a 9% ‘success’ rate, but in their first experiment the average correct responses across all tasks was 30.5%, a figure which was further increased in later versions (2000).

Evans (2007a), and Kahneman and Frederick (2002) have each suggested that the frequency format facilitates normative reasoning by making clear to participants the nested sets within a task (see also Neace, Michael, Bolling, Deer & Zecevic, 2008). For instance, the number of blue cabs involved in an accident is a nested set within the greater set of all blue cabs, and in this way participants are primed to use the analytic system 2, rather than the heuristic system 1. Looking again at the Bayesian data presented in Chapter 8, it can be seen that although it did not facilitate normative reasoning, format remained a clear predictor of the type of response made, and similarly in Chapter 9 it did significantly affect responses given to the disease task. Focussing on the cabs task in Chapter 8, many participants did respond to what Birnbaum (2004) characterises as being based either on the base rate only (the number of blue cabs in that city), revealing base rate neglect (the witness accuracy rate only) or a multiplication of these two values. Birnbaum also suggests that the normative answer is the one given least often, as was the case in the current data with no participant reaching this answer. Discriminant function analysis, and subsequent stepwise analysis, revealed that those showing base rate neglect or responding with the base rate only were more likely to be completing tasks in the probability format, while those who attempted some calculation and gave an answer that, while inaccurate, did take into account both of these pieces of evidence by multiplying them together, were more likely to be in the frequency condition. Indeed, all 17 participants (12% of those who completed the task) who gave this response were completing the frequency version. This may be evidence that the frequency format was recruiting the analytical thinking system, by making it clearer that each of these pieces of information, the base rate and witness accuracy rate, are needed to derive the correct answer. However, just because the analytical process has been cued does not automatically mean that the Bayesian calculation will be correctly applied. Again, the fact that no participant got the normatively correct answer for this task did mean that it is not possible to identify which of the predictors may be most strongly associated with normative reasoning, but also serves to illustrate how much more constructive it is to consider participants' responses in terms of the different categories, rather than the reductionist 'correct' or 'incorrect' response. It can also be seen that the probability condition in each task predicted the very highest response values, these being the test accuracy rate of 95 and the number of blue cabs 80, both of which indicate base rate neglect, and an over reliance on the case specific information of test and witness accuracy. This lends support to the suggestion that the frequency condition

is leading to some awareness of nested sets, priming participants to appreciate that the very low base rate is of relevance, and that the number which satisfies the condition of being the right colour cab (or a correct positive test result) is likely to be a smaller number than the number suggested by looking at the witness (or test accuracy) out of context.

In the disease task used in Chapter 8 up to 11% of participants did obtain the correct answer, but the analysis was particularly poor at categorising these individuals, with only 25% being correctly categorised (in the data presented in Chapter 8), while many more (38%) were incorrectly assigned in the analysis to the base rate neglect response category. The other possible categories (further to 'correct' and 'base rate neglect') were to give the base rate only, the false positive rate of the test, or a value of 1. The last of these was felt to be an indication of some level of manipulation of the values involved, in terms of an appreciation that the very low base rate would lead to a very low positive test rate, despite the actual test's high level of accuracy but to be different enough from the normative answer (approximately 2) as to be classified as incorrect. If the pattern observed in the cabs task were to be followed, it would be expected that this category should be associated with the frequency format of the task. This was not the case, however, with the response of 1 and base rate neglect being associated with the probability format, while those completing the frequency versions were more likely to show either base rate only or the false positive rate. Crucially, and contrary to previous findings (Gigerenzer & Hoffrage, 1996; Evans, 2003) those giving the normative answer were equally likely to have completed a probability or a frequency task, which again illustrates that while the frequency format may be making some of the complexity of the task easier for participants to understand, the facilitating effect is not apparent in the same way as in the conjunctive and disjunctive tasks.

Chapter 9 compared the probability format with natural frequency format, rather than a normalised frequency format. The latter use a consistent reference class, usually of 100, while the former use a reference class containing information regarding the relevant base rates (see Gigerenzer & Hoffrage, 1995). While the natural frequency formats did not show a great increase in normative reasoning, they and the reworded probability versions did produce higher numbers of base rate only and base rate neglect responses, with fewer apparently random answers. This may suggest that the participants were attending to the content of the tasks in greater detail, and in doing so engaging system 2

processes to a certain extent. While thinking styles were not related to responses by this sample, the task format clearly was, and this suggests that it was not individual differences that led to the priming of system 2 analytical processes, but the external prompt of the task itself.

In the disease task only, there was a significant effect of format on response value, with the frequency format leading to significantly lower estimates, and a mean that was far closer to the actual normative possibility of the individual having the disease. Again, as in Chapter 8, most of the base rate only responses were in the probability condition, and also in this case all of the normative responses were in the frequency condition.

Gigerenzer, Gaissmaier, Kurz-Milecke, Schwartz and Woloshin (2008) looked at the impact of poor statistical literacy on decisions made by both physicians and health professionals, finding that wild overestimations are frequently made. At best this leads to unnecessary worry, and at worst it can lead to avoidable fatalities. When put in this context, that the presentation of data in natural frequencies leads to a reduction in the amount of overestimation can be usefully applied to the presentation of health and other advisory materials.

10.1.2 Lack of Age Effect and Measures of Individual Difference

An older group of participants was used in this study in order to better understand the cognitive processes and individual differences involved in the reasoning performance of a more heterogeneous sample. It was anticipated that they would provide a greater variance in the individual differences data, and specifically in terms of the thinking styles, due to the possibility that older participants may be more likely (than their younger counterparts) to rely on heuristic and intuitive approaches to the tasks.

It was felt that the lack of an age effect on the conjunctive and disjunctive tasks in Chapter 7 might have been due to the fact that the tasks were not sufficiently cognitively demanding. For example, Chasseigne *et al.* (2004) found that older people performed significantly worse than their younger counterparts only on the more complex tasks that were administered. This possibility was addressed in Chapters 8 and 9, which used the more difficult Bayesian tasks. However age was not a significant predictor of response category in either of the tasks, providing support for the absence of an ageing effect in this particular form of reasoning.

The age effect was absent despite older participants' significantly slower processing speeds (Chapter 7), indicating that this aspect of cognitive performance was not associated with a probabilistic reasoning deficit in the present research. In one instance, the error data collected from the conjunction tasks, both information processing speed and verbal intelligence were significant predictors of performance. Higher scores on the Mill Hill Vocabulary Scale and the Information Processing Speed Task were both associated with lower levels of error, but this did not translate into a significant effect when examined through ANCOVA. Although no time limits were placed on the completion of the set of materials, Salthouse (2000) found that older persons' slower information processing speed impaired performance even on tasks where no time limit was imposed. Nonetheless the current results show that slower processing speed does not inevitably result in impaired performance, suggesting that neither the limited time nor the simultaneity mechanism is an issue in this case (Salthouse, 1996). In a decision making task, Johnson (1993) found that although an overall age effect was not evident, younger people appeared to be making more use of the information available to them, rechecking the materials more frequently than older persons.

Similarly, Mata, Schooler and Rieskamp (2007) found that although both younger and older persons were reasonably flexible in responding to tasks that demanded either 'information intensive' or 'information-frugal' strategies, the older group did have a stronger tendency towards using simpler, less intensive and cognitively demanding strategies. They also looked up less of the available information and took longer to process the material. Mata *et al.* (2007) also found a significant relationship between processing speed and information searching, whereby the faster the speed, the greater the amount of information searched for. Riis and Schwarz (2003) found that a greater attention to detail may actually be a disadvantage in such tasks as were presented here, stating that a "detail-orientated processing style" actually negatively affected performance on this style of conjunctive task. It is therefore feasible that a slower processing speed may actually present an advantage for this aspect of the task, potentially balancing out the disadvantages suggested by Salthouse (2000) and Wareing *et al.* (2007) and discussed in Chapter 4.1. This is in contrast to Fisk and Sharp (2002), who found that controlling for older participants' slower IPS removed an age effect on syllogistic reasoning ability. This would lend further support to the proposition that probabilistic reasoning, and inductive reasoning in general, is dependent on different cognitive processes from syllogistic, deductive, reasoning.

Lastly, one of the main reasons for looking at probabilistic reasoning in an older age group was that they were expected to show different thinking styles from the younger cohort, as found by Klaczynski and Robinson (2000), and that this would enable a greater examination of the outcomes associated with the different styles. In other words, the lack of relationship between thinking styles and reasoning in Chapter 6 may have been due to the low levels of variance in the thinking styles in that sample. Chapter 7 did find that the age groups differed on many thinking styles, with the older group showing stronger levels of Absolutism and Social Desirability, but weaker levels of Dogmatism and Superstitious Thinking/Luck. They were also lower on the Faith in Intuition and Need for Cognition scales. The current research also went on to find differences between those participants in Chapter 7 and Chapter 9, despite their being sampled from the same populations (students and U3A members). Although the older group did again show stronger levels of Absolutism and Social Desirability, they no longer showed differences in Faith in Intuition, Dogmatism, or Superstitious Thinking, and did show a new difference in their stronger levels of Categorical Thinking. The lower levels of Need for Cognition were replicated, however, while the opposite effect was observed by Klaczynski and Robinson (2000), with the older group in their sample showing greater levels of NFC. In the current research, these differences in thinking style did not appear to be related to any differences in reasoning performance.

10.1.3 Interaction Between Age Group and Task Format

The significant interaction between age group and task format on the disease task in Chapter 9 revealed that while younger participants were susceptible to the effect of format (with the natural frequencies leading to responses showing less overestimation) older participants did not appear to be affected by format to nearly the same degree. They instead showed very little difference between the different conditions, with their estimates in the probability condition being lower than the younger group, and their estimates in the frequency condition being higher than the younger group. While such an interaction was observed in Chapter 7, on the reasoning fallacy data, it had not been a significant interaction.

Previous research (Johnson 1993; Mutter & Pliske, 1994) suggests that older participants are not responding to salient information when it is presented, or that they

are not fully integrating it (Mutter & Plumlee, 2009). Fontaine and Pennequin (2000) have also found that older participants not only fail to integrate all information, but that they are actually less able to extract the relevant information in the first place, and have difficulty in inhibiting irrelevant information. The current data similarly suggest that the older participants are not responding to – either ‘taking in’ or able to utilise – the base rate information contained within the natural frequencies (Gigerenzer & Hoffrage, 1995). However, they are equally not as misled or distracted by the probability information as the young group are. In fact, in a task looking at the likelihood of disease, whereby the common error is to wildly overestimate the likelihood of infection, they can actually be seen to be reasoning more accurately. That they are less susceptible to the effect of natural frequencies lends support to the theory that older participants do have established methods of making such decisions, and may have a level of experience that leads them to avoid wild overestimations of likelihoods. Evans (2010a) has suggested that our ability to use intuition to our own advantage does come from our previous experience, and naturally older participants have greater experience of making judgements of likelihood. While it had been anticipated that the natural frequencies would produce a *greater* advantage to older participants, based on their experiences of using such frequencies during their lives (as suggested in Section 9.1), this has clearly not been the case. It may be that they were considering them all as needing to be considered using percentages, as Brase (2008) suggests that it is the participants’ interpretation of the task that is most important, rather than the investigators intentions.

A further explanation may be that, as suggested by Chen and Sun (2003) and Baltes (1997) as well as Fisk (2005), older participants are using differing, less cognitively demanding, strategies, and are doing so regardless of the framing of the tasks. In this case, that the set structure has been made more apparent may be of less use, if participants are already pre-disposed to tackle the problems without taking the set-structure information into account.

10.1.4 Thinking Styles

Performance on many of the thinking styles measures was related to reasoning performance, but none of these relationships were consistent across all of the data. Regarding the Rational Experiential Inventory subscales, neither Faith in Intuition nor Need For Cognition showed any significant effect or relationship with any of the

reasoning tasks. This may indicate either that a tendency to reason heuristically does not affect an individual's answers on these tasks, or that the REI has limitations as a measure of the propensity for heuristic and analytic reasoning.

In relation to the Thinking Disposition Questionnaire, a number of the subscales did load onto the same function as did the task format, within the Bayesian 'disease task'. These were Categorical Thinking, Dogmatism, Absolutism and Flexible Thinking. This suggests that the propensity to reason analytically, as measured by these scales, may map onto the same underlying factor as is represented by the format variable.

Absolutism predicted exclusive disjunction errors, with those showing weaker tendencies towards the disposition also showing lower levels of error. and Superstitious Thinking/Luck and Social Desirability were both significant predictors of inclusive disjunction fallacies, with weak levels of these dispositions being associated with fewer fallacies. In the case of ST, this would suggest that analytical thinking *was* associated with more normative reasoning, in support of similar findings by Dagnall *et al.* (2007) who found that participants with strong levels of paranormal belief were less able to correctly complete probabilistic reasoning tasks.

There is one variation between the use of the TDQ scales in the present research and the way that it has been used by other researchers. In their original use of the scales, Kokis *et al.* (2002) created the actively open minded thinking (AOT) composite scale from the FT, BI, A, D and CT scales. In their data, the mean correlation between them was .32. In the current data however, the correlation was 0.19, and the composite scale was therefore deemed inappropriate. Other studies using the AOT, both prior and subsequent to Kokis *et al.* (2002) have been less clear about their own data, in terms of average correlations. Stanovich and West (2007), for instance, use the AOT without referring to the composite subscales, or their correlations. The scale does show high levels of internal consistency from their data, although as a 41 item scale this may not reflect high correlations. Conversely, if the scales do correlate highly, alpha is no indication of unidimensionality (Field, 2009), and can not necessarily provide reassurance that the AOT is a valid composite.

It should also be noted that although the subscales were not strongly correlated, all relationships were in the expected direction, e.g. Flexible Thinking showed a negative

relationship with both Absolutism and Categorical Thinking, and Belief Identification positively correlated with Absolutism (full table available in Appendix 15).

10.1.5 Differences Between Exclusive and Inclusive Disjunctions

Carlson and Yates (1989) looked at the possibility that participants misunderstand the term 'or' in disjunctions, investigating the suggestion that it is incorrectly read as meaning 'x or y but not both' (an exclusive disjunction), rather than the logically correct interpretation of *or* as '*x or y or both*' (an inclusive disjunction). Carlson and Yates (1989) found little evidence consistent with this hypothesised form of linguistic misunderstanding in their own research, since making the inclusive disjunction more explicit (i.e. 'x or y or both') did not lead to any reduction in reasoning errors. The current findings, in both Chapter 6 and 7, might imply that participants are actually assuming that all disjunctions are inclusive in that when producing their component probability estimates the values chosen tend to be subadditive. More likely – given the wordings of the tasks themselves making the exclusive nature clear – is that participants are responding on the basis of representativeness, and judging each component to be likely/representative, leading to subadditivity (see Fisk, 2002). In the inclusive case the consequences of this are reduced since the joint (conjunctive event) is deducted when computing the normative disjunctive value. In the exclusive case no such deduction is made (the conjunctive event has a value of zero, as a logical impossibility) and so the subadditivity is more readily apparent in the normative estimate. Although the component estimates were scaled down to adjust for this subadditivity (see 6.2.2 for details), other things being equal this would have left the errors generally larger in the exclusive case. This can be seen from the larger amounts of error made on the exclusive tasks, despite the fact that fallacies on this task remained (relatively) low.

