Judgement Analysis of Patient Management: General Practitioners' Policies and Self-Insight.

by
CLARE HARRIES

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

Department of Psychology
Faculty of Human Sciences

September 1995

Judgement Analysis of Patient Management: General Practitioners' Policies and Self-Insight.

Clare Harries

In this thesis judgement analysis (multiple linear regression techniques) was used to look at both GPs' decisions to prescribe certain types of drug for patients and their judgements of patients' risk of coronary heart disease. All of these were idiographic analyses in that decision making by each GP was modelled separately. Judgement analysis (paramorphically) describes a subject's judgement or decision making policy in terms of the relative influence of different pieces of information.

The amount of information subjects could take into account was limited. For all types of judgement or decision doctors were influenced on average by only four of the thirteen or twelve cues available.

The decision to prescribe one of the types of drug was modelled not only in terms of the individual effects of cues (judgement analysis) but also in terms of the influence of the doctor's assessment of the patient's risk. Doctors agreed more about judgements of risk and the factors influencing this than about prescription. Doctors only prescribed to patients they rated as at high risk but factors such as for example smoking behaviour led some doctors not to prescribe to individuals in this group.

Judgement and decision making policies (explicit policies) were also elicited verbally from doctors. These showed greater agreement than the policies captured using judgement analysis (tacit policies) did. When these explicit policies were compared to tacit policies a moderate amount of correspondence was found. However, doctors tended to over-rate the importance of certain cues. A number of explanations for this pattern of self-insight were investigated including the possibilities that doctors have self-insight but are unable to state it and that the pattern was an artefact of linear modelling. Both of these hypotheses were rejected. Subjects' explicit policies were found to resemble the pattern of selection of information more than the pattern of its use. Both the hypotheses that subjects' explicit policies were based on phenomenal knowledge and that they are based on some ideal model (influencing which cues are selected) were supported.

List of contents

Glossary of Terms vii

List of Tables ix

List of Illustrations x

List of Appendices xii

Acknowledgement xiv

Author's Declaration xv

Chapter 1 Introduction 1

Decision making by General Practitioners 2
Analysis of decision making in General Practice 4
Issues being addressed 8
Summary overview of the chapters 10

Chapter 2 Analysis of Judgement and Decision making 12

Introduction 12
Decision making, judgement making and problem solving 13
Distinctions between approaches 16
The decision theory approach 18
Problem solving analysis - Process Tracing techniques 24
Social Judgement Theory (SJT) 26
Errors in Judgement and Decision Making 28
Overview 32

Chapter 3 Judgement Analysis and its Application in Medicine 36

Introduction 36
Study design and real versus 'paper' cases 39
Linear modelling and consistency 42
Number of cues used 46
Individual Differences 48
Subjective models and self-insight 49
The Lens model and Social Judgement Theory 50
Practical applications 54
Conclusions 57

Chapter 4 Self-Knowledge 58

Introduction 58
The "unknown" known 58
Self-knowledge: introspection vs. Self-hypothesising 60
Types of self-knowledge: Separate phenomena vs. Levels 61
How to test self-knowledge 62
Knowledge of causes and processes 64
Self-insight in Judgement Analysis 66
Other evidence for self-insight in JA 73
General judgements of contingent or correlational relationships 74
Conclusions 75

Chapter 5 Study 1: Capturing GPs' Prescribing Policies 79

Introduction 79
Method 83
Results and Discussion 94
Conclusions 115

Chapter 6 Study 2: Information Selection Study 119

Introduction 119
Method 125
Results and Discussion 130
Conclusions 146

Chapter 7 Study 3: Policy Recognition 150

Introduction 150
Method 155
Results and Discussion 157
Conclusions 162

Chapter 8 Study 4: Risk 165

Introduction 165
Method 170
Results and Discussion 176
Conclusions 204

Chapter 9 The Confounded Rating Hypothesis Revisited 206

Introduction *206*
Evidence in the literature for non-linear policy use *208*
Stated non-linear policy use and self-insight *210*
Evidence from Studies 1 to 4 *211*
Testing the hypothesis *212*
Study 4 (Chapter 8) reanalysed *213*
Conclusions *219*

Chapter 10 General Discussion 220

Introduction *220*
On the capturing of GPs' policies *220*
On limits to information processing capacity *223*
On self-insight *224*
Implications for communication *231*
On analysis of judgement and decision making *234*
Future Research *235*

References 240

Appendices 251

Glossary of Terms

These, sometimes quite context specific, terms are explained on their first occurrence in this thesis but for easy reference the explanation is repeated here.

Additive	The influence of (a function of) each cue is independent of the value of other cues.
Calibration	The nearness of judgements to the criterion they are trying to estimate. Measurement in terms of difference between values and the criterion rather than correlation.
CHD	Coronary Heart Disease. Blood supply to the heart muscle (supplied by the coronary artery) reduced.
CJA	Clinical Judgement Analysis. A subsection of Judgement Analysis and of Social Judgement Theory (see Chapter 3). The analysis of judgements made by health professionals in their field of practice.
Cognitive control	Consistency with which a particular policy is used.
Compensatory	A compensatory policy is one in which the value of one cue and its impact on the judgement or decision can be compensated for by the value of another cue, which will also, independently, influence the judgement or decision. Linear models assume compensatory behaviour.
Configural	Judgement making is influenced by the relationships between cues.
Consistency	The degree to which a subject makes the same decision about the same case on a second presentation. Measured using correlation (r_a) of judgements made on two presentations of same set of cases.
Cue	Piece of information available (describing a hypothetical patient or case).
Cue use	A cue is defined as being used when its index of relative importance (standardised regression coefficient in these studies) is significantly different from 0.
Cue selection	In Study 2 (Chapter 6) subjects were only aware of cues they selected to reveal on each case. Cue selection policies used in the analysis in Study 2 consisted of sets of cue weights each of which indicated the percentage of times that cue was selected over a series of cases.
Explicit knowledge	Knowledge that is accessible to consciousness (and is communicable).
Explicit learning	The subject is aware of (trying) to learn and explicitly held strategies are invoked.
Explicit policy	A subjective model of a doctor's decision or judgement making policy. Sets of subjective ratings (stated by the doctor after the task) indicating the relative importance of different cues on decision making.
HRT	Hormone Replacement Therapy. Also the task used in Study 1 (Chapter 5) in which subjects made judgements about their likelihood of prescription of HRT for hypothetical patients. Response on 0-100 scale.
Idiographic	Analysis of the judgements and policy of an individual. See also Nomothetic.
IHD	Ischaemic Heart Disease. Reduction of blood flow to the heart.
Implicit Knowledge	Knowledge a subject is shown to have: it is manifested in their behaviour. But it is not accessible to consciousness or

	behaviour. But it is not accessible to consciousness or communicable.
Implicit learning	Subject learns of relationship between cues (stimuli) or between cues and behaviour without being aware that they are learning.
IS	Information Selection task of Study 2 (Chapter 6) in which subjects made judgements about their likelihood of prescription of lipid lowering agents for hypothetical patients but did this only on the basis of the cues they had selected. Response on 0-100 scale.
JA	Judgement Analysis. The analysis of judgement (or decision) making in which static statistical models are formed of the relationship between cues and judgements.
Judges	People making judgements.
LIPID	Task used in Study 1 (Chapter 5) in which subjects made judgements about their likelihood of prescription of lipid lowering agents for hypothetical patients. Response on 0-100 scale.
Linearity	The relationship between every cue and the judgement is linear when a regular increment (or detriment) of any cue leads to a regular increment or detriment of the criterion. See also configural.
MIGRAINE	Task used in Study 1 (Chapter 5) in which subjects made judgements about their likelihood of prescription of prophylaxis for migraine for hypothetical patients. Response on 0-100 scale.
Nomothetic	Based on judgements of several individuals and trying to ascertain the underlying pattern. See also Idiographic.
Objective policy	Same as tacit policy.
Paramorphic	A representation having the same shape. Used to describe the models formed in judgement analysis: the psychological representation of the policy may be different but its general shape (in terms of its implementation) is captured.
Policy	The judgement or decision policy of a subject is a (static) description of the relationship between cues and judgements (over a set of cases).
PRESCRIBE	Task used in Study 4 (Chapter 8) in which subjects made judgements about their likelihood of prescription of lipid lowering agents for hypothetical patients. Response on "very low" to "very high" anchored scale.
Relative Importance	The influence a cue has on judgement making. Different measures have been used for this but most amount to a measure of the contribution that a cue has made to the total variance in judgements.
RISK	Task used in Study 4 (Chapter 8) in which subjects made judgements about the risk of coronary heart disease for hypothetical patients. Response on "very low" to "very high" anchored scale.
Self-insight	A subject's ability to state the relative importance of pieces of information in their decision making. See Chapter 4.
SJT	Social Judgement Theory. Term coined by Hammond to refer to the approach in judgement analysis whereby the probabilistic nature of the real world is taken into account. See Chapter 3.
Stated policy	Same as explicit policy.
Subjective policy	Same as explicit policy.
Tacit policy	An objective model of a doctor's decision or judgement making policy. Here sets of standardised regression coefficients indicate the relative importance of different cues on decision making.

List of Tables

Table 4.1 Examples of Measurement of Self-Insight in JA (p.70-71)
Table 5.1 Tasks completed by each doctor (p.85)
Table 5.2 Cues and their ranges on the LIPID, MIGRAINE and HRT tasks (p.89)
Table 5.3 Concordance between doctors on different tasks (p.96)
Table 5.4 Mean Consistency and Linear fit of models for each task (p.97)
Table 5.5 Correlations of R[1] and consistency with standard deviation (p.98)
Table 5.6 Effects of inconsistency in judgement making on the first 30 cases (p.99)
Table 5.7 Average Latencies on the LIPID, MIGRAINE and HRT tasks (p.100)
Table 5.8 Subjective ratings LIPID task (N = 33) (p.103)
Table 5.9 Subjective ratings MIGRAINE task (N = 34 doctors) (p.103)
Table 5.10 Subjective ratings HRT task (N = 12 doctors) (p.103)
Table 5.11 Standardised regression coefficients LIPID task (N = 33 doctors) (p.104)
Table 5.12 Standardised regression Coefficients MIGRAINE task (N = 34 doctors) (p.104)
Table 5.13 Standardised regression Coefficients HRT task (N = 12 doctors) (p.104)
Table 5.14 Square of correlation of subjective predictions and actual decision values (R^2_S) (p.107)
Table 5.15 Correlation of regression coefficients with subjective ratings of relative importance (p.108)
Table 5.16 Differences between the number of cues doctors rate above five and their number of significant cues (p.111)
Table 5.17 Differences in means and standard deviations of regression coefficients of cues dichotomized in terms of high and low subjective ratings (p.112)
Table 5.18 Two way analysis of variance of subjective weight-regression coefficient differences (p.114)
Table 6.1 Agreement, linear fit and average judgement on the LIPID and IS tasks (p.130)
Table 6.2 Standardised regression coefficients last 100 cases of LIPID task (N = 30 doctors) (p.134)
Table 6.3 Standardised regression coefficients IS task (N = 30 doctors) (p.134)
Table 6.4 Summary of within-doctor correlations over the 100 cases on the IS task (p.137)
Table 7.1 Results for all doctors on the Policy Recognition task (p.158)
Table 8.1 Cues and their ranges on the PRESCRIBE and RISK tasks (p.174)
Table 8.2 Standardised regression coefficients PRESCRIBE task (N = 36 doctors) (p.179)
Table 8.3 Subjective ratings PRESCRIBE task (N = 36) (p.179)
Table 8.4 Mean Consistency and Linear fit of models for each task (p.180)
Table 8.5 Standardised regression coefficients RISK task (N = 36 doctors) (p.184)
Table 8.6 Subjective ratings RISK task (N = 36) (p.184)
Table 8.7 Correlations between Cue Weights on the PRESCRIBE and RISK tasks (p.187)
Table 8.8 Two way analysis of variance of subjective weight-regression coefficient differences (p.192)
Table 8.9 Cues and ranges used in the Merke, Sharp and Dohme Coronary Risk Calculator (p.196)
Table 8.10 Table of Lens model characters RISK task (p.200)

[1] Correlation coefficients used as data here are transformed to Fisher's Z.

List of Illustrations

Figure 1.1 Example of a model of decision making by General Practitioners (Fox, 1985) (p.4)

Figure 2.1 Decision Tree for a pneumonia diagnosis problem:
Wright (1984), adapted from Christensen-Szalanski and Busheyhead (1979) (p.20)

Figure 2.2 Diagram of Hammond's six modes of enquiry, taken from Hamm (1988) (p.33)

Figure 3.1 Diagram of the univariate Lens Model showing the relationship between the criterion (Y_e), the linear estimate of it (Y'_e) from the cues (X_i), the judgement (Y_s) and the linear estimate of the judgement (Y'_s) from the cues (X_i). For key to the indices, see the Lens Model Equation in the text. (p.51)

Figure 4.1 Plot of average or median fit of subjective linear model to judgement data (R_s^2) with average or median linear fit of data (R^2). Data from 21 studies marked in bold on Table 4.1. (p.68)

Figure 4.2 Hypothetical graphs showing (a) an equal distribution of cue weights, (b) a peaked distribution, typical of actual cue use and (c) a flatter distribution of cue weights typical of subjective ratings. (p.69)

Figure 5.1 (a) Example of a case from the LIPID task (p.90)

Figure 5.1 (b) Example of a case from the MIGRAINE task (p.91)

Figure 5.1 (c) Example of a case from the HRT task (p.92)

Figure 5.2. Plot of Subjective ratings against standardised regression coefficients for (a) the LIPID task (N = 33), (b) the MIGRAINE task (N = 34), and (c) the HRT task (N = 12). (p.109)

Figure 5.3. Plot of Subjective weights against Standardised Regression Coefficients on the (a) LIPID task (N = 33), (b) MIGRAINE task (N = 34), and (c) the HRT task (N = 12) (p.110)

Figure 5.4. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the LIPID task (N = 33). (p.113)

Figure 5.5. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the MIGRAINE task (N = 34). (p.113)

Figure 5.6. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the HRT task (N = 12). (p.113)

Figure 5.7. Measurement of self-knowledge (p.117)

Figure 6.1 The screen seen at the start of a case in the Information Selection task. (p.127)

Figure 6.2 The screen seen during information selection on a case in the Information Selection task. (p.128)

Figure 6.3 The screen seen after information selection and during decision making on a case in the Information Selection task. (p.129)

Figure 6.4 Cue Standardised Regression Coefficients for all doctors calculated on the last 100 cases of the LIPID task (Study 1) and the IS task. (p.133)

Figure 6.5 Average Standardised Regression Coefficients calculated from judgements on the IS task and on the LIPID task of Study 1 (last 100), N = 30. (p.135)

Figure 6.6 Average percentage of cue selection for each cue over all doctors on the Information Selection task (p.141)

Figure 6.7 Average cue selection rating for doctors on the Information Selection task (13 = first selected cue) (p.141)

Figure 6.8 Average absolute standardised regression coefficients on the Information Selection task (p.142)

Figure 6.9 Average absolute subjective rating of cue importance (from the LIPID task, Study 1) (p.142)
Figure 8.1 An example of a case from the PRESCRIBE task (p.171)
Figure 8.2 An example of a case from the RISK task. (p.172)
Figure 8.3 Average standardised regression coefficients for doctors on the PRESCRIBE task (N = 36). (p.178)
Figure 8.4 Scatter plot showing the relationship between 36 doctors' judgements of risk of Coronary Heart Disease (RISK) and their likelihood of prescribing a lipid lowering drug (PRESCRIBE) for 100 hypothetical patients. (p.181)
Figure 8.5 Average standardised regression coefficients for doctors on the RISK task (N = 36). (p.183)
Figure 8.6 Plots of subjective ratings against standardised regression coefficients for doctors on (a) the PRESCRIBE task (N = 30) and (b) the RISK task (N = 36) (p.190)
Figure 8.7 Subjective weights plotted against standardised regression coefficients for doctors on the (a) PRESCRIBE (N = 30) and (b) RISK tasks (N = 36). (p.191)
Figure 8.8 Average standardised regression coefficients and average subjective weights for each cue on the PRESCRIBE task (N = 30). (p.194)
Figure 8.9 Average standardised regression coefficient and average subjective weight for cues on the RISK task (N = 36). (p.195)
Figure 8.10 Diagrammatic representation of the lens model analysis of risk judgements (p.199)
Figure 8.11 Bar graph showing the standardised regression coefficients of the coronary risk calculations over 100 cases. (p.202)
Figure 8.12 Bar graph showing the standardised regression coefficients of the coronary risk calculations over 100 cases. (p.203)

List of Appendices

Appendix 1 Examples of JA studies. (p.251)

Appendix 2 Instructions for the LIPID task (Study 1, Chapter 5). (p.260)

Appendix 3 Instructions for the MIGRAINE task (Study 1, Chapter 5). (p.261)

Appendix 4 Instructions for the HRT task (Study 1, Chapter 5). (p.262)

Appendix 5 Inter-cue correlations on the LIPID task (p.263)

Appendix 6 Inter-cue correlations on the MIGRAINE task (Study 1). (p.264)

Appendix 7 Inter-cue correlations on the HRT task (Study 1). (p.265)

Appendix 8 Social Class groupings and job titles for Occupation cue used in the LIPID, MIGRAINE and HRT tasks of Study 1, Chapter 5. (p.266)

Appendix 9 Sample sheet to remind doctors of the cues available on the LIPID task whilst rating the cues' relative importance (Study 1, Chapter 5). (p.267)

Appendix 10 Standard sheet used by the experimenter to note down subjective ratings in Study 1, Chapter 5. (p.268)

Appendix 11 Mathematical explanation for the conversion of subjective ratings to subjective weights (ψ_i). (p.269)

Appendix 12 Indices on the LIPID task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency. (p.270)

Appendix 13 Indices on the MIGRAINE task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency. (p.271)

Appendix 14 Indices on the HRT task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency. (p.272)

Appendix 15 Explicit policies: Subjective ratings on the LIPID task (Study 1) (p.273)

Appendix 16 Explicit policies: Subjective ratings on the MIGRAINE task (Study 1) (p.274)

Appendix 17 Explicit policies: Subjective ratings on the HRT task (Study 1) (p.275)

Appendix 18 Tacit policies: Standardised Regression Coefficients (β_i) on the LIPID task (Study 1) (p.276)

Appendix 19 Tacit policies: Standardised Regression Coefficients (β_i) on the MIGRAINE task (Study 1) (p.277)

Appendix 20 Tacit policies: Standardised Regression Coefficients (β_i) on the HRT task (Study 1) (p.278)

Appendix 21 Correlation of cue weights LIPID task, Study 1. (p.279)

Appendix 22 Correlation of cue weights MIGRAINE task, Study 1. (p.280)

Appendix 23 Correlation of cue weights HRT task, Study 1. (p.281)

Appendix 24 Between task correlation of Relative Importances of cues, Study 1. (p.282)

Appendix 25 Subjective Weights (ψ_i) on the LIPID task, Study 1. (p.283)

Appendix 26 Subjective Weights (ψ_i) on the MIGRAINE task, Study 1. (p.284)

Appendix 27 Subjective Weights (ψ_i) on the HRT task, Study 1. (p.285)

Appendix 28 Instructions for the IS task (Study 2, Chapter 6). (p.286)

Appendix 29 Inter-cue correlations on the IS task (Study 2). (p.287)

Appendix 30 Comparison of indices on the LIPID task of Study 1 and the IS task of Study 2. (p.288)

Appendix 31 Tacit policies: Standardised Regression Coefficients based on decisions made on last 100 cases of the LIPID task of Study 1. (p.289)

Appendix 32 Tacit policies: Standardised Regression Coefficients on the IS task, (Study 2).

(p.290)
Appendix 33 Percentage cue selection on IS task (Study 2). (p.291)
Appendix 34 Average cue selection positions on IS task (Study 2). (p.292)
Appendix 35 Instructions for the Policy Recognition task (Study 3, Chapter 7). (p.293)
Appendix 36 Key to aid interpretation of LIPID policy bar charts (Study 3). (p.294)
Appendix 37 Key to aid interpretation of MIGRAINE policy bar charts (Study 3). (p.295)
Appendix 38 Key to aid interpretation of HRT policy bar charts (Study 3). (p.296)
Appendix 39 Example LIPID policy bar chart (Study 3). (p.297)
Appendix 40 Policy Recognition scores (Study 3). (p.298)
Appendix 41 Inter-cue correlation in Study 4, Chapter 8. (p.299)
Appendix 42 Instructions for the PRESCRIBE task (Study 4, Chapter 8). (p.300)
Appendix 43 Instructions for the RISK task (Study 4, Chapter 8). (p.301)
Appendix 44 Verbal instructions to ascertain self-insight on Study 4. (p.302)
Appendix 45 Sample sheet seen by GP when rating relative importance of cues in Study 4. (p.303)
Appendix 46 Indices on the PRESCRIBE task: linear fit, consistency, mean judgement and mean latency. (p.304)
Appendix 47 Indices on the RISK task: linear fit, consistency, mean judgement and mean latency. (p.305)
Appendix 48 Tacit policies: Standardised Regression Coefficients on the PRESCRIBE task (Study 4). (p.306)
Appendix 49 Tacit policies: Standardised Regression Coefficients on the RISK task (Study 4). (p.307)
Appendix 50 Correlations between cue values and decisions about prescription for cases with risk judgement greater than the median risk judgement (for that doctor). If N = 50, significant correlations ($p < 0.05$) are > 0.231 (one-sided). (p.308)
Appendix 51 Explicit policies: Subjective ratings on the PRESCRIBE task (Study 4). (p.309)
Appendix 52 Explicit policies: Subjective ratings on the RISK task (Study 4). (p.310)
Appendix 53 Subjective Weights on the PRESCRIBE task (Study 4). (p.311)
Appendix 54 Subjective Weights on the RISK task (Study 4). (p.312)
Appendix 55 Relationship between RISK and PRESCRIBE. (p.313)
Appendix 56 Summary of Comments made by GPs on non-linear cue use on the PRESCRIBE task. (p.314)
Appendix 57 Summary of Comments made by GPs on non-linear cue use on the RISK task. (p.316)
Appendix 58 Published article based on Study 1 (British Journal of General Practice). (p.317)

Acknowledgement

I undertook the research described in this thesis because I was interested in examining General Practitioners' decision making. The grant for the project was given by Polytechnics and Colleges Funding Council.

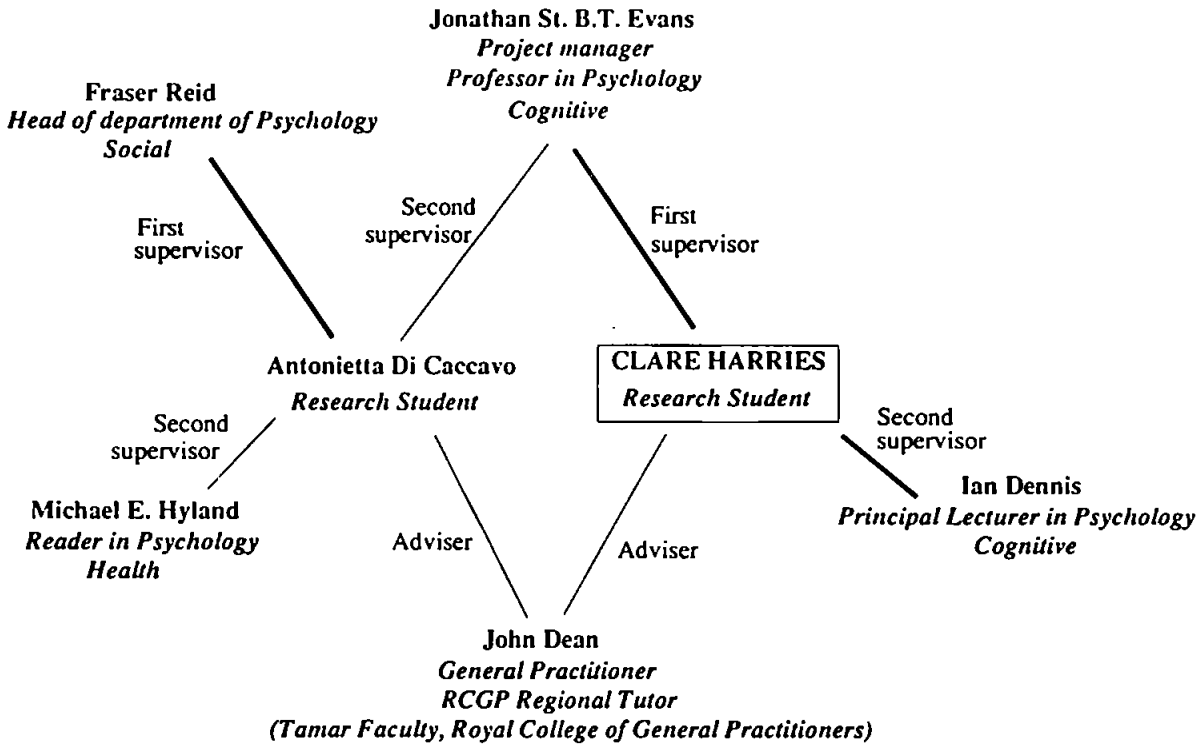
I acknowledge and am grateful for the benefit gained from discussions with both my supervisors Jonathan Evans and Ian Dennis and with other members of the project team (Fraser Reid, Michael Hyland, John Dean and Toni Di Caccavo).

I wish to thank Toni Di Caccavo for her friendship, her outlook, personality and determination that have made the last three years enjoyable.

I wish to thank also the 71 doctors in this study who were paid to participate and Drs John Dean, Janet Longworth, Josephine Harries and Mark Harries who gave consultations free of charge.