Noveck *et al.*, (2002) found that participants had a tendency to assume disjunctions to be inclusive when exclusivity is not made explicit. The current tasks were specifically designed to give concrete examples of each type of disjunction. For instance, that a student might get grade A in maths or grade A in biology is inclusive, as they are studying two separate subjects, while getting a grade A in maths or failing maths must be exclusive, as only one grade is awarded for the course. It is possible, that the lack of the absolutely explicit 'or both' or 'but not both' on the end of the inclusive and exclusive propositions respectively may have meant that the participants were not primed to think in those terms. Alternatively, participants may simply have thought that

each of the given alternatives were quite possible (that is, representative) and assigned them each an estimate of over 0.5 (or 50 out of 100). This would lead to an increased likelihood of subadditivity and larger errors in the exclusive tasks, as was found to be the case in Chapter 7 in particular.

While there was no significant interaction between task type and format, the data do indicate that it is only in the frequency condition that greater error is made in the exclusive disjunctions, as compared to the inclusive disjunctions. This could suggest that the frequency format is making the inclusive nature of those tasks – that the sets involved may overlap – particularly clear to the participants, and enabling them to make judgements which are closer to the normative.

Verbal intelligence was also related to the effect of task type. When scores on the Mill Hill scale were controlled for (in Chapter 7) the error made by the older group on each type of task – inclusive disjunction, exclusive disjunction and conjunction – remained unchanged, suggesting that the greater verbal intelligence which is typical in this age group was not in fact an advantage in terms of the propensity for error on each task. The younger participants' adjusted means did indicate a small (but statistically significant) change, indicating that the younger group showed *greater* levels of error on the exclusive disjunction tasks, once verbal intelligence had been controlled for. However, these differences were very small, and with large amounts of missing data these findings should be taken only to suggest that there was no improvement, and not that there was a great decline. One possibility is that those who had greater verbal intelligence were over confident in their performance, and had an expectation of being able to perform on such tasks which led to a lower level of application and cognitive effort, counter balancing the benefits of greater verbal intelligence. Equally, it could be that those with greater verbal intelligence paid more attention to each detail in the vignette, something that Riis and Schwarz (2003) found is linked to poorer conjunctive reasoning, as discussed above. Gigerenzer and Brighton (2009) and Gigerenzer and Gaissmaier (2011) have also discussed the phenomenon known as the 'less-is-more' effect, whereby not utilising some of the available information can actually lead to more accurate predictions than when taking on board every single cue available. While fast and frugal methods (see 2.5.8) may be of benefit here, leading to more accurate answers from those who were less 'bogged down' in the details of the task, it should be noted that in the Bayesian tasks this could not be seen to be the case. In the Bayesian tasks

presented here, it appeared that most participants were selecting (or focusing on) only a single cue, often the base rate or the additional evidence, with this selective focus leading to erroneous judgements. This indicates that such tasks are one of the exceptions to the rule that ‘simple’ heuristics (Gigerenzer & Brighton, 2009) lead not just to adequate judgements, but to the most accurate ones.

10.1.6 Dual Process Theory

To summarise the above in terms of evidence for a dual-process theory of reasoning, from the individual differences measures employed here, there is no direct evidence that suggests that a predisposition to either thinking style, analytic or heuristic, is directly associated with greater accuracy in probabilistic reasoning. However, there is clear evidence that analytical thinking can be primed by the format of the task, which therefore offers support from these measures for the two process system.

When looking at the effect of format, there is a strong facilitative effect consistent with the increased utilisation of the analytic system 2 rather than the heuristic system 1, thereby resulting in participants making fewer fallacies on inclusive disjunction, exclusive disjunction and conjunction tasks (see also Sprenger & Dougherty, 2006) whilst also showing lower levels of absolute error. Within the Bayesian reasoning tasks presented in Chapters 8 and 9, the relationship was more complex. While it cannot be clearly stated that the frequency format – either normalised or natural frequencies – had a facilitative effect, the former were able to predict response group in both the cab and disease task (see Chapter 8), while the latter significantly affected the size of judgements given in the disease task (Chapter 9).

Natural frequencies did lead to more participants clearly attending to one or other of the available cues from within the task as suggested by Evans' (2007a), but they continued to fail to fully integrate all of the information available. In fact, the greater levels of base rate neglect and base rate fallacy in the natural frequencies version of the tasks actually suggests that participants made *less* effort to manipulate the information, and were more inclined to simply pick out the value that was most salient to them. Stanovich (2004) is clear that The Autonomous Set of Systems (system 1) are indeed autonomous, and it may be that the natural frequency format is not strong enough, as a prompt, to lead the reasoner to over ride them.

Recent developments to dual process theories suggest a third component, which may be defined as being a reflective system (Stanovich, 2009). It is this reflective system that is thought to trigger analytical processes when necessary, with necessity perhaps being defined by the lack of a Feeling of Rightness (FOR, Thompson, 2009). In the context of the Bayesian tasks, it is probable that a FOR would often be present even when a great overestimation has been made. When receiving a positive test result from a medical practitioner the general assumption is that this test indicates the presence of the condition, not least because the patient has been prompted to take the test (e.g. by showing symptoms) and already has reasonable evidence for suspecting its presence. While the disease task presented in the current research made no mention of such prior evidence, only the base rate and the test accuracy rate, the participants may still have felt a strong FOR in responding that if a test indicates the presence of the disease, it is more than likely that the disease is indeed present. Similar to the ‘if only’ effect (Epstein *et al.* 1992) the participants were allowing what they would think *if it happened to them* affect their judgements, and failing to fully engage analytical reasoning to overcome this bias.

That the different presentations affected responses but did not lead to increases in normative Bayesian reasoning can also be examined from the perspective of each ‘system’ being made up of a number of systems or processes (Evans (2008;2009;2010a;2010b) Saunders and Over (2009) Stanovich 2009). It may be that the different presentations are prompting not different *systems* to engage with the task in terms of ‘system 1’ or ‘system 2’, but are instead prompting different processes within system 1 to engage, leading to quite clearly observable differences in the judgements made, but no greater levels of normative reasoning. This leads to the question of what might prompt these different processes, and why such different processes may have evolved. If it is advantageous to us to sometimes focus almost exclusively on the base rate *even though there is additional evidence available* and vice versa (a ‘less is more’ approach, see Gigerenzer and Gaissmaier, 2011), why should it be that sometimes it is advantageous to focus on the base rate, and sometimes advantageous to focus on the additional evidence. Great care was taken in the current research to make the tasks equivalent in terms of the prominence given to each piece of information.

Bonnefon, Eid, Bautier and Jmel (2008) have suggested that the dual process theory, as it stands, is an oversimplification of the processes, and they suggest that different mechanisms operate within both systems 1 and 2, with these mechanisms being capable of yielding differing answers from within each system. As such, differing responses should not be assumed to indicate a different system. While the current research has addressed this up to a point, by examining individual differences, 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 may also affect the choice of mechanism within the analytic system. Stanovich has frequently used SATs results as a measure of cognitive ability, for instance Stanovich and 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. As mentioned above, however, Stanovich (2009) has placed more importance on the existence of a reflective mind, which uses system 2 processes but is more closely linked to thinking style than to measures of intelligence. The latter are instead more closely linked to what he terms the algorithmic mind.

The thinking style measures showed little relation to reasoning performance in the current study. It may be that the heuristics used by participants on these particular tasks (for instance, the 'fast and frugal' approach, Gigerenzer & Goldstein, 1996, and representativeness, Tversky & Kahneman, 1983) were not captured by these thinking styles. Despite finding significant correlations between many types of reasoning task and thinking dispositions, West *et al.* (2008) found no such correlation with conjunctive reasoning, and while they did observe significant correlations between thinking dispositions and disjunctive reasoning and non-casual base rate problems, such as those used here, there were of a magnitude of $r < .1$. As such, thinking dispositions accounted for very small levels of variance in reasoning performance, as was the case in the current data. It seems instead that problem format, probability or frequency, is a far greater predictor of performance on such probabilistic reasoning tasks. However, it should be noted that while West *et al.* (2008) did find highly significant correlations between Bayesian reasoning and thinking dispositions, the tasks used in that case used an updating technique, making it more explicit to participants that they should update their judgements on the basis of the additional evidence provided.

Evans, (2007b) has also suggested that any conflict between the systems will not necessarily lead to the analytic system 2 over riding the heuristic system 1. Just because the frequency format has made the possibility of analytical reasoning more transparent, this does not necessarily mean that participants will then go on to utilise that system.

10.2 Methodological Limitations

Three main methodological problems were identified in the first study, which were then addressed throughout the following research. Firstly, the Linda task has been identified as having limitations as a measure of probabilistic reasoning ability. It was used in the first study here as it makes our results easily comparable with other studies that have used the same tasks, but it was later decided that the limitations presented by Donovan and Epstein (1997) were such that the advantage of comparability with previous research were outweighed, and it was omitted from the remainder of the research. Donovan and Epstein (1997) describe tasks as being either concrete or abstract (real life examples or abstract, letter based examples) and either natural or unnatural (heuristic thinking arrives at the right answer, or does not), and the Linda task is concrete but unnatural. They state that all concrete-unnatural tasks lead to high levels of error, but that the Linda task does so over and above other concrete-unnatural tasks. This was supported in the current research, with the Linda task showing a greater level of fallacies than the other conjunctive task used in Chapter 5.

Similarly, the Venus Williams disjunctive task was not used after the initial study. It was felt that the task led participants to believe that their general knowledge may have an impact on their ability to ‘correctly’ solve the problem. Two participants did note on their answer sheets that they did not know who Venus Williams was, in justification of their (perceived) poor performance on that particular task. On reflection, the task also refers to the likelihood that Venus will ‘lose the first set’, and the term ‘set’ may not be one that is known to all participants in this context, and may again prime the participants to feel unable to tackle the task. Whilst Tversky and Kahneman (1983) used similar materials in their own research, it is suspected that their audience may have been more familiar with their protagonist, Bjorn Borg, than the current sample was with Venus Williams, in part because Venus is an American. With regards to the anomalous result found with this task – that participants did significantly better in the probability condition – it was felt that the task did not translate well from the probability to the frequency format. When asking about ‘100 top ranked female tennis players’ it is clear

that not all 100 can have the top ranking at any one time, which again is an issue which should not, in logical terms, impact on a person's ability to complete the task correctly, but may have led to the task not being as concrete (that is, possible) in the frequency condition as in the probability condition.

This first study also failed to present many of the tasks in a manner that allowed for the normative estimates of the conjunctions/disjunctions to be calculated. This is discussed in full in Chapter 5, and essentially limited the data that could be analysed, and could therefore be used to address one of the main aims of the research – to be able to quantify the magnitude of participants' reasoning errors. A further issue with this measure is that as it was calculated from the participants' own judgements of the components, the implicit 'normative' disjunctive estimate could be over 100, which is clearly a logical impossibility. This was again addressed in the following chapters, where judgements were scaled down to take account of subadditivity so that no implicit normative outcome was an impossible value.

This method of calculating the error also has its own limitations, which are to an extent seemingly unavoidable. The error measure, whether originally negative or positive, is converted to an absolute value, so that over and underestimates are treated similarly. This was done in order to calculate more meaningful mean errors scores for each conjunctive or disjunctive judgement. In taking a participant's mean inclusive disjunction score (for instance), if they had on one task overestimated the disjunction and obtained a positive error score, but on another underestimated it and obtained a negative score, a mean of these would result in the total amount of error committed by that participant on those tasks being under represented. This current method of measurement was useful in allowing the researcher to look not just at whether a fallacy had been committed, but also at the *amount* by which they tend to be committed. Furthermore, focussing on fallacies totally ignores the tendency to underestimate conjunctive probabilities and overestimate disjunctive ones while the use of absolute errors captures these tendencies. This was successful in its main aim of gaining more detail about performance than the categorical 'fallacy' or 'no fallacy', as illustrated by the findings of Chapters 6 and 7. The use of the scale data did allow a clearer illustration of the differences between performance on the two types of induction task, and also in many cases showed stronger effects than the fallacy data.

There is clearly room for improvement in this measure, not least due to the fact that in disjunctive tasks, scaling down to take subadditivity into account leads to some of the variance in the data being lost (see 6.2.2). Only those that do show subadditivity (those participants whose implied normative is greater than 1) are subject to this process. Those that do not show subadditivity are not scaled down, which can mean that those who reasoned almost ‘correctly’ can end up with an error measure which is the same as one who has reasoned less ‘correctly’.

For instance, participant A gives component values of .5, and .6, leading to a ‘normative’ exclusive disjunction of 1.1. They give as their own estimate of the exclusive disjunction a value of 1. This leads to an error of .09. Participant B gives component values of .5 and .5, leading to a normative of 1, but gives the exclusive disjunction a value of .9. In this case participant B now has an error value of .10, greater than participant A’s error value of .09 – despite the facts that a) both were .1 away from their implied normative, and b) participant A’s judgements showed subadditivity, which in itself indicates a lack of understanding of the nature the calculation involved and/or the logical impossibility of a disjunctive value that exceeds 1. As such, some of the information about the participants’ judgements has been lost in this process – although far less than is lost through the still more reductionist approach of simply labelling all disjunctive responses as either committing the fallacy or not doing so.

An alternative to the system adopted in this research was to use as an error score the ratio of actual response to normative, again in order to account for subadditivity. This however led to a range of extreme values. In the problem regarding the student who might ‘get an A in Maths or an F in maths’, for instance, the majority of the errors calculated in this way are clustered around values of 2 or less (showing a relatively small overestimate, a correct answer, with a ratio of 1, or an underestimate with a value of less than one) while there are a small – but noteworthy – number of values that are far greater than this, from values of around 3 to up to 16. These naturally lead to the data being greatly positively skewed, whereas the method described above and used in the current study produced a scale with no outliers, and a distribution far closer to normal.

Potential improvements to the measure would need to take into account the problem identified above, that the error measure is absolute, and so both underestimates and

overestimates are just given as a value of absolute error, leading to a mean score that may be hiding the true nature of the data (although this can be examined in greater detail by looking at each task's individual mean, as in Chapters 6 and 7). This would certainly be the case where a participant had overestimated conjunction values on one task, but underestimated them on another, something which was only occasionally apparent in the current data. This leads us to investigate the utility of treating each conjunctive task separately, so that instead of each participant having one absolute conjunctive error score, they had a number of scores. Creating a group mean of this one score would then be self-defeating – if half the participants greatly overestimated, and half underestimated, then the resulting mean would be close to zero, giving the false impression that error rates were low. This would not have been an issue in the current data (very few responses did not fit the general trend of over or under estimating), but its very possibility helps to illustrate the value of continuing to use the fallacy measure to provide a more detailed picture of the way participants are responding.

A further possible innovation would be to split them into fallacy groups and calculate how much the non-fallacy group were away from the normative, and how much the fallacy group were away from the normative. Within the fallacy group, this would remove the problem of the over and under estimates cancelling each other out, and would slightly enrich the fallacy data, by indicating the average amount by which *all* participants were making errors, even those who avoided the actual fallacy. However, this does not address the issue of subadditivity, and nor does it deal with the issue of over and underestimates within the non-fallacy group.

Clearly this new method has its flaws, in terms of losing some of the detail in the data due to being an absolute measure, and being scaled down. However, it is also clear that it does add a large amount of detail which is lost when only the binomial 'fallacy or no fallacy' approach is used. Also worthy of consideration is the way in which Bayesian tasks are scored. In the current research they were categorised on the basis of the cues in the tasks on which they were based, or as attempts at calculation that clearly utilised two such cues, or they were used as raw data. Both of these methods have their limitations, with the categorisation of the responses being to some extent subjective (albeit based on the previous research of, for instance, Birnbaum, 2004), and the raw responses leading to a broader brushstroke of showing a tendency to over or underestimate the response. Previous research has used verbal protocol analysis to

understand how participants have made their judgements (Yates & Carlson, 1986; Ericsson & Simon, 1993), but with the analytical system 2 being known to be frequently involved in confabulation, justifying decisions made as being rational when in reality they were highly intuitive, the value of such protocols is highly questionable (Evans, 2008; see also Yates & Carlson, 1986, for an acknowledgement that such protocols can be justifying the given response). DeNeys and Glumicic (2008) also found that participants' verbal protocols were not always reliable, as they would not state that they were aware of the base rates, despite their responses to further tasks suggesting that they actually had processed the relevant information.

One of the measures used in this study was the paper and pencil measure of information processing speed, based on a computerised task developed and used by Fisk (2005), Fisk and Sharp (2002), and Salthouse and Babcock (1991). Participants were presented with sets of letters and asked to identify whether they were identical or had one letter different in each case, circling 'D' for different or 'S' for same, as appropriate. The paper and pencil version was developed due to previous research finding that older people (defined as over the age of 60, as in the current research) remain less familiar with, and more anxious about, using personal computers than their younger counterparts (Czaja, Charness, Fisk, Hertzog, Nair, Rogers, *et al.*, 2006). While this may not be a problem when all tasks are being conducted on a computer, it would have been an extraneous variable in this case, where the remainder of the data was completed on pencil and paper.

Within all research looking at effects of ageing, and taking a 'snapshot' of two or more separate age groups, there is likely to be a cohort effect, with different age groups having life experiences that are, to an extent, unique. In the current research, the two groups were matched, as far as practicable, by recruiting older participants through the University of the Third Age. This ensured that despite many of them being retired from full time employment, they remained physically and mentally active, and were as interested in learning and taking part in research as the younger group of university students. There was also very little difference between the mean length of education for each group, with the one year difference being attributable to the change in statutory school leaving age in the United Kingdom. The different cohorts will have inevitably experienced different educations however, with curriculums and teaching methods changing and developing over the years. As such, it remains a possibility that the

participants in each cohort may have arrived at the study with age-related differences stemming from their previous training on such tasks which would attenuate for any other aging related differences (however, see also section 4.4).

10.3 Conclusions

The research presented here introduced a novel way of assessing probabilistic reasoning performance, with the error measure providing a value which describes the distance from the 'normative' answer which potentially provides greater insight into how people are completing such tasks. This was particularly the case for inclusive and exclusive disjunctions, with evidence being presented that suggests participants are greatly underestimating exclusive disjunctive values, having assigned a large probability value to each of the components. This is possibly due to approaching them as if they inclusive tasks (although it is very unlikely that this is a conscious decision), and failing to appreciate the differences in calculation required. This lead to large amounts of subadditivity in the exclusive tasks in particular. With inclusive disjunctions the same tendency may exist in terms of judging each of the individual statements to be very likely, but because they are not mutually exclusive, the conjunction can be meaningfully defined and can be used to adjust the normative value downward so that subadditivity is avoided.

Although some links between thinking style were found through the use of disjunction error data, and the categorisation of responses to Bayesian tasks, the research as a whole did not find any consistent associations between thinking styles and reasoning ability. In order to overcome some of the problems presented by ageing research, and specifically the education related cohort effect that may well particularly influence performance on tasks involving judgements of probability, it is advisable for future research to include more background measures, to obtain more detail about participants' mathematical achievements, experiences and abilities. However, while it is often desirable to include additional measures, it is crucial that this is done without generating high levels of fatigue thereby compromising performance on the cognitively demanding probabilistic reasoning tasks.

The findings of the effect of format presented here generally support the assertion that the frequency format primes the analytic system 2 (Sloman, 2002; Sprenger & Dougherty, 2006) in the case of conjunctive, inclusive disjunctive and exclusive

disjunctive reasoning tasks, and that presenting data in natural frequencies can significantly reduce the overestimates made in a Bayesian task. There are also some findings here that link thinking styles with performance on reasoning tasks, but crucially they do not provide clear cut evidence that such thinking styles are directly linked to instances of each of the dual processes of analytical and heuristic reasoning being applied to the reasoning tasks. This suggests two main propositions. First, that the dual process theory may be something of an oversimplification, as suggested by Cleermans and Jimenez (2002) Osman, (2004), Stanovich (2009) and Thompson (2009). Second, that the measures used are not adequately tapping into the processes of interest, possibly due to limitations inherent in the measures themselves, and their ability to directly assess propensities for ‘analytic’ and ‘heuristic’ styles of reasoning.

References

- Agnoli, F., & Krantz, D. H. (1989). Suppressing Natural Heuristics by Formal Instruction: The Case of the Conjunction Fallacy. *Cognitive Psychology*, 21, 515-550.
- Akerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52, 39-37.
- Baltes, P. B. (1997). On the incomplete architecture of human ontogeny: Selection, optimization, and compensation as foundation of developmental psychology. *American Psychologist*, 52, 366 - 380.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241-297.
- Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: a disjunction fallacy in probability judgements. *Journal of Personality and Social Psychology*, 65, 1119 - 1131.
- Baratgin, J. (2002). Is the human mind definitely not Bayesian? A review of the various arguments. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 21, 653 - 680.
- 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*. Hove, UK: Psychology Press.
- Birnbaum, M. H., Anderson, C. J., & Hynan, L. G. (1990). Theories of bias in probability judgment. In J. P. Caverni, J. M. Fabre & M. Gonzalez (Eds.), *Cognitive Biases* (pp. 477-498). Amsterdam: North Holland.
- Blanchard-Fields, F. (1996). Emotion and everyday problem solving in adult development. In S. H. McFadden & C. Magai (Eds.), *Handbook of emotion, adult development, and aging*. (pp. 149-165). San Diego, CA, US: Academic Press.
- Bonnefon, J. F., Eid, M., Vautier, S., & Jmel, S. (2008). A mixed Rasch model of dual-process conditional reasoning. *The Quarterly Journal of Experimental Psychology*, 61, 809-824.
- Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin and Review*, 15, 248-289.

- Burke, D. M., & Osborne, G. (2007). Aging and Inhibition Deficits: Where are the Effects? In D. S. Gorfein & M. Macleod (Eds.), *Inhibition in Cognition*. Washington DC: American Psychological Association.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116-131.
- Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, B. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197-253.
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306-307.
- Carlson, B. W., & Yates, J. F. (1989). Disjunction Errors in Qualitative Likelihood Judgment. *Organizational Behaviour and Human Decision Processes*, *44*, 368-379.
- Chapman, G. B., & Liu, J. (2009). Numeracy, frequency and Bayesian reasoning. *Judgment and Decision Making*, *4*, 34-40.
- 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.
- Chasseigne, G., Mullet, E., & Steward, T. R. (1997). Aging and multiple cue probability learning: the case of inverse relationships. *Acta Psychologica*, *97*, 235-252.
- Chen, Y., & Sun, Y. (2003). Age differences in financial decision-making: using simple heuristics. *Educational Gerontology*, *29*, 627-635.
- Chiesi, F., Gronchi, G. & Primi, C. (2008) Age-trend-related differences in tasks involving conjunctive probabilistic reasoning. *Canadian Journal of Experimental Psychology*, *62*, 188-191.
- Christensen, I. P. (1979). Distributional and Non-distributional Uncertainty. In C. R. Bell (Ed.), *Uncertain Outcomes*. Lancaster: MTP Press.
- Clark, E., Gardner, M. K., & Brown, G. (1990). Changes in analogical reasoning in adulthood. *Experimental Aging Research*, *16*, 95-99.
- Cleermans, A., & Jimenez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleermans (Eds.). *Implicit Learning and Consciousness: An Empirical, Philosophical and Computational Consensus in the Making*. New York: Taylor and Francis.

- Cobos, P. L., Almaraz, J., & Garcia-Madruga, J. A. (2003). An associative framework for probability judgment: An application to biases. *Journal of Experimental Psychology*, 29, 80-96.
- Copeland, D. E., & Radvansky, G. A. (2004). Working memory and syllogistic reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 27A, 1437-1457.
- 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.
- Costello, F. J. (2009) Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making*, 22, 235-251.
- Craik, F. I. M. (1994). Memory changes in normal aging. *Current Directions in Psychological Science*, 3, 155-158.
- Crisp, A. K. & Feeney, A. (2009) Causal conjunction fallacies: The roles of causal strength and mental resources. *The Quarterly Journal of Experimental Psychology*, 62, 2320-2337.
- Croskerry, P. (2009) Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in Health Sciences Education*, 14, 27-35.
- Crupi, V., Fitelson, B., & Tentori, K. (2008) Theoretical note: Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning*, 14, 182-199.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., et al. (2006). Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and Aging*, 21, 333-352.
- Dagnall, N., Parker, A., & Munley, G. (2007). Paranormal belief and reasoning. *Personality and Individual Differences*, 43, 1406-1415.
- De Beni, R., & Palladino, P. (2004). Decline in working memory updating through ageing: Intrusion error analyses. *Memory*, 12, 75-89.
- De Neys, W. (2006a). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *The Quarterly Journal of Experimental Psychology*, 59, 1070-1100.
- De Neys, W. (2006b). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, 17, 428-433.

- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking *Cognition*, *106*, 1248-1299.
- Donovan, S., & Epstein, S. (1997). The Difficulty of the Linda Conjunction Problem Can Be Attributed to Its Simultaneous Concrete and Unnatural Representation, and Not to Conversational Implicature. *Journal of Experimental Social Psychology*, *33*, 1-20.
- Epstein, S. (1973). The self-concept revisited. *American Psychologist*, *May*, 404-416.
- Epstein, S. (2003). Cognitive-Experiential Self-Theory of Personality. In T. Millon & M. J. Lerner (Eds.), *Handbook of Psychology Volume 5: Personality and Social Psychology*. New Jersey, US: John Wiley & Sons.
- Epstein, S., Lipson, A., Holstein, C., & E., Huh. (1992). Irrational reactions to negative outcomes: evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *62*, 328-339.
- Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in Intuitive-experiential and analytical rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390-405.
- Ericsson, K. A. & Simon (1993). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts, Massachusetts Institute of Technology,
- Erwin, T. D. (1983). The scale of intellectual development: Measuring Perry's scheme. *Journal of College Student Personnel*, *24*, 6-12.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, *13*, 378-395.
- Evans, J. St. B. T. (2007a). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. New York: Psychology Press.
- Evans, J. St. B. T. (2007b). On the resolution of conflict in dual process theories of reasoning. *Thinking and Reasoning*, *13*, 321-339.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.
- Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans & Frankish, K. (Eds.) *In Two Minds: Dual Processes and Beyond*. New York, Oxford University Press
- Evans, J. St. B. T. (2010a). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*, 313-326.
- Evans, J. St. B. T. (2010b). *Thinking Twice: Two Minds in One Brain*. New York, Oxford University Press.

- Evans, J. St. B. T., Barston, J. L. & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295-306.
- Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning*, *11*, 382-389.
- Evans, J. St. 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.
- Evans, J. St. B. T., & Over, D. (1996). A Dual Process Theory of Thinking Rationality and Reasoning *Rationality and Reasoning*. Hove, UK: Erlbaum.
- Evans, J. St. B. T., Over, D., & Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition*, *49*, 165-187.
- Fabre, J. M., & Caverni, J. P. (1995). Causality does influence conjunctive probability judgments if context and design allow for it. *Organizational Behaviour and Human Decision Processes*, *63*, 1-5.
- Fantino, E., & Stolarz-Fantino, S. (2005). Decision-making: Context matters. *Behavioural Processes*, *69*, 165-171.
- Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Attitudes and Social Cognition*, *91*, 797-813.
- Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*, 123-129.
- Field, A. (2009) *Discovering Statistics Using SPSS*. London, SAGE
- Finucane, M. L., Slovic, P., Hibbard, J. H., Peters, E., Mertz, C. K., & MacGregor, D. G. (2002). Aging and decision making competence: an analysis of comprehension and consistency skills in older versus younger adults considering health plan options. *Journal of Behavioural Decision Making*, *15*, 141 - 164.
- Fisk, J. E. (1996). The conjunction effect: Fallacy or Bayesian inference? *Organizational Behaviour and Human Decision Processes*, *67*, 76-90.
- Fisk, J. E. (2002). Judgments under uncertainty: Representativeness or potential surprise? *British Journal of Psychology*, *93*, 431-449.
- Fisk, J. E. (2004). Conjunction Fallacy. In R. F. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment, and Memory*. London, UK: Psychology Press.

- 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.
- Fisk, J. E., Bury, A. S., & Holden, R. (2006). Reasoning about complex probabilistic concepts in childhood. *Scandinavian Journal of Psychology*, 47(497-504).
- Fisk, J. E., & Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, 94, 1-20.
- Fisk, J. E., & Pidgeon, N. (1997). The conjunction fallacy: The case for the existence of competing Heuristic strategies. *British Journal of Psychology*, 88, 1 - 27.
- Fisk, J. E., & Pidgeon, N. (1998). Conditional Probabilities, Potential Surprise, and the Conjunction Fallacy. *The Quarterly Journal of Experimental Psychology*, 51A, 655-681.
- Fisk, J. E., & Sharp, C. (2002). Syllogistic reasoning and cognitive ageing. *The Quarterly Journal of Experimental Psychology*, 55A, 1273-1293.
- Fisk, J. E., & Slattery, R. (2005). Reasoning About Conjunctive Probabilistic Concepts in Childhood. *Canadian Journal of Experimental Psychology*, 59, 168-178.
- Fisk, J. E., & Warr, P. (1996). Age-related impairment in associative learning: The role of anxiety, arousal and learning self-efficacy. *Personality and Individual Differences*, 21(5), 675-686.
- Fiske, S. T. (1991). *Social Cognition* (2nd ed.). New York: McGraw-Hill.
- Fontaine, R., & Pennequin, V. (2000). Effect of aging on inferential reasoning about class inclusion. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 19, 453-482.
- Franco, R. (2009) The conjunction fallacy and interference effects. *Journal of Mathematical Psychology*, 53, 425-422.
- Franssens, S. & De Neys, W. (2009) The effortless nature of conflict detection during thinking. *Thinking and Reasoning*, 15, 105-128.
- Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and Conjoint Probability. *Journal of Personality and Social Psychology*, 61, 181-194.
- Gaynor, S. T., Wahio, Y., & Anderson, F. (2007). The conjunction fallacy: a derived stimulus relations conceptualization and demonstration experiment. *The Psychological Record*, 57, 63-85.
- Gigerenzer, G. (1996a). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making*, 16, 273-280.