Author's Declaration

The work described in this thesis was completed as part of a larger project looking at Medical Decision Making in Primary Care. This was funded by a grant from the Polytechnics and Colleges Funding Council and ran from 1992 to 1995. There were seven members of the project research team, including two research students (one of whom was the current author), both of whom were registered for Ph.D.s. The whole project was overseen by Jonathan Evans who also acted as director of studies for the author and second supervisor for the second research student. The role of the different members of the team can be seen in the Figure below. The research programmes of the two students remained separate throughout the entire project. However, one group of subjects (those in Studies 1, 2, and 3 in this work) were used by both researchers and their recruitment was a joint endeavour, also involving the time of other members of the project team. Recruitment of subjects for Study 4 was carried out by the author (with the help of Jonathan Evans and John Dean).



MEDICAL DECISION MAKING IN PRIMARY CARE RESEARCH PROJECT
1992-1995
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF PLYMOUTH

The computer programs run during the research were written by Jonathan Evans (Case Generation, Study 1; Case presentation program, Studies 1, 2, and 4; concept keyboard adaptation program, Study 2), Clare Harries (Subroutine of case presentation program allowing for a pause in its presentation, Studies 1, 2, and 4) and Ben Rood (Program to randomly assign numbers to doctors' policies and to pick sets of these for presentation, Study 3).

In the course of this research work I observed consultations by Dr. John Dean and Dr. Janet Longworth. I arranged and attended team meetings both with national researchers in the field of general practice decision making (with Dr. Whitfield of Bristol University; meeting in Bristol) and international researchers (with Dr. Rolf Wahlström, Karolinska

Institutet, Stockholm, Sweden; meeting in Plymouth). I have also corresponded in writing with a number of researchers active in the field of medical decision making.

I attended a number of conferences and workshops and courses:-

- **European Summer School on Reasoning and Decision Making (including Workshop on Mental Model and Mental Simulation),** Sienna, Italy, June 1995
- **16th Annual Meeting of Society for Medical Decision Making (including workshops on Social Judgment Theory and its medical applications and Issues in Analysis of Clinical Judgment: Method and Application),** Cleveland, USA, October 1994
- **British Psychological Society (BPS) Cognitive Section Annual Meeting,** Cambridge, September 1994
- **BPS Social Section Annual Meeting,** Oxford, September 1993
- **BPS Postgraduate Workshop,** Lancaster, May 1993
- **Reasoning and Decision Making module of MSc Intelligent Systems** University of Plymouth, 1993.

I have also presented results of this thesis both to other members of the project team and to other groups:-

General Practitioners' insight into their own decision making policies: What are they telling us? British Psychological Society Cognitive Section conference, 1994; Departmental Seminar (10/94).
Social Judgement Theory & lens model analysis. Warnings compliance research group, Department of Psychology, University of Plymouth.

Two papers have so far been written up for publication:-

Evans, J. St. B.T., Harries, C., Dennis, I. & Dean, J. (1995). ***General Practitioners' tacit and stated policies in the prescription of lipid lowering drugs.*** **British Journal of General Practice 45, 15-18.** Presents results of one task of Study 1.
Harries, C., Dennis, I., Evans, J. St. B.T. & Dean, J. (accepted). ***A clinical judgement analysis of prescribing decisions in general practice.*** **Le Travail Humain.** Presents results of Study 1.
A copy of the former is included at the back of this volume as Appendix 58.

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

Clare Harries
Plymouth,
Thursday, August 10th, 1995

Chapter One
Introduction

This thesis has as its main substance the examination of patient management decision-making and patient judgement by General Practitioners (GPs). The nature of general practice, which will be outlined in this introductory chapter, is such that certain decision analytic methods are more applicable and profitable than others (see Chapter 2). Judgement Analysis, introduced in Chapter 2 and further described in Chapter 3, is the main method used here to ascertain the factors affecting GPs' judgement and decision making.

Patient management is what the GP decides to do with the patient (or suggest that they themselves should do). The decision is made during the course of a consultation. Examples of patient management include prescription (when the GP hands you a script at the end of the consultation saying something along the lines of "This should clear it up."); referral ("..make an appointment with the nurse on the way out." or "You should get a letter from the hospital in a few days.") or temporizing ("Lets leave this a few days and see if it gets any better by itself."). In this thesis, the patient management strategy focussed on is prescription. This is a common patient management strategy and makes up the largest item on GPs' budgets (Audit Commission, 1994). Here, decision making by GPs on a few different types of prescription is analysed in terms of the influential factors. The influence of pieces of information is looked at both individually and in terms of combined contribution to judgements about a patient's risk which itself influences decision making.

Decision making by GPs is also used to look at evidence for two other phenomena. The first is the dissociation between explicit and implicit knowledge. Evidence from a number of areas of research suggest that there are limits to self-knowledge: Explicit knowledge is accessible to consciousness and can be communicated. Implicit knowledge can be exhibited but the subject is not aware of what they know and cannot state it. Literature on self-knowledge and self-insight is reviewed in Chapter 4. In the studies presented in this work, GPs' ability to state the factors affecting their decision making is looked at and an explanation is sought as to the basis of this explicit knowledge. The other

phenomenon which will be looked at in relation to decision making and judgement making by GPs is the limit to information processing capacity. This is introduced in Chapter 3. Limited information processing capacity - limited working memory capacity - has been used to explain a number of aspects of human behaviour. Most notably, it is used to explain our suboptimal performance in a number of settings. The implications of limited information processing capacity for medical decision making are discussed.

Throughout the following chapters a number of terms will be used that may be medical, may be specific to the work described here or may be unusual for some other reason. Although these will all be defined at their first point of introduction, a brief definition will also be found in the Glossary of Terms.

Decision making by General Practitioners

The structure of the National Health Service (NHS) is such that, nasty accidents aside, the person who plays the largest part in determining the re-establishment or maintenance of your health is your General Practitioner (GP). Their role has been long-established as the gatekeeper to the resources of the NHS. Over the last few years, changes in funding have, in this environment of private enterprise, led to considerable changes in the role and provisions of different parts of the NHS, General Practice in particular. Increasingly, emphasis is placed on the General Practice team. Different primary care team members are encouraged to take on their own caseload. Nevertheless, GPs still differ from hospital doctors in the ways briefly outlined by Brooke and Sheldon (1985) and described below.

Patient management by General Practitioners is unique amongst doctors in the NHS. GPs see a vast array of case types of grossly differing severities. Instead of having a clinically specific area of expertise defined by a specific and largely identifiable structure of knowledge, GPs encounter a vast array of problems whose treatment requires the use of a vast and expanding knowledge base. It is the identifiable knowledge base that has made several other areas of medical expertise appropriate subjects for the development of expert systems.

Problems encountered by hospital specialists have already been defined somewhat

before reaching the expert. This is not just in terms of the type of case. All cases must be non-trivial to merit referral. On the other hand, patients seen by GPs may range from having no problem in anyone's eyes, to having essentially trivial problems which are of concern to the patient, to having life-threatening and urgent problems. Brooke and Sheldon (1985) indicate that an increase in the number of psychosocial problems being presented is also a feature of general practice. GPs are the first line management. They sort through cases, dealing with many problems themselves and referring a huge number of problems to other agents such as nurses, physiotherapists, counsellors, psychologists, consultants, even social workers.

For a long time GPs have been under incredible pressure at work. The number of patients on practice lists is large, the average GP seeing 140 patients a week just during surgery (Audit Commission, 1994). The time that can be spent on each consultation is necessarily limited. Average consultations have been found to last 5.5 or 6.6 minutes (Morrell, Evans, Morris and Roland, 1986). Time pressure and the desire to bring the consultation to an end have been cited as leading to greater than otherwise prescribing (*e.g.* Audit Commission, 1994, p.17). At the same time a person is the patient of a particular doctor for a long time: for as long as they are living in that area. This contrasts with hospital medicine where each consultation may be longer but an individual is the patient of a particular doctor for a period of time defined by their treatment. In General Practice patients can come back and their progress can be observed over a period of time. The doctor often has personal experience of the patient outside the role caused by the currently presented problem.

What this combination of factors described in the last few paragraphs leads to is the emphasis in primary care of patient management rather than diagnosis led treatment. With the wide variety of cases the development of expert systems and decision making theories for General Practice is problematic (Fox, 1985). Medicine in general, but general practice in particular has been described as an art: Decisions are often made intuitively rather than with a full and explicit understanding of the problem and its solution. Management decisions are made on the basis of evidence available at the time and it is rare that a diagnosis will be complete prior to treatment of the problem. Treatment and assessment of

the problem are often built up side by side over a period of consultations.

So decisions made in general practice are not necessarily based on the models explicitly taught in medical schools. These advocate data collection, diagnosis, therapy planning and management. Models of the general practice consultation are less straight forward. For example, Figure 1.1 shows a model quoted by Fox (1985) where a number of management strategies are available after a certain degree of problem definition. Actual patterns of practice are built up over a period of time. Although doctors in primary care can go back to first (biomedical) principles to decide what to do, immediate patient management decisions are more likely to be based on practical experience and judgements about the best person and way of handling the problem.

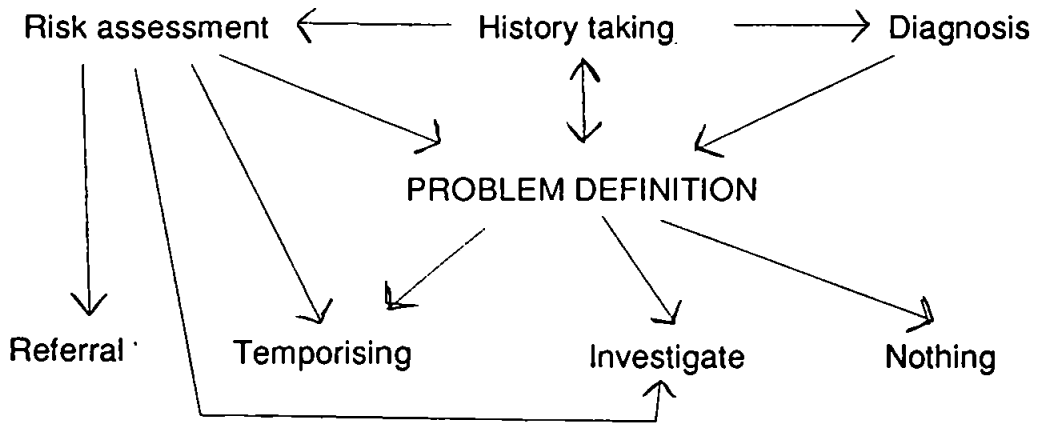


Figure 1.1 Example of a model of decision making by General Practitioners (Fox, 1985)

Analysis of decision making in General Practice

Research into decision making in General Practice, has tended to follow two distinct lines. One method of investigation is based on epidemiological type of analyses influenced by the principles of health economics or survey or questionnaire techniques. The other approach gathers information from the GP in the form of verbal protocols to analyse the consultation, medical problem solving, knowledge base underlying reasoning, or the decision making process. These two approaches will be summarised in the next two paragraphs. There are, however, other ways of analysing decision making which will also be briefly outlined here.

The influence of health economics on research reflects an emphasis on expenditure in the NHS. The emphasis of cost benefit analyses and the search for good quality of life per year measures continues. GPs are provided with feedback in the form of the Prescribing Analysis and CosT (PACT) system. They are set target levels of prescription which are agreed with their FHSA. Much research involves looking at outcomes of prescription decisions and comparing this to global differences in population. For example, in the Office of Health Economics' symposium on factors influencing clinical decisions in general practice; John Griffin discusses the influence of factors such as the number of Catholics in a country, consulting rates by ethnic minorities and the unemployment levels of the population on drug prescription[1] (Griffin, 1990). There are trends in terms of epidemiology and population features. But decisions taken are analysed in groups. Individual differences are rife in general practice as elsewhere in medicine. Decisions are made on the basis of factors such as the knowledge of the patient and of resources quite apart from an understanding of the principles of medicine. There may be several different types of decision being grouped together. The emphasis of this sort of research is on the patient and environment and on when prescription (or referral *etc.*) occurs - one end point of decision making. This sort of approach to decision or judgement making - where individuals responses are grouped for analysis - is known as nomothetic analysis.

The other approach to medical decision making in general practice analyses the sequential process of diagnosis and decision making during the consultation. This leads to the imposition of the kind of consultation framework that Fox (1985) gave an example of (see Figure 1.1 above). The framework is the whole consultation - information gathering, risk assessment, diagnosis and problem definition *etc.* These may be idealistic rather than based on what actually goes on during the consultation where two people are involved in the interaction, and questions of resources, patient attitude, psychosocial problems *etc.* all play a part. Brooke and Sheldon (1985) criticise several existing models of General Practice decision making (of this type) for their emphasis on the somewhat redundant process of diagnosis. One line of research has concentrated on the whole consultation, on

[1] There is, incidentally, a strong positive correlation between a country's percentage spend of gross national product on drugs and the number of Catholics; there is increased prescribing in areas of high unemployment and there are ethnic differences in rates of consultation.

doctor patient communication and on doctor-centred versus patient-centred decision making (see McWhinney, 1985).[2] Many of these models are able to generalise across several types of case which is of course an intrinsic part of general practice. However, again much reliance is on (the experimenter's interpretation of) doctors' explicit representations of the consultation.

Some models looking at medical decision making in general terms have attempted to describe the strategies being used in expert medical problem solving (see Elstein Shulman and Sprafka, 1990). Early studies were influenced by Newell and Simon's (1972) approach to analysing simpler problem solving (Elstein, Shulman and Sprafka, 1978) and there is emphasis on problem definition or diagnosis rather than patient management. Verbal protocol analysis has lead to the identification of strategies such as problem closure (hypothetico-deductive method), pattern recognition or pattern directed search, as well as of course abundant use of heuristics (see also Brooke and Sheldon, 1985). The hypothetico-deductive model of decision making, suggests that hypotheses are formed early in the process, and then evidence for their confirmation sought. Patel and Groen (1986) have suggested that on common problems in their domain of expertise experts are more likely to use a forward chain of reasoning to reach a diagnosis. Grant and Marsden (1987) cite the recognition of forceful features as a prime part of diagnosis and others have pointed out that much of the data could be explained in terms of pattern matching.

As well as descriptions of general strategies, models of particular decision making or problem solving and of knowledge structures have also been formed from analyses of verbal protocols. A decision maker's verbal protocol or their responses to probes can be analysed in terms of the underlying propositional knowledge (Patel and Groen, 1986), hypotheses being formed or causal networks of production rules. In this case a structure is formed, identifying logical connections between information considered by an individual and the end judgement or decision. Fox (1985) emphasises the importance of the explicit formulation of decision processes in forming expert systems. He advocates the construction of rules to represent a GP's general policy or their network of decisions as at

[2] Being patient-centred is characterised by being aware of the patient's point of view and attitude towards their illness and taking their wishes into account. Doctor-centred approaches focus on the illness itself, its identification, and how best to overcome it.

least an educational exercise. In contrast to the general models described in the last paragraph this allows the formation of models specific to a type of encounter and specific to a doctor. There are however dangers here. One is that if decision making in a particular case is being discussed, there may be overemphasis on the features of that particular case. Factors that are not obvious in the case may be neglected. Secondly, several models of hierarchical rules can be analysed from the same data. For example, Groen and Patel (1985) cite Clancey (1984) as showing with his Neomycin model the possibility of analysing several models from the same protocols.

This problem solving approach, modelling the problem solving process, is one of three methods for looking at medical decision making discussed in Chapter 2. Chapter 2 introduces three methods for the analysis of an *individual's* judgement or decision making (idiographic analysis) - decision analysis, information processing analysis and judgement analysis - and compares them. All have different emphases and are more or less appropriate for different settings. Decision analysis again shows the influence of economics. This concentrates on the decision rather than on decision making. The decision is structured in terms of options and outcomes with the aim of identifying the ideal for that individual. Behavioural decision making theories and decision analysis tend not to be appropriate for the analysis of General Practice decision making. Reasons for this will be discussed in Chapter 2.

The information processing approach is described above. The information under examination can be identified in ways other than verbal protocol analysis. However, much is left to the interpretation of the analyst and again the influence of particular features of a case may be distorted. In addition to that, if verbal protocols are to be used then the subject may be required to talk aloud during the task. The appropriateness of this for general practice is discussed in Chapter 2.

Judgement analysis has been applied to decision making by GPs and is the method used in this thesis. Judgement analysis also puts a structure on decision making. This time however, the structure is a more objective, statistical one. The effects of information processing over time are lost but the relative influence of cues on the decision making can be seen and a policy is captured. These different approaches are compared in Chapter 2. In

Chapter 3 the studies applying judgement analysis in the medical context (Clinical Judgement Analysis) are reviewed.

Issues being addressed

One of the reasons judgement analysis is so useful in capturing GPs' policies is because it provides an objective measure of how information is combined. There is no reliance on subjects' ability to state what they are doing. [There is also little reliance on interpretation by the experimenter.] Self-knowledge is discussed in Chapter 4. Generally subjects' ability to state causal, process or combinatorial information seems limited. The dissociation of explicit and implicit knowledge in the field of medicine has implications for how it is taught and how performance is evaluated. Explicit knowledge cannot be relied upon to give information about implicit knowledge.

One model of the development of expertise in medicine can give some insights as to how the dissociation between known and practiced behaviour might come about. Schmidt and Boshuizen (1993) hypothesize that as expertise is aquired, *functionally* different patterns of knowledge are developed. After aquisition of large amounts of disease related causal knowledge, upon contact with clinical experiences this is transposed to what Schmidt and Boshuizen describe as illness scripts. These are context based models of causal patterns of disease. After increased contact with clinical cases, they describe the impact of episodic memories of cases on practice. The implications of this last stage are that some sort of system of recognition system must have been developed for new cases to be recognised as similar to old. This sort of knowledge only comes with contact with numerous cases. An analogy can be made to learning any skill. A person can be told verbally what to do. But it is only through practice that they are able to translate this into a process. However, if a skilled person is asked how they do what they do they are likely to give the same description that they were initially given as instruction. For example, a doctor may have a good explicit understanding of the biomedical principles underlying clinical practice and may have good explicit procedural knowledge but may act intuitively, reflecting implicit knowledge.

The type of procedural knowledge that is expressed explicitly is only one aspect of

the procedure. It is this that is used to convey to a person some of the fundamentals of a procedure. But it does not convey what an expert is doing. It is this sort of procedural knowledge that is tapped during verbal protocols. Einhorn, Kleinmuntz and Kleinmuntz (1979) have described process tracing and policy capturing as two aspects of the same phenomenon. Here I wish to argue that they are two phenomena within the same process. That processes are taught or explicitly described in terms of phenomenal information - information attended to. How cues are combined is discovered after attending to that information. When subjects are asked to state combining information they rely simply on the type of process knowledge that they *can* state - the type that they were taught and the type of which they are aware. Evidence for this is looked at over all the studies, described in Chapters 5 to 9.

If there is evidence for the dissociation of explicit and implicit knowledge in general practice as elsewhere, then there are a number of possible explanations. The possibility that the tacit policies captured in judgement analysis do not adequately capture behaviour is explored and discussed. Judgement analysis captures static policies. Dynamic or otherwise non-linear policies may apparently be mostly captured with a linear model but the variance left unexplained may be the key to their existence. Insight may appear to be lacking where tacit and explicit policies are forced into static linear structures. Evidence for this is explored in Chapters 6 and 9. Similarly some other feature of the task setup such as orthogonality of cues may effect the policy captured. The effects of task design on apparent behaviour is discussed in the chapter on judgement analysis (Chapter 3) and in the final discussion in Chapter 10. If the lack of self-insight on policy capturing tasks is not task created, then the basis for these explicitly stated policies and this discrepancy between stated and tacit policies can be explored. This is done in Chapter 6 in relation to phenomenal knowledge.

The other theme that is explored in relation to GPs' decision making is the limits to information processing capacity. In the studies described differences in policy are measured in terms of the amount of as well as which pieces of information GPs took into account. It was hypothesized that, like any information processing, the number of cues that GPs can take into account when decision making would be limited. With experience,

information processing capacity is still limited in terms of the number of chunks of information that can be held in working memory. However, within the domain of expertise these chunks take on different amounts or groups of information. Theoretically then GPs should be able to use more information in their domain of expertise than in other areas since it will be grouped together in chunks. However, the ability to process more information relies on being able to chunk it. This does not mean that more is attended to. It means that patterns are seen rather than individual features. Medical diagnosis has been likened to pattern matching. Grant and Marsden (1987, also Gale and Marsden, 1985) describe the importance of the recognition of forceful features. But the advantage gained by the knowledge of patterns is lost when the material seen does not fit those patterns. As will be seen, the sets of cases presented did not fit those patterns.

Summary overview of the chapters

The next three chapters give an overview of literature that has a bearing on themes discussed in this thesis. Chapter 2 provides a rationale for the use of judgement analysis on GP decision making. Chapter 3 outlines some of the findings of judgement analyses and its applications in medicine, some design variations and their effects on findings. Chapter 4 introduces the literature on self-knowledge, including an outline of the degree of self-insight shown in judgement analysis studies. The following chapters are all experimental ones.

In the first study, described in Chapter 5, GPs' policies in different types of decision making are captured. As stated before the essence of general practice lies in the sheer variety of cases that may be seen. Here, more than one type of case was presented and doctors' performances on these were compared. Doctors' explicit stated use of cues was also noted and was compared to their captured policies. Study 1 then is used to look at the possibilities for policy capturing in general practice decision making, at doctors' ability to use cues and which cues are used on different decision making tasks, to compare stated and actual policies and to describe the nature of the relationship between the two. The next chapter (Chapter 6) tests one of the hypotheses concerning self-insight - namely that what subjects can describe is what they have attended to. Chapter 7 looks at the degree of self-

knowledge shown by subjects when asked to pick out rather than state their policies. This forced choice task may prove useful for analysing whether meta-knowledge exists but cannot be stated.

One of the tasks presented in Chapter 5 and also in Chapter 6 involved making decisions that would usually be based on assessments of risk factors. Chapter 8 gave doctors the task of making judgements about a patient's risk. Doctors also made managerial decisions on the same set of cases. The difference in policies for the judgement and for the decision were compared. However, in addition the degree of self-insight shown into judgement making on risk was of interest. If subjects are taught to assess risk as part of the management decision making under examination, then the explicit knowledge might have a bearing on the self-insight shown in both the management task and the risk judgement task. Chapter 9 looks at an alternative explanation for the findings of self-insight in judgement analyses. It addresses the issue that the findings are a product of the task design or analysis used. Chapter 10, the final discussion, pulls together the findings from the experimental studies and literature.

Chapter Two
Analysis of Judgement and Decision Making

Introduction

The purpose of this chapter is to look at three approaches used to examine the judgement and decision making process in individuals and to compare their appropriateness for the examination of patient management decision making by GPs. The nature of decision making in General Practice was introduced in Chapter 1. To recap, in an environment where problems are ill-defined and may or may not be severe, time per consultation is limited but knowledge of and contact with the patient is long term, GPs tend to be oriented towards (immediate) patient management decision making. This contrasts with the ideals of hospital medicine where therapeutic plans are set up after comprehensive identification of the disease and more time is allowed for consultation.

In keeping with the findings of other disciplines, there are individual differences in judgement and decision making in general practice. Thus one of the important features of a method for looking at GP decision making is that it should take decision making by different GPs separately. The study of an individual's policy in decision making is an idiographic analysis (see Hammond, McClelland and Mumpower, 1980, Chapter 10). This contrasts with nomothetic analyses that group decision making by different subjects together and assume that different subjects' policies are, bar random error, essentially the same. The three theoretical approaches that will be discussed here - decision theory, problem solving and social judgement theory - can all look at individual's policies. Although nominally different, the types of behaviour looked at in the study of decision making, problem solving and judgement making are often the same. This will be discussed later. However, the three approaches have very different ways of analysing behaviour - through decision analysis (utility or risk based models), through process tracing models and through judgement analysis (often linear regression models) respectively. This chapter will outline the characteristics and benefits of the different approaches with reference to their suitability for analysing decision making behaviour in general practice.

Decision making, judgement making and problem solving

In the laboratory, problems used in the study of problem solving are such that the goal is clear and predefined. In practice in the real world, problems are seldom like this. However, the information processing approach of Newell and Simon (1972) influenced Elstein, Shulman and Sprafka (1978) in their analysis of medical problem solving (see Elstein, Shulman and Sprafka, 1990). The problems under consideration were diagnostic problems. Even when physicians' performances on paper-presented patient management problems were analysed, the focus was on the diagnostic process rather than patient management. Human problem solving methodology was applied to the complex field of medical judgement and decision making. Subjects were encouraged to think aloud and their verbal protocols, as well as their observed behaviour, were analysed in terms of the cues gathered and the hypotheses explicitly generated. The phenomenon Elstein et al were referring to under the title "medical problem solving" is diagnostic judgement and decision making. The goal of the problem would be achievement of the ideal decision or the correct judgement.

Just as "problems" in the real world are not the clearly defined phenomena of the laboratory, the distinction between judgement and decision making too can become hazy in the real world. They are fundamentally interlinked but the distinction between them is itself slim (Arkes and Hammond, 1986). Judgements have to be made in order to make decisions. Judgement making can be thought of as a decision to opt for a particular judgement. Decision making can be thought of as a judgement to choose a particular option. For both judgement and decision making certain information about a situation is available. There is a range, or a number, of possible options to choose from. One of these is selected. Where this is acted upon it can be considered a decision. Where it is assimilated into the subject's view of the world a judgement has been made.

Judgements may be either categorical or continuous and medicine includes both. Diagnostic judgements are categorical. Judgements of severity may be continuous or categorical. Judgements of risk may again be expressed continuously (in terms of numerical probability) or in categorical terms of moderate or high risk etc. Decisions, including patient management decisions, are in general categorical: A choice is made

between the options available. Decision analyses particularly involve framing the decision in terms of the relative merits of a number of choices. However, choice may be made with more or less certainty (a continuous variable) and some cases may fall into diagnostic categories more easily than others. The congruence of the phenomena being observed in theories of decision making, of problem solving or of judgement making in the real world, means that the same real world domain can be examined using different methods and different approaches.

There are different types of judgements and decisions and these may be more or less suitable to analysis by the different approaches. Maule and Svenson (1993) distinguish between static, sequential and dynamic decision or judgement situations. Static situations are those where only one judgement or decision is being made, on the basis of the information available at the time. A sequential situation is made up of a series of static decisions or judgements each of which is affected by the previous one in the sequence. A dynamic situation is one in which decisions are made in the context of an external environment that is changing continuously. In fact decisions in medicine are most likely to be of the latter type when the situation is viewed over a period of time. However, decisions during one consultation can be viewed as sequential. Several static decisions or judgements can also be identified within this. Even static decisions may vary. They may be complex, important decisions (about which the decision maker may feel anxious). They may be (frequently occurring) easy and apparently automatic. Both of these sorts of static decision or judgement may occur during the general practice consultation. Both decision analyses and judgement analyses are really most suited to the analysis of static decisions. Where the decision situation is sequential or dynamic either a series of analyses may be done or a process tracing technique may be used.

Related to this, and discussed further in Chapter 4, there are some judgement processes of which the subject may be able to say little. Recognition and automatic processes are examples given by Ericsson and Simon (1980, 1984) as being unsuitable for protocol analysis. The distinction between these and complex, considered processes need not be rigid: Through practice decisions or judgements that were once complex and

effortful may become automatic (Schneider and Shiffrin, 1985)[1]. Learning to categorize cases on the basis of diagnosis may be an example of automization: At first diagnosing may be a slow carefully considered process but after a while the type of case may be known automatically. Some short cut technique may be used or the process may truly become automatised. One step diagnosis by experts may resemble a recognition process. In the problem solving framework, after practise the achievement of diagnosis may occur by insight as well as through trial and error or some sort of hypothetico-deductive method.

Alternatively judgement or decision processes may have been learnt implicitly (Reber, 1989; Berry and Dienes, 1993; Berry, 1994). In this case the workings of the judgement or decision process would not have been made explicit. But they would have been learnt in the course of dealing with real clinical cases. Schmidt and Boshuizen's (1993) theory of the development of medical expertise in terms of functionally changing knowledge structures might be classified thus. They hypothesize that as skill develops, the pattern of behaviour calls upon different types of knowledge. Initially, knowledge is learnt in terms of abstract patterns of diseases. Then models or scripts of clinical problems in the form of actual patient encounters are learnt as the student doctor is transferred to the clinical setting. Finally, the process becomes one of matching prototypes. Memories of actual previous encounters are accessed in making decision about how to treat the current patient. If the treatment behaviour is learnt on the job, whilst explicitly thinking about the pattern of disease, it might be considered an implicitly learnt process.

Both of the ways of looking at expert decision making described in the last two paragraphs suggest that there is one type of decision making that is automatic or has become a process of recognition. In this type of static situation process tracing techniques may be of questionable value. Certainly protocol analysis has been identified as being inappropriate (see earlier). Other process tracing techniques such as information boards may or may not prove useful. Decision analyses would also appear to be of doubtful use in this type of situation. If a diagnosis or decision has occurred to someone in an automatic or semi-automatic way it is unlikely that they have explicitly analysed all the different options

[1] One interpretation of the automisation phenomenon is that the process becomes made up of short cut methods based on recategorisation (Cheng, 1985). However, Schneider and Shiffrin (1985) argue that this does not explain all the findings.

and will be able to outline them. Judgement analysis on the other hand may be appropriate in that it does not interfere with behaviour or rely on the subject's interpretation of their behaviour.

Distinctions between approaches

Approaches to decision and judgement making can be found to differ along several lines. Firstly, they differ in terms of their original motivating discipline. Much of decision theory was originally based in the principles of economics. The influential problem solving work evolved within the domain of cognitive science and artificial intelligence (AI). Social judgement theory rests firmly within the realms of psychology. Hammond, McClelland and Mumpower (1980) distinguish between economic and psychologically based approaches to human judgement and decision making although they do not discuss problem solving analysis approaches. In the first chapter of their book "Time pressure and stress in human judgement and decision making" Maule and Svenson (1993) outline other essential distinctions between approaches to the analysis of judgement and decision making. They distinguish between structural approaches and process approaches, between riskless and risky decision making and between normative, descriptive and prescriptive approaches.

Structural analyses differ fundamentally from those in process terms. Structural approaches analyse the relationship between the information available for basing the judgement or decision on and the decisions or judgements made. Process approaches try to get at the underlying, sequential cognitive processes. Both behavioural decision theory models and models of judgement analysis are structural models. Analyses of decision making problems on the other hand tend to be process oriented.

Riskless decision making is defined by Maule and Svenson as decision making in the situation where all outcomes are known with certainty. They describe this situation as realistically unlikely. However, making choices between known alternatives would be example of this (see the section on the decision theory approach). The group of decisions Maule and Svenson place under the title "risky" is subdivided into uncertain decision situations and risky decision situations by Fishburn (1988). Risky decisions are those

where, although the outcome is not known, the probability of each occurrence can be estimated. Much behavioural decision theory has been based on this sort of well defined gamble. In uncertain decision situations, the probability of different outcomes is not known. Medical judgement and decision making is intrinsically bound up with both uncertainty and risk. The effect of risk or uncertainty on decision making can be incorporated in different ways.

Decision making theories can be normative, descriptive or prescriptive. A descriptive decision making theory or model captures the essence of an individual's actual behaviour. Decision making models are prescriptive if they define the ideal option and how the decision maker can reach it. A normative model of decision making just defines the optimum decision. Some authors choose to talk of normative theorems as if this were the opposite of description (e.g. von Winterfeldt and Edwards, 1986). In general terms a normative model is one that sets out what the norms or laws of a situation are and is usually expressed in terms of a formula. Laws in society are prescriptive in that they set out a mode of behaviour that should be followed. Laws of nature are simply (mathematical) formulations of how a phenomenon occurs and are therefore in some way descriptive. However, in the realms of decision, judgement, thinking and associated research, a normative model is one whereby the ideal solution is calculated without reference to the method of thinking or cognitive capacity of the subjects but with reference to the environment (Baron, 1988; Maule and Svenson, 1993).

All normative decision making or judgement making theories will refer to a type of decision or judgement. Some of these are designed to structure the decision making or judgement making process as it is being carried out (a priori analyses). This is the form of decision analysis. Other normative decision making or judgement theories take sets of decisions or judgements and form an a posteriori model of the effects on judgement of different facts (Arkes and Hammond, 1986). Social Judgement Theory and Judgement analysis are just such models (see Hammond, Stewart, Brehmer and Steinman, 1975).

The aim here was to describe GPs' decision making as it is, not to capture what GPs might consider to be the ideal mode of decision making. Therefore a descriptive model of decision making is necessary. Much uncertainty abounds in general practice decision

making. Therefore the method used need to be able to be applied to risky decision making or decision making under uncertainty. The description of the behaviour in terms of a process or a structural model is perhaps a matter of taste.

The decision theory approach

The main characteristic of decision analytic techniques is that it is the decision, and the individual's perception of the decision, that is analysed. Because of this the majority are structural rather than process approaches: The structure of the decision is analysed. The decision is seen as a choice between different options which lead to a number of differing outcomes, depending on interim uncertain events. The decision analysis is carried out by initially identifying the options and outcomes. Use is often made of some measure of the (perceived) probability of an event (outcome) and its utility. An option's utility is its usefulness or worth. Some models have incorporated other psychological aspects of decision making such as regret and rejoicing.

There are several riskless decision analytic models. These consider how choices are made between differently valued *known* options. Options are compared on the basis of their overall utility. Examples of these are:-

- Multiattribute Utility Theory (MAUT) in which options are described in terms of the utility of their different attributes. Attributes themselves are weighted in comparison to each other and the final choice is based on the sum of the weighted attribute utilities (Edwards and Newman, 1982);
- Elimination by Aspects in which options are seen in terms of a scalable set of aspects (attributes) and the option is chosen that is left having eliminated options scoring badly on important aspects (described in Wright 1984, p 105-107);
- Minimax strategy where the option that minimises the maximum losses is chosen (Lee, 1971 p.33-3) can be equated in principle to the Maximin strategy where the choice producing maximising the minimum utility would be opted for (Wright 1984 p.12). A maximax strategy selects the choice with the maximum (possible) utility.

Decision theories however, also explicitly take into account the uncertainty involved in a decision. Some of the above strategies (such as MAUT and maximin) can when expected attribute utilities are used instead of attribute utilities. These sorts of models are most relevant to decision making in general practice, which is rife with uncertainty. The original utility based model of decision making under uncertainty was expected utility theory which was formalised by two economists - von Neumann and Morgenstern (1947). This included uncertainty simply in terms of probability attached to the particular outcome's utility and is consequently a model of risky rather than uncertain decision making. According to this normative model, sums of probability weighted utilities are compared for each option and the one with the best expected utility should be chosen. Although this type of expected utility model could be useful in normative terms, it does not deal with the subject's expectations about the likelihood of events.

In Subjective Expected Utility (SEU) theory it is the subject's expectations (or view of the probability of an event) rather than an objective probability that is included (see Wright, 1984 for a discussion). The perceived utility and perceived likelihood of these is used to calculate the best option to be taken. However, SEU too is descriptively inadequate even for the decision making on the simple gambles used in laboratory experiments. Expected Utility theories rest on certain axioms.

- Cancellation (also the sure thing principle): any outcome that is not affected by the decision taken should not be allowed to influence the decision making.
- Transitivity: If option A is preferred to option B and option B is preferred to option C then option A should also be preferred to option C.
- Dominance: If option A is better than option B in one condition and at least equal to B in all other respects, then option A should be the decision taken.
- Invariance: The same decision should be preferred no matter how the decision options are represented.

However, these axioms are found to be violated by subjects' behaviour (Tversky and Kahneman, 1988). For example, subjects' favouring of certainty is such that the axiom of cancellation is inappropriate. Two otherwise equal gambles are assessed differently depending on whether they both additionally have a certain probability of an outcome of nothing or something (Allais' Paradox, 1953 - see Kahneman and Tversky, 1979 p.265). The principle of invariance is broken when subjects respond differently to decisions

described in terms of gains rather than losses (Kahneman and Tversky, 1979).

SEU has been found to be lacking as a descriptive model. Its status as a normative model has also been questioned (see Wright, 1984, p.64). However, decision analysis which is based on the principles of SEU has been found to be useful as a normative, prescriptive model when the decision is complex (Pollitser, 1991). Much of its use depends on a decomposition of the decision in terms of options and outcomes, probability and utility. These are attached to nodes and branches in a schematic representation of the decision known as the decision tree. It may be the aid in identification of options and possible outcomes, rather than the calculation of expected utilities *per se*, that makes decision analysis useful in complex situations but of questionable value in simple ones (Pollitser, 1991). An example of a simple decision tree on the assignment of a diagnosis of pneumonia can be seen in Figure 2.1. Decision analysis has been widely used in medicine (see any volume of the journal Medical Decision Making). For example, Barrett, Parfrey, Foley and Detsky (1994) compared choice of contrast media in the diagnostic cardiac catheterization.

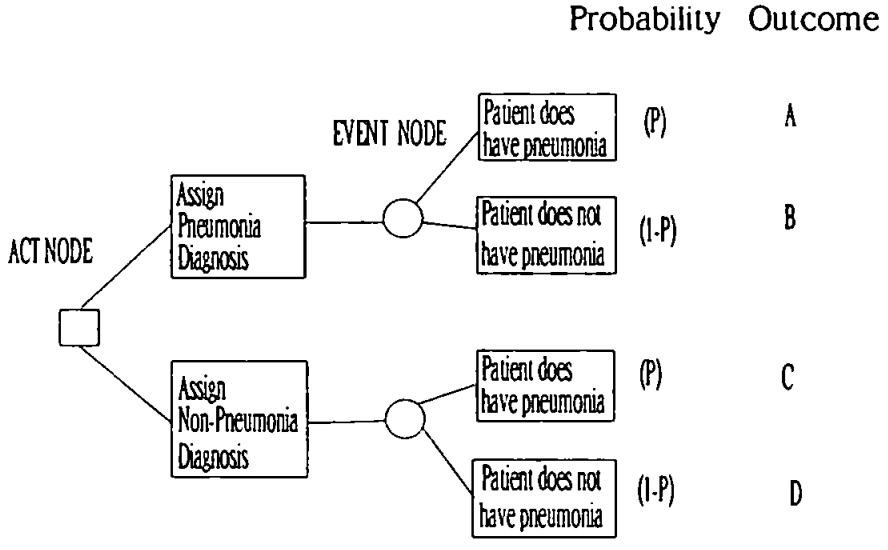


Figure 2.1 Decision Tree for a pneumonia diagnosis problem:
Wright (1984), adapted from Christensen-Szalanski and Busheyhead (1979)

Regret theory models decision making in terms of the regret or rejoicing that different outcomes would elicit having made a choice as opposed to if they had occurred anyway (*e.g.* Loomes and Sugden, 1982). Violations of the principle of transitivity are accounted for in terms of comparative assessments of options rather than their independent

valuation. However, Regret theory maintains the axioms of invariance and dominance and thus, for example, does not account for the different risk seeking and risk aversive behaviour when problems are presented differently (Tversky and Kahneman, 1988). Hynes, Levine, Littenberg and Nease (1994) recently developed a measure of regret by comparing the utility of outcomes of avoidable compared to unavoidable blindness.

Prospect theory (Kahneman and Tversky, 1979) goes a long way towards giving a descriptive account of decision making. Decision making is seen to have two phases - a framing and editing phase and an evaluation phase. In this psychological theory the difference in perception or framing of the decision in terms of losses and gains is taken into account. In simple gambles the observed risk aversion with gains and risk seeking behaviour with losses is predicted by the theory. The effect of framing on decision making has also been seen in complex real world decisions. For example, McNeil, Pauker, Sox and Tversky (1982, also described in McNeil, Pauker and Tversky, 1988) presented subjects with statistical information about two treatments for cancer. Therapy preferences were dramatically affected by whether the information for both was presented in terms of mortality or in terms of survival. Similarly, Verhoef, de Haan and van Daal (1994) found prospect theory predicted healthy women's patterns of risk aversion and risk seeking with respect to gambles with years of life. Most subjects were risk seeking in the short term (up to the lifetime they aspire to) but risk averse in the long term. However, although Prospect theory explains general patterns of decision making and can be used in both nomothetic and idiographic analyses, its use for describing medical decision making is questionable. In the majority of circumstances, decision making in general practice is plagued with uncertainty and not just risk. Analyses looking at behaviour in terms of prospect theory have presented subjects with clearly defined gambles. Probabilities are set to describe outcomes. The statistics are to be taken as presented. In general practice (and indeed the rest of medicine) where statistics exist they are often debated.

Ford (1987) objects to the use of probability *per se* in theories of decision making: Probability is established by pseudo-identical repetitions of events whereas each decision is a unique situation. Probability density functions (p.d.f.) are of no help in determining what will happen in one particular instance. In addition, p.d.f. are based on the inclusion of

all hypothesized rival outcomes. If an outcome has been omitted from this, the probabilities of other outcomes would numerically change. Subjective assessments of probability (subjective probability), however, do not function like this. It is not necessarily true that a decision maker would feel the likelihood of original outcome had changed. According to Ford (1987) both his own Perspective theory and Shackle's theory (Shackle, 1952, 1961) incorporate the perception of uncertainty in terms of 'potential surprise'. Gains and losses are included in terms of an 'ascendancy function' and, in the case of Perspective theory, as a utility base measure of the 'perspective index'. Portfolio theory (Coombs, 1975) discusses preferences between options in terms of risk preferences and maximising expected utility.

Support theory (Tversky and Koehler, 1994) gives an account of subjective probability in terms of the relative support a subject attributes to an event. Probabilities are attached to hypotheses rather than to events and the same event may be described by more than one hypothesis. The probabilities subjects ascribe to the same event have been found to differ depending on how it is described. This violates the principle of extensionality which is upheld by objective probability (whereby the probability of the same event or events has the same probability of occurrence). Support theory accounts for this. The probability of implicit disjunctions (whereby a description implicitly encompasses several types of event) are assessed with a global judgement based on support for the most representative and salient events making up the disjunction. When the disjunction is made explicit (referred to as "unpacking"), the subject may be reminded of possibilities they had previously overlooked, or the explicit mention of an event may increase its salience and the sum of subjective probabilities exceeds that of the implicit disjunction. Tversky and Koehler (1994) give several examples of this, however, the phenomenon has also been demonstrated in medicine.

Redelmeier, Koehler, Liberman and Tversky (1995) showed that house officers[2] estimating the probability of three possible diagnoses for a case assigned lower probabilities for the "none of the above" category than house officers for whom "none of the above" had been further specified in terms of two other diagnoses plus "none of the

[2] These are qualified doctors.

above". When another sample of doctors was asked to assess the probability of one of four possible prognoses for a case, each subject gave a relatively high probability of the outcome they were specifically asked to rate and the sum of probabilities was 164% (objective probabilities should sum to 100%). This unpacking phenomenon is also seen in terms of decision making. When several possible diagnoses were explicitly suggested on a case description the patient management decision by fourth year medical students changed compared to when only one diagnosis was explicitly suggested. In this latter case the possibility suggested on its own appeared more likely than when other alternative possibilities were additionally mentioned. Decisions are affected by discounting of unspecified possibilities.

Decision theory based methods that might potentially do well descriptively have a number of problems. Firstly, some models break up the decision in terms of poorly defined components such as the risk preference referred to in Coombs' Portfolio theory (1975). Secondly, there may be difficulty measuring even clearly defined components such as relate to utility and probability in complex real life decision making. One of Janis and Mann's (1977) criticisms of many decision making theories is that they are applied to important, meaningful, stress causing decisions but are often tested on insignificant or hypothetical decisions such as clearly laid out simple gambles. This difference is important because they hypothesize that the fact that the decision environment is a stressful or risky one plays an important part in the way the decision is made. Indeed there is evidence that decision making under time pressure and stress does affect the underlying decision making or judgement making behaviour (Edland and Svenson, 1993).

The use of decision analytic techniques such as described above have additional problems in the realm of GP decision making. All are a priori methods in that they analyse the decision as it is being carried out and before a decision is reached (see Arkes and Hammond, 1986). There are several problems associated with this. Firstly, just doing a concurrent analysis may prove difficult when the consultation is carried out under time pressure anyway. Secondly, what the subject says may not be a reliable description of their thoughts or attitudes but may be created for that analysis (see Chapter 4 on Self-knowledge). Thirdly, related to this, between them the subject and analyst need to have a

good understanding of the decision situation in order to be able to identify all options and outcomes. This may also prove problematic in the prescriptive application of decision analysis to general practice decision making. Finally, explicitly breaking up the process may change the subject's behaviour, making the approach of little descriptive use. Although some experts perceive SEU as a formalised version of their reasoning, the fact that it is formalised leads to a change of effect (Kassirer, Kuipers and Gorry, 1982). In general practice, as stated in the previous chapter much decision making is carried out intuitively. Although some different decision options and their outcomes will be considered, the explicit outline of all possibilities may change the GP's behaviour.

Problem solving analysis - Process Tracing techniques

Process tracing techniques try to describe the decision making process rather than the perception (and evaluation) of the decision. The use of process tracing techniques originated in the investigation of problem solving (Newell and Simon, 1972). Here a subject's exploration of the problem space of relatively simple problems was described. A process rather than structural description is created. The analysis done is *a posteriori* (see Arkes and Hammond, 1986) but the data is gathered during the decision making process. In process tracing techniques the decision making or judgement process is mapped through time in terms of the pieces of information attended to. The most widely used process tracing technique for data gathering is protocol analysis. However, subjects' eye movements, information seeking and response time can also be used as sources of information about the decision making or judgement process (Maule and Svenson, 1993).

In protocol analysis, verbal protocols are elicited as the subject carries out a decision making or judgement task. It has been argued that although time taken may increase, concurrent thinking aloud by subjects does not lead to any change in decision or judgement making behaviour (Ericsson and Simon, 1980, 1984). These verbal protocols do not provide reliable information about the way decisions are made but may be useful sources of data about information attended to, in the way that retrospective accounts would not. Ericsson and Simon (1980, 1984) suggest that it is the content of working memory that is elicited. Thinking aloud is only possible where there is a process rather than an instant

conclusion. In these cases subjects' descriptions of the contents of their consciousness gives more information than the final judgement or decision. Even Ericsson and Simon (1980, 1984) argued that there were some mental processes (*e.g.* recognition processes) for which introspection was of limited use and which might be subject to interference. It may be that many intuitively made judgements or decisions are just such one step processes (see Chapter 4 on self-knowledge). As discussed earlier, many (well practised) medical decisions are intuitive or automatic. Process tracing techniques will not be of use in describing this behaviour.

Although elicitation of a verbal protocol may not interfere with the decision making process and, like other process tracing techniques, will yield useful data about the information being attended to, these data then need to be used by the analyst to describe the process. Relevant information may not have been made explicit. For example, a person thinking aloud about prescribing HRT for a menopausal woman may omit to make explicit the fact that it is important that the case is female and of menopausal age because these things may seem so obvious to the decision maker. The analyst would need to give the subject a male case in order to find out the relevance of gender to the decision process.

The final description of the process is usually in terms of production (IF... THEN...) rules. The analyst may need to have some knowledge of the decision situation in order to identify these rules. Patel and Groen (1986) suggest that the use of Newell and Simon's (1972) process tracing methodology may be difficult in verbally complex situations. However, their use is possible. For example, Boreham (1989) used process tracing techniques to map out a doctor's decisions on drug dosage in the control of epileptic fits. However, although he felt a significant amount of expertise had been readily expressed as production rules, these rules were occasionally overridden. Since the rules were formed on the basis of decision making on only three cases there may be many other types of cases they would not fit. Einhorn, Kleinmuntz and Kleinmuntz (1979) also identified the difficulty of identification of rules from verbal protocols. Both reliability of the protocols and the rules were questionable. Protocols may not elicit all the information attended to. Rules were difficult to formulate. Process tracing of decisions by novices (on cereal preference) fared better than decisions by experts (on MMPI profiles).

When knowledge has been successfully elicited from an expert it has been combined with the fruits of artificial intelligence research and used to form the knowledge base of expert systems and other decision simulators or decision aids (Kassirer, Kuipers and Gorry, 1982; Fox, 1984; Lundsgaarde, 1987). Several descriptive studies have relied on protocol analysis of patient management problems (*e.g.* Elstein, Shulman and Sprafka, 1978). The importance of the knowledge base in this sort of expert decision making is now well accepted. Elstein, Shulman and Sprafka (1990) identified the need to look at decision making on several types of decision in order to identify expert novice differences. However, the results of Boreham's (1989) study also show the importance of using several different cases to model the rules used in one decision.

Process tracing techniques allow the subject to express the final decision or judgement as they would usually - in categorical or continuous terms. This is certainly an advantage in terms of interference with the decision making process. The focus is on the information aquisition phase of decision or judgement making. Information combination is inferred by the decision maker and included implicitly in the rules formed (Einhorn, Kleinmuntz and Kleinmuntz, 1979). The use of process tracing techniques in general practice is problematic. As mentioned earlier, the intuitive judgements and decisions (those that were implicitly learnt or have become automatic) will be difficult to describe. Where decision making is less automatic, however there are also difficulties. One of the key features of general practice is the interaction between patient and doctor. Talking aloud techniques may well interfere with a behaviour in which talking is often involved. The use of hypothetical patients presented on, for example, an information board would be a possibility. Again, the formation of rules from data about the information gathered would be necessary.

Social Judgement Theory (SJT)

The remit of social judgement theory (Hammond, Stewart, Brehmer and Steinman, 1975) and of judgement analysis is very different from the other two methods discussed above. Decision analysis describes the decision situation (or the subject's perception of it) and process analysis describes the process undergone. Although again it is essentially a

structural approach, judgement analysis, however, describes the judgements or decisions taken in terms of the information available. The subject's response is compared to the stimulus of the cues (or information available) and the relationship between the two is modelled statistically. In this form it is essentially a descriptive approach. However, it also has prescriptive applications. Chapter 3 will discuss judgement analysis in more detail.

In order to carry out judgement analysis, both the response and the cues or information available must be clearly measurable. The analysis (usually multiple linear regression or analysis of variance) is carried out after a number of judgements have been recorded on a set of situations defined along the same dimensions. In the case of clinical judgements these would be cases about which the same pieces of information are known. Subjects' responses must be expressed along a one dimensional scale. Decisions and categorical judgements may be able to be fitted into one dimension if options are mutually exclusive. However, often in medicine decisions include options that can be mutually inclusive *e.g.* any combination of prescribing, referring, giving advice on lifestyle *etc.* Each option or combination of options can be considered separately: Responses can be recorded for these in terms of likelihood or confidence *e.g.* Poses, Cebul, Collins and Fager (1985). Complex decisions can be analysed by comparing models showing the factors affecting the choice of each options on a type of case (*e.g.* the multivariate lens model analysis of Cooksey and Freebody, 1985).

Linear models have been found to be a good fit for much human judgement making (Dawes and Corrigan, 1974). However, it may be the case that the linear fit of the model of the judgements or decision making is moderate but not good. This may be either because the judge is using a non-linear model or being inconsistent in decision making. But even non-linear models can be approximated by linear ones so that important cues can be identified.

Both intuitive or carefully and explicitly thought through decisions and judgements can be modelled using judgement analysis. However, no self-knowledge is required of the subject and the analyst only needs to collect the data and see how input (a description of the case to be considered) varies with output (the decision or judgement made). Judgement analysis has both a descriptive and a prescriptive role to play: the decision or judgement

making by the subject is modelled (in the context of the predictability of the real world) giving a description. Descriptions of a subject's model can be fed back to them along with descriptions of colleagues policies, to gain greater agreement, or the ideal policy, to gain greater accuracy. This cognitive feedback can be used prescriptively by the subject.