- Gigerenzer, G. (1996b). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, *103*, 592-596.
- Gigerenzer, G. & Brighton (2009). Homo heuristics: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107-143.
- Gigerenzer, G. & Gaissmaier (2011) Heuristic decision making. *Annual Review of Psychology*, *62*, 451 – 482.
- 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.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650-669.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.
- 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.
- Gigerenzer, G. & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behavioral and Brain Sciences*, *30*, 264-267.
- Gilinsky, A. S., & Judd, B. B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging*, *9*(356-371).
- Goldsmith, R. W. (1978). Assessing probabilities of compound events in a judicial context. *Scandinavian Journal of Psychology*, *19*, 103-110.
- Handley, S. J., Evans, J. S. B. T., & Thompson, V. A. (2006). The negated conditional: A litmus test for the suppositional conditional? *Journal of Experimental Psychology*, *32*, 559-569.
- Hertwig, R., Benz, B., & Krauss, S. (2008) The conjunction fallacy and the many meanings of *and*. *Cognition*, *108*, 740-753.
- Hertwig, R., & Chase, V. M. (1998). Many Reasons or Just One: How Response Mode Affects Reasoning in the Conjunction Problem. *Thinking and Reasoning*, *4*. 319-352.
- Hess, T. M. (1990). Aging and Schematic Influences on Memory. In T. M. Hess (Ed.), *Aging and Cognition: Knowledge organization and utilization*. Oxford, England: North-Holland.

- Hess, T. M., Osowski, N. L. & Leclerc, C. M. (2005). Age and experience influences on the complexity of social inferences. *Psychology and Aging, 20*, 447-459.
- 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.
- 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.
- Hogarth, R. M. & Karelaia, N. (2006) "Take-the-best" and other simple strategies: Why and when they work "well" with binary cues. *Theory and Decision, 61*, 205-249.
- Hogarth, R. M. & Karelaia, N. (2007) Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733-758.
- 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*(2), 245-263.
- Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007) Paradoxical effects of base rates and representation in category learning. *Memory and Cognition, 35*, 1365-1379.
- Johns, M. W. (1992). Reliability and factor analysis of the epworth sleepiness scale. *Journal of Sleep Research and Sleep Medicine, 15*, 376-381.
- Johnson, M. M. S. (1993). Thinking about strategies during, before and after making a decision. *Psychology and Aging, 8*, 231-241.
- Johnson-Laird, P. N. (1983). *Mental Models: Towards a cognitive science of language, inference and consciousness*. Cambridge: Cambridge University Press.
- Johnson-Laird, P. N. (2005). Mental Models and Thought. In R. G. Morrison & K. J. Holyoak (Eds.), *The Cambridge Handbook of Reasoning and Thinking*. New York: Cambridge University Press.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*, 62-88.
- Juslin, P. & Persson, M. (2002) PRObabilities from Exemplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563-607.
- Juslin, P., Nilsson, H., & Winman, A. (2009) Probability theory, not the very guide of life. *Psychological Review, 116*, 856-874.

- Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, 78 (1, Whole No. 578).
- 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.
- Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In R. G. Morrison & K. J. Holyoak (Eds.), *The Cambridge Handbook of Reasoning and Thinking*. New York: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology*, 3, 430 - 454.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237 - 251.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582-591.
- Kariyazono, A. (1991). The study of conjunction fallacy in the probability judgment task of compound event. *Japanese Journal of Psychonomic Science*, 10, 57-64.
- Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability and epistemological beliefs: adult age differences in everyday reasoning biases. *Psychology and Aging*, 15, 400-416.
- Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, 83, 26-52.
- Krampe, R. T. & Charness, N. (2006). Aging and expertise. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, Cambridge University Press.
- Krynski, T. R. & Tenenbaum, J. B. (2007) The role of causality in judgment under uncertainty. *Journal of Experimental Psychology*, 136, 430-450.
- Lieberman, M. D. (2007). The X- and C-systems: The neural basis of automatic and controlled social cognition. In E. Harmon-Jones & P. Winkielman (Eds.) *Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior*. New York, Guilford Press.

- Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences, 17*, 115-127.
- Manktelow, K. I. (2004) Reasoning and Rationality; The pure and the practical. In K.I. Manktelow & Chung, M. C. (Eds.), *Psychology of Reasoning: Theoretical and Historical Perspectives*. New York; Psychology Press
- Manktelow, K. I. & Over, D. E. (1990). *Inference and Understanding: A Philosophical and Psychological Perspective*. London; Routledge
- Markus, H., & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology* (3rd Edition ed., Vol. 1). New York: Lawrence Erlbaum Associates.
- Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging, 22*, 796-810.
- Mellers, B. & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review, 106*, 417-424.
- Mickler, C. & Staudinger, U. M. (2008). Personal wisdom: Validation and age-related differences of a performance measure. *Psychology and Aging, 23*, 787-799.
- Morrier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin, 10*, 243-252.
- Moutier, S., & Houdé, O. (2003). Judgement under uncertainty and conjunction fallacy inhibition training. *Thinking and Reasoning, 9*, 185-201.
- Mutter, S. A. (2000). Illusory correlation and group impression formation in young and older adults. *Journal of Gerontology, 55B*, 224-237.
- Mutter, S. A., & Goedert, K. M. (1997). Frequency discrimination vs. frequency estimation: Adult age differences and the effect of divided attention. *Journal of Gerontology, 52B*, 319-328.
- Mutter, S. A., Haggblom, S. J., Plumlee, L. F., & Schirmer (2006). Aging, working memory, and discrimination learning. *The Quarterly Journal of Experimental Psychology, 59*, 1556-1566.
- Mutter, S. A., & Pliske, R. M. (1994). Aging and illusory correlation in judgments of co-occurrence. *Psychology and Aging, 9*, 53-63.
- Mutter, S. A. & Plumlee, L. F. (2009). Aging and integration of contingency evidence in causal judgment. *Psychology and Aging, 24*, 916-926.

- Mutter, S. A., & Williams, T. W. (2004). Aging and the detection of contingency in causal learning. *Psychology and Aging, 19*, 13-26.
- Neace, W. P., Michaud, S., Bolling, L. Deer, K. & Zecevic, L. (2008). Frequency Formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making, 3*, 140-152.
- Nilsson, H. (2008) Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making, 21*, 471-490.
- Nilsson, H. & Andersson, P. (2010) Making the seemingly impossible appear possible: Effects of conjunction fallacies in evaluations of bets on football games. *Journal of Economic Psychology, 31*, 172-180.
- Nilsson, H., Juslin, P., & Olsson, H. (2008) Exemplars in the mist: The cognitive substrate of the representativeness heuristic. *Cognition and Neurosciences, 49*, 201-212.
- Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunctions. *Thinking and Reasoning, 8*, 297-326.
- Oechssler, J., Roider, A., & Schmitz, P. W. (2009) Cognitive abilities and behavioral biases. *Journal of Economic Behavior and Organization, 72*, 147-152.
- Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia, 36*, 369-376.
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review, 11*, 988-1010.
- Over, D. (2005). Naive probability and its model theory. In V. Girotto & P. N. Johnson-Laird (Eds.), *The Shape of Reason: Essays in Honour of Paulo Legrenzi*. New York: Psychology Press.
- Park, D. C., Willis, S. L., Morrow, D., Diehl, M., & Gaines, C. L. (1994). Cognitive function and medication usage in older adults. *The Journal of Applied Gerontology, 13*, 39-57.
- Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex, 11*, 954-965.

- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.
- Peters, E., Finucane, M. L., MacGregor, D. G., & Slovic, P. (2000). The bearable lightness of aging: Judgment and decision processes in older adults. In C. o. F. D. f. C. R. o. A. National Research Council, P. C. Stern & L. L. Carstensen (Eds.), *The aging mind: Opportunities in cognitive research* (pp. Appendix C, 144-165). Washington DC: National Academy Press.
- Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29-46.
- Phillips, L. D. & Edwards, W. (1966) conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346-354.
- Piaget, J. (Ed.). (1952). *The Child's Conception of Number*. London: Routledge and Kegan Paul.
- Raven, J. C., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 3, The standard progressive matrices*. Oxford: Oxford Psychologists 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.
- Reyna, V. F. & Brainerd, C. J. (2007) Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89-107.
- Riis, J., & Schwarz, N. (2003). Approaching and avoiding Linda: Motor signals influence the conjunction fallacy. *Social Cognition*, *21*, 247-262.
- Roberge, J. J. (1976). Effects of negation on adults' disjunctive reasoning abilities. *The Journal of General Psychology*, *94*, 23-28.
- Rogers, P., Davis, T. & Fisk, F. (2009) Paranormal belief and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, *23*, 524-542.
- Rybash, J. M., Hoyer, W. J. & Roodin, P. A. (1986). *Adult Cognition and Aging: Developmental Changes in Processing, Knowing and Thinking*. New York, Pergamon Press.
- Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*, 497-510.

- Saczynski, J. S., Willis, S. L., & Schaie, K. W. (2002). Strategy use in reasoning training with older adults. *Aging Neuropsychology and Cognition*, 9(1), 48-60.
- Salthouse, T. A. (1993). Speed mediation of adult age differences in cognition. *Developmental Psychology*, 29, 722-738.
- Salthouse, T. A. (1998). Independence of age-related influences on cognitive abilities across the life span. *Developmental Psychology*, 34, 851-864.
- Salthouse, T. A. (2000). Item analyses of age relations on reasoning tests. *Psychology and Aging*, 15, 3-8.
- Salthouse, T. A. (2005). Effects of Aging on Reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York, NY: Cambridge University Press.
- Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, 27, 763-776.
- Saunders, C. & Over, D. (2009). In two minds about rationality? In J. St. B. T. Evans & Frankish, K. (Eds.) *In Two Minds: Dual Processes and Beyond*. New York, Oxford University Press
- Sedlmeier, P., & Gigerenzer, G. (2001). Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology*, 130, 380-400.
- Shackle, G. L. S. (1969). *Decision, order and time in human affairs*. Cambridge: Cambridge University Press.
- Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, 30, 191-198.
- 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.
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296-309.
- Sprenger, A., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behaviour and Human Decision Processes*. 99, 202-211.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NF: Lawrence Erlbaum Associates Inc.

- Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago, Chicago University Press.
- Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & Frankish, K. (Eds.) *In Two Minds: Dual Processes and Beyond*. New York, Oxford University Press
- Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342-257.
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161-188.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645 - 726.
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning, 13*, 225 - 247.
- Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Personality Processes and Individual Differences, 94*, 672-695.
- Stanovich, K. E., West, R. F., & Sá, W. C. (1999). Thinking dispositions and decontextualized reasoning. In K. E. Stanovich (Ed.), *Who is Rational: Studies of Individual Differences in Reasoning*. New Jersey: Lawrence Erlbaum Associates.
- Staudinger, U. M. (1999). Older and wiser? Integrating results on the relationship between age and wisdom-related performance. *International Journal of Behavioural Development, 23*, 641-664.
- Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. *The American Journal of Psychology, 116*, 15-34.
- Stolarz-Fantino, S., Fantino, E., & Van Borst, N. (2006) Use of base rates and case cue information in making likelihood estimates. *Memory and Cognition, 34*, 603-618.
- Stuart-Hamilton, I. (2006). *The Psychology of Ageing: An Introduction*. Gateshead: Jessica Kingsley Publishers
- Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th edn). Boston MA: Allyn and Bacon.

- Teigen, K. H. (2004). Judgements by representativeness. In R. F. Pohl (Ed.), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment, and Memory*. London, UK: Psychology Press.
- Teigen, K. H. (2005). The proximity heuristic in judgments of accident probabilities. *British Journal of Psychology*, *96*, 423-440.
- Teigen, K. H., Brun, B., & Frydenlund, R. (1999). Judgments of Risk and Probability: The Role of Frequentistic Information. *Journal of Behavioural Decision Making*, *12*, 123-139.
- Teigen, K. H. & Keren, G. (2007) Waiting for the bus: When base-rates refuse to be neglected. *Cognition*, *103*, 337-357.
- Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, *28*, 467-477.
- Tentori, K., Osherson, D., Hasher, L., & May, C. (2001). Wisdom and aging: irrational preferences in college students but not older adults. *Cognition*, *81*, B87-B96.
- Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. St. B. T. Evans & Frankish, K. (Eds.) *In Two Minds: Dual Processes and Beyond*. New York, Oxford University Press.
- Thornton, W. J. L., & Dumke, H. A. (2005) Age differences in everyday problem-solving and decision-making effectiveness: A meta-analytic review.
- Thuring, M., & Jungermann, H. (1990). The conjunction fallacy: Causality vs. event probability. *Journal of Behavioural Decision Making*, *3*, 61-74.
- Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *94*, 197-209.
- Troldahl, V. C., & Powell, F. A. (1965). A short-form dogmatism scale for use in field studies. *Social Forces*, *44*, 211-215.
- Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in Social Psychology*. New Jersey: Erlbaum.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking and Reasoning*, *11*, 239-278.
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational Integration, Inhibition, and Analogical Reasoning in Older Adults. *Psychology and Aging*, *19*, 581-591.

- Wareing, M., Fisk, J. E., Montgomery, C., Murphy, P. N., & Chandler, M. D. (2007) Information processing speed in ecstasy users. *Human Psychopharmacology*, 22, 81-88.
- Wedell, D. H. & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 107, 105-136.
- Wells, G. L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition*, 3, 266-279.
- West, R. F., Toplak, M. E. & Stanovich, K. E. (2008) Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100, 930-941.
- Wolford, G., Taylor, H. A., & Beck, J. R. (1990). The conjunction fallacy? *Memory & Cognition*, 18, 47-53.
- Wyer, R. S. (1970). Quantitative prediction of belief and opinion change: A further test of a subjective probability model. *Journal of Personality and Social Psychology*, 16, 559-570.
- Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology*, 50, 97-106.
- Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: evidence for multiple judgment procedures, including 'signed summation'. *Organizational Behavior and Human Decision Processes*, 37, 230-253.
- Yates, J. F., & Patalano, A. L. (1999). Decision Making and Aging. In D. C. Park, R. W. Morrell & K. Shifren (Eds.), *Processing of medical information in aging patients: Cognitive and human factors perspectives*. Mahwah, NJ: Erlbaum.
- Zizzo, D. J. (2003). Verbal and behavioral learning in a probability compounding task. *Theory and Decision*, 54, 287-314.

Appendices

Appendix 1 – Bob task (Chapter 5)

Probability version:

Bob works five days a week. Most days he drives to work but occasionally he takes the bus when his wife uses the family car for shopping. He usually buys a sandwich at the local delicatessen (deli) but sometimes he eats at the local public house (pub) with his work mates. On an average day how likely is it that:

Bob takes the bus to work and buys a sandwich at the deli

Bob takes the bus to work

Bob buys a sandwich at the deli

Bob buys a sandwich at the deli given that he takes the bus to work.

Frequency version:

Imagine a man who works five days a week. Most days he drives to work but occasionally he takes the bus when his wife uses the family car for shopping. He usually buys a sandwich at the local delicatessen (deli) but sometimes he eats at the local public house (pub) with his work mates. Now imagine that there are exactly 100 men fitting this description. Out of these 100, how many do you think, on an average day at work:

Take the bus to work and buy a sandwich at the deli

Take the bus to work

Buy a sandwich at the deli

Of those that buy a sandwich at the deli, how many also take the bus to work?

Appendix 2 – Venus task (Chapter 5)

Probability version:

If Venus Williams plays in NEXT YEAR'S opening match at Wimbledon how likely is it that:

She will lose the first set or break her racquet

She will break her racquet

She will lose the first set

She will play with only one racquet for the entire match

She will lose the first set, given that she breaks her racquet

She will break her racquet or she will only use one racquet for the whole match

Frequency version:

Imagine there are 100 top ranked female tennis players in a tennis tournament. How many of these players will, in their first match:

Lose the first set or break their racquet

Break their racquet

Lose the first set

Play with only one racquet for the entire match

Break their racquet or only use one racquet for the whole match

Of those who lose the first set, how many will also break their racquet?

Appendix 3 – Participant Instructions (Chapter 5)

Probability version:

Before you begin, please fill in your PARTICIPANT NUMBER, your AGE in years, and whether you are MALE or FEMALE.

PARTICIPANT	AGE	MALE	FEMALE

In this questionnaire you will be presented with a number of statements. You will be asked to judge how likely each statement is. Please make your judgements by specifying how many chances in 100 there are that the statement is or will be true. An example is provided on the following page. You may choose any number between 0 and 100. If you feel that a statement is impossible or that it is definitely untrue you would enter the number 0 in the box provided. On the other hand if you feel that a statement is definitely true or certain to occur then you would enter the number 100.

Sometimes you may be very confident about your prediction. Next to each estimate there is another box where you can indicate just how confident you are. Enter the number 10 if you are totally confident in your prediction. Enter the number 0 if you have absolutely no confidence in your estimate. If you are neither confident nor unconfident about your estimate then you would enter the number 5. Of course, you may also use all of the other numbers in between.

On each of the pages that follow you will be given some back ground information followed by a series of statements for you to evaluate. In making your judgement make use of the information provided and of your own personal knowledge.

In each case remember to indicate how confident you are in your prediction by entering a number between 0 and 10. If you make a mistake or change your mind cross out the incorrect estimate and replace it with your amended judgement. If you do make any changes please try to ensure that it is clear which number is your actual estimate.

HOWEVER ONCE YOU HAVE TURNED THE PAGE PLEASE DO NOT GO BACK AND MAKE ANY CHANGES TO THE PREVIOUS ONE.

Please read the example on the following page before continuing.

EXAMPLE

Peter is an athlete who is training to run the mile in an important regional competition. In three of his last 4 races, Peter has run the mile in less than 4 minutes. His best time to-date has been 3 minutes 58 seconds. For his next race how likely is it that:

	How many chances in 100. (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
Peter will run the mile in under 4 minutes	81	5
Peter will run the second half mile in under 2 minutes	90	8
Peter will run the mile in 3 minutes 55 seconds or less	39	8
Peter will take over 4 minutes 5 seconds to run the mile	35	4
Peter will run the first half mile in less than 2 minutes 15 seconds OR will lose the race	90	9
Peter will run the first half mile in less than 1 minute 50 seconds OR will win the race	16	3

Now please provide judgements for each group of statements that follow. Complete the questionnaire as quickly as possible, but without working so fast that you make mistakes.

Some of the questions will not have boxes for you to fill in your answer, but will have a line instead. Make sure you read all of the instructions thoroughly so that you understand what each individual problem is asking of you.

Please do not confer with anyone else while completing the questionnaire.

Frequency version:

Before you begin, please fill in your PARTICIPANT NUMBER, your AGE in years, and whether you are MALE or FEMALE.

PARTICIPANT	AGE	MALE	FEMALE

In this questionnaire you will be presented with a number of statements. You will be asked to judge how likely each statement is. Please make your judgements by specifying how many chances in 100 there are that the statement is or will be true. An example is provided on the following page. You may choose any number between 0 and 100. If you feel that a statement is impossible or that it is definitely untrue you would enter the number 0 in the box provided. On the other hand if you feel that a statement is definitely true or certain to occur then you would enter the number 100.

Sometimes you may be very confident about your prediction. Next to each estimate there is another box where you can indicate just how confident you are. Enter the number 10 if you are totally confident in your prediction. Enter the number 0 if you have absolutely no confidence in your estimate. If you are neither confident nor unconfident about your estimate then you would enter the number 5. Of course, you may also use all of the other numbers in between.

On each of the pages that follow you will be given some back ground information followed by a series of statements for you to evaluate. In making your judgement make use of the information provided and of your own personal knowledge.

In each case remember to indicate how confident you are in your prediction by entering a number between 0 and 10. If you make a mistake or change your mind cross out the incorrect estimate and replace it with your amended judgement. If you do make any changes please try to ensure that it is clear which number is your actual estimate. **HOWEVER ONCE YOU HAVE TURNED THE PAGE PLEASE DO NOT GO BACK AND MAKE ANY CHANGES TO THE PREVIOUS ONE.**

Please read the example on the following page before continuing.

EXAMPLE

Peter is an athlete who is training to run the mile in an important regional competition. In three of his last 4 races, Peter has run the mile in less than 4 minutes. His best time to-date has been 3 minutes 58 seconds. For his next race how likely is it that:

	How many chances in 100. (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
Peter will run the mile in under 4 minutes	81	5
Peter will run the second half mile in under 2 minutes	90	8
Peter will run the mile in 3 minutes 55 seconds or less	39	8
Peter will take over 4 minutes 5 seconds to run the mile	35	4
Peter will run the first half mile in less than 2 minutes 15 seconds OR will lose the race	90	9
Peter will run the first half mile in less than 1 minute 50 seconds OR will win the race	16	3

Now please provide judgements for each group of statements that follow. Complete the questionnaire as quickly as possible, but without working so fast that you make mistakes.

Some of the questions will not have boxes for you to fill in your answer, but will have a line instead. Make sure you read all of the instructions thoroughly so that you understand what each individual problem is asking of you.

Please do not confer with anyone else while completing the questionnaire.

Appendix 4 – Participant Information Sheet (Chapter 6)

Participant Information Sheet

Name of experimenter: Rosey Stock

Supervisory team: Dr John Fisk, Dr Phil Brooks and Dr Cathy Montgomery

Title of study/project: Age and Probabilistic Reasoning Skills

Purpose of study:

To investigate the way that reasoning ability and techniques may alter across the lifespan.

Procedures and Participants' Role:

As a participant you will be asked to complete a number of tasks with a pen and paper. These will take place at Liverpool John Moores University, and you will need to attend the University on two separate occasions, for around an hour and a half at a time. Each of these tasks will be explained to you in detail, and no experience of any of them is necessary. The tasks will be measuring various cognitive skills, such as how quickly you can process information, and the way that you think. There is no time limit on the majority of the tasks, and you may take a break in between tasks if you wish to. You will also sign a consent form, agreeing to take part in the above tasks, and you have the right to withdraw from this study at any point. This right is not affected by having signed the consent form.

All of the information you provide will be kept strictly confidential. You will be asked to sign a consent form which will be kept separate from your responses, and you will be given a participant number, to identify your responses, rather than needing to put your name on any of the actual response sheets. All other information you provide will be kept secure at all times. If you decide to withdraw from the study at any time, you can also request that information you have provided be destroyed immediately.

The results of this research will be included in my PhD thesis, and may also be published in an academic journal. You will not, at any time, be identified in these reports.

Please Note:

All participants have the right to withdraw from the project/study at any time without prejudice to access of services which are already being provided or may subsequently be provided to the participant.

Appendix 5 – Conjunctive Tasks (Chapter 6)

Probability versions:

Bob task³

Bob works five days a week. Most days he drives to work but occasionally he takes the bus when his wife uses the family car for shopping. He usually buys a sandwich at the local delicatessen (deli) but sometimes he eats at the local public house (pub) with his work mates. On an average day how likely is it that:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Bob takes the bus to work and buys a sandwich at the deli.		
2.	Bob takes the bus to work.		
3.	Bob buys a sandwich at the deli.		
4.	Bob buys a sandwich at the deli given that he takes the bus to work.		

Helen task

Helen is a sociable person who thrives on human company and seeks out exciting activities. She is restless, impulsive and optimistic. Given Helen's description please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Helen loves going to nightclubs and parties and works in the local library.		
2.	Helen loves going to nightclubs and parties.		
3.	Helen works in the local library.		
4.	Helen works at the library given that she loves going to nightclubs and parties.		

³ Participants were presented with each task on a separate page, and were not given the title. The current presentation is for the convenience of the reader.

Morris task

Morris is an orderly person, restrained and serious. He prefers the company of books to people and his hobbies include crossword puzzles and photography. Given this description of Morris please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Morris collects stamps for a hobby and plays team sports.		
2.	Morris collects stamps for a hobby.		
3.	Morris plays team sports.		
4.	Morris plays team sports given that he collects stamps for a hobby.		

Jim task

Jim is in his final year at secondary school. He has Maths almost every day. He particularly enjoys Physical Education, which takes place once or twice a week. Usually Jim takes the school bus but occasionally his parents pick him up after school. On any given day, given Jim's description please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Jim has maths and games.		
2.	Jim has maths.		
3.	Jim has games.		
4.	Jim has games given that he has maths.		

Marbles task

Suppose that two jars containing different coloured marbles are set on a table in front of you. The first jar contains 90 marbles - 10 green and 80 yellow. The second jar contains 25 marbles - 20 orange, and 5 purple. Imagine that without looking, you draw one marble from each jar, write down its colour and replace it in the appropriate jar. How likely is it that:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	You draw a green marble from the first jar and an orange one from the second jar.		
2.	You draw a green marble from the first jar.		
3.	You draw an orange marble from the second jar.		
4.	You draw an orange marble from the second jar given that you draw a green marble from the first jar.		

Frequency versions:

Bob task

Imagine a man who works five days a week. Most days he drives to work but occasionally he takes the bus when his wife uses the family car for shopping. He usually buys a sandwich at the local delicatessen (deli) but sometimes he eats at the local public house (pub) with his work mates. Now imagine that that we assemble a very large number of men fitting this description.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these men at random on any given day, how many do you think take the bus to work and buy a sandwich at the deli?		
2.	If we select 100 of these men at random on any given day, how many do you think take the bus to work?		
3.	If we select 100 of these men at random on any given day, how many do you think buy a sandwich at the deli?		
4.	From this group, if we select 100 men who have all taken a bus to work on a certain day, how many of this number do you think buy a sandwich at the deli?		

Helen task

Imagine a woman who is a sociable person who thrives on human company and seeks out exciting activities. She is restless, impulsive and optimistic. Now imagine that we assemble a very large number of women fitting this description.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these women at random, how many do you think love going to nightclubs and parties, and work in the local library?		
2.	If we select 100 of these women at random, how many do you think love going to nightclubs and parties?		
3.	If we select 100 of these women at random, how many do you think work in the local library?		
4.	From this group, if we select 100 women all of whom love going to nightclubs and parties, how many of this number do you think work in the local library?		

Morris task

Imagine an orderly person, restrained and serious. He prefers the company of books to people and his hobbies include crossword puzzles and photography. Now imagine that we assemble a very large number of such men.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these men at random, how many do you think collect stamps for a hobby and play team sports?		
2.	If we select 100 of these men at random, how many do you think collect stamps for a hobby?		
3.	If we select 100 of these men at random, how many do you think play team sports?		
4.	From this group if we select 100 men, all of whom collect stamps for a hobby, how many of this number do you think play team sports?		

Jim task

Jim is in his final year at secondary school. He has Maths almost every day. He particularly enjoys Physical Education, which takes place once or twice a week. Usually Jim takes the school bus but occasionally his parents pick him up after school. Now imagine that that we assemble a very large number of secondary school pupils all fitting this description.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	On any given day, if we select 100 of these pupils at random, how many do you think have maths and games?		
2.	On any given day, if we select 100 of these pupils at random, how many do you think have maths?		
3.	On any given day, if we select 100 of these pupils at random, how many do you think have games?		
4.	On any given day, if we select 100 pupils from this group who all have maths, how many do you think have games?		

Marbles task

Suppose that two jars containing different coloured marbles are set on a table in front of you. The first jar contains 90 marbles - 10 green and 80 yellow. The second jar contains 25 marbles - 20 orange, and 5 purple. Imagine that without looking, you draw one marble from each jar, write down its colour and replace it in the appropriate jar. Now imagine that you do this 1000 times or more.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these draws at random, how many times do you think you will draw a green marble from the first jar and an orange one from the second jar?		
2.	If we select 100 of these draws at random, how many times do you think you will draw a green marble from the first jar?		
3.	If we select 100 of these draws at random, how many times do you think you will draw an orange marble from the second jar?		
4.	Imagine that we focus on those draws in which you draw a green marble from the first jar. If we select 100 of these draws at random how many times would you draw a orange marble from the second jar?		

Appendix 6 – Disjunctive tasks (Chapter 6)

Probability version:

James task

James is in secondary school and taking GCSE examinations. He is to sit papers for eight different subjects. His best subjects are biology and chemistry and he has consistently achieved excellent results in these. His worse subjects are mathematics and physics. He did not want to study these but his parents had insisted. While he has put a lot of work into these two subjects and his performance has improved somewhat he is not optimistic concerning his prospects in the exam. How likely is it that James:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Achieves a grade A in mathematics or a grade A in biology.		
2.	Achieves a grade A in mathematics.		
3.	Achieves a grade A in biology.		
4.	Fails mathematics.		
5.	Achieves a grade A in biology given that he achieves a grade A in mathematics.		
6.	Achieves a grade A in mathematics or fails mathematics.		