Although judgement analysis does not have a probabilistic index of uncertainty, the basis of social judgement theory is to take into account the predictability of the environment. The ideal judgement or decision may follow a non-linear or inconsistent pattern. Lens model analysis (Brunswik, 1952), from which social judgement theory was developed (Hammond, Stewart, Brehmer and Steinman, 1975), analyses a subject's judgements in terms of the information used (judgement analysis) and compares this with the ideal judgements and how well they can be modelled in terms of the information available. A study by Tape, Heckerling, Ornato and Wigton (1991) illustrates the importance of seeing the success of judgements in context. The accuracy of physicians estimates of pneumonia was found to differ in the states of Illinois, Nebraska and Virginia. However, the predictability of pneumonia in patients in Illinois, Nebraska and Virginia, on the basis of the information available, was found to vary in a similar manner.

In that judgement analysis can be used whether judgements are intuitive or thought out it is useful to analyse general practice decision making. However, because of the diversity of cases seen in general practice, collecting large enough samples of the same type of decision problem would be difficult for several problems. Identification of the information available about a patient is also problematic in a clinical relationship characterised partly by its longevity. The GP may be aware of many things about the patient that are not explicitly discussed during the consultation. However, the use of hypothetical patients (with hypothetical histories) gives a solution to both of these problems.

Errors in Judgement and Decision Making

One of the odd things about judgement and decision making is that it is taught in terms of the process involved, but is often judged in terms of the outcome. This is especially true in the area of medicine: It is difficult to imagine complaints about the

practice of a doctor when patients are doing well. However, even when the result is death or debilitation the doctor's behaviour or practice may not be at fault. One maxim might be that the best decision is not always the one with the best outcome. Errors in judgement and decision making can however occur for different reasons. For example, humans may make errors due to the degree of uncertainty involved in the judgement or decision making. Related to this, they may make non-systematic errors or may make mistakes in applying any strategy for other reasons. Humans may also show systematic biases, and may actually be using an erroneous or non-ideal strategy. But the identification of an error relies on identification of an ideal process. If this is not identified then it might be the probabilistic world that is to blame for a poor outcome rather than a suboptimal decision making process.

Judgement and decision making is permeated with uncertainty at all levels. The actual environment in which the judgement is being made may be probabilistic in itself: the relationship between variables and the variable being judged is not completely determinable. In the context of decision making the subject does not know exactly how the choice made will affect the outcome: Other undetermined factors may come into play once the decision has been made. This does not necessarily mean that the possibilities are undetermined in themselves, rather that the judge or decision maker has not yet determined what they are and cannot calculate what they will be. Within this probabilistic environment, the subject may incorrectly guess the value of many unknown variables. Several types of decision analysis aim to quantify or at least conceptually capture this uncertainty in terms of probability or subjective probability (SEU - see Von Winterfeldt and Edwards, 1986) or risk (Portfolio Theory - Coombs, 1975) or some measure of expected loss and gain (Prospect theory, Kahneman and Tversky, 1979) or potential surprise (Shackle's theory, 1952, 1961). Judgement Analysis on the other hand does not explicitly incorporate uncertainty. Where the environment is modelled, Social Judgement Theory takes its probabilistic nature into account (see Chapter 3).

In the context of medicine especially there is much uncertainty. Diagnoses are categorical judgements based on information about a patient. This information often relates in some probabilistic rather than deterministic way to diagnostic categories. Complications

arise and diseases may follow different and unpredictable paces of development in different individuals. Where processes of disease development are understood outcomes and interventions may be predictable. However, often there are so many things to be taken into account when considering the best option that, even if the information was all readily available, the physician (in common with other decision makers) simply cannot hold it all at once.

It is these limits of human information processing that underlie the theory of 'bounded rationality': Humans are limited in their ability to reason perfectly, to problem solve, to make ideal judgements or ideal decisions not least because of the limits on the amount of information that they can process at any one time. To make the best decisions or judgements humans must be selective about and adapt the information available into simple frameworks. But limits to information processing may lead to two types of error (1) the inadequate use of any strategy with frequently occurring but unsystematic errors or (2) the use of heuristics that lead to *systematic* biases (see below).

The importance of our cognitive capacity is widely accepted: many theoretical approaches have acknowledged information processing limits. For example, Janis and Mann (1977) discuss how stress may lead to different types of decision making in terms of the effectiveness of information processing. Simon put forward the idea of a "satisficing" principle: Humans pick an option that is good enough rather than hold on till the ideal choice is found (Wright, 1984 p.104-105). Decision analysis attempts to put some structure on the task to ease handling of information (See Von Winterfeldt and Edwards, 1986). Indeed most decision aids are based on the principle of "divide and conquer" (Slovic, Fischhoff and Lichtenstein, 1977), aiming to limit the amount of information that must be attended to at any one time. The problem is not just a matter of how we cope with uncertainty given our limitations: Models of choice based on trying to weigh up fully known alternatives against each other also take limitations into account. Tversky's notion of "elimination by aspects" discusses decision making in terms of looking at a few pieces of information at a time (see Wright 1984, p.105-107 for a brief description).

Where errors are unsystematic such as when a subject has no real policy, is inconsistent, or makes mistakes, again different approaches to analysing decision making

and judgement may be beneficial. Decision analyses, described above, may help to clarify exactly which pieces of information are important to the decision maker and may help to take them all into account. Process tracing techniques such as verbal protocol analysis may help because the subject is led to think about a policy and be consistent in its application. Judgement analysis allows ascertainment of the subject's policy. 'Bootstrapping' is the process whereby this policy is then applied to a new set of data with better results than the error-prone human from whom it was derived (see Chapter 3).

Limits to our information processing capacity have been put forward as reasons for our use of heuristics (Tversky and Kahneman, 1974). The point of a heuristic or "rule of thumb" is that it usually yields an appropriate formulation of an idea in an easier way than other, more thorough methods of information processing would. However the use of heuristics may lead to *biases*. It has often been through these biases or systematic errors that the particular heuristics have been educed *e.g.* representativeness and availability. Kahneman and Tversky showed the use of heuristics in subjects' judgements of likelihood of events or categorizations. Where decisions are being restructured in terms of the subjects' likelihood judgements, this is obviously relevant to decision making. However, there are other heuristics such as Discounting (Kelley, 1972; described in Heller, Saltzstein and Caspe, 1992) and other biases, *e.g.* confirmation bias, that have been identified that clearly affect judgement making (Janis and Mann, 1977, p.82-85). Many of these may derive from limits on our cognitive capacity. For a fuller discussion and review of biases in judgement see Evans (1992) or with reference to medicine see Heller, Saltzstein and Caspe (1992). Boreham (1989) suggested that the production rules identified in process tracing were capturing the heuristics used in expert decision making.

A systematic bias shows use of an erroneous strategy. To overcome this sort of error in judgement and decision making, prescriptive models can be applied: during decision analysis the ideal decision can be mapped out using the subject's utilities or cognitive feedback can be used in the context of judgement analysis. In judgement analysis algebraic models look at which items of information are used in relation to each other to come to a judgement (Hursch, Hammond and Hursch 1964, Tucker 1964). Where access can be gained to the criterion being estimated, its relationship to the information can also

be calculated and this can be fed back to the subject (see the section below on cognitive feedback). With process tracing techniques, models formed from the protocols of experts can be used to guide more naive individuals through the decision making process[3].

So, errors that are systematic biases or or mistaken policies can be rectified by use of a prescriptive model of one sort or another, *e.g.* decision analysis or cognitive feedback. However, systematic errors caused by the use of heuristics may be eliminated by decision aids or other systems to aid the handling of information. Unsystematic errors can also be rectified by the use of a decision aid or through bootstrapping to identify the core policy to be implemented.

Overview

As stated earlier the three different approaches described briefly in this chapter each have their pros and cons. One obvious difference is the mode of response. Decision analytic methods necessitate a categorical response. Although analysis in terms of production-rules may be easier if the response is categorical, process tracing techniques may allow the subject to define the response being given. Having defined the judgement or decision being made judgement analysis necessitates a response along a scale (for regression analyses) or in terms of categorical, mutually exclusive options (for analyses using ANOVA). This is in accordance with several sorts of judgement anyway (*e.g.* judgements of severity or risk) but also allows the subject to express categorical judgements or decisions in terms of degrees of confidence or likelihood.

The type of decision or judgement being made will have a strong bearing on which method is suitable. Although judgement analysis can be applied to both decisions that are carefully thought through and 'instant' ones, both decision analysis and protocol analysis may rely on subjects having some ability to verbally express their perception of decision or judgement making. The analytic demands on the subject are intellectually fewer in process tracing techniques than in decision analysis, but this relies on decision or judgement making occurring in a series of (verbalizable) steps. Where decisions or judgements are made quickly, the elucidation of the steps in the process may not be possible. Much of

[3] These models can also be used as the basis of an expert system.

General Practice and medicine has been described as intuitive. GPs spend on average seven minutes with their patient in a consultation and many judgements and management decisions are made about the patient. Information is presented to them in the consultation on top of previous knowledge they have amassed. In an interview study, many general practitioners, most of whom also participated in Studies 1, 2 and 3 reported here, explicitly admitted lack of knowledge of the mental processes that affect their decision making. Others described the process of negotiation with the patient once they themselves had worked out courses of action; another group focussed on medical experience, and others discussed the issue in decision theoretic terms (Di Caccavo and Reid, 1995). Thus in this study, judgement analysis, which does not require metacognition, or any sort of insight on the part of the subject seemed an appropriate method.

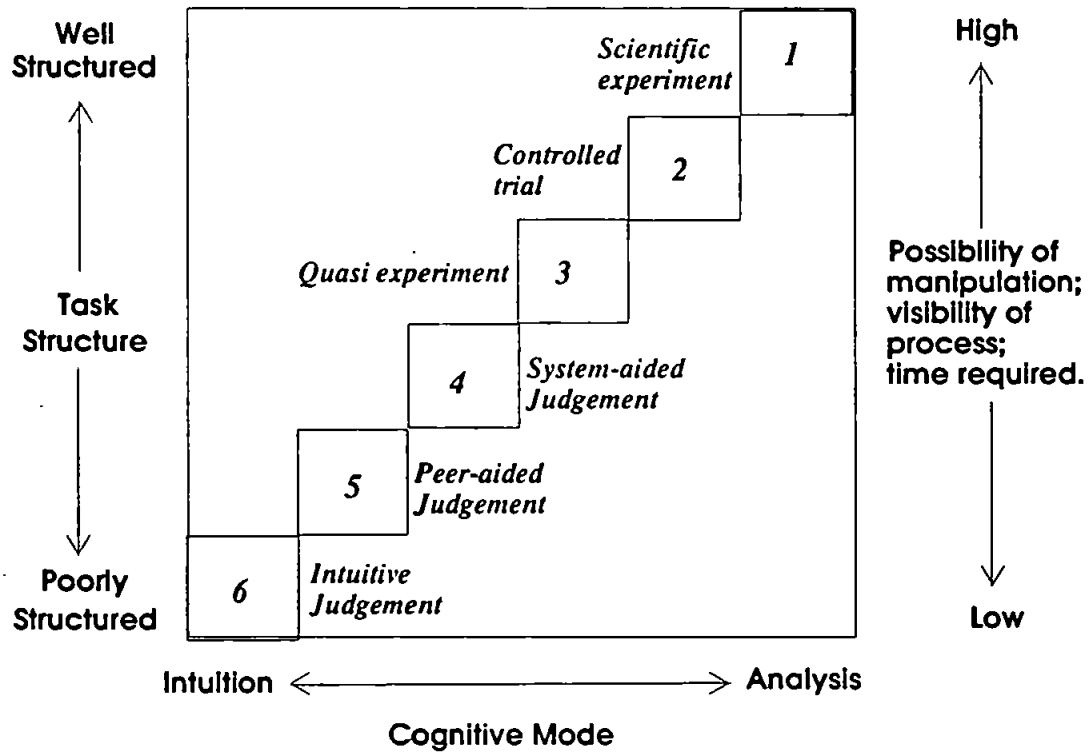


Figure 2.2 Diagram of Hammond's six modes of enquiry, taken from Hamm (1988)

Hammond's Cognitive Continuum shown in Figure 2.2 also provides a good argument for the advantage in the clinical setting of judgement analysis over the other approaches discussed here (see Hamm, 1988). The cognitive continuum is a framework of thinking. Analytical thinking occurs at one end and intuition at the other end of a

continuum. Most thinking, being quasi-rational, lies somewhere in the middle. The degree of conscious awareness of the process decreases as the process becomes more intuitive. Hammond argues that the appropriateness of the type of thinking depends on how well the task is structured. Where tasks are well structured, any of the judgement and decision analysis approaches are useful. At the analytical, well structured end of the continuum, judgement analysis may require a more complicated model than a simple linear one. However, as tasks become less well defined judgement analysis has an advantage over other methods. Hamm (1988) argues that thinking in the clinic usually involves intuitive or peer-aided judgement. He refers to the work of Dreyfus and Dreyfus (1986) who argue that more and more intuition is involved in the judgement and decision making process as greater expertise is gained. Thus clinical judgement and decision making by experienced doctors would best be modelled using judgement analysis.

The purpose of the analysis is also important in considering which method to use. Descriptive applications of decision analytic techniques have been found wanting. Prescriptively they are widely used in medicine and elsewhere for decision support but there are problems even with this application (see Doubilet and McNeil (1985) for a discussion of decision analysis in medicine). The descriptive capabilities of Prospect theory, whilst impressive in a risky decision making setting, are of less use in a situation characterised by uncertainty rather than risk. Process tracing techniques can be both descriptive and prescriptive and have been used to develop expert systems and other computer based decision aids. Descriptively and prescriptively they are useful if (and this may sound tautological) the process is one that is identifiable through time. Some pioneering work in medical decision making and judgement was done using protocol analysis (Elstein, Shulman and Sprafka, 1978). The emphasis in this sort of research is on data aquisition rather than interpretation or use. The assumption is that if data is collected or attended to consciously then it must have a bearing on the decision. But Elstein, Shulman and Sprafka (1978) found no correlation between data collected and accuracy of its interpretation (see Elstein, Shulman and Sprafka, 1990). If a subject is aware of the process and the information they are attending to process tracing will be useful for training but of little use to the individual concerned. If the subject is not aware then process tracing

techniques can be of little descriptive or prescriptive use. Although the model may be a paramorphic rather than a representative one (Hoffman, 1960), judgement analysis is clearly useful descriptively.

Each of these three approaches has a different role to play. This is much the conclusion reached by Einhorn, Kleinmuntz and Kleinmuntz (1979) who discuss the relationship between linear models (the basis of judgement analysis) and process tracing models of judgement. The difference between these two is in terms of differing levels of analysis and of emphasis within the same behavioural phenomenon. Kaplan (1975) pointed out that our understanding of any phenomenon is dependent on the approach taken. This becomes particularly important when, as in a probabilistic environment, a subject's behaviour and policy is evaluated not by its outcome but by comparison with some ideal. In these studies the decision is not easily analysed, whether intuitive quasi-rational or analytical. General practitioners do not necessarily have the time or the ability to talk us through the process. They have become experts who are working under time constraints and the method of judgement analysis that interferes least and describes best that process seems the most efficacious.

Chapter Three
Judgement Analysis and its Application in Medicine

Introduction

The basic principle of judgement analysis was introduced in Chapter 2. Its aim is to capture judgement or decision making policies in terms of the available pieces of information. The decisions or judgements could be those of several people (a nomothetic analysis) or of individuals (an idiographic analysis). In judgement analysis, judgements (or decisions) are made over a number of cases or situations which are made up of pieces of information (cues). The judgements are described statistically in terms of the information that is available (see Stewart, 1988). This chapter will outline the findings and uses of judgement analysis and some of its applications in medicine but first a word about terminology. Where judgement analysis is applied to clinical judgements it is known as clinical judgement analysis (CJA). Social Judgement Theory (SJT) provides a framework within which judgement analysis is usually carried out. Social Judgement Theory (SJT) is based on the psychology of Egon Brunswik (1952) and places the ability to make correct judgements within the framework of the probabilistic real world environment (see Brehmer and Joyce, 1988). This contrasts with the deterministic and process oriented view of much of science (and medicine). Another term often used to refer to judgement analysis is "Policy capturing". However, this term has been objected to for two reasons[1]. Firstly, it assumes that there is something there that is being captured. In fact, as is pointed out later, the description of judgement policy is a paramorphic representation of behaviour (Hoffman, 1960). Secondly, it is not immediately obvious that the term policy refers to the judgement policy. The term has been used in other contexts to refer to some ideal or prescribed mode of behaviour. Throughout this thesis the term "policy" will refer (descriptively) to judgement policies.

There are a number of ways of carrying out judgement analysis. The decision or judgement and the information pertinent to it must be identified. But then the decisions or judgements to be analysed can be real ones or made on hypothetical cases. A subject's policy is described in terms of the relative importance of the cues available in his or her

[1] Tom Stewart makes this point in a message on the electronic mail Brunswiklist, 8/95.

judgement or decision making. Although multiple linear regression is commonly used as a basic analysis, non-linear terms can be included, though add little to the fit of the model to the judgements (Brehmer and Brehmer, 1988). Having identified the relative importance of each piece of information in the decision making or judgement making this can be compared with a subjective measure of each cue's relative importance in an examination of self-insight. It can be compared with other subjects' relative importance weights of cues or it can be compared with the ideal relative importance weights. In either of these cases the models produced can form the basis of a "lens model analysis" of achievement or agreement. This will be described later.

The widespread use of judgement analysis, linear modelling and social judgement theory in medicine can be seen in the overviews of the area (Stewart and Joyce, 1988; Wigton, 1988a and 1988b and Engel, Wigton, La Duca and Blacklow, 1990). Work has been carried out looking at diagnostic judgements (*e.g.* Tape, Heckerling, Ornato and Wigton, 1991; Poses, Cebul, Collins and Fager, 1985; Centor, Witherspoon, Dalton, Brody and Link, 1981; Wigton, Hoellerich and Patil, 1986), judgements of severity (*e.g.* Kirwan, Chaput de Saintonge, Joyce and Currey, 1983b; Fisch, Gillis and Daguet 1982; Fisch, Hammond Joyce and O'Reilly, 1981; Chaput de Saintonge, Kirwan, Evans and Crane, 1988; Kirwan, Chaput de Saintonge and Joyce, 1990), of risk or morbidity (*e.g.* Tape, Kripal and Wigton, 1992; Rovner, Rothert, Holmes, Ravitch, Holzman and Elstein, 1985), and other judgements about the patient or other subjects within medicine (*e.g.* Rothert, 1982; Chaput de Saintonge, Crane, Rust, Karadia and Whittam, 1988; Speroff, Connors and Dawson, 1989). Several studies have looked at patient management decisions such as prescription (*e.g.* Holzman, Ravitch, Metheny, Rothert, Holmes and Hoppe, 1984; Elstein, Holzman, Belzer and Ellis, 1992), referral (*e.g.* Rovner, Rothert, Holmes, Ravitch, Holzman and Elstein, 1985; Rothert, Rovner, Elstein, Holzman, Holmes and Ravitch, 1984) or other, more specific types of management (*e.g.* Smith and Wigton, 1983 analysed decisions about tube feeding in seriously ill patients).

As can be seen from the repetition in the above references, some of these studies have collected more than one type of judgement about the same case (*e.g.* Rovner, Rothert, Holmes, Ravitch, Holzman and Elstein, 1985; Fisch, Hammond Joyce and O'Reilly, 1981).

Judgement analysis has also been used on decisions about applicants for medical courses and posts (*e.g.* Young, Woodiscroft and Holloway, 1986, cited in Wigton, 1988a) and on clinical decision making by other professionals apart from doctors and medical students (*e.g.* Holzemer, Schleutermann, Farrand and Miller, 1981, Ullman, Egan, Fielder, Jurenec, Pliske, Thompson and Doherty, 1981; Ullman and Doherty, 1984). Judgement analysis can only use judgements on a univariate scale. Judgements that are normally forced to be categorical can be expressed in probabilistic terms (*e.g.* Tape, Heckerling, Ornato and Wigton, 1991; Poses, Cebul, Collins and Fager, 1985) or in terms of certainty (*e.g.* Fasoli, Lucchelli, Blasi, Tosi and Colombini, 1992; Roose and Doherty, 1976) or as degrees of severity (*e.g.* Kirwan, Chaput de Saintonge, Joyce and Currey, 1983a,b,c; La Duca, Engel and Chovan, 1988).

Judgement analysis identifies policies being used. The proliferation of individual differences in medical decision making has been well documented (*e.g.* Slovic, Rorer and Hoffman, 1971, cited in Wigton, 1988a; Fisch, Gillis and Daguet, 1982; Kirwan, Chaput de Saintonge, Joyce and Currey, 1983; Fasoli, Lucchelli, Blasi, Tosi and Colombini, 1992). As a consequence idiographic analyses, taking decision making by one individual, may be more useful than nomothetic analyses where judgements from several individuals are grouped to identify trends in behaviour. [Judgement analyses can of course be done on either.] However, caution is needed in interpretation of inter-individual comparisons. When cues are correlated, different policies of cue use may lead to the same judgements or decisions (Brehmer and Brehmer, 1988). Although capturing actual differences in policies, studies, as will be seen later, some aspects of task design such as the degree of inter-cue correlation may exaggerate differences in individuals' practices. Judgement analysis can however be used prescriptively to reduce individual differences. Cognitive feedback or feedback of policy, which will be described later in this chapter, has been found to help both agreement between medics and their achievement of correct judgements (see Wigton, 1988b).

As the above overview has shown, clinical judgement analysis has usefully captured individuals' policies in a variety of clinical decision or judgement making situations. The findings in medicine however, also fit the general conclusions reached by

Brehmer and Brehmer (1988):

(1) that linear models adequately describe the judgement process,

(2) that judges are inconsistent,

(3) that there are wide interindividual differences,

(4) that judges use few cues,

(5) that judges have little self-insight.

Brehmer and Brehmer examine the validity of these conclusions and point out that design aspects of judgement analysis which may have affected results have rarely been controlled for. Different design aspects of judgement analysis are discussed below.

Study design and real versus 'paper' cases

There are a few factors that clearly affect the results of judgement analyses. The cues that are presented, the range they take and their intercorrelations can be seen to affect subjects' apparent consistency, the fit of the linear model generated, agreement between subjects and the number of cues used.

The number of cues presented will affect the subject in several ways. Firstly if only a few cues are presented where more would be available in real life the task may seem less realistic. If non varying information is given as a filler then the results are less generalisable. For example some medical studies might describe several different cases, all of whom were white 50 year old males. If more cues are varied then the number of cases needed for analysis increases dramatically.

The range of cues presented is again important. The relative importance of the cue seen will of course be affected by its range in the cases presented. If a cue varies greatly it may account for most of the variance in decision making and the role of the other cues in decision making may be obscured. If different sets of cases are presented to the same or different subjects different policies will result: Cue weights will be different.

Inter-cue correlations can have several effects. Firstly, subjects who alternate use of highly correlated cues between cases may appear to have greater consistency of policy than is actually the case. Consequentially they may appear to have a better fitting linear model.

Where cues intercorrelate it is difficult to identify the actual cues leading to variance in decision making. Subjects may appear to use fewer cues than they actually are. For example, Phelps and Shanteau (1978) found subjects used a maximum of three cues where an intercorrelated cue design was used and a step-wise regression was run as normal. But when the factors were put into one-variable regression analyses seven to eleven cues were significant and the same judges used between nine and eleven cues where cues were orthogonal. Reilly and Doherty (1992) ended up measuring the influence of twelve correlated cues in terms of five factors. Finally, intercorrelated cues may exaggerate the agreement between subjects. Although Brehmer and Brehmer (1988) make this point the other way around: tasks with orthogonal cue design may exaggerate the disagreement between subjects the point is the same one. Differences of policy which show up best with an orthogonal cue design may disappear when the relevant cues are highly correlated as they often are in real life.

Judgements for analysis can be made on real or hypothetical cases, presented as real or paper 'patients'. 'Paper' patients are written descriptions of the case which may appear on cards or paper (*e.g.* Holmes, Rovner, Rothert, Schmitt, Given and Ialongo, 1989), on computer screens (*e.g.* Wigton, Patil and Hoellerich, 1986) or could in theory just be described to the subject. Wigton (1988a) gives a summary of use of paper and actual cases in clinical judgement analyses. Judgements are by necessity given as a univariate response but using real patients maximises validity so that generalisations can be made to real life (*e.g.* Poses, Cebul, Collins, and Fager, 1985; Tape, Heckerling, Ornato, and Wigton, 1991). However, there are a number of disadvantages to this approach too. For example, there are problems of interpretation of cue use where cues naturally correlate, the range of cases presented cannot be controlled, measurement of independent variables may be difficult, and the same cases cannot be presented to more than one subject so that intersubject comparisons may be confounded. Using paper patients can overcome some of these problems but a question hangs over the validity and therefore generalizability of results.

The validity of paper cases, or their ability to elicit judgement and decision making similar to real life, has only been examined in a few studies. Morrell and Roland (1990)

who concluded that the validity of paper cases varied and couldn't be guaranteed can be regarded as correct. However, in their analysis they fail to examine the differences between those studies where validity was good and those in which it was deficient and they fail to make any assessment of the means of measurement of validity. For example all the studies they cite as examples of where poor validity has been shown are total task studies such as patient management problems rather than judgement analysis studies (*e.g.* Goran, Williamson and Gonella, 1973; Page and Fielding, 1980; Norman and Feightner, 1981). The validity of the former, measured in terms of the whole behaviour, including the information selected, tend to have low validity. The total task may be made up of a series of judgements rather than just the end judgement or decision and so may be more open to inconsistencies.