Smiths task

The Smiths are extremely fond of their pedigree dog and take him for walks on the nearby Common every day weather permitting. Their other interests include card games and particularly bridge, which they play on average once a week usually in the evening. Being retired they walk to the local shops nearly every day to buy a few grocery items, and usually they take their dog with them. They are also keen gardeners and visit the garden centre on average once or twice a week during the spring and summer. On an average day, given the above description please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	The Smiths go to the shops with the dog or play bridge in the evening.		
2.	The Smiths go to the shop with the dog.		
3.	The Smiths play bridge in the evening.		
4.	The Smiths go to the shops without taking their dog.		
5.	The Smiths play bridge in the evening, given that they have taken the dog to the shops.		
6.	The Smiths take their dog to the shops or they go to the shops without taking the dog.		

Bill task

Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In college, he was strong in mathematics but weak in social studies and literature. Given Bill's description please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Bill is an accountant or plays jazz for a hobby.		
2.	Bill is an accountant.		
3.	Bill plays jazz for a hobby.		
4.	Bill is unemployed.		
5.	Bill plays jazz for a hobby given that he is an accountant.		
6.	Bill is an accountant or is unemployed.		

Mr F task

A health survey was conducted in a representative sample of adult males of all ages and occupations. Mr F. was included in the sample. He was selected by chance from the list of participants. Given this description of Mr F., please indicate how likely each of the following statements are:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	Mr. F. is short sighted or has had one or more heart attacks.		
2.	Mr. F. is short sighted.		
3.	Mr. F. has had one or more heart attacks.		
4.	Mr. F. has perfect vision.		
5.	Mr F. has had one or more heart attacks given that he is short sighted.		
6.	Mr. F is short sighted or has perfect vision.		

Marbles task

Suppose that two jars containing different coloured marbles are set on a table in front of you. The first jar contains 25 marbles - five red, and 20 blue. The second jar contains 90 marbles - 80 black, and 10 white. Imagine that without looking, you draw one marble from each jar, write down its colour and replace it in the appropriate jar. How likely is it that:

		How many chances in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	You draw a blue marble from the first jar or a black marble from the second jar.		
2.	You draw a blue marble from the first jar.		
3.	You draw a black marble from the second jar.		
4.	You draw a red marble from the first jar.		
5.	You draw a black marble from the second jar given that you draw a blue marble from the first jar.		
6.	You draw a blue marble or a red marble from the first jar.		

Frequency version

James task

Imagine a boy who is in secondary school and taking GCSE examinations. He is to sit papers for eight different subjects. His best subjects are biology and chemistry and he has consistently achieved excellent results in these. His worse subjects are mathematics and physics. He did not want to study these but his parents had insisted. While he has put a lot of work into these two subjects and his performance has improved somewhat he is not optimistic concerning his prospects in the exam. Now imagine that we assemble a very large number of students fitting this description.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics or a grade A in biology?		
2.	If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics?		
3.	If we select 100 of these students at random, how many do you think will achieves a grade A in biology?		
4.	If we select 100 of these students at random, how many do you think will fail mathematics?		
5.	From this group, if we select 100 students, all of whom achieve a grade A in mathematics, how many of this number will achieve a grade A in biology?		
6.	If we select 100 of these students at random, how many do you think will achieve a grade A in mathematics or fail mathematics?		

Smiths task

The Smiths are extremely fond of their pedigree dog and take him for walks on the nearby common every day, weather permitting. Their other interests include card games and particularly bridge, which they play on average once a week, usually in the evening. Being retired they walk to the local shops nearly every day to buy a few grocery items, and usually they take their dog with them. They are also keen gardeners and visit the garden centre on average once or twice a week during the spring and summer. Imagine that we observe the behaviour of this couple on a large number of days.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 days at random, on how many days do you think the Smiths go to the shops with the dog or play bridge in the evening?		
2.	If we select 100 days at random, on how many days do you think the Smiths go to the shop with the dog?		
3.	If we select 100 days at random, on how many days do you think the Smiths play bridge in the evening?		
4.	If we select 100 days at random, on how many days do you think the Smiths go to the shops without taking the dog?		
5.	If we select 100 days on which the Smiths take their dog to the shops, on how many of these days do they play bridge in the evening?		
6.	If we select 100 days at random, on how many days do you think the Smiths take their dog to the shops or they go to the shops without taking the dog?		

Bill task

Imagine a man who is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In college, he was strong in mathematics but weak in social studies and literature. Now imagine that we assemble a very large number of men fitting this description.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these men at random, how many do you think are employed as accountants or play jazz for a hobby?		
2.	If we select 100 of these men at random, how many do you think are employed as accountants?		
3.	If we select 100 of these men at random, how many do you think play jazz for a hobby?		
4.	If we select 100 of these men at random, how many do you think are unemployed?		
5.	From this group, if we select 100 men who are employed as accountants, of this number how many play jazz for a hobby?		
6.	If we select 100 of these men at random, how many do you think are employed as accountants or are unemployed?		

Mr F task

A health survey was conducted in a representative sample of adult males of all ages and occupations.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these adult males at random, how many do you think are short sighted or have had one or more heart attacks?		
2.	If we select 100 of these adult males at random, how many do you think are short sighted?		
3.	If we select 100 of these adult males at random, how many do you think have had one or more heart attacks?		
4.	If we select 100 of these adult males at random, how many do you think have perfect vision?		
5.	If we select 100 adult males, all of whom are short sighted, how many of this number will have had one or more heart attacks?		
6.	If we select 100 of these adult males at random, how many do you think are short sighted or have perfect vision?		

Marbles task

Suppose that two jars containing different coloured marbles are set on a table in front of you. The first jar contains 25 marbles - five red, and 20 blue. The second jar contains 90 marbles - 80 black, and 10 white. Imagine that without looking, you draw one marble from each jar, write down its colour and replace it in the appropriate jar. Now imagine that you do this 1000 times or more.

		How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
1.	If we select 100 of these draws at random, how many times do you think you will draw a blue marble from the first jar or a black marble from the second jar.		
2.	If we select 100 of these draws at random, how many times do you think you will draw a blue marble from the first jar.		
3.	If we select 100 of these draws at random, how many times do you think you will draw a black marble from the second jar.		
4.	If we select 100 of these draws at random, how many times do you think you will draw a red marble from the first jar.		
5.	Imagine that we focus on those draws in which you draw a blue marble from the first jar. If we select 100 of these draws at random how many times would you draw a black marble from the second jar?		
6.	If we select 100 of these draws at random, how many times do you think you will draw a blue marble or a red marble from the first jar.		

Appendix 7 – Instructions to Participants (Chapter 6)

Probability version

Before you begin, please fill in your PARTICIPANT NUMBER, your AGE in years, and whether you are MALE or FEMALE.

PARTICIPANT	AGE	MALE	FEMALE

In this questionnaire you will be presented with a number of statements. You will be asked to judge how likely each statement is. Please make your judgements by specifying how many chances in 100 there are that the statement is or will be true. An example is provided on the following page. You may choose any number between 0 and 100. If you feel that a statement is impossible or that it is definitely untrue you would enter the number 0 in the box provided. On the other hand if you feel that a statement is definitely true or certain to occur then you would enter the number 100.

Sometimes you may be very confident about your prediction. Next to each estimate there is another box where you can indicate just how confident you are. Enter the number 10 if you are totally confident in your prediction. Enter the number 0 if you have absolutely no confidence in your estimate. If you are neither confident nor unconfident about your estimate then you would enter the number 5. Of course, you may also use all of the other numbers in between.

On each of the pages that follow you will be given some back ground information followed by a series of statements for you to evaluate. In making your judgement make use of the information provided and of your own personal knowledge.

In each case remember to indicate how confident you are in your prediction by entering a number between 0 and 10. If you make a mistake or change your mind cross out the incorrect estimate and replace it with your amended judgement. If you do make any changes please try to ensure that it is clear which number is your actual estimate. **HOWEVER ONCE YOU HAVE TURNED THE PAGE PLEASE DO NOT GO BACK AND MAKE ANY CHANGES TO THE PREVIOUS ONE.** Please read the example on the following page before continuing.

EXAMPLE

Peter is an athlete who is training to run the mile in an important regional competition. In three of his last 4 races, Peter has run the mile in less than 4 minutes. His best time to-date has been 3 minutes 58 seconds. For his next race how likely is it that:

	How many chances in 100. (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
Peter will run the mile in under 4 minutes	81	5
Peter will run the second half mile in under 2 minutes	90	8
Peter will run the mile in 3 minutes 55 seconds or less	39	8
Peter will take over 4 minutes 5 seconds to run the mile	35	4
Peter will run the first half mile in less than 2 minutes 15 seconds OR will lose the race	90	9
Peter will run the first half mile in less than 1 minute 50 seconds OR will win the race	16	3

Now please provide judgements for each group of statements that follow. Complete the questionnaire as quickly as possible, but without working so fast that you make mistakes.

Some of the questions will not have boxes for you to fill in your answer, but will have a line instead. Make sure you read all of the instructions thoroughly so that you understand what each individual problem is asking of you.

Please do not confer with anyone else while completing the questionnaire.

Frequency version

Before you begin, please fill in your PARTICIPANT NUMBER, your AGE in years, and whether you are MALE or FEMALE.

PARTICIPANT	AGE	MALE	FEMALE

In this questionnaire you will be presented with a number of statements. You will be asked to judge how often certain events might occur, or how many people may fit a certain description. Please do this by specifying how many (out of 100) events or people will fit the description in the statement. An example is provided on the following page. You may choose any number between 0 and 100. If you feel that none of the 100 events or people would fit the statement, you would write the number 0 in the box provided. On the other hand if you feel all of them fit then you would enter the number 100.

Sometimes you may be very confident about your prediction. Next to each estimate there is another box where you can indicate just how confident you are. Enter the number 10 if you are totally confident in your prediction. Enter the number 0 if you have absolutely no confidence in your estimate. If you are neither confident nor unconfident about your estimate then you would enter the number 5. Of course, you may also use all of the other numbers in between.

On each of the pages that follow you will be given some back ground information followed by a series of statements for you to evaluate. In making your judgement make use of the information provided and of your own personal knowledge.

In each case remember to indicate how confident you are in your prediction by entering a number between 0 and 10. If you make a mistake or change your mind cross out the incorrect estimate and replace it with your amended judgement. If you do make any changes please try to ensure that it is clear which number is your actual estimate. **HOWEVER ONCE YOU HAVE TURNED THE PAGE PLEASE DO NOT GO BACK AND MAKE ANY CHANGES TO THE PREVIOUS ONE.**

Please read the example on the following page before continuing.

EXAMPLE

Imagine an athlete who is training to run the mile in an important regional competition. In three of his last 4 races, this athlete has run the mile in less than 4 minutes. His best time to-date has been 3 minutes 58 seconds. Now imagine that a very large number of men fitting this description will be running a mile race this weekend.

	How many in 100 (Enter a number between 0 and 100)	Confidence in your prediction. (Enter a number between 0 and 10)
If we select 100 of these men at random, how many do you think will run the mile in under 4 minutes?	81	5
If we select 100 of these men at random, how many do you think will run the second half mile in under 2 minutes?	90	8
If we select 100 of these men at random, how many do you think will run the mile in 3 minutes 55 seconds or less	39	8
If we select 100 of these men at random, how many do you think will take over 4 minutes 5 seconds to run the mile	35	4
If we select 100 of these men at random, how many do you think will run the first half mile in less than 2 minutes 15 seconds or will lose the race	90	9
If we select 100 of these men at random, how many do you think will run the first half mile in less than 1 minute 50 seconds or will win the race	16	3

Now please provide judgements for each group of statements that follow. Complete the questionnaire as quickly as possible, but without working so fast that you make mistakes.

Some of the questions will not have boxes for you to fill in your answer, but will have a line instead. Make sure you read all of the instructions thoroughly so that you understand what each individual problem is asking of you.

Please do not confer with anyone else while completing the questionnaire.

Appendix 8 – TDQ and REI as presented to participants
Thinking Dispositions Questionnaire

For each of the statements below, please indicate to what extent you agree with that statement. If you strongly agree, please circle the 1 after the question. If you agree, but not so strongly, please circle the 2. If you disagree a little please circle the 3, and if you strongly disagree please circle the 4.

Please keep this scale in mind as you rate each of the statements below:

1 = strongly agree

2 = agree

3 = disagree

4 = strongly disagree

1.	If everybody in a group has too many different ideas, the group will break up.	1	2	3	4
2.	A good person usually does what they are told to do.	1	2	3	4
3.	It's really good when kids believe in the same things as their parents.	1	2	3	4
4.	It's ok to hang out with people who don't share my values.	1	2	3	4
5.	I like to do things that I've learned well over and over, so that I don't have to think about it anymore.	1	2	3	4
6.	Changing your beliefs shows that you are a strong person.	1	2	3	4
7.	It's OK to be undecided about some things.	1	2	3	4
8.	It is better to simply believe in a religion than to be confused by doubts about it.	1	2	3	4
9.	I sometimes tell lies if I have to.	1	2	3	4
10.	Nobody can change my mind if I know I am right.	1	2	3	4
11.	It's better to learn about useful things instead of ideas.	1	2	3	4
12.	The things you believe in come from inside you rather than from what happened to you.	1	2	3	4
13.	There are times I have taken advantage of someone.	1	2	3	4
14.	People make bad choices when they listen to lots of different opinions.	1	2	3	4
15.	It's fantastic when someone famous believes in the same things as me.	1	2	3	4
16.	The number 13 is unlucky.	1	2	3	4
17.	Changing your mind is a sign of weakness.	1	2	3	4
18.	I like to be in charge of a problem that needs lots of thinking.	1	2	3	4
19.	It really makes me angry when someone can't say they are wrong.	1	2	3	4
20.	It is bad luck to have a black cat cross your path.	1	2	3	4
21.	I like to do jobs where I don't have to think at all.	1	2	3	4
22.	I like everyone I meet.	1	2	3	4
23.	The way to fix a problem is to think about the best answer – not stand around and wait for the problem to fix itself.	1	2	3	4
24.	I really hate some people because of the things they stand for.	1	2	3	4

25.	I do not believe in superstitions.	1	2	3	4
26.	Feelings are the best guide in making decisions.	1	2	3	4
27.	Horoscopes can be useful in making personality judgments.	1	2	3	4
28.	Most people just don't know what's good for them.	1	2	3	4
29.	I like to spend a lot of time and energy thinking over something.	1	2	3	4
30.	I always obey rules even if I probably won't get caught.	1	2	3	4
31.	It's really cool to figure out a new way to do something.	1	2	3	4
32.	I don't believe in luck.	1	2	3	4
33.	Right and wrong never change.	1	2	3	4
34.	Often, people who criticize me don't know what they are talking about.	1	2	3	4
35.	There is one right way and lots of wrong ways to do most things.	1	2	3	4
36.	It's a good idea to look at your horoscope every day.	1	2	3	4
37.	I have things that bring me luck.	1	2	3	4
38.	Wise people make fast decisions.	1	2	3	4
39.	There are basically two kind of people in this world, good and bad.	1	2	3	4
40.	I never change what I believe in – even when someone shows me that my beliefs are wrong.	1	2	3	4
41.	It's important to change what you believe after you learn new information.	1	2	3	4
42.	I think people are either with me or against me.	1	2	3	4
43.	Mostly, I know everything I need to know.	1	2	3	4
44.	A person should always consider new possibilities.	1	2	3	4
45.	I try to avoid problems that I have to think about a lot.	1	2	3	4
46.	People should always consider evidence that goes against their beliefs.	1	2	3	4
47.	I like to do jobs that make me think hard.	1	2	3	4
48.	I'm not interested in learning new ways to think.	1	2	3	4
49.	I like hard problems instead of easy ones.	1	2	3	4
50.	If I think longer about a problem I will be more likely to solve it.	1	2	3	4
51.	Opening an umbrella inside can bring you bad luck.	1	2	3	4
52.	I have said something bad about a friend behind his or her back.	1	2	3	4

Participant Number:
Rational-Experiential Inventory Scales

For each of the statements below, please indicate to what extent the statement is characteristic of you. If the statement is extremely uncharacteristic of you (not at all like you) please circle the '1' beside the question; if the statement is extremely characteristic of you (very much like you) please circle the '5' next to the question. Of course, a statement may be neither extremely uncharacteristic nor extremely characteristic of you; if so, please use the number in the middle of the scale that describes the best fit. Please keep the following scale in mind as you rate each of the statements below; 1= extremely uncharacteristic; 2 = somewhat uncharacteristic; 3 = uncertain; 4 = somewhat characteristic; 5 = extremely characteristic.