The appearance of validity of judgement analyses depends partly on how validity is measured. If global comparisons are made and behaviour is compared over sets of decisions or judgements it is important that the case mixes presented in real life and on paper cases is the same. Otherwise cue use and judgements made could vary considerably for one individual's behaviour (e.g. Holmes *et al,* 1989) or between groups. For example, Morrell and Roland (1990) found a poor correlation between doctors' referral rates on a set of 21 paper cases and in real life. This may be because the case mix was not matched with any of the doctors' case loads. Similarly, although Holmes *et al* (1989) set up their cases to reflect epidemiologic data, the real patients included in the study aged from 16 years whereas the youngest paper 'patients' were 25 years. Fenichel, Murphy, Wigton and Schwartz (1984) found that cue use was much the same on paper and real cases but they do not mention controlling for a similar case mix. When the distribution of cases has been carefully controlled for correlations between behaviour in real life and on paper cases validity has been good. For example, Rovner, Rothert, Holmes, Given and Ialongo (1986) found that the number of tests physicians ordered on paper cases correlated highly with the number they ordered on actual cases. The correlation was slightly higher where cues were present in realistic proportions and the same cues seemed to affect the decision as affect real life decisions. However, on the real life cases, physicians used slightly less tests.

Case by case comparisons of validity compare the judgement made by an

individual on a real case and on a paper version of the same. However, some account must be taken of the level of consistency shown and this is generally measured with a repeated presentation of cases. Kirwan, Chaput de Saintonge, Joyce and Currey (1983a) found a high correlation (mean r = 0.9) between rheumatologists' judgements of disease activity on real patients and paper cases made from these. However, consistency was also extremely high on these paper cases (mean r = 0.97). Although quoted as an example of the validity of paper cases in antibiotic prescription for Otitis Media (Chaput de Saintonge and Hattersley, 1985), Chaput de Saintonge and Hathaway (1981) showed a significant but not brilliant level of agreement between diagnosis reached in real life and that shown on paper version and paper version plus photograph of the ear in question ($\kappa = 0.45$). Agreement here may have been reduced since decisions in three conditions rather than two were being compared. In both the Chaput de Saintonge and Hathaway (1981) and the Kirwan, Chaput de Saintonge, Joyce and Currey (1983a) studies the real life decision making could have been made more artificial and perhaps more similar to the paper case decision making by the use of proforma lists to record patient data. Although behaviour may be warped just in the clear collection of cues and the univariate expression of judgement analysis, the paper presentation of cases does not seem to make this any worse.

Linear modelling and consistency

In making a judgement, proximal and distal variables are put together to form a "cognitive map" (Tolman, 1948). This map can be statistically modelled. The representation of the judge's map, through linear regression of the judgement on the various variables (cues) or through analysis of variance techniques, is a "paramorphic" one (Hoffman, 1960). This is not necessarily what is going on inside the judge's mind but is one way of representing it. The ability of the model to describe the judge's decision making behaviour is measured by its fit. This is a measure of the amount of variance in the judgement or decision making that can be explained by the variables available in the model. In the case of a linear model this **coefficient of multiple determination** of a model (R^2) is the square of the multiple correlation between the cues and the judgement (R).

If the cases are described in terms of independent variables (X_i) and the judgement

made on a case is the dependent variable (Y_S) then the linear estimation of Y_S is Y'_S:

$$Y'_S = b_{s0} + b_{s1}X_1 + b_{s2}X_2 + b_{s3}X_3 + ... + b_{si}X_i$$

The linear fit of the model shows the subject's consistency of policy use or cognitive control. Hammond, Stewart, Brehmer and Steinman (1975) define this as the similarity between an individual's judgements and the predictions of a specific model (see also Hammond and Summers, 1972). Of course this is affected by the **consistency** of judgement or decision making as well as the appropriateness of the model for the judge's behaviour. The index used for consistency, or reliability, is the correlation (r_u) between judgements on repeated cases.

Where consistency has been measured in studies it has varied from person to person. For example, Einhorn (1974) found consistency ranged from r = 0.19 to r = 0.93. Averages, typically around r = 0.7, indicate that not all variance in judgements could be accounted for by one static model (see Appendix 1). Consistency of categorically expressed judgements can be expressed as a percentage agreement in responses. In these circumstances consistency appears to be generally good. For example, Chaput de Saintonge and Hattersley (1985) found consistency varied between 79 and 100% agreement. Where only a few cases are repeated initial judgements may actually be remembered and consistency may be inflated. Bech, Haaber, Joyce and the Danish University Antidepressant group (1986) is cited by Brehmer and Brehmer (1988) as an example of this.

Studies using linear models have found an interesting phenomenon known as **bootstrapping.** In a comparison of the performance of judges with the performance of their linear models on new sets of data (a 'hold out' sample), the linear model is superior or as good as the judge (Meehl, 1954). In a review of the one study in which the judge was clearly superior to the model, Meehl (1965) came to the conclusion that the clinicians in the study had actually had access to "signs" that were not included in the actuarial models and stayed with his initial viewpoint. In fact even improper models can also be found to outperform judges if the correct variables are included (see Dawes, 1979 for a discussion). The explanation usually put forward for bootstrapping is that subjects lack **consistency**.

Random errors produced by the subject will lead to different decisions being made on different occasions. Using a formula eliminates the error caused by the inconsistency in the judge's implementation of his or her own policy (see Baron, 1988).

Measurement of consistency is important in determining whether the model used has captured all the explicable variance. A linear fit may be low because the subject is using a non-linear model or because they are failing to consistently apply any model. Generally there is a significant correlation between linear fit and consistency (Brehmer and Brehmer, 1988). Linear fit varies between individuals and often where the linear fit is low the subject is inconsistent. However, linear fit is not a measure of the judge's consistency unless all consistent behaviour has been captured by the model.

Although it may seem as if judgement making is a complicated, non-linear process, in practice, a linear model approximates the data very well (Brehmer and Brehmer, 1988; Maule and Svenson, 1993). Slovic and Lichtenstein (1971) indicate that multiple correlations of artificial tasks tend to be in the 0.80s and 0.90s and they tend to be in the 0.70s where the cases are more complex realistic ones. Although different indices have been used such as linear fit, multiple correlation, percentage agreement or a Fisher's z value of any of these many studies since then have shown a similar pattern (see Appendix 1).

Despite the usual adequacy of linear models, a judge may be using a non-linear model. A model formed by multiple regression assumes linear additive behaviour. Stewart (1988) outlines alternative possibilities. For example, if subjects were using a conjunctive model, in which case all cues would have to have high values to elicit a high response, or a disjunctive model, in which case a high value on any of the cues would lead to a high response, the model would not be additive. Non-linear models can be created by the inclusion of curvilinear or configural (interactive) functions in the multiple regression equation or by fitting other equations (see Ganzach and Czackes, 1995; Einhorn, 1970). Where terms are simply added to the original linear equation the fit would improve in any case. Usually little is added by the inclusion of non-linear components (Hoffman, Slovic and Rorer, 1968). However, this little may be a significant amount and some subjects may have configural policies (Summers, Talioferro and Fletcher, 1970; Wiggins and Hoffman,

1968; see also Goldberg, 1968). Unusually Einhorn (1972) found a conjunctive model was the best fit for all three of his subjects). But even if the actual judgemental behaviour is non-linear it can still be well approximated by a linear equation. Rorer (1971) and Dawes (1968) apparently show that a high degree of fit can be seen when looking at the correlation between various non-linear models and the output of the linear approximation to these models (see Dawes and Corrigan, 1974). Indeed it may be that the little predictive power gained by the addition of terms to the regression function is due to overfitting of the model to the data set rather than a better capturing of the judge's policy (Hammond, Stewart, Brehmer and Steinman, 1975, Goldberg, 1968).

From Slovic and Lichtenstein (1971) it would be expected that clinical judgement analyses, consisting of complex realistic judgements would tend to have lower linear fits. In fact a number of aspects of the task have been seen to affect linear fit. For example, Slovic and Lichtenstein (1971) cite studies that show multiple correlation coefficients decrease as the correlation between cues decreases although the pattern of cue use remains much the same. Similarly they gave evidence to suggest that as the number of cues increases linear fit decreases. Einhorn (1971) suggested that as more cues are introduced, subjects may use more complex, nonlinear policies. He pointed out the distinction between complexity of mathematical descriptions of models and their cognitive difficulty. However, consistency has also been seen to decrease with more cues (Hoffman and Blanchard, 1961, cited in Slovic and Lichtenstein, 1971).

However, the study by Reilly and Doherty (1992) produced results which support only the first of these findings and do not support the second. Multiple correlation coefficients for subjects given six cues were no different to those given twelve but subjects whose cues were orthogonal had worse multiple correlation coefficients than those whose cues were representative and were thus correlated. However, Cook and Stewart's (1975) subjects do tend to have better linear fits on the three cue task than on the seven cue task. The studies cited by Slovic and Lichtenstein varied the number of cues presented between two and seven. The difference between studies that find differences and those that do not may lie in the number of cues being presented: there is a detrimental effect when six or seven cues are presented as opposed to two or three but there is no more worsening when

more cues are added beyond that. Although subjects may be relying on non-linear policies when more cues are available they may also change which cues they rely on to make decisions. When cues are intercorrelated this would not have so drastic an effect as when cues are orthogonal. Where there are more cues more different ones may be relied upon and consistency may decrease.

Number of cues used

In multiple linear regression analyses several indices have been used for the estimation of the relative importance of cues (Darlington, 1968; Stewart, 1988). This is the amount of influence a cue has on the judgement making or in other words how much the judgement or decision will change as the value of the cue changes. Indices that can be used for this are the standardised regression coefficient or beta weight (*e.g.* Tape, Heckerling, Ornato and Wigton, 1991), the 'squared validity' or the square of the correlation between a cue and the judgement (r_i^2), Darlington's (1968) usefulness index (*e.g.* Reilly and Doherty, 1992), Hoffman's relative weights (Hoffman, 1960) and Darlington (1968) also cites a measure used by Englehart (1936)[2]. When cues are intercorrelated these different indices lead to different sets of weights (Schmitt and Levine, 1977; Darlington, 1968). When cue intercorrelations are zero all these indices become equivalent (Darlington, 1968) and cues can be defined as used when the index of relative importance is significant. Although multiple regression is of value in ascertaining the relative importance of each cue in the judgement or decision making, it is not the only statistical method used in judgement analysis. The analysis of variance (ANOVA) allows easier ascertainment of cue interactions where cues are expressed categorically (*e.g.* Hoffman, Slovic and Rorer, 1968; Phelps and Shanteau, 1978). This may of course add artificiality to the task. In an ANOVA approach the relative importance of cues can also be measured (Hoffman, Slovic and Rorer, 1968). Cues showing a significant main effect can be viewed as influencing

[2] Darlington's usefulness index is a measure of the amount the fit R^2 drops if that cue is not included in the linear equation. Hoffman's relative weights ($W_i = \frac{\beta \times r}{R^2}$) which sum to 1 are similar: Hoffman refers to them as the independent contribution of each cue to the linear model. However, Darlington (1968) points out that this is only true when cues are orthogonal. Engelhart's (1936) measure of the contribution to variance of the cue and its combination with other cues is similarly criticised by Darlington.

decision making significantly, those with an insignificant main effect do not influence decision making and have not been used.

In both types of analysis where cues have a significant impact on the decision making it indicates they are being used. Brehmer and Brehmer (1988) stated the common finding that judges use few cues. The majority of studies are more interested in which cues rather than how many cues were used. Some idea of the range in numbers of cues and cases presented can be seen in Appendix 1. Many studies only present a few cues. For example, Rovner *et al* (1985) varied four cues when looking at factors affecting doctors' decisions to refer obese patients. However, even when only a few cues are presented, they are not all used (*e.g.* Holzman, Ravitch, Metheny, Rothert, Holmes and Hoppe, 1984). Where more cues are presented the number actually used is still limited. Brehmer and Brehmer (1988) report that when 64 cues were available six to nine cues were used (Roose and Doherty, 1976) and when 19 cues were available, one to six cues were used (Ullman and Doherty, 1984). In a more recent study judgements were based on between one and ten out of 19 cues available (Reilly and Doherty, 1989). An average of five (range two - seven) were used when Chaput de Saintonge, Kirwan, Evans and Crane (1988) presented rheumatologists with ten cues. In the study by Fisch, Hammond, Joyce and O'Reilly (1981) subjects used on average four out of eight standard Hamilton Depression Rating Scale cues, but used an average of five cues when they chose the eight available. In the condition where 12 correlated cues were presented, subjects in Reilly and Doherty's (1992) study used one to four cues. As stated previously, it is less clear which or how many cues have been used when cues are intercorrelated (*e.g.* Poses *et al*, 1985; Phelps and Shanteau, 1978). Altogether, even when many cues are available the average number of cues that have a significant affect on subjects' decision or judgement making can be counted on one hand with some subjects necessitating the use of a second.

This finding is interesting from the point of view of determining the limits of human information processing. Simon (1974) re-estimated the number of chunks that can be held in working memory as not seven (plus or minus two) as Miller (1956) had suggested, but nearer four or five. However, the chunk capacity of working memory does seem to decrease slightly as the size of chunks increases. Experts are supposed to have the

same working memory capacity as novices but may chunk the information differently in their domain of expertise (Simon, 1974). For example, Simon (1974) describes a study by de Groot (1978) who found that experts' memory for impossible chess game patterns was no better than novices and the positions of about six pieces could be matched up. But when the pattern was a possibility they had a considerable advantage and could reproduce the positions of 20-25 pieces. In medicine, Elstein *et al* (1990, p.10) report that successful physicians do not generate nor store more hypotheses than unsuccessful physicians do. Shanteau (1992) found that the number of cues used by experts was equal to, if not fewer than, the number used by novices. However, judgement analysis measures the cues rather than chunks that influence behaviour. It would be expected that if more expert subjects are able to group information that more cues would be found to influence their decision making than influences that of novices. However, two of the studies cited by Shanteau (1992) are judgement analyses and in the more general studies cited above the number of cues used is still limited.

One explanation for this limit to the number of cues used is that the advantage gained by experts in grouping cues which correlate in the real world may be lost when orthogonal cues are used. The chunking or pattern recognition hypothesized as being used automatically by experts may be actually based on the identification of a few key features in a correlated group. Indeed the use of forceful features in diagnosis has been shown by Gale and Marsden (1985; also Grant and Marsden, 1987). Thus where cues are correlated patterns are recognised on the basis of one or two cues, in orthogonal cases just those two cues will be used. If a representative design was used (cues were correlated as they are in real life) the actual use of only a few cues would be blurred by the apparent use of other correlating cues. The use of orthogonal cues allows this distinction.

Individual Differences

Despite the fact that subjects use relatively few cues it is a different few cues that they are using. The ability to distinguish relevant from irrelevant information, rather than the amount of information used, can be seen as the main distinction between experts and novices (Shanteau, 1992). But even within groups of experts considerable differences in

cue use is shown (see Wigton, 1988; Engel *et al*, 1990). Orthogonal cues will of course show policy differences between judges that may be of little significance where cues are intercorrelated in real life (Brehmer and Brehmer, 1988). However, if correlated cues were used these policy differences would be difficult to ascertain. There is some evidence that subjective descriptions of policy show greater agreement than tacitly calculated ones (Chaput de Saintonge and Hattersley, 1985).

Subjective models and self-insight

It was mentioned in Chapter 2 that one of the advantages of judgement analysis over other methods is the lack of reliance on what the subject is able to explicitly state about his or her judgement or decision making. Subjects' verbal reports may be unreliable when certain types of information are requested (White 1988), or may not be useful because of the nature of the task (*e.g.* the automatic processes described by Ericsson and Simon, 1980, 1984), or where the process is verbalizable much effort may be required on the part of the experimenter to develop models to fit it. Ericsson and Simon (1980, 1984) amongst others argue that certain types of knowledge about certain processes are verbalizable. The issue here is not really about what is known, although one cannot assume that which is verbalised is 'known'[3]. The main issue is that if subjects cannot verbalise their policy then they cannot communicate it to others. Therefore, in order to discover a person's policy, it is not good enough just to ask them.

In keeping with the idea that people find it difficult to verbalise their policies, several judgement analysis studies have also collected data about the subjects' assessments of the relative importance of cues. In comparison with the analysis of their tacitly held policies these have been found wanting (Slovic and Lichtenstein, 1971). There is some evidence that self-insight is affected by the correlations between cues, although perhaps surprisingly, the number of cues presented does not seem to be an influencing factor. The few studies looking at recognition of policies have also found moderate but significantly positive results (Reilly and Doherty, 1989, 1992; Wigton, personal communication). These significant results have been taken as an indication that subjects do have insight into their

[3] It may not be true.

policies (Reilly and Doherty, 1992). However, none of these studies took account of the similarity of policies being presented. See Chapter 4 for greater discussion of self-knowledge.

The Lens model and Social Judgement Theory

One type of judgement analysis, associated now with the framework of Social Judgement Theory, is lens model analysis. Egon Brunswik (1952) called his approach the Lens Model because of the similarity between a diagrammatic representation of the workings of a lens and his diagrammatic representation of the workings of human judgement in a probabilistic setting (see Figure 3.1). Although Brunswik labelled the various phenomena associated with perceptual judgement in a probabilistic setting, Hammond and others extended the model to form a framework for looking at general judgement and decision making, such as social judgement, under uncertainty (see Hursch, Hammond and Hursch 1964, Brehmer, 1988). The crux of the Lens model is to see how accurate judgements are whilst taking into account the predictability of the real phenomenon being judged. For example, it may be that with the information available a person is able to make a judgement that is correct eighty percent of the time. They may look for ways of improving their accuracy. However, analysis of the actual phenomenon being judged may show a probabilistic element: using that information it is only predictable eighty-three percent of the time in any case.

The focus is the interplay between two systems - a person's judgement and the environment[4]. Both systems are modelled, commonly using policy capturing procedures such as linear regression. The use of other modelling procedures for example fuzzy logic is also being explored (Cooksey, in preparation). Then any systematic variance, not captured by the models, is compared. The linear estimation of the judgement made (Y_S) is Y'_S, as described earlier. Similarly, if the criterion being judged is actually Y'_e, its linear estimation is $Y'_e = b_{e0} + b_{e1}X_1 + b_{e2}X_2 + b_{e3}X_3 + ... + b_{en}X_n$, where X are the cue values and b are regression coefficients.

[4] In this case the 'environment' refers to the criterion being judged or an ideal decision trying to be attained.

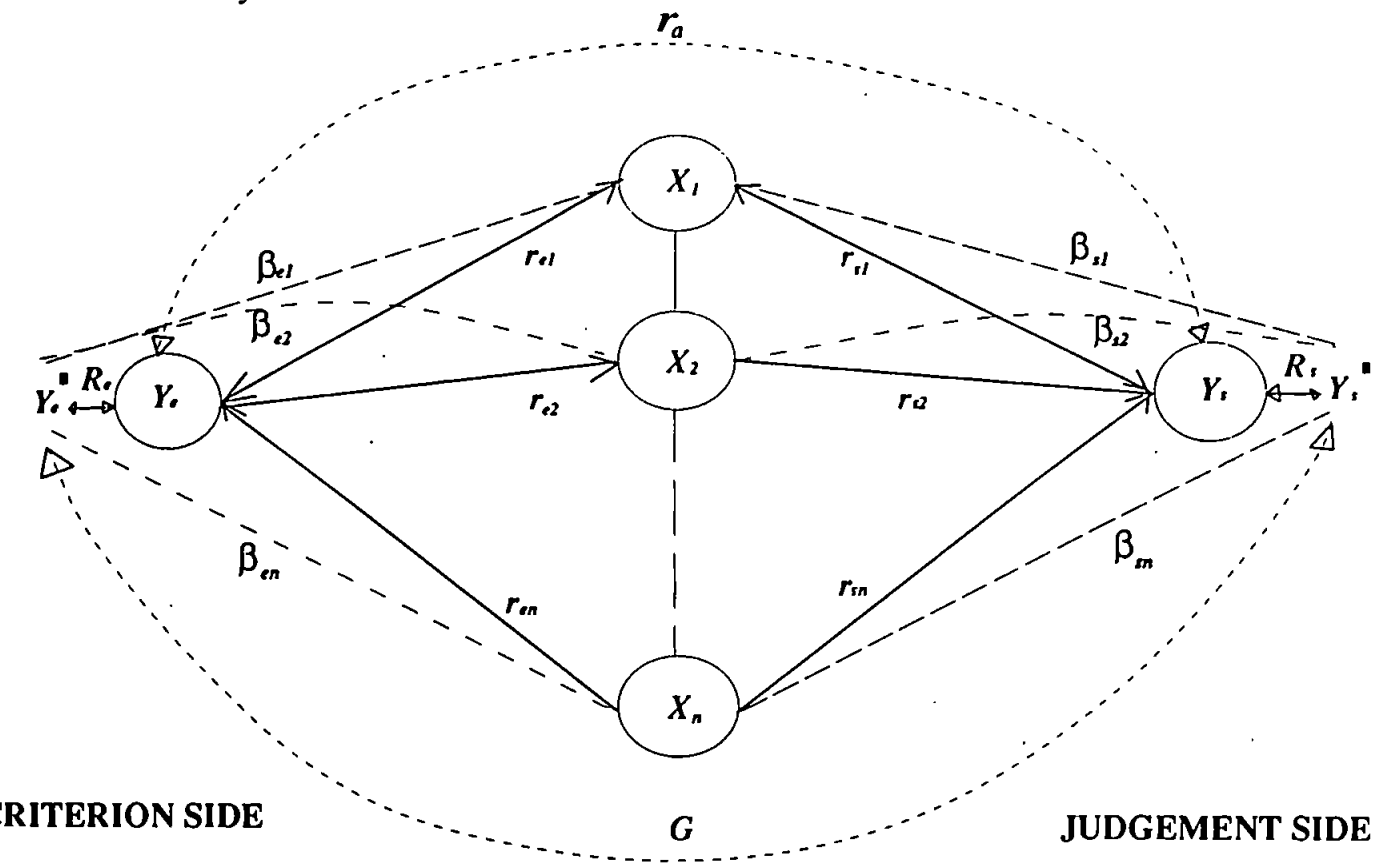


Figure 3.1 Diagram of the univariate Lens Model showing the relationship between the criterion (Y_e), the linear estimate of it (Y'_e) from the cues (X_i), the judgement (Y_S) and the linear estimate of the judgement (Y'_S) from the cues (X_i). For key to the indices, see the Lens Model Equation in the text.

The section on linear models in this chapter discussed the possibility of non-linear behaviour by the subject. The environmental criterion may also have a non-linear relationship with the cues available. So $Y_s = Y'_s + Z_s$ and $Y_e = Y'_e + Z_e$, where Z_s and Z_e are residuals each made up of a consistent non-linear component and a random error component.

The lens model equation (Hursch, Hammond and Hursch, 1964; Tucker, 1964) describes the level of accuracy of the judgement (the achievement) in terms of the relationship between the two linear models of the systems, the similarity between the non-linear systematic variance of judgement and environment and the amount both judge and environment are unpredictable.

The Lens Model Equation

$$r_a = R_s R_e G + C\sqrt{(1 - R_s^2)(1 - R_e^2)}$$

r_a **(achievement** /accuracy) = correlation between judgements (Y_s) and criterion (Y_e)

R_s **(cognitive control)**[5] = correlation between predictions from the linear model of the judge (Y'_s) and actual judgements from which it was calculated (Y_s) = coefficient of multiple correlation between cues and judgements.

R_e **(linear predictability** of criterion)[6] = correlation between predictions from the model of the environment (Y'_e) and the environmental criterion (Y_e) = coefficient of multiple correlation between cues and criterion.

G **(matching** /r_m) = correlation between predictions based on the two models (Y'_s and Y'_e).

C **(configural /non-linear matching)** = correlation between the components of the environment and the judge's behaviour that are not accounted for by the two respective

[5] R^2_s = the coefficient of multiple determination of the judgements
= the fit of model to the judgements
= the variance in the judgements explained by the model
$= \beta_{s1} r_{s1} + \beta_{s2} r_{s2} + \beta_{s3} r_{s3} + \ldots + \beta_{sn} r_{sn}$
where r_{sl} = correlation between the judgement and cue X_l = **utilization validity.**

[6] R^2_e = the coefficient of multiple determination of the criterion
= the fit of model to the criterion
= the variance in the criterion explained by the model.
$= \beta_{e1} r_{e1} + \beta_{e2} r_{e2} + \beta_{e3} r_{e3} + \ldots + \beta_{en} r_{en}$
where r_{el} = correlation between the criterion value and cue X_l = **validity.**

models (Z_s and Z_c).

Achievement can be seen to vary with different values of the lens model equation indices. For example, if environment follows a perfectly linearly modellable pattern, R_E = *1* and the level of achievement can be seen as being affected by the amount of correlation between the two models and the correlation between the judge and the model of the judge. The second component of the equation

$$(C\sqrt{(1 - R_E^2)(1 - R_S^2)})$$

tends towards zero as R_E tends towards one. Alternatively, if the judge is perfectly linearly modellable the level of achievement is affected by correlation between the two models and the linear predictability of the criterion. Again the second component has tended towards zero. If both the linear models are perfect (both R_E and R_S are both approximately one) the second component has tended towards zero but the first component has tended towards *G* or the correlation between the two perfect models. In another scenario suppose that one or both of the linear regression models does not estimate the criterion or judgement (respectively) very well. In this case R_E or R_S or both would be close to zero and the first component of the equation would be close to zero. $1\text{-}R^2_E$ or $1\text{-}R^2_S$ or both would be close to one and therefore the second component of the equation would be significant. Achievement (R_a) would be mainly affected by *C* which can be thought of as a measure of the extent to which Y_E and Y_S diverge from their respective linear models in the same manner. So, as the goodness of fit of the models varies and R_E and R_S vary between zero and one, different amounts of weight are given to *G* and *C* in determining the level of achievement.