1.	I can typically sense right away when a person is lying.	1	2	3	4	5
2.	I believe I can judge character pretty well from a person's appearance.	1	2	3	4	5
3.	I don't reason well under pressure.	1	2	3	4	5
4.	I often have clear visual images of things.	1	2	3	4	5
5.	I generally prefer to accept things as they are rather than to question them.	1	2	3	4	5
6.	I don't like to have the responsibility of handling a situation that requires a lot of thinking.	1	2	3	4	5
7.	I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	1	2	3	4	5
8.	I would prefer complex to simple problems.	1	2	3	4	5
9.	I am quick to form impressions about people.	1	2	3	4	5
10.	My initial impressions of people are almost always right.	1	2	3	4	5
11.	I believe in trusting my hunches.	1	2	3	4	5
12.	I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something.	1	2	3	4	5
13.	I feel relief rather than satisfaction after completing a task that required a lot of mental effort.	1	2	3	4	5
14.	I tend to set goals that can be accomplished only by expending considerable mental effort.	1	2	3	4	5
15.	The notion of thinking abstractly is not appealing to me.	1	2	3	4	5
16.	I have a very good sense of rhythm.	1	2	3	4	5
17.	The idea of relying on thought to make my way to the top does not appeal to me.	1	2	3	4	5
18.	I can usually feel when a person is right or wrong even if I can't explain how I know.	1	2	3	4	5
19.	I find little satisfaction in deliberating hard and for long hours	1	2	3	4	5

20.	Simply knowing the answer rather than understanding the reasons for the answer to a problem is fine with me.	1	2	3	4	5
21.	It is enough for me that something gets the job done, I don't care how or why it works.	1	2	3	4	5
22.	I am a very intuitive person.	1	2	3	4	5
23.	I prefer my life to be filled with puzzles that I must solve.	1	2	3	4	5
24.	I have difficulty thinking in new and unfamiliar situations.	1	2	3	4	5
25.	I trust my initial feelings about people.	1	2	3	4	5
26.	When it comes to trusting people, I can usually rely on my "gut feelings".	1	2	3	4	5
27.	Thinking is not my idea of fun.	1	2	3	4	5
28.	I am good at visualising things.	1	2	3	4	5
29.	I prefer to talk about international problems rather than to gossip or talk about celebrities.	1	2	3	4	5
30.	Learning new ways to think doesn't excite me very much.	1	2	3	4	5
31.	I would rather do something that requires little thought than something that is sure to challenge my thinking abilities	1	2	3	4	5

Appendix 9 – Cronbach's alpha values for all thinking style measures

		Chronbach's alpha	Number of items
REI	FI	.76	12
	NFC	.80	19
TDQ	FT	.58	10
	A	.33	5
	D	.46	6
	CT	.46	3
	ST/LC	.86	8
	sfNFC	.72	9
	SD	.56	5
	BI	.28	6

Appendix 10 – Descriptive statistics for all data from Chapter 6 tasks

Short titles for each problem are presented here for brevity, for full tasks as presented see Appendices 5 and 6.

Conjunctive tasks:

Problem	Mean (SD) responses to each element						
	Conjunction	1 st component	2 nd component	Conditional statement	Normative	Error	Absolute error
Bob (takes the bus and buys a sandwich)	39.85 (24.29)	27.07 (19.06)	61.29 (23.28)	44.75 (23.28)	12.54 (12.47)	27.29 (21.23)	28.28 (19.87)
Helen (nightclubs and library)	31.31 (27.07)	86.79 (14.33)	16.14 (15.58)	15.50 (19.30)	13.19 (16.59)	18.00 (22.69)	19.06 (21.78)
Morris (stamps and team sports)	17.89 (18.70)	52.49 (31.59)	9.42 (12.40)	9.93 (12.94)	6.41 (9.65)	11.39 (15.74)	12.34 (15.01)
Jim (has maths and game)	51.74 (24.06)	85.51 (12.72)	38.21 (21.17)	41.06 (26.20)	35.78 (25.24)	15.95 (26.53)	22.91 (20.73)
Marbles (green first and orange second)	34.17 (21.54)	14.90 (11.99)	68.19 (25.29)	45.03 (28.78)	6.56 (5.83)	27.64 (19.94)	27.64 (19.93)

Values for each of the disjunctive tasks are presented overleaf.

Disjunctive tasks:

Problem	Mean (SD) responses to each element			
	1 st component	2 nd component	3 rd component	Conditional
James (grade A maths or biology / fails maths)	32.45 (23.47)	76.37 (21.56)	30.51 (21.95)	54.28 (33.90)
Smiths (dog or bridge / no dog)	83.78 (15.67)	40.36 (27.54)	20.19 (23.12)	39.67 (27.46)
Bill (accountant or jazz / unemployed)	62.78 (27.35)	18.60 (19.73)	22.71 (20.30)	14.17 (14.32)
Mr F. (short sighted or heart attack / perfect sight)	34.60 (18.16)	21.58 (18.68)	39.24 (22.79)	15.71 (15.48)
Marbles (blue first or black second / red first)	72.14 (20.28)	79.15 (16.32)	18.89 (10.76)	60.44 (27.82)

Problem	Mean (SD) responses to each element			
	Inclusive disjunction	Inclusive disjunction normative	Inclusive disjunction error	Inclusive disjunction absolute error
James (grade A maths or biology)	65.26 (22.10)	90.73 (26.37)	-24.17 (23.71)	27.36 (19.90)
Smiths (dog or bridge)	72.32 (25.42)	90.38 (27.41)	-15.32 (30.73)	25.96 (22.33)
Bill (accountant or jazz)	45.35 (24.18)	71.40 (31.58)	-24.95 (25.85)	26.75 (23.95)
Mr F. (short sighted or heart attack)	34.96 (20.55)	49.92 (27.25)	-14.67 (19.28)	17.61 (16.60)
Marbles (blue first or black second)	74.23 (20.03)	106.48 (27.94)	-26.56 (19.19)	28.03 (16.93)

Problem	Mean (SD) responses to each element			
	Exclusive disjunction	Exclusive disjunction normative	Exclusive disjunction error	Exclusive disjunction absolute error
James (grade A maths or fails maths)	43.00 (24.64)	62.96 (29.24)	-20.12 (34.16)	31.81 (23.49)
Smiths (dog or no dog)	70.11 (30.99)	103.97 (26.46)	-28.60 (33.39)	33.19 (28.76)
Bill (accountant or unemployed)	56.03 (26.86)	85.49 (36.60)	-26.30 (26.86)	29.44 (23.32)
Mr F. (short sighted or perfect sight)	62.57 (27.01)	73.83 (27.19)	-10.70 (29.25)	23.72 (20.02)
Marbles (blue first or red first)	81.66 (20.07)	91.03 (25.04)	-8.73 (24.61)	14.99 (21.33)

Appendix 11 – Full beta weights for regression analysis (Chapter 6)

Conjunction Fallacy Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		2.68**		1.50		1.28
Flexible Thinking			.11	0.75	.10	0.67
Absolutism			-.22	-1.71	-.24	-1.77
Dogmatism			-.05	-.36	-.05	-0.36
Categorical Thinking			.16	1.113	.16	1.14
Superstitious Thinking/Luck Composite			-.024	-.19	-.03	-0.21
short form Need For Cognition			-.02	-.16	-.01	-0.05
Social Desirability			-.16	-1.10	-.17	-1.15
Belief Identification			.09	-.62	.09	0.65
Mill Hill Vocabulary Scale	-.01	-.11			.07	.49
R squared increment		.00		.10		.10
Entered at Step 2:	Beta	t test	Beta	t test	Beta	t test
Constant		3.15**		1.92		1.71
Flexible Thinking			.09	0.67	.08	0.60
Absolutism			-.23	-1.82	-.24	-1.86
Dogmatism			-.07	-0.49	-.07	-0.49
Categorical Thinking			.14	1.07	.15	1.09
Superstitious Thinking/Luck Composite			-.12	-0.95	-.12	-0.96
short form Need For Cognition			-.12	-0.88	-.11	-0.75
Social Desirability			-.10	-0.68	-.10	-0.72
Belief Identification			.14	1.03	.14	1.04
Mill Hill Vocabulary Scale	-.01	-0.09			.05	0.42
Task Format	-.32	-2.69**	-.35	-2.70**	-.35	-2.67**
R squared increment		.10**		.10**		.10
R squared total and		.10		.20		.20
significance of the regression		F(2, 65)=3.63*		F(9,58)=1.58		F(10,57)=1.42

Exclusive Disjunction Fallacy Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		2.21*		1.94		2.03*
Flexible Thinking			.01	0.11	.03	0.19
Absolutism			-.15	-1.20	-.13	-1.03
Dogmatism			.08	0.56	.08	0.56
Categorical Thinking			-.29	-2.19*	-.30	-2.20*
Superstitious Thinking/Luck Composite			-.00	-0.01	.00	0.03
short form Need For Cognition			.14	1.07	.12	0.91
Social Desirability			.08	0.58	.09	0.64
Belief Identification			-.21	-1.57	-.22	-1.61
Mill Hill Vocabulary Scale	-.10	-0.78			-.08	-0.61
R squared increment		.01		.19		.19
Entered at Step 2:						
	Beta	t test	Beta	t test	Beta	t test
Constant		2.37*		1.98		2.07*
Flexible Thinking			.01	.08	.02	0.17
Absolutism			-.15	-1.20	-.13	-1.02
Dogmatism			.07	0.53	.07	0.52
Categorical Thinking			-.29	-2.18*	-.30	-2.20*
Superstitious Thinking/Luck Composite			-.02	-0.14	-.01	-0.11
short form Need For Cognition			.12	0.90	.10	0.73
Social Desirability			.09	0.65	.10	0.71
Belief Identification			-.20	-1.47	-.21	-1.51
Mill Hill Vocabulary Scale	-.10	-0.80			-.08	-0.63
Task Format	-.14	-1.14	-.06	-0.46	-.07	-0.50
R squared increment		.02		.00		.00
R squared total and		.03		.19		.20
significance of the regression		F(2,65)=.95		F(9,58)=1.51		F(10,57)=1.39

Inclusive Disjunction Fallacy Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		2.36*		-0.32		-0.35
Flexible Thinking			.17	1.27	.17	1.23
Absolutism			.14	1.13	.13	1.06
Dogmatism			-.06	-0.48	-.06	-0.48
Categorical Thinking			.03	0.21	.03	0.21
Superstitious Thinking/Luck Composite			.09	0.75	.09	0.73
short form Need For Cognition			.32	2.45*	.32	2.40*
Social Desirability			-.28	-2.03*	-.28	-2.02*
Belief Identification			.01	0.04	.01	0.05
Mill Hill Vocabulary Scale	-.02	-0.15			.02	0.15
R squared increment	.00		.21		.21	
Entered at Step 2:						
	Beta	t test	Beta	t test	Beta	t test
Constant		3.16**		0.05		.05
Flexible Thinking			.16	1.19	.16	1.17
Absolutism			.14	1.17	.14	1.12
Dogmatism			-.08	-0.65	-.08	-0.65
Categorical Thinking			.02	0.14	.02	0.13
Superstitious Thinking/Luck Composite			.00	0.01	.00	0.01
short form Need For Cognition			.23	1.76	.23	1.71
Social Desirability			-.22	-1.68	-.22	-1.66
Belief Identification			.06	0.44	.06	0.44
Mill Hill Vocabulary Scale	-.03	-0.23			.00	0.02
Task Format	-.43	-3.79***	-.33	-2.69**	-.33	-2.67**
R squared increment	.18***		.09**		.09**	
R squared total and significance of the regression	.18 F(2,65)=7.21**		.30 F(9,58)=2.71*		.30 F(10,57)=2.40*	

Conjunction Error Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		4.75***		.81		1.09
Flexible Thinking			-.01	-.07	.01	.08
Absolutism			-.31	-2.49*	-.28	-2.19*
Dogmatism			.09	.64	.09	.65
Categorical Thinking			.14	1.03	.13	.96
Superstitious Thinking/Luck Composite			.14	1.12	.15	1.17
short form Need For Cognition			.11	.79	.07	.52
Social Desirability			-.04	-.29	-.02	-.16
Belief Identification			.13	.97	.12	.89
Mill Hill Vocabulary Scale	-.24	-1.97			-.14	-1.10
R squared increment		.06		.16		.18
Entered at Step 2:						
	Beta	t test	Beta	t test	Beta	t test
Constant		6.17***		1.62		2.00
Flexible Thinking			-.03	-.29	-.01	-.09
Absolutism			-.32	-2.98**	-.28	-2.61*
Dogmatism			.06	.55	.07	.57
Categorical Thinking			.12	1.03	.11	.94
Superstitious Thinking/Luck Composite			-.01	-.13	-.01	-.07
short form Need For Cognition			-.05	-.41	-.09	-.74
Social Desirability			.06	.52	.09	.70
Belief Identification			.22	1.84	.20	1.76
Mill Hill Vocabulary Scale	-.23	-2.26*			-.16	-1.49
Task Format	-.51	-4.98***	-.55	-4.97***	-.56	-5.08***
R squared increment		.26***		.25***		.26***
R squared total and		.32		.41		.43
significance of the regression		F(2,65)=15.02***		F(9, 58)=4.52***		F(10, 57)=4.38***

Exclusive Disjunction Error Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		3.40**		1.67		1.79
Flexible Thinking			.05	.38	.06	.47
Absolutism			-.24	-2.02*	-.22	-1.81
Dogmatism			.17	1.30	.17	1.29
Categorical Thinking			-.19	-1.45	-.19	-1.48
Superstitious Thinking/Luck Composite			.15	1.29	.16	1.33
short form Need For Cognition			.183	1.43	.16	1.23
Social Desirability			.00	.01	.01	.08
Belief Identification			-.16	-1.22	-.17	-1.27
Mill Hill Vocabulary Scale	-.14	-1.10			-.09	-.69
R squared increment		.02		.23*		.24
Entered at Step 2:						
	Beta	t test	Beta	t test	Beta	t test
Constant		3.80**		1.80		1.93
Flexible Thinking			.04	.32	.06	.42
Absolutism			-.24	-2.03*	-.22	-1.80
Dogmatism			.16	1.24	.16	1.23
Categorical Thinking			-.19	-1.48	-.20	-1.51
Superstitious Thinking/Luck Composite			.12	.96	.12	.99
short form Need For Cognition			.15	1.11	.12	.91
Social Desirability			.02	.17	.04	.25
Belief Identification			-.14	-1.06	-.15	.27
Mill Hill Vocabulary Scale	-.14	-1.17			-.09	-.74
Task Format	-.26	-2.16*	-.13	-1.01	-.13	-1.04
R squared increment		.07*		.01		.01
R squared total		.08		.25		.25
and significance of the regression		F(2, 65)=2.97		F(9, 58)=2.11*		F(10, 57)=1.94

Inclusive Disjunction Error Regression Analyses * p<.05; ** p<.01; *** p<.001

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3	
	Beta	t test	Beta	t test	Beta	t test
Constant		3.56*		-.08		.03
Flexible Thinking			.26	1.87	.27	1.89
Absolutism			.07	.55	.08	.61
Dogmatism			-.08	-.59	-.08	-.58
Categorical Thinking			-.04	-.30	-.04	-.32
Superstitious Thinking/Luck Composite			.12	1.00	.13	1.02
short form Need For Cognition			.18	1.39	.17	1.26
Social Desirability			-.15	-1.07	-.14	-1.01
Belief Identification			-.01	-.09	-.02	-.12
Mill Hill Vocabulary Scale	-.04	-.33			-.05	-.38
R squared increment	.00		.18		.18	
Entered at Step 2:						
	Beta	t test	Beta	t test	Beta	t test
Constant		3.64**		.44		.60
Flexible Thinking			.24	1.87	.25	1.92
Absolutism			.07	.57	.08	.69
Dogmatism			-.10	-.84	-.11	-.84
Categorical Thinking			-.05	-.44	-.06	-.47
Superstitious Thinking/Luck Composite			.01	.05	.01	.08
short form Need For Cognition			.06	.50	.05	.34
Social Desirability			-.08	-.58	-.07	-.51
Belief Identification			.06	.44	.05	.39
Mill Hill Vocabulary Scale	-.05	-.45			-.07	-.61
Task Format	-.50	-4.62***	-.44	-3.64**	-.44	-3.65**
R squared increment	.25***		.15***		.16***	
R squared total and significance of the regression	.25 F(2, 65)=10.73***		.33 F(9, 58)=3.18**		.33 F(10, 57)=.286**	

Appendix 12 – Information Processing Speed Measure⁴

Participant number: 1st or 2nd time

Please do not open this booklet until asked to do so.

This task requires you to look at two lines of random letters, one line on top of the other, and compare them. You must decide if two lines of letters are the same, or different. If they are both exactly the same, you would circle the letter S. If you think that the lines are at all different, you would circle the D. You should try and do this as quickly and as accurately as you can.

Here is an example of what you'll be asked to complete – try completing it now. Circle S if the two lines of letters are the same, and D if they are different.

RCM RCG	D	S
JTC JTC	D	S
GFW GNW	D	S
JMF JMF	D	S
NVC NVC	D	S

You should have circled D for the first one, S for the second, D for the third, and then S for the remaining two.

Remember to work as quickly and as accurately as you can – you will have 30 seconds to answer as many as possible.

Turn over the page *when you are asked to* – NB this is a double sided booklet, so page 2, where the task begins, is on the other side of this page.

⁴For brevity, one page only of each level of complexity is shown here, the participants were actually presented with 4 pages of the sets of 3 letters, and 3 pages each of the 6 and 9 letter sets.

NYS NFS	D	S
DCK DCK	D	S
JZN JZN	D	S
QTV QTV	D	S
QJX QNX	D	S
PYW PYR	D	S
PKJ PKF	D	S
KMZ KMZ	D	S
FYT FYT	D	S
TYG TYG	D	S
BYL BXL	D	S
DGM DGM	D	S
WLJ BLJ	D	S
NSC NSC	D	S
GFH XFH	D	S
QKR QKT	D	S

Please now wait until you are told that to start working on the next section of this task.

In a moment you will be asked to turn the page and begin comparing more lines of letters – this time the sets of letters will all be longer, but you will still only have 30 seconds to answer as many as possible, as accurately as possible.

SKVXWQ SKVXWQ	D	S
YJLHMB YJLHXB	D	S
NPCMKW NPCMYW	D	S
THPGFC THPGFQ	D	S
KFJVRX KFJVRX	D	S
ZJGCTB ZJGYTB	D	S
CWJLMP CWJLMP	D	S
HXMJGQ HXMJFQ	D	S
ZQKXSC ZQKXSC	D	S
QJVFLM QJVFLM	D	S
GYFHBD GYFHJD	D	S
KDHXZP KDHXZP	D	S
ZJMFQT ZJMFQT	D	S
CDRNMT CDRNMT	D	S
DPBHKT QTKFWB	D	S
WMTZBF WMTZBF	D	S

Please now wait until you are told that to start working on the last section of this task.

In a moment you will be asked to turn the page and begin comparing more lines of letters – this time the sets of letters will be even longer than the sets you have just seen, but you will still only have 30 seconds to answer as many as possible. as accurately as possible

FGQMDTWRN FGJMDTWRN	D	S
FYMNCRTBV FYMNCRTBV	D	S
DHGVZSWMY DHGVZSWMY	D	S
HSXTFWPZR HSXTFWPZN	D	S
GNUPLBJWQ GNUPLBJWD	D	S
WBVTFGPNK WBVTFGPNK	D	S
LDHKZRJXG LDHKZRJXG	D	S
QNVYGZMBS QNVYGZMBS	D	S
JGCHLFDPM JGCHLFDPM	D	S
WPNCJTKGF WPNCJTKGF	D	S
BLMNQTDCW SLMNQTDCW	D	S
WPJYTBCKH WPJYTBCKD	D	S
ZKMQNVGBY ZKMQNVGBL	D	S
BMTCHYLVK BMTCHYLVK	D	S
XWJZDGQYL XWJZDGQYL	D	S
TPJBYZFSK TPJBYZFSK	D	S

Appendix 13 – Beta weights for regression (Chapter 7)

Conjunction Fallacy

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		21.24***		5.16***		3.12**		19.78**		3.16**
Task Format	-.03	-.37							-.03	-.33
Information Processing Speed			-.12	-1.13					-.08	-.66
Mill Hill Vocabulary Scale			-.16	-1.56					-.11	-.71
Flexible Thinking					-.02	-.14			-.02	-.14
Absolutism					-.15	-1.41			-.13	-1.20
Dogmatism					-.10	-.95			-.11	-1.02
Categorical Thinking					-.13	-1.26			-.12	-1.08
Superstitious Thinking/Luck Composite					-.08	-.77			-.10	-.80
short form Need For Cognition					.05	.48			.03	.32
Social Desirability					.04	.37			.03	.24
Belief Identification					-.05	-.52			-.04	-.43
Age Group							-.04	-.40	.04	.22
R squared increment		.00		.02		.10		.00		
R squared total								.11		
and significance of the regression								F(12, 115) = 1.15, p>.05		

* p<.05; ** p<.01; *** p<.001

Exclusive Disjunction Fallacy

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		13.59***		4.09***		1.16		11.50***		1.85
Task Format	-.11	-1.28							-.12	-1.34
Information Processing Speed			-.22	-2.16*					-.27	-2.05*
Mill Hill Vocabulary Scale			-.09	-.93					-.04	-.26
Flexible Thinking					.01	.12			-.00	-.01
Absolutism					.02	.15			.06	.54
Dogmatism					-.01	-.14			-.03	-.26
Categorical Thinking					-.07	-.58			-.03	-.27
Superstitious Thinking/Luck Composite					-.04	-.42			-.13	-1.08
short form Need For Cognition					.10	.91			.06	.60
Social Desirability					-.09	-.88			-.09	-.85
Belief Identification					.02	.15			.028	.28
Age Group							.05	.61	-.03	-.17
R squared increment		.01		.04		.03		.00		
R squared total								.08		
and significance of the regression								F(12, 114) = 0.88, p>.05		

* p<.05; ** p<.01; *** p<.001

Inclusive Disjunction Fallacy

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		17.54***		5.12***		2.86**		16.258*		3.78***
Task Format	-.17	-2.00							-.14	-1.56
Information Processing Speed			-.20	-1.90					-.322	-2.72**
Mill Hill Vocabulary Scale			-.22	-2.15*					-.05	-.34
Flexible Thinking					.03	.28			.01	.09
Absolutism					.08	.83			.08	.71
Dogmatism					-.09	-.89			-.08	-.81
Categorical Thinking					-.10	-.96			-.06	-.54
Superstitious Thinking/Luck Composite					-.18	-1.86			-.22	-2.05*
short form Need For Cognition					.14	1.41			.10	1.01
Social Desirability					-.20	-2.09*			-.24	-2.50*
Belief Identification					-.08	-.86			-.04	-.47
Age Group							-.11	-1.27	-.21	-1.17
R squared increment		.03		.04		.17		.01		
R squared total								.24		
and significance of the regression								F(12, 112) = 2.90, p<.01		

* p<.05; ** p<.01; *** p<.001

Conjunction Error

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		16.60***		5.78***		2.07		14.28		2.77**
Task Format	-.12	-1.39							-.14	-1.54
Information Processing Speed			-.24	-2.39*					-.11	-.94
Mill Hill Vocabulary Scale			-.28	-2.84**					.39	-2.66**
Flexible Thinking					.04	.32			.04	.36
Absolutism					-.14	-1.34			-.10	-.90
Dogmatism					-.03	-.29			-.05	-.49
Categorical Thinking					-.17	-1.56			-.14	-1.31
Superstitious Thinking/Luck Composite					.03	.27			-.00	-.01
short form Need For Cognition					.07	.64			.02	.16
Social Desirability					.08	.81			.06	.59
Belief Identification					-.10	-1.00			-.07	-.68
Age Group							.00	.04	.26	1.42
R squared increment		.02		.07		.11		.00		
R squared total								.18		
and significance of the regression								F(12, 113) = 2.02, p<.05		

* p<.05; ** p<.01; *** p<.001

Exclusive Disjunction Error

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		18.54***		4.26***		1.99*		14.13** *		2.51*
Task Format	-.28	-3.26**							-.30	-3.31**
Information Processing Speed			-.19	-1.83					-.19	-1.54
Mill Hill Vocabulary Scale			-.06	-.60					-.15	-1.04
Flexible Thinking					.03	.23			.02	.19
Absolutism					.02	.15			.11	.97
Dogmatism					-.09	-.88			-.10	-1.00
Categorical Thinking					-.02	-.21			-.01	-.07
Superstitious Thinking/Luck Composite					.01	.14			-.09	-.80
short form Need For Cognition					.03	.27			-.03	-.27
Social Desirability					-.11	-1.02			-.07	-.71
Belief Identification					-.07	-.70			-.05	-.47
Age Group							.09	1.05	.17	.93
R squared increment		.08		.03		.03		.01		
R squared total								.15		
and significance of the regression								F (12, 114) = 1.63, p>.05		

* p<.05; ** p<.01; *** p<.001

Inclusive Disjunction Error

Entered at Step 1:	Sequence 1		Sequence 2		Sequence 3		Sequence 4		Full Model	
	Beta	t test	Beta	t test	Beta	t test	Beta	t test	Beta	t test
Constant		22.32***		2.90**		2.60*		18.10** *		2.56
Task Format	-.40	-4.78							-.34	-3.85***
Information Processing Speed			.05	.50					-.08	-.68
Mill Hill Vocabulary Scale			-.03	-.32					.12	.85
Flexible Thinking					-.06	-.49			-.09	-.78
Absolutism					.03	.31			-.00	-.04
Dogmatism					-.09	-.88			-.04	-.46
Categorical Thinking					.03	.25			.03	.27
Superstitious Thinking/Luck Composite					-.10	-.95			-.09	-.86
short form Need For Cognition					.18	1.75			.16	1.66
Social Desirability					-.14	-1.39			-.14	-1.41
Belief Identification					-.15	-1.53			-.09	-.98
Age Group							-.14	-1.61	-.25	-1.43
R squared increment		.16		.01		.10		.02		
R squared total and significance of the regression								.23		
								F(12, 112) = 2.78, p<.01		

* p<.05; ** p<.01; *** p<.001

Appendix 14 – Linear and quadratic curve estimations of thinking styles and Bayesian task responses

Cab Task

Thinking style	Linear Estimation (df = 1, 87)		Quadratic estimation (df = 2, 86)	
	R ²	F	R ²	F
FI – Faith in Intuition	.00	.37	.02	.93
NFC – Need for Cognition	.00	.10	.00	.09
FT – Flexible thinking	.04	3.91	.05	2.17
A - Absolutism	.00	.25	.01	.45
D - Dogmatism	.03	2.34	.03	1.37
CT – Categorical thinking	.03	2.53	.04	1.64
ST/LC – Superstitious Thinking/Luck Composite	.03	2.81	.03	1.41
sfNFC – Short form need for cognition	.02	1.97	.03	1.13
SD – social desirability	.00	.22	.02	.75
BI – belief identification	.02	1.81	.03	1.14

* p < .05

Disease Task

Thinking style	Linear Estimation (df = 1, 83)		Quadratic estimation (df = 2, 82)	
	R ²	F	R ²	F
FI – Faith in Intuition	.00	.13	.07	3.08
NFC – Need for Cognition	.00	.00	.01	.48
FT – Flexible thinking	.00	.03	.00	.15
A - Absolutism	.01	.52	.02	.83
D - Dogmatism	.00	.03	.02	.96
CT – Categorical thinking	.00	.01	.00	.12
ST/LC – Superstitious Thinking/Luck Composite	.09	8.12**	.09	4.03*
sfNFC – Short form need for cognition	.00	.02	.00	.04
SD – social desirability	.00	.11	.00	.06
BI – belief identification	.01	.44	.01	.23

* p < .05, ** p < .01

Appendix 15 – Correlations of AOT subscales (Pearson's R)

	Belief Identification	Absolutism	Dogmatism	Categorical Thinking
Flexible Thinking	-.26**	-.23**	.01	-.31**
Belief Identification		.23**	.14*	.15*
Absolutism			.021	.32**
Dogmatism				.27**

* p<.05; ** p<.01; *** p<.001