The success or achievement of a person's judgement or decision making is defined by the relationship between their judgements or decisions and the ideal judgement or decision (the criterion). Lens model analysis has an important role to play in capturing the different components of judgement making. If a judgement is inaccurate it may be because the judge is using the wrong method or is not using any method at all, or because the thing being judged is an evasive phenomenon and no model can be found to describe it. The importance of the lens model approach is well illustrated in a study by Tape, Heckerling, Ornato and Wigton (1991) in which the accuracy of physicians estimates of pneumonia

were found to differ in Illinois, Nebraska and Virginia. However, an analysis of the environmental system found that predictability differed similarly. The information available was less good an indicator of whether the patient has pneumonia in Illinois than in Virginia or Nebraska.

The environmental criterion does not have to be known to do a Lens model analysis. An expert's judgements can be taken as the criterion. A consensus between several judges can be taken as the criterion. Another subject's judgements can also be put on the other side of the Lens model and agreement can be analysed in a similar way to achievement. Usually however, a one-sided lens model is performed. This analysis without an environmental side of the lens is judgement analysis or policy capturing. However, the assumptions of Social Judgement Theory of the probabilistic nature of the world are borne in mind. In any of these types of analyses the relative importance of different cues can be compared.

Practical applications

One of the difficulties of judgement analysis in general, as described so far, is the need for the details of the judgement or decision process to be expressed on a unidimensional scale. Many judgements and decisions are very complicated affairs and the decision options are not usually expressible in a solely scalar form. In most cases there are a number of options which may each be expressed in a continuous fashion but which may have specific discrete possibilities. Each of these options may be independent or it may be possible to chose more than one at a time. The lens model described so far is the univariate one: it allows analysis of decisions or judgements that can be expressed on one continuous scale. Cooksey and Freebody (1985) also describe a multivariate lens model that permits the analysis of "complex human inference tasks". The multivariate analysis relies on finding the largest canonical correlation between the judgement and the environment (standardised cue weights are used). Canonical weight vectors are obtained and then the multivariate achievement is decomposed into the various lens model indices. An alternative is to express the judgement or decision in a continuous scalar form.

Judgement analysis is obviously useful in identifying a subject's policy in

circumstances in which it is difficult for them to express the process involved: where generalisations rather than specific cases are being discussed. For example, an expert's policy can be elicited just as they perform a task in everyday life. Real or hypothetical cases can be used. If hypothetical cases are used comparisons can be made between different subjects since performance on the same cases can be compared. This might be useful even where subjective descriptions of policy are available. Subjective policies elicited have tended to show greater agreement than was shown by tacit policies (*e.g.* Chaput de Saintonge and Hattersley, 1985; Kirwan, Chaput de Saintonge, Joyce and Currey, 1983c). If real cases are used and some other measure of the criterion is available or the ideal decision can be calculated a lens model analysis can show the exact source of disagreement or lack of achievement. A second use of the tacit policy might be in eliminating a subject's random errors through bootstrapping (discussed earlier in the chapter) or using a computer run algorithm of the policy to make judgements.

Although subjects' policy descriptions have been unreliable this should not be taken as proof that they do not know the policy they are using (see Chapter 4). However, a third increasingly realistic use of judgement analysis is in training through cognitive feedback. This has been found to change subjects' performance on tasks although again this does not prove whether they are receiving new information about their own policies or not. The phenomenon of cognitive feedback has been much explored in judgement analysis, especially in the context of medical judgement and decision making (Wigton, 1987; Poses, Cebul Wigton and Collins, 1986; Tape, Kripal and Wigton, 1989; Tape, Kripal and Wigton, 1992; Wigton, Patil and Hoellerich, 1986; Wigton, Poses, Collins and Cebul, 1990). Information about the accuracy or success of decision making or judgement making is rarely received. Feedback is only normally gained (if at all) on positive decisions made. For example, the employer never finds out if a person rejected for an post would have been successful at it, an adoption panel never finds out if the family they did not allow to adopt a child would have brought it up successfully. The doctor never finds out if a treatment option not plumped for would have been beneficial. A dissatisfied patient may simply go to another doctor rather than risk the worsening of their complaint. A dead patient will obviously not return and give the clinic feedback. A doctor may not find out about the

affects of his or her decisions or the accuracy of his or her judgements for several reasons.

However, there is evidence to show that even if we do have information about the correct judgement or decision (outcome feedback) it does not lead to any change in behaviour. A number of authors have shown that outcome feedback (giving the correct results of each judgement), which is not feedback in the strict sense at all, makes no difference to judgement making (Steinman, 1974; Doherty and Balzer, 1988 and Balzer, Doherty and O' Connor, 1989). Perhaps biases such as hindsight bias and confirmation bias add to our lack of awareness of any shortcomings of our judgements: Things we definitely know we feel would have been likely to happen. When our hypothesis is confirmed or our judgement proves right we notice (see Evans, 1989).

Applications of cognitive feedback have been increasing that have lead to greater specification as to which parts of feedback lead to an improvement in performance (Balzer, Doherty and O' Connor, 1989) or greater agreement between subjects (Luckett and Hirst, 1989; Chaput de Saintonge and Hattersley, 1985 and Kirwan *et al*, 1983b). Cognitive feedback can include any of *cognitive information* about the relationship between judgements made and the cues available (or the subject's policy), *task information* about the relationship between the ideal judgement and the cues available (or the ideal policy) and *functional validity information* about the relationship between judgements and criterion values and between the model of the criterion and the model of judgements. The lens model, which includes all of these types of information is clearly of use here. Performance is improved because of increased matching of ideal and subject policies rather than an increase in consistency (Balzer, Doherty and O' Connor, 1989). However, a more recent study has demonstrated that although task information can lead to dramatic changes in behaviour, the addition of cognitive information and functional validity information gains nothing (Balzer, Sulsky, Hammer and Sumner, 1992), see Chapter 4.

Any of these applications would be difficult without a computerised version of the lens model or judgement analysis such as POLICY PC (Executive Decision Services Inc., Albany, NY). Use of computers for training in medicine has been slowly increasing (Frisse, 1992). Greatest success in the introduction of information technology and computer decision aids seems to lie within the realms of general practice. General

Practitioners are increasingly used to using computers. In an area where intuition is rife the possibility of learning to modify behaviour on the basis of instant, computer presented feedback generated within the framework of Social Judgement Theory seems a real one.

Conclusions

This chapter gave an overview of the applications of judgement analysis and SJT to medicine. The probabilistic nature of the world means that the best decision does not necessarily lead to the best outcome. In medicine in particular the SJT approach is useful. It refrains from focussing on the decision making process or aspects of the decision itself and instead focuses on the information available to make the decision and on how this shapes the decision made. Aspects of judgement analysis study design and their implications for findings were discussed. Interpretation of results of studies presented in later chapters will be made with these design variables in mind. In Study 1 (Chapter 5) three different types of GP decision making are analysed using judgement analysis techniques. This amounts to an analysis on one side of a lens model analysis: No ideal or criterion decision is identified. However, in Study 4 (Chapter 8) GPs' judgements are analysed in a (two sided) lens model analysis. This takes into account the probabilistic relationship between pieces of information in the real world. It is this complex probabilistic environment that characterises the uncertainty of medical decision making.

Chapter Four
Self-knowledge

Introduction

This chapter reviews the literature on self-knowledge. Judgement analysis, the method used in studies reported throughout this thesis, is an unusual method of investigation of judgement or decision making in that there need be no reliance on the subject's interpretation of their own behaviour. As was seen in the brief review of these methods in Chapter 2, both process tracing techniques and decision analysis approaches often rely on subjects' stated interpretations of their own cognitive processes (process tracing) or of the decision making situation. Where methods other than verbal protocol analysis are used in process tracing much reliance is then placed on the experimenter's analysis of the situation. As this chapter will show, a considerable amount of research questions the validity of explicit statements about our self-knowledge. Thus the avoidance of reliance on metacognition is an important aspect of judgement analysis. There is a second benefit of a review of the literature on self-knowledge at this point, apart from justification of the use of judgement analysis. One question running through the thesis is whether subjects' statements about their behaviour could be relied upon at any level. Here the pattern of self-knowledge seen in the literature is reviewed.

The "unknown" known

The issue of self-knowledge appears in a number of different areas of research. For example, research into implicit learning (*e.g.* Reber, 1989; Berry and Dienes, 1993), perception without awareness (*e.g.* Merikle, 1992), expertise and automatic processing (*e.g.* discussion in Cheng, 1985 and Schneider and Shiffrin, 1985), reasoning and biases (*e.g.* Evans, 1989 chapter 5), problem solving and insight (*e.g.* Schooler, Ohlsson and Brooks, 1993), memory, judgement and decision making (*e.g.* Reilly and Doherty 1992), attribution, emotions and attitudes (*e.g.* Nisbett and Ross, 1980 chapter 9) plus a whole host of work on introspection and verbal reports (*e.g.* Ericsson and Simon, 1980, 1984) have all involved discussion of self-awareness or self-knowledge.

Many researchers have put forward dichotomous theories about knowledge based on the statable and therefore explicitly known, *versus* the unstatable (un)known. The idea of a distinction between conscious and unconscious processing has become part of our culture: it is generally accepted that there are things about ourselves and our mental processes of which we may not be aware. The division point of the dichotomy has varied from author to author. Berry and Dienes (1993) distinguish between implicit and explicit knowledge and between implicit and explicit learning. Implicit learning differs from explicit learning in that the environment is 'understood' without (or despite) the use of explicitly held conscious strategies. Explicit knowledge is accessible to consciousness and is communicable. Implicit knowledge is knowledge that the subject can be shown to have in that it is manifested in their behaviour but is less accessible to consciousness and communication.

Other authors have distinguished between automatic and controlled processing. With practice a behaviour or process that previously required conscious effort and was controlled becomes automatic and unconscious. It has been argued (Cheng, 1985) that this automatization is often a task specific change of tactic rather than a more general transfer to a different processing domain. Schneider and Shiffrin (1985) argue that recategorisation does not explain all the findings and that some tasks are automatised. The idea is still there that the process becomes less accessible; less conscious steps are involved. Certain behaviours become automatised with practise. If expertise develops with practice automatization may be one of the features of expertise. Policy capturing tasks focus on experienced subjects since novices have no policy to be captured. If the process has become automatised or rationalised, as happens with practice, self-insight may be reduced.

In Ericsson and Simon's (1984) discussion of the appropriateness of verbal protocols to analyse underlying mental processes, such automatised processes are one type they identify as being unreportable. The other type of mental process for which verbal reports are redundant are recognition processes. Research on expertise has moved away from searching for different ways of thinking and has realised the importance of the knowledge base in expert behaviour (see Gilhooly, 1990; Gilhooly and Simpson, 1992). As mentioned in the previous chapter, experts and novices do not differ in the amount of

information they tend to use. With this emphasis on the importance of domain knowledge, expert behaviour has been likened to pattern or schema recognition (Gilhooly, 1990; Gilhooly and Simpson, 1992). Patel and colleagues have also suggested that rather than a hypothetico-deductive process physicians tend to use something more akin to pattern recognition, or production rules in making diagnoses in their domain of expertise (see Patel and Groen, 1986, Groen and Patel, 1985). Concurrent verbal reports can however give insight into the information that is attended to (from which thought processes can then be inferred) in other situations.

Evans (1984) makes a distinction between heuristic and analytic reasoning. However, this is not a conscious-unconscious distinction of the type previously discussed. He argues that although the bias-invoking heuristics are preconscious, and can be compared with Ericsson and Simon's recognition processes, some of the analytic process that may also cause (more rectifiable) biases may be preconscious. Analytic processes then may be of the automatised or controlled type or may be the implicitly learned or explicitly learned type (see Evans, 1989 for a discussion).

Self-knowledge: introspection vs. self-hypothesising

According to Flanagan (1991), Dennett (1982) has made a distinction between autophenomenology or comment on a psychological system from the inside and heterophenomenology or comment on a psychological system from the outside. Theoretically, an individual is in a position to give both of these types of comment: she can observe her own behaviour and make hypothetical judgements about it or she can introspect and describe the phenomena of which she is aware. Some combination of the two may also occur. Even those that argue that much of self-knowledge is based on hypothesizing, such as Nisbett and Ross (1980 chapter 9) argue that there is still some information to which the individual subject may have privileged access. Self-knowledge would be shown when the individual's description of themselves or their behaviour is equal to or better than that given by an outsider with full knowledge of the environment and behaviour.

From a practical point of view, where self-knowledge has been shown to be correct,

it doesn't really matter whether it has been arrived at through introspection or through hypothesizing with good knowledge of the situation. However, the method of elicitation does matter if it is to be taken as representative of self-knowledge. Self-reports have frequently been found to be erroneous (*e.g.* Nisbett and Wilson, 1977). If there are these two types of self-knowledge possible (direct and inferred), some testing methods may encourage hypothesis testing even where direct introspection is possible.

Types of self-knowledge: Separate phenomena vs. levels

To have self-knowledge is to know (things about) oneself as a psychological system *e.g.* what you know, how you know it, how you came to a conclusion, how you feel *etc.* Although all examine self-knowledge in a broad sense, research areas such as these mentioned above focus on different phenomena. It might be argued that these separate phenomena actually represent different levels of self-knowledge. Flanagan (1991) distinguishes five forms of self-knowledge ranging over basic awareness of the existence of one's thoughts and mind (Simple Cartesianism), of one's mental state, of the content of intentional states, of the causes of those states, and of the functioning of internal mental processes. With this structure of breakdown, to have a certain form of self-knowledge might imply that you have certain other forms. For example, it is difficult to conceive of having access to the contents of one's mental states without having knowledge of the mental states and knowledge of the existence of one's own mind.

Difficulty arises if an assumption is made that subjects have a certain level of self-knowledge and subsequent levels are being tested. Of the different types of self-knowledge, some can be seen to fall into levels and are accessible through 'introspection' *i.e.* there are some of which one can be directly aware. This might be referred to as phenomenal knowledge. Other types of self-knowledge one cannot be directly aware of but can only hypothesize about. Causal self-knowledge and process self-knowledge are of this latter type.

How to test self-knowledge

Generally speaking the distinction between the things of which we are aware and unaware is measured in terms of our ability to state them. If knowledge is then shown in some other way this is taken to mean that we were not aware of what we knew. In other words a subject's self-knowledge is measured in terms of how well their elicited description of their behaviour or reasoning fits the observed facts. If, for example, causal reports are inaccurate in terms of covariation observed by the experimenter the assumption is that the subject's self-insight was erroneous.

There are problems associated with taking any sort of verbal report as a measure of what a subject knows consciously (White, 1988; Evans, 1989). Not only is an 'incorrect' account not proof of lack of self-knowledge; but a 'correct' account is no proof of the existence of self-knowledge. Firstly, what people say may not be what they know or are aware of i.e. they may deliberately give a false account. Secondly, the relevant facts as perceived by the experimenter may not be the relevant facts as perceived by the subject. For example, it may be the case that subjects give a 'false' account of their behaviour because they were not aware of the manipulations of the experimenter or because they interpret things in different ways from the experimenter. Thirdly, even if experimenter and subject accounts match, this is no proof of knowledge rather than similar hypothesizing: It may be that the subject is 'rationalising' about his or her own behaviour in the same way as the experimenter is.

Retrospective and concurrent verbalisations each have their pros and cons. Ericsson and Simon (1980,1984) argued that although retrospective reports may be open to error from forgetting and subsequent hypothesizing, concurrent verbal reports may be more accurate. They argue that concurrent reporting of behaviour may slow but not cognitively alter performance and can therefore be a useful source of data.

The crux of the difference seems to be that where concurrent verbalisations are useful 'as sources of data' the type of self-knowledge elicited is of a phenomenological nature[1] . Concurrent verbalisation allows comment on which pieces of information are being attended to at which time - information that may be forgotten before retrospective

[1] The phenomena consciously experienced are described and no interpretation is given.

reports. Ericsson and Simon (1980) were frank about the limits on the use of verbal reports to get at what lies in 'short term memory'. This sort of introspection is only of use if the processes being described are more than one-step processes, unlike recognition and automatic processes (including implicitly learned processes). They were wary of the influence of the instructions given to the subjects which should not draw their attention to things that would otherwise not be attended to. Kellogg (1982) also argued that thinking about one's mental processes has value in yielding information about hypothesis testing but not about automatic processes. However, this argument is tantamount to saying that where processes are automatic and unconscious (as measured by their verbalisability) introspection (verbal description of mental processes) yields little. Similarly where processes are verbalisable, verbalisation is useful.

Asking subjects to verbalise self-knowledge other than just the focus of attention may lead to alterations of the patterns of behaviour that would not be found with retrospective verbalisation. For example, Wilson and Schooler (1991) found naive judges, asked to analyse why they made judgements (study one) or to evaluate the attributes used and reasons for judgements (study two), agreed less with experts than control subjects who just made judgements. However, Schooler, Ohlsson and Brooks (1993; Experiment 3) also found that *non-directive* concurrent verbalisation impaired results in insight[2] as opposed to non-insight problem solving. Insight problem solving would however, be classified by Ericsson and Simon as unsuitable for concurrent verbalisation because it is a 'one-step' process.

Thinking aloud seems to come easier to some subjects than to others and elicits information of varying richness. For example, Elstein, Holzman, Belzer and Ellis (1992) obtained verbalisations from subjects making judgements about the likelihood of prescribing a steroid. Transcripts varied in length from one sentence to three pages. There is no guarantee that all pieces of information being attended to can be verbalised.

Concurrent verbal reports will focus on the case being undertaken rather than generalising across any range of cases. Subjects giving retrospective accounts are in a better position to make judgements about their behaviour over several cases. Causal and

[2] Insight problem solving is characterised by a sudden realisation of the answer.

process accounts of information processing falls into this category: They are not phenomenological so much as interpretative descriptions of behaviour.

Knowledge of causes and processes

Research on self-insight into *causes* of behaviour and judgements has typically found that subjects' reports are inaccurate. Classic studies by Nisbett and Wilson (1977) and also Wilson and Nisbett (1978) are summarised by Nisbett and Ross (1980) as finding that effects of influential factors are missed and effects of non-influential factors are fabricated in causal reports. Nisbett and Wilson concluded that reports were based on "a priori, implicit causal theories, or judgements about the extent to which a particular stimulus is a plausible cause of a given response".

A number of criticisms have been made of these studies. Ten years after the publications, White (1988) reviewed many of the subsequent criticisms such as the questionable relationship between verbal reports and self-knowledge. White discusses the distinction between the processes whose accessibility is under question, and their accessible contents and products that are described in phenomenological reports. Obviously the *accessibility* of content and process is not a valid method of distinguishing between them when accessibility is precisely what is then being measured.

White refers to the work of Sabini and Silver (1981) who make an analogy between human and machine information processing. A distinction can be made between the program specifying conditionals, the 'trace' or relationships between specific inputs and outputs. White's point is that the trace can be observed without being able to specify conditional relationships. Similarly, the programme can be known and *conditional relationships specified* without knowledge of the trace. It may be that Nisbett and Wilson were giving one description of the trace or relationships between variables, whereas the subjects were giving a different type of description.

The point that is important here is that there is more than one way to skin a cat: There is more than one way to discuss information processing. One way is by describing the conditional process relationship, another is by describing correlations between variables - or describing the trace in some way. Another is by describing the contents or

subject of the trace or the inputs and outputs. It seems that causal descriptions of behaviour and process descriptions of behaviour are not different levels of self-knowledge rather than separate phenomena. To be able to describe the process does not imply an ability to describe the trace, nor the trace-specific inputs and outputs. To be able to describe the trace does not imply description of the process.

What this amounts to is that there may be different ways of explicitly knowing. If asked for one sort of explicit knowledge hypothesizing may be necessary to develop the currently held explicit knowledge or awareness. Rich (1979) proposes that content answers a "What?" question and process answers a "How?" question. Hixon and Swann (1993, experiment 3) found that the wording of the question affected whether or not an increase in time to 'self-reflect' led to an increase in 'self-insight' by subjects with negative self-concepts. Given a 'why' focus, subjects rated the more favourable descriptions of a pair of self-concepts as more accurate whether they had 45 or 10 seconds to evaluate a pair of personality descriptions supposing to represent themselves. Given a 'what' focus, subjects given longer periods (45 seconds) tended to rate the less favourable descriptions of their personality as more accurate, whereas subjects given less time (10 seconds) rated the more favourable descriptions as more accurate. Hixon and Swann were hypothesizing that differences in the instructions given to subjects was the reason for the different findings by studies looking into changes in self-insight with increased time to self-reflect. The explanation given by Hixon and Swann is that the 'what' focus facilitated introspection, the 'why' focus disrupted it. Another explanation is that the 'what' focus allowed introspection of the type that fitted with the evaluation being given whereas the 'why' focus led to a different type of introspection or 'self-hypothesising'.

Nisbett and Wilson (1977) and Nisbett and Ross (1980) argue that there cannot be direct 'introspection' of the type being discussed: At this causal (and process for that matter) level, only theories or hypotheses could be produced about either the behaviour of self or others and that the only difference between hypothesizing about self and hypothesizing about others is the source of the basic information. Different types of explicit knowledge may be constructed as there are different theoretical ways of structuring information processing.

Self-insight in Judgement Analysis

"Self-insight" was the term used by Reilly and Doherty (1989) to refer to subjects' causal knowledge of the relative importance of various pieces of information in their judgement making.

Judgement analysis (JA), as the name suggests is an analysis of how a subject makes judgements. To recap from Chapter 3, in a typical judgement analysis, an objective model of the subject's policy, made up of indices of relative importance for each cue available, is calculated using multiple linear regression. Hoffman (1960) pointed out that this method gives a statistical "paramorphic" representation of the behaviour rather than describing the psychological process of decision or judgement making. The description is at a somewhat causal level: co-variation of cues and judgements is described without indication of the 'program' underlying that co-variation. Cues are identified that play a part in the process but *exactly* how they play that part is left open.

The subject's elicited model, whether linear and directly comparable to the objective model, or non-linear, has consistently been found to be inferior to the objective, linear model (Slovic and Lichtenstein, 1971; Reilly and Doherty, 1992). The degree of this inferiority varies considerably between subjects within studies: Some individuals show good self-insight (Chaput de Saintonge and Hattersley, 1985; Ullman and Doherty, 1984). Interpretation of the measure of self-insight is also difficult. For example, the same correlation may be a sign of self-insight because it is significantly non-zero or a sign of poor insight because it is sub-optimal (Reilly and Doherty, 1992). The use of different indices makes comparisons of self-insight across studies difficult. The fit (R^2) of subjective and objective models of the data may be compared (*e.g.* Cook and Stewart, 1975). Fits of subjective models may be compared to those of equal weight or random weight models (*e.g.* Kirwan, Chaput de Saintonge, Joyce, Holmes and Currey, 1986; Fisch, Hammond, Joyce and O'Reilly 1981). Multiple correlation coefficients (R and R_s[3]) of two models may be compared (*e.g.* Reilly and Doherty, 1989). Subjective cue weights may be directly

[3] Here, R refers to the correlation between predicted judgements from an objective model and actual judgements. R_S refers to the correlation between predicted judgements from a subjective model and actual judgements. R^2 gives a measure of the amount of variance explained by the model in question.

correlated with objective weights indicating relative importance (*e.g.* Reilly and Doherty, 1992). Predictions from the two models may also be correlated. With these between study and between individual differences Brehmer and Brehmer (1988) were reluctant to draw conclusions about the *degree* of self-insight demonstrated in JA studies. However, bearing in mind that average self-insight is generally found to be significant but sub-optimal it seems fair to describe it as moderate.

Subjective weights, comparable to linearly based objective relative importance weights, have been elicited in a number of different ways: Hoffman (1960) (and also apparently Martin (1957) according to Slovic and Lichtenstein, 1971) originally proposed the now widespread method of dividing up 100 points between all the cues available (*e.g.* Kirwan, Chaput de Saintonge, Joyce, Holmes and Currey, 1986; Rothert, 1982; Ullman, Egan, Fielder, Jurenec, Pliske, Thompson and Doherty, 1981; Chaput de Saintonge, Kirwan, Evans and Crane, 1988; Mear and Firth, 1987; Reilly and Doherty, 1992; Summers, Talioferro and Fletcher, 1970; Cook and Stewart, 1975). Sections of pie charts have also been used as a method of allocation (Fisch *et al*, 1981). Subjects have rated cues to indicate their relative importance (Reilly and Doherty, 1989; Cook and Stewart, 1975). Roose and Doherty (1976) changed these subjective ratings into ranks. Paired comparisons, ratio techniques and judgements of the number of times a cue was influential over fifty judgements have also been used (Cook and Stewart, 1975).

With these subjective weights, average self-insight is moderate no matter how it is measured. As expected, when subjective weights are put into a linear equation the degree of variation in judgements explained by the model is less than that explained by the line of best fit (the objective model). However, this is even true when cross-validating: The degree of variation explained by subjective models is less than that explained by objective models when both are used to predict judgements on a new data set (Roose and Doherty, 1978; Reilly and Doherty, 1992; Reilly and Doherty, 1989). As the studies in Table 4.1 illustrate, the average degree of variation explained by subjective models has varied ($0.34 < R^2_S < 0.76$). Clinical Judgement Analysis (CJA) studies typically show lower R^2_S than studies involving other subjects. One reason for looking at the degree of variation explained by the subjective model in the context of that explained by the objective model is that it may be

that the amount of variation that *can* be explained is limited. Linear fit decreases when a subject is being inconsistent in their decision making or when using a non-linear policy. When the 21 sets of data from studies in Table 4.1 giving both linear fit and subjective model fit are plotted in Figure 4.1, increased fit of the data by subjective models tends to coincide with an increased linear describability of data anyway. Worse performance by subjective linear models seems to be related to worse performance by any linear model.

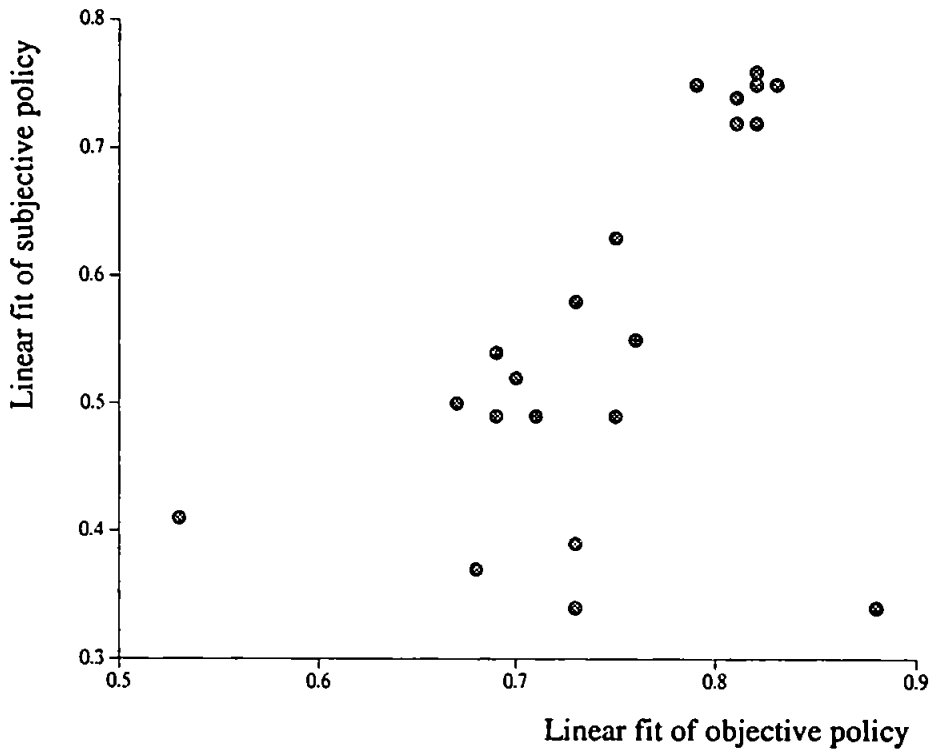


Figure 4.1 Plot of average or median fit of subjective linear model to judgement data (Rs^2) with average or median linear fit of data (R^2). Data from 21 studies marked in bold on Table 4.1.

Greater insight is gained as to why subjective weights give a sub-optimal description of data by comparing subjective and objective weights directly. Table 4.1 shows that average correlations between objective and subjective weights for sets of data vary between r = 0.2 (Reilly and Doherty, 1989) and Fisher's z = 0.92 (Reilly and Doherty, 1992). Subjective weights given over cues show relatively flatter distributions, with virtually no zero weights being assigned (Ullman and Doherty, 1984; Mear and Firth, 1987; Reilly and Doherty, 1989). Subjects generally report using more cues than they are shown to use from objective models (Ullman and Doherty, 1984; Slovic and Lichtenstein, 1971). For example, Summers *et al* (1970) show that although only 26% of subjects in their sample actually used all four cues available 98% reported doing so. Slovic and

Lichtenstein (1971) point out that subjects tend to overestimate the importance of minor cues and underestimate the importance of a few major cues.

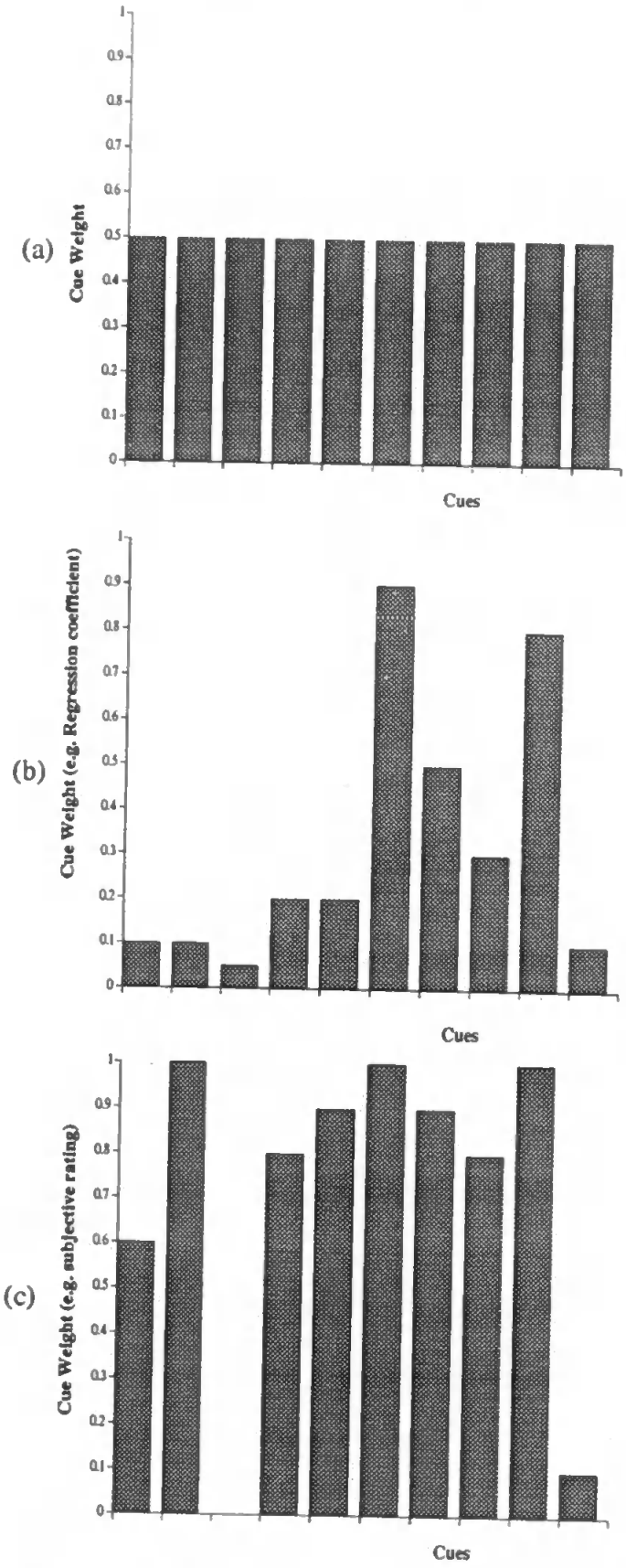


Figure 4.2 Hypothetical graphs showing (a) an equal distribution of cue weights, (b) a peaked distribution, typical of actual cue use and (c) a flatter distribution of cue weights typical of subjective ratings.

Table 4.1 Examples of Measurement of Self-Insight in JA

Reference	r	R_S^2	R	R^2	R^2_E	Cues	Cases	S	Measure	Judgement
Chaput de Saintonge & Hattersley 1985				0.8 - 1.0		24	50	6	I → decn rules	Antibiotics for OM
Chaput de Saintonge *et al* 1988				> 0.8 (15φ)		10	30	48φ	100	Changes in R.A.
Cook & Stewart 1975		0.75		0.79	See	3	50	21	100	Financial aid applications (FA)
		0.58		0.73	text	7	50	21	100	Graduate school applications. (GA)
		0.76		0.82	"	3	50	19	Rate (1-100)	FA
		0.49		0.69	"	7	50	19	Rate (1-100)	GA
		0.72		0.82	"	3	50	17	Paired comp.	FA
		0.49		0.71	"	7	50	17	Paired comp.	GA
		0.74		0.81	"	3	50	29	Ratio	FA
		0.54		0.69	"	7	50	29	Ratio	GA
		0.75		0.82	"	3	50	21	Times influential	FA Faculty members and graduate st.
		0.55		0.76	"	7	50	21	Times influential	GA
		0.72		0.81	"	3	50	16	ICR I + 100	FA
		0.50		0.67	"	7	50	16	ICR I + 100	GA
		0.75		0.83	"	3	50	20	ICR II	FA
		0.52		0.70	"	7	50	20	ICR II	GA
Elstein *et al* 1992						2	12		I + P. A. →d. rules	HRT prescription
Fisch *et al* 1981				0.6		8	49	15φ		Depression assess and prescribe
		0.49		0.75	0.25	8			16 seg. pie	Cues chosen by subjects
Holzman *et al* 1984						4	24	24φ	Qu.	HRT
Kirwan *et al* 1986		0.39		0.73	0.41	10	50	89φ	100	R.A. assessment
		0.34		0.88				4	I → prod. rules	
Mear & Firth 1987	0.30	0.34		0.73	0.31	9	30	38	100	F.A. risk & return
Reilly & Doherty 1989	0.20		0.7			19	115	40	Rate (1-9) before	Job assessment by students
	0.44	$R_S = 0.34$	0.7						& after task	
Reilly & Doherty 1992	0.43		0.73			6 orth	100 (70)	19	100	Room mate assessment
	0.66		0.84			6 rep	100 (70)	21	100	by students
	0.69		0.73			12 orth	100 (70)	19	100	Cross validated Obj & subj wts
	0.92		0.81			12 rep	100 (70)	18	100	on 30 cases.

Table 4.1 continued on next page. Key on next page.

Table 4.1 continued.

Reference	r	R_S^2	R	R^2	R^2_E	Cues	Cases	S	Measure	Judgement
Roose & Doherty 1976			0.75			66 (→9)	360 (200)	16	Rate → rank	Salesmen selection
Rothert 1982		0.37		0.68		4	27	30 φ	100	Compliance with hypertensive
		0.41		0.53		4	27	30 pats	100	regimen judgement
Rothert *et al* 1984				0.75		4	24	45	I	Endocrine disorder + obesity
Ullman *et al* 1981	0.7*	0.63*		0.75*		19	80	16	100	Ψ judgements of hyperactivity
Ullman & Doherty 1984				0.67		7	52		100 comp UI	Ψ judgements of hyperactivity
				0.66 (75%)		19	80		100 comp UI	Ψ judgements of hyperactivity
Summers *et al* 1970		$R_S = 0.6$*		0.75*		4	175	131	100 + P.A.	Nation assessment by students

Key:- r = average correlation between statistical and subjective cue weights, R_S^2 = average linear fit of equations using subjective weights, R = average coefficient of multiple correlation of the statistical weights, R^2 = average linear fit of the regression equation, Mean R^2_E = average variance explained by a model of equal cue weights, Cues = number of variables available, Cases = number of cases over which the statistical weights were calculated, S = the number of subjects, Measure = method used to assess insight (I = interview, 100 = distn 100 points, pie = allocation of segments of a pie diagram, P.A. = protocol analysis, Qu. = questionnaire, rate = assign each cue a number (options in brackets), Times influential = number of cases on which cue was influential [turned to proportion], paired comp. = pairs of cues rated for equality of importance cue ratings summed over all pairings, ratio = moderate cue assigned 100 points and other cues assigned ratios of that to indicate their relative importance, ICR I = most likely judgements of each value of a cue taken separately, ICR II = non-linear indication of cue use: each level of each cue given an importance rating between 1 and 9 regardless of the value of other cues), Judgement = type of judgment or decision being measured, * = median value; φ = physicians, ψ = psychologists, pat = patients; orth = orthogonal cues, rep = realistically correlated cues.

Although subjective weights tend to form a less sharply peaked graph than objective weights do (see Figure 4.2), their distribution still tends to be closer to that of objective weights than an equal distribution of weights would be (Mear and Firth, 1987). Studies comparing the degree of variation in judgements explained by subjective models with that explained by equal weight models generally show subjective weight models to be superior (Mear and Firth, 1987; Fisch *et al*, 1981; Cook and Stewart, 1975). This need not indicate any self-knowledge at a causal or process level but may just indicate an awareness of which cues were and were not looked at. One study found that subjective models were actually worse than equal weight models for 49 out of 89 subjects (Kirwan *et al*, 1986).

Subjective distributions tend generally to be flat and indicate the use of what are insignificant cues. Differences in self-knowledge shown may actually relate more to the way the subject used the cues rather than a special self-knowledge on their part. For example, if all cues that were attended to played an equal part in the judgement, then a subject giving almost equal weights to all cues to which he or she attended would apparently show good insight. Adelman, Sticha and Donnell (1984) found that differences in peakedness of the ideal distribution of cue importance and use on both a five cue and a nine cue task led to different distributions of subjective and multiple regression based weights.

It may be that the peakedness of the objective distribution of cue use is what affects apparent variance in self-knowledge with changing numbers of available cues. As the number of cues available increases it may be that unused cues of which the subject is aware are added. If subjects generally tend to give flat distributions of subjective weights, in tasks where more of the cues available would be used insight should be better than where just one or two of the cues stand out as being important. If several unused cues are available on a task worse insight would be shown. Two studies using changing numbers of cues apparently show differing results: Cook and Stewart showed an decrease in self-insight on a seven cue task as opposed to a three cue task whereas Reilly and Doherty (1992) showed an increase in self-insight in subjects with twelve rather than six cues available. One explanation for this discrepancy is that there was a dramatic increase in the number of insignificant cues between the six and twelve cue tasks. Indeed, Reilly and

Doherty comment that there were a smaller proportion of zeros in the six cue task. Many fewer cues may have been unused on the three and seven cue tasks and indeed the insight shown is unusually high.

A number of methods have been used to elicit non-linear subjective policies. However, the same pattern of limited self-knowledge emerges as is seen with linear subjective models. In interviews leading to a series of decision rules (Chaput de Saintonge and Hattersley, 1985); in protocol analyses and in questionnaires subjects have indicated using cues that have been found to be statistically unimportant (Holzman, Ravitch, Metheny, Rothert, Holmes and Hoppe, 1984). Average fits of these other sorts of subjective models have also consistently been found to be worse than linear fits of the same data (*e.g.* Kirwan *et al*, 1986). However, studies have differed as to whether there is an improvement (Summers *et al*, 1970) in prediction from using non-linear as opposed to linear subjective models or not (Cook and Stewart, 1975).

Other evidence for self-insight in JA

There are two areas of research within judgement analysis that might challenge the idea that subjects only have moderate self-insight. Two studies by Reilly and Doherty (1989, 1992) have shown that subjects were able to pick out their own policies from all other subjects' objective and then subjective policies at much greater than chance rates. These results were also shown in the clinical field by Wigton (personal communication). However, because of the confounding factor of the similarity between policies the sub-optimal but significantly positive results shown here do not allow any further conclusions than subjective ratings did.

Secondly the effects of cognitive feedback, discussed in Chapter 3, indicate that when subjects are given information about the ideal policy their performance can be changed, although the addition of information about their own policy makes no difference to this improvement. In a discussion on the psychological validity of the different indices of relative importance, Schmitt and Levine (1977) suggest comparing change of behaviour on feedback of policy using different indices. They also suggest training subjects to make judgements that conform to their subjective weights. But the fact that cognitive

information (CI) does nothing to change the subject's behaviour is important. It is cognitive information that subjects are asked to give as subjective policies and it is this that they do not do well in. Although unable to express it in terms of subjective weights, one explanation is that subjects already 'know' the information being given to them in terms of objective weights. Subjective ratings then, as interpreted by the experimenter, are not a measure of what the subject actually knows. Again this is not conclusive evidence of knowledge: It may be that performance can be improved with task information simply because the attention paid to cues can be changed or some other strategy used. No understanding is needed as to the way weights describe contingent connections. The subject need not have any great self-knowledge other than of the cues they attend to. In accordance with this, Luckett and Hirst (1989) found that although agreement with the official policy weighting system was improved with feedback, there was no relationship between degree of self-insight and feedback. Groups of subjects received no feedback, outcome feedback, task information or both task information and outcome feedback. However, Luckett and Hirst credit the similarity between groups to the generally good self-insight shown by all groups (mean correlation between subjective and objective weights = 0.78, 0.76, 0.86, and 0.80 for the different groups used).

General judgements of contingent or correlational relationships

The fact is that in general, never mind with reference to one's own behaviour, humans have difficulty in judging correlations (see Baron, 1988 chapter 14; Nisbett and Ross, 1980 chapter 5). The scales used to express relationships are at best non-linear functions of correlation coefficients. When "theory free" data is presented subjects' perception of covariation is insensitive: correlations of 0.2 and 0.4 are not perceived at all. Correlations of 0.6 and 0.8 are rated as being small. Subjects are quite good at identifying correlations around 0 and around ± 1, but have difficulty quantifying moderate correlations and in differentiating low correlations. However, more important than this problem of scaling is the systematic biases that we are prone to. These can be classified into two types that are really fundamentally related.

Subjects show an attentional bias whereby it seems to be the proportion of times

two dichotomous variables are both positive that subjects use to measure their relationship. At best account is additionally taken of one or both of the positive negative cells. Account is rarely taken of when both are negative which is also necessary to calculate correlation. This bias occurs whether the data are arranged in a two by two contingency table or are perceived sequentially as they are experienced in real life (Nisbett and Ross, 1980 chapter 5). Prior beliefs can also be said to have an effect on the perception of correlation. Nisbett and Ross present several studies in which subjects, including practising clinicians, perceive the covariations they expect to be in the data and do not perceive counter intuitive covariations that actually do exist. Subjects' statements about relations between variables are based on prior expectations or hypotheses. Although data driven estimates of correlation might be percieved as conservative, theory driven estimates (where the material is meaningful in the real world) are not.

The illusion of control can be linked to the attentional bias: subjects think they have control over a task because often what they want to happen happens when they behave in a certain way. However, this is also related to the effect of prior beliefs on our judgements. Subjects tend to find the correlations they think will be there not only through attending to the co-occurences that they are expecting but also through selective interpretation of evidence (Baron, 1988). Examples can be found where subjects both choose evidence to test particular hypotheses and ignore evidence against them (Baron, 1988).

Conclusions

Although the general conclusion from JA studies is that subjects have a significant, moderate but sub-optimal, degree of self-insight when measured by eliciting subjective weights or through policy recognition, the negligible effect of feedback of Cognitive information (CI) on policy change must indicate that either subjects are receiving no new knowledge, or that they are not using this new knowledge (see earlier in this chapter and also Chapter 3). If the former is the case then the apparent lack of insight shown by elicited subjective weights must be a feature of report rather than knowledge. There are different possible reasons for the problem of self-knowledge lying at the level of report. It may be the form of report that is problematic. It may be experimenter's

that is problematic. Again it may be that generally verbal reports should not be equated with conscious knowledge or awareness (White, 1988). If the latter is the case: new knowledge is not being used it may be because it is not meaningful to the subject in terms of actually carrying out the task. If subjects are not using the new knowledge they have aquired from cognitive feedback then one explanation lies in the discrepancy between explicit and implicit knowledge.

It appears from all this that, when making judgements about causes of one's own behaviour, humans are prone to the same problems as when making judgements about the behaviour of others. The problem may be a scaling one similar to the problems shown in general by subjects making assessments of relationships between variables. For example, Doherty and Balzer (1988) argue that the lack of self-insight shown in studies eliciting subjective weights may well be misleading. They comment that the construct validity of subjective weights has not been established and reiterate that subjective-objective weight correspondence is imperfect but not absent. Brehmer and Brehmer (1988) also comment that the apparent lack of complete self-insight may occur at the point of interpretation of the investigator. They cite as evidence a study by Ekegren (1983) who found that subjects could communicate with fellow subjects about their policies in a way that allowed reproduction of judgements. Both of these challenges to the use of subjective reports are feasible. However, in a study on implicit learning by Mathews, Buss, Stanley, Blanchard-Fields, Cho and Druhan (1989) described by Berry and Dienes (1993 p. 41) yoked subjects following the explicit descriptions of implicit learners had greater than chance and improving performance. But this was always worse than that of their instructors. The suggestion by Brehmer and Brehmer also seems questionable given that some subjects show good insight given the same form of reporting and interpretation as others.

It may be that the little added in descriptions of judgement making when using a non-linear as opposed to linear model is significant when trying to distinguish between good and moderate self-insight. It may be that individuals have used cues in a non-linear way to different degrees and so appear to have different degrees of self-insight when linear comparisons are made. However, if this were the case although weights of relative importance do not match up, non-linear models of judgements should be able to account

for more variance than their linear counterparts. The few studies using non-linear models do not show that this is the case.

The inability to describe causal or process aspects of judgement behaviour may have a parallel in research in implicit learning. The ability to express rules explicitly lags behind but may eventually match implicit learning of those rules. Some researchers have suggested that explicitly held mental models of the situation are gradually formed as more trials are completed. Evidence for this comes from McGeorge and Burton (1989) who developed computer simulations from subjects' instructions for others (described in Berry and Dienes, 1993 p.24). Several subjects were able to describe their learnt behaviour in a modellable way. Explicit knowledge is gradually built up on the basis of implicitly learnt behaviour. This explicit knowledge is used to help subsequent learning. The suggestion is that these explicit verbalisations encapsulate the subject's developing mental model. Causal or process descriptions of judgement making may be the explicitly held hypotheses. Hypotheses will be based on salient features of the task in hand and may be affected by the way subjects have discussed or explicitly thought about the topic before. In medicine the way a hypothesis is expressed about a type of case will be influenced by how that case has been explicitly learnt about. But a doctor's intuitive behaviour may have been learnt more or less implicitly in the hospital.

A comparison can be made with the biases caused by prior beliefs about correlations (see earlier). It may be that the cues subjects are rating are those that they believe affect their decisions, regardless of whether they have done on this occasion. For example, in judgement analysis, reports may be based on something with which the subject is more familiar generally such as real life parameters or ideal textbook models or may be based on salient features of the task which might be those to which they attended. Some studies seem to provide fuel for the idea that subjects are giving prior belief driven hypotheses. For example, Chaput de Saintonge and Hattersley (1985) found the expressed policies of their doctor subjects suggested agreement that was not apparent in models of their diagnostic policies. Goldstein and Mitzel (1992) also found that although subjects used relative importance ratings to infer preferences in particular data sets they could give more general relative importances. However, subjective ratings will vary if collected

before or after a task and so are not simply biased by real life parameters or general hypotheses (Reilly and Doherty, 1989).

In line with differences in types of self-knowledge, causal hypotheses may be based on phenomenological experience. For example, Reilly and Doherty (1992) suggested that subjective weights might be reflective of the focus of the subject's attention. The subject has a certain level of self-knowledge in that they are aware of the features of the task that they paid attention to but not how these had a bearing on the decision. Attentional focus may be influenced by the theory held about what should affect decisions, regardless of whether or not that is how decisions were affected.

To summarise, it appears that certain types of description, namely causal and process descriptions, of behaviour must necessarily be based on hypotheses rather than on direct awareness of behaviour. The degree of self-knowledge shown when forming these hypotheses increases as greater familiarity with the task is shown: implicit learners and novices to tasks were able to show some level of self-knowledge. However, too great a familiarity with a task and explicit learning of it perhaps leads to a reliance on explicitly held knowledge that interferes with assessment of current behaviour.

Chapter Five
Study One: Capturing GPs' Prescribing Policies

Introduction

The first study in this project is somewhat explorative. Judgement analysis techniques, outlined in Chapter 3, were used to capture General Practitioners' patient management decision making policies on different tasks. Two main areas were of interest, apart from whether GPs' policies could be captured. These were the number of cues subjects used on the different tasks and subjects' ability to state the relative importance of cues.

One of the obvious benefits of judgement analysis is that it does not rely on subjects' descriptions of their own judgement making behaviour. There are circumstances in which analysing verbal protocols can be of use but the ascertainment of the relative impact of cues is not one of them (see Chapter 4). The circumstances in which protocol analysis may be of use is if the spiel is concurrent. This in itself may alter the process of capturing a general practitioner's decision making process since, in a normal encounter, communication plays a large part. Where a think aloud method is used, either with video presented or written material, the protocol may focus on aspects of the one or two cases presented rather than on the general decision making situation. Although this enables a clear account of the thought processes for that case, comparison is lacking and generalisations cannot be made about behaviour. At best, think aloud methods will elicit which cues are attended to during the judgement or decision making and perhaps in what way they were influential. From this the experimenter can map out the process, and work out the influence of cues over a number of cases, in a process tracing technique. Subjects' stated judgement or decision making policies give poor descriptions in comparison to those derived statistically (see Chapter 4). Slovic and Lichtenstein's (1971) review and subsequent studies report our inability to state the relative importance of cues and our tendency to both underestimate the importance of influential cues and overestimate the importance of uninfluential cues.

Both previous policy capturing studies and clinical problem solving or decision

making studies have failed to give subjects any range of tasks (Elstein, Shulman and Sprafka, 1990; Brehmer and Brehmer, 1988). The focus has often been on differences between experts and sub-experts (*e.g.* Patel, Groen and Arocha, 1990). Individual and expert-novice differences have been found to depend on differences in knowledge structure that are task dependent. Inter-task comparisons are obviously important in capturing other aspects of expertise (Elstein *et al*, 1990). Other researchers have looked at more than one type of judgement made on the same case (*e.g.* Rovner, Rothert, Holmes, Ravitch, Holzman, and Elstein, 1985; Fisch, Hammond, Joyce, and O'Reilly, 1981). The judgements are usually made at the same time and may influence each other. Comparisons between different types of decision making must usually be made across studies and therefore across subjects.

In General Practice, which is characterised by the sheer range of possible complaints that may appear, the presentation of more than one type of task might be considered to be particularly important (Essex, 1985, p. 183). This first study examines the ability to capture policies in three areas of decision making. These areas are related only in that they are all within the realm of the General Practitioner and they involve decisions on the prescription of prophylactics: lipid lowering treatment, prophylaxis for migraine and Hormone Replacement Therapy (HRT).

Lipid lowering drugs are prescribed to reduce the cholesterol or fat levels in the blood and reduce the progression of atherosclerosis, myocardial infarctions (heart attacks) and of Coronary Heart Disease (CHD) in general in those with raised blood cholesterol levels (Heller, 1987; British National Formulary, September 1994, pp.106-107). If the treatment proved satisfactory *i.e.* side affects were bearable and cholesterol levels were reduced, the patient would take the medication, daily, for a long time. Prophylactics can be prescribed to prevent migraine headaches. These are chronic headaches often accompanied by blurred vision and nausea and can be quite disabling (MacGregor, 1993). Their frequency and duration varies from person to person although women are two to three times more likely to have migraine attacks than men (Grant, 1992). Again tablets are taken every day but even if successful the need for treatment is reviewed every six months (British National Formulary, September 1994, p. 192). Hormone Replacement Therapy

(HRT) adjusts the hormone levels in a menopausal female to relieve symptoms associated with menopause and assist prevention of osteoporosis (Grant, 1992, pp. 329-332). Administration may take several forms and the initial period of treatment where there is an intact uterus is about a year (British National Formulary, September 1994, p. 286).

The three tasks differ in certain ways. Firstly, they differ as to the amount of agreement shown explicitly in their use. The use of lipid lowering drugs has been under dispute for some time. The evidence is only now coming out in favour of their use (*e.g.* Oliver, 1987; Marmot, 1994) and there is considerable variation in the availability of lipid clinics (Laker, Reckless, Betteridge, Durrington, Miller, Nicholls, Shepherd, and Thompson, 1991). In stark contrast, use of prophylaxis for migraine is not subject to disputes in the national medical press and use of HRT, although not being pushed by every General Practitioner in Britain, only tends to be discussed in terms of the pros and cons of its different forms. It was thought therefore that greater disagreement, in judgement and policy would be shown on the LIPID task.

Secondly, the tasks differ in their effects on the patient. Although all three types of treatment are preventative to some extent, use of a lipid lowering drug will not make any immediate difference to the quality of the patient's life. Prophylaxis for migraine will however prevent previously endured migraine headaches and HRT will alleviate menopausal symptoms as well as being beneficial in preventing osteoporosis and reducing the risk of heart problems. As a consequence of this it might be expected that the tasks would differ in terms of the amount GPs were influenced by the patient's attitude to treatment: Doctors might be more influenced by patients' requests for treatment where the patient will actually feel the result and will obtain symptom relief.

Thirdly, it was hypothesized that the tasks would differ in the amount of information that would ideally be used. This is introduced here without certainty since there was no criterion ideal on any of the tasks. Several pieces of information are reported in the medical literature as related to risk of CHD and therefore to whether to treat with lipid lowering therapy. There are a number of pieces of information that are relevant to prophylactic treatment of migraine too. However, the framing of the decisions in both of these cases were such that fewer of the independent variables presented in the MIGRAINE

task were relevant than in the LIPID task. To confirm this the number of cues explicitly thought to be of relevance were looked at and it was expected that doctors would state that more cues were important on the LIPID task than on the MIGRAINE task.

The greater number of relevant cues, the more cognitively demanding the task. If a task is more cognitively demanding less cognitive control would be expected (see Chapter 3). Cognitive control is seen in the fit of the model to the judgements. Therefore on the LIPID task, where more cues were thought to be relevant and there would therefore be less cognitive control, lower fits were expected than on the MIGRAINE task.

Despite differing numbers of relevant cues it was also expected that no more cues would actually be used on the LIPID task than were used on any of the other tasks and that the number used would be limited. Other studies have shown that there are limits to our cognitive capacity and that subjects tend to take into account around 5 or 6 cues (see Chapter 3). For example, experts have been found to use no more cues than novices (Shanteau, 1992).

The number of cues that are relevant to a decision is important in that it affects the peakedness of the distribution of importance weights. This phenomenon is discussed in Adelman, Sticha and Donnell (1984) (see Chapter 3 and Figure 4.2 in Chapter 4). Where several cues are important the pattern will be generally flatter than where fewer cues are important. In this case the pattern of cue weights for MIGRAINE, where fewer cues were thought to be relevant, should be more peaked than that for LIPID. This is the expected pattern of explicitly generated weights as discussed earlier. However, if as hypothesized the number of cues actually used on the tasks did not differ, the peakedness of the objective importance weights of cues should be similar between tasks. This leads to a prediction about self-insight on these two different tasks. Distributions of subjective weights have been found to be flatter generally than those of objective weights (see Chapter 4). The expected explicit pattern of subjective weights on the LIPID task is expected to be flatter than that on the MIGRAINE task since more cues are thought to be relevant. But tacit patterns of cue use are expected to be the same on the two tasks. Two things could happen here: If subjects have no self-insight there is more likely to be a match between explicit and tacit policies if more cues are indicated as being used on the explicit

policy. In this case self-insight will appear to be better on the LIPID task. On the other hand, if some degree of self-insight is shown and the cues used on the tacit policy are a subset of those stated as being used in the explicit policy then self-insight will be worse where there is more of a discrepancy between explicit and tacit policies. In this case self-insight would appear to be better on the MIGRAINE task. The explicit policies of the LIPID task would be relatively flatter, whereas objective weights would be equally peaked.

A number of findings are expected: In line with other studies it was expected that policies on the three tasks would be able to be captured using multiple linear regression and that cue use would be limited and be much the same between tasks. It was expected that the number of cues thought to be influential in explicit policies would differ between tasks. If this is the case it was thought that cognitive control (linear fits) would be less on tasks where more information was thought to be relevant. Self-insight was expected to be suboptimal in general: the importance of uninfluential cues would be overestimated and that of influential cues would be underestimated. Differences in self-insight between tasks would depend on the general degree of self-insight being shown.

Method

Subject recruitment

Recruitment of subjects for Studies 1-3 (Chapters 5, 6 and 7) was done within recruitment for a larger project consisting of these plus two additional studies[1]. A notice about the project was put in the national medical press. Then 202 doctors, in the PL postcode region[2] of Devon, were contacted by post. They were asked to respond, through a prepaid envelope or by telephoning, if they required more information about the project. Eighty-nine doctors (44%) replied affirmatively and they were sent further details in which they were asked to contact us again if they wished members of the project team to visit them to discuss the project or if they wished to participate in one or all of the studies of the

[1] A PCFC funded project looking at Medical Decision Making in Primary Care comprising of a total of 5 studies. There were 7 people participating in this project. Two of these, myself included are Research Students. One member is a General Practitioner and lecturer for the RCGP. The remaining four members of the team are permanent staff in the Department of Psychology at the University of Plymouth and supervised the work of the Research Students. See the Author's Declaration at the beginning this thesis.

[2] In our sample of 35 subjects the doctors who were furthest north of Plymouth were in Tavistock and the furthest east of Plymouth was in Kingsbridge.

project. All doctors who had been sent further information but who had not replied were additionally contacted by telephone and were asked whether or not they wished to participate in the project. Thirty-five of the 89 doctors sent further information participated in this study (39%).

Doctors were given a financial incentive to participate in the study in the form of monetary payment at a rate of £25 *per* hour for every hour spent on the tasks. They were also promised eventual feedback of findings which might provide them with a better understanding of their own decision making.

For anonymity all participating doctors were allocated project specific codes. These were used in all studies.

Subjects

Thirty-five general practitioners, recruited as part of a larger project, participated in this study. Thirty-two of these doctors participated in a follow up test of consistency ten months (range 8 to 13 months) after the original tasks[3].

The average age of the participating doctors was 39 (ranging from 31 to 55, one unknown). The average period since qualifying was 15.4 years (ranging from 7 to 31 years) and the average period of work in general practice was 9.3 years (ranging from 6 months to 27 years). There were four females in the sample and 31 males. Doctors were recruited from different types of practice setting including health centres and one single-handed practice and the range of practices included those in the city centre and in villages outside Plymouth.

Tasks and instructions

There were three possible computer tasks for each doctor to complete, each requiring judgements about 130 separate cases presented in turn. After each task the doctor was asked a series of questions aimed at establishing what they felt had been affecting their own decision making on that task, and how it might relate to anything they do in real life.

[3] Two doctors withdrew due to increased work loads (GP8 and GP35). One withdrew for other reasons not related to the project (GP10).

These tasks will be referred to as the original tasks and this study as the original study. An average of 10 months later (range 8-13 months) doctors were tested with a precise repetition of the first 30 cases of each computer task to see how consistent their judgements of decisions were. Table 5.1 shows which original tasks and consistency tasks were completed by each doctor.

Table 5.1 Tasks completed by each doctor

GP code	LIPID	Consistency	MIGRAINE	Consistency	HRT	Consistency
GP1	Y	Y	Y	Y	-	-
GP2	Y	Y	Y	Y	-	-
GP3	Y	Y	Y	Y	-	-
GP4	Y	Y	Y	Y	-	-
GP5	Y	Y	Y	Y	-	-
GP6	Y	Y	Y	Y	-	-
GP7	Y	Y	Y	Y	-	-
GP8	Y	-	Y	-	-	-
GP9	Y	Y	Y	Y	Y	Y
GP10	Y	-	Y	-	-	-
GP11	Y	Y	Y	Y	-	-
GP12	Y	Y	Y	Y	-	-
GP13	Y	Y	Y	Y	Y	Y
GP14	Y	Y	Y	Y	-	-
GP15	Y	Y	Y	Y	-	-
GP17	Y	Y	Y	Y	Y	-
GP19	Y	Y	Y	Y	-	-
GP20	Y	Y	Y	Y	-	-
GP21	Y	Y	Y	Y	-	-
GP22	Y	Y	Y[4]	Y	-	-
GP23	-[5]	-	Y	Y	-	-
GP24	Y	Y	Y	Y	Y	-
GP25	Y	-	Y	-	-	-
GP26	Y	Y	Y	Y	*	-
GP27	*[6]	-	Y	Y	Y	Y
GP28	Y	Y	Y	Y	Y	-
GP29	Y	Y	Y	Y	-	-
GP30	Y	Y	Y	Y	Y	-
GP31	Y	Y	Y	Y	Y	Y
GP32	Y	Y	Y	Y	-	-
GP33	Y	Y	Y	Y	Y	Y
GP34	Y	Y	-	-	Y	Y
GP35	Y	Y	Y	Y	Y	-
GP36	Y	Y	Y	Y	Y	-
GP37	Y	Y	Y	Y	-	-
Total	**33**	**30**	**34**	**31**	**12**	**6**

All computer tasks were presented on the same portable Acorn A4 computer and in most cases the doctor was visited in his or her practice surgery. Five of the doctors

[4] This doctor only complete the first 100 cases of this task. Analysis was based on this.
[5] This GP served as a pilot for this task.
[6] * indicates subjective policies were elicited but decision data failed to save.

participated at their home (GP15, GP17, GP19, GP21, GP24) and two doctors came to the Department of Psychology at Plymouth University to participate (GP1, GP26).

For the main tasks two hours of the doctors' time was booked in one sitting and they were told that there were up to three tasks that could be done depending on the time left after each one. It was emphasised that they should make decisions at whatever speed they felt was appropriate to the task. Most doctors were able to complete both tasks on one occasion. However, GP29 and GP37 each had a week between the first and the second task. GP25 had just over a week between the two. GP22 ran out of time and only completed 100 of 130 cases on the second task, though did answer questions about these afterwards. GP35 and GP36, volunteering to give more time and do an extra task, had about one month between the second and the third task.

The consistency tasks were run in conjunction with Studies 2 and 3 (see Chapters 6 and 7). Limits on the doctors' time meant that not all were able to complete all consistency tasks (see Table 5.1).

Each of the main tasks had an initial set of instructions followed by sequential presentation of 130 cases which differed on 13 distinct cues. The instructions indicated that the cases to be presented were hypothetical and that there were certain assumptions that could be made regarding the 'patient's' previous treatment. On all of the tasks it was made clear that the option to refer patients to a consultant was not available and the treatment decision lay in the hands of the GPs themselves. The instructions for the original tasks are shown in Appendices 2, 3 and 4.

On the LIPID task doctors were told that dietary advice had been given, the cholesterol level had been tested six months previously and where appropriate cessation of smoking had been recommended. They were told to assume that all modifications to the patient's life style that were likely to occur had been made and that the cholesterol level reflected any such changes. The doctors were asked to indicate the likelihood that they would prescribe lipid lowering drugs for each patient.

On the MIGRAINE task they were told that advice had been given on lifestyle modification in order to avoid trigger factors and it was indicated that there was a high degree of confidence that this was migraine headache. The doctors were required to

indicate the likelihood of them prescribing some prophylactic treatment.

On the HRT task doctors were told that there was no reason why the patient should already be receiving HRT, they had a normal pelvic examination and there was no other relevant medical history. It was indicated that the decision was on the likelihood of prescribing HRT in whatever particular form that would take for each patient.

Generation of items and presentation of the tasks

Table 5.2 shows the 13 cues used in each of the three tasks in the fixed order presented to subjects (left hand column) and also the range of values available. Both the cues and their ranges were selected with the assistance of a non-participating GP, who was a member of the project team. Cues were arranged such that where a general consensus of medical opinion could be identified in the literature, use of this consensus policy would lead to a *positive* regression coefficient (see Table 5.2 for directionality). Ranges were set so that more than one cue would need to be considered. For example, the blood cholesterol level was not allowed to be set at such a low level that no doctor would ever consider treating with lipid lowering drugs or so high that treatment would be automatic. This was done with the advice of Dr. John Dean, the General Practitioner on the Medical Decision Making project team (see Author's Declaration). Some of the cues (*e.g.* Age, Occupation, Gender, Attitude to treatment) were presented on more than one task so that their use could be compared between tasks. Both clinically relevant and irrelevant information was provided.

The 130 cases were created for each task thus. The range and settings which would be suitable for each of the 13 cues were allocated. These were entered into a random generation program[7] in which categorical data was represented as ordinal integer numbers. The program used to generate the cue values was separate from that which presented the task to the subjects. The generation program computed a correlation matrix between cue values after an initial generation and then repeated the process generating a number of sets of values in an attempt to minimise the maximum intercorrelation. For all tasks, the final

[7] As noted in the Author's Declaration, the majority of the software used in these studies was written by Jonathan Evans.

set of values chosen was such that the largest intercorrelation of any two was less than 0.20. In fact the maximum correlations between cues over the cases generated were 0.185, - 0.189 and 0.192 on the LIPID, MIGRAINE and HRT tasks respectively. Appendix 5, 6 and 7 show inter-cue correlations on the three tasks. Once generated, the cue values for the task were stored in a data file read by the program presenting the experiment proper.

'Occupation' was initially allocated in terms of social class. During the execution of the program this was then allocated a specific job title picked randomly by the program from a prespecified set of about ten possibilities for each class, obtained from OPCS Classification of occupations, 1980. [These are shown in Appendix 8.] Thus all doctors saw an identical set of cases except for the specified occupation.

The program which presented the task to the doctors showed one screen of instructions followed by a separate screen for each patient. In each case a set of cue labels appeared on the left hand side of the screen, in the order and form shown on the left of Table 5.2. To the right of each label was printed the particular value assigned to that case. Figures 5.1 a, b and c show an example of a case from each of the tasks.

Data recorded

Doctors were asked to indicate the *degree of likelihood* of their prescribing the treatment specified on each task. Doctors registered their responses using a mouse[8] to indicate their likelihood of prescribing. Movement of the mouse pointer was restricted within a fixed rectangle on the screen (see Figure 5.1). The rectangle was anchored with 0% at one end and 100% at the other. Likelihood was shown as a dark area drawn from 0 to the current mouse position and also by a number printed next to the rectangle. Both were updated as the mouse was moved. The initial mouse position for each case was set at 50%.

All doctors completed at least two of the tasks, with the LIPID first. Doctors completing all three tasks did the MIGRAINE task second and then the HRT task. All but one of the doctors who completed only two tasks did the LIPID and the MIGRAINE tasks: One (GP34) opted to do the HRT task rather than the MIGRAINE task.

[8] When a doctor was unfamiliar with computers brief verbal instructions on how to use the mouse were given.

Table 5.2 Cues and their ranges on the LIPID, MIGRAINE and HRT tasks

Cue	CUE WEIGHT: NEGATIVE ← → POSITIVE Range
	LIPID
CHOLESTEROL LEVEL	*6.5-8*
HYPERTENSION	*No/Yes, well controlled/Yes, poorly controlled*
AGE	*30-60*
GENDER	*Female/Male*
OCCUPATION	*I/II/IIInm/IIIm/IV/V*
EVIDENCE OF ARTERIOSCLEROSIS	*No/Yes*
SMOKES	*No/Occasionally/Regularly/Heavily*
DIABETES	*No/Yes, well controlled/Yes, poorly controlled*
COMPLIANCE WITH ADVICE ON DIET	*No/Some/Yes*
WEIGHT	*Under/Normal/Over/Obese/Very obese*
ATTITUDE TO TREATMENT	*Opposed/Cautious/Open to advice/Requesting treatment*
FAMILY HISTORY I.H.D.	*No/2nd degree relative/1st degree relative*
PERSONALITY	*Co-operative/Passive/Demanding*
	MIGRAINE
DURATION OF ATTACK	*1-24*
FREQUENCY OF ATTACK	*About every six months/About every three months/ Monthly/Fortnightly/Weekly*
AGE	*30-60*
GENDER	*Female/Male*
OCCUPATION	*I/II/IIInm/IIIm/IV/V*
MISSES WORK	*Yes/No*
SMOKES	*No/Occasionally/Regularly/Heavily*
NAUSEA	*Absent/Mild/Severe/Disabling*
VISUAL DISTURBANCE	*Absent/Mild/Severe/Disabling*
WEIGHT	*Under/Normal/Over/Obese/Very obese*
ATTITUDE TO TREATMENT	*Opposed/Cautious/Open to advice/Requesting treatment*
RESPONSE TO ACUTE TREATMENT	*Good/Some response/No response*
PERSONALITY	*Co-operative/Passive/Demanding*
	HRT
MENSTRUATION	*No periods/Infrequent periods/Normal periods/Irregular periods*
HOT FLUSHES	*No/Yes*
AGE	*40-55*
OCCUPATION	*I/II/IIInm/IIIm/IV/V*
MOOD STATES	*Normal/Somewhat irritable/Highly irritable*
SMOKES	*No/Occasionally/Regularly/Heavily*
LIBIDO	*Normal/Some loss/Major loss*
VAGINAL DRYNESS	*No/Yes*
WEIGHT	*Under/Normal/Over/Obese/Very obese*
ATTITUDE TO TREATMENT	*Opposed/Cautious/Open to advice/Requesting treatment*
FAMILY HISTORY I.H.D.	*No/2nd degree relative/1st degree relative*
FAMILY HISTORY B.C.	*No/2nd degree relative/1st degree relative*
PERSONALITY	*Co-operative/Passive/Demanding*

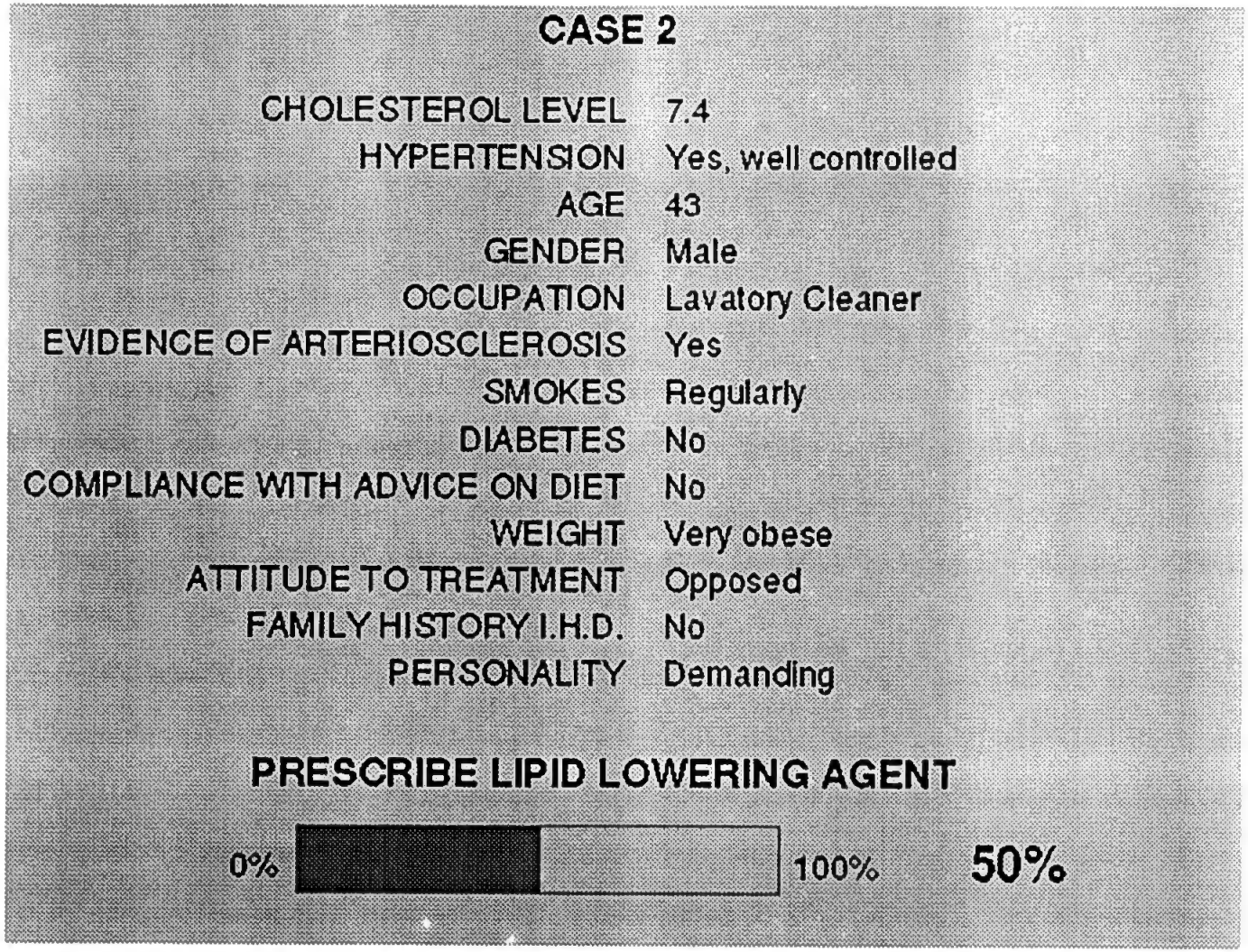


Figure 5.1 (a) Example of a case from the LIPID task

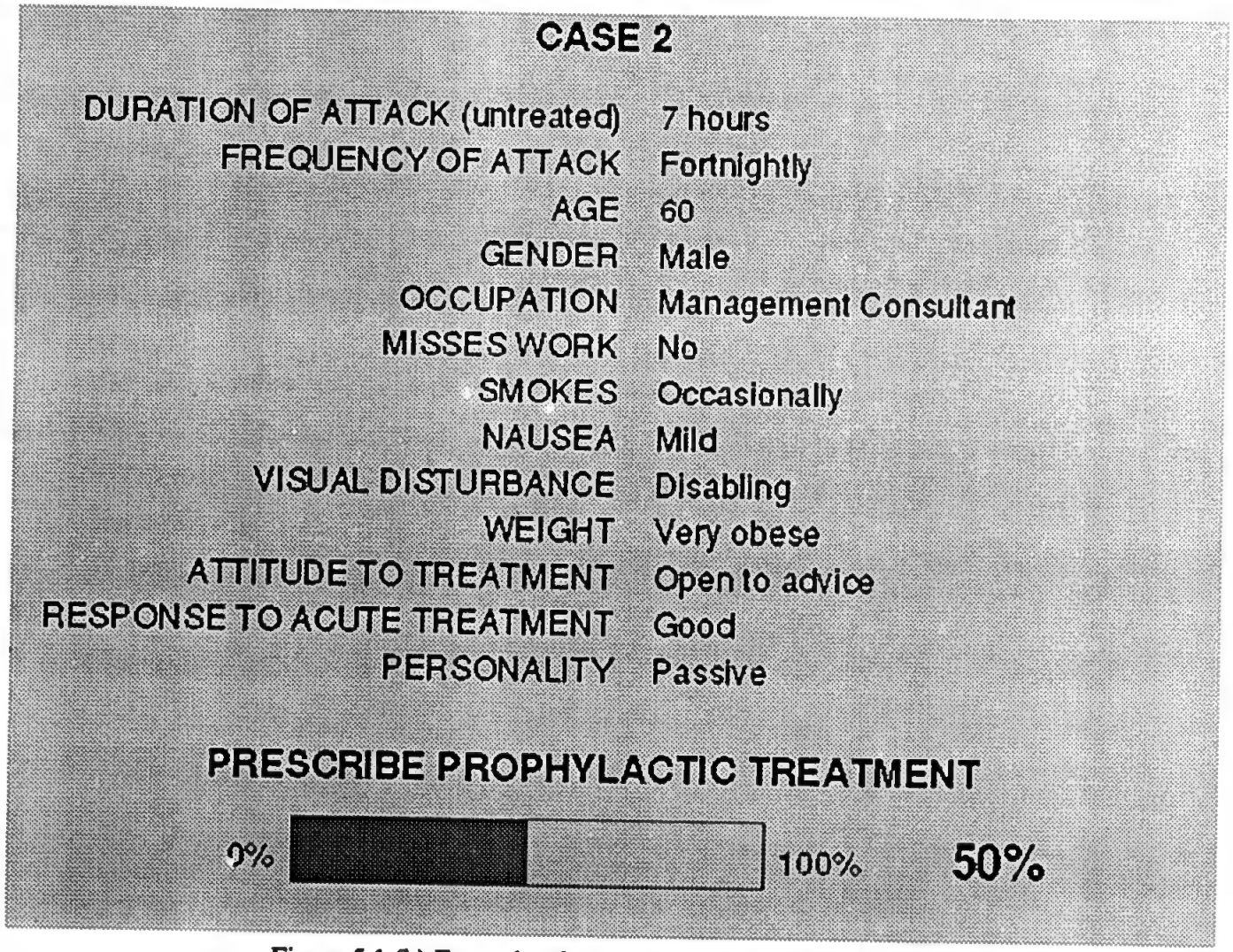


Figure 5.1 (b) Example of a case from the MIGRAINE task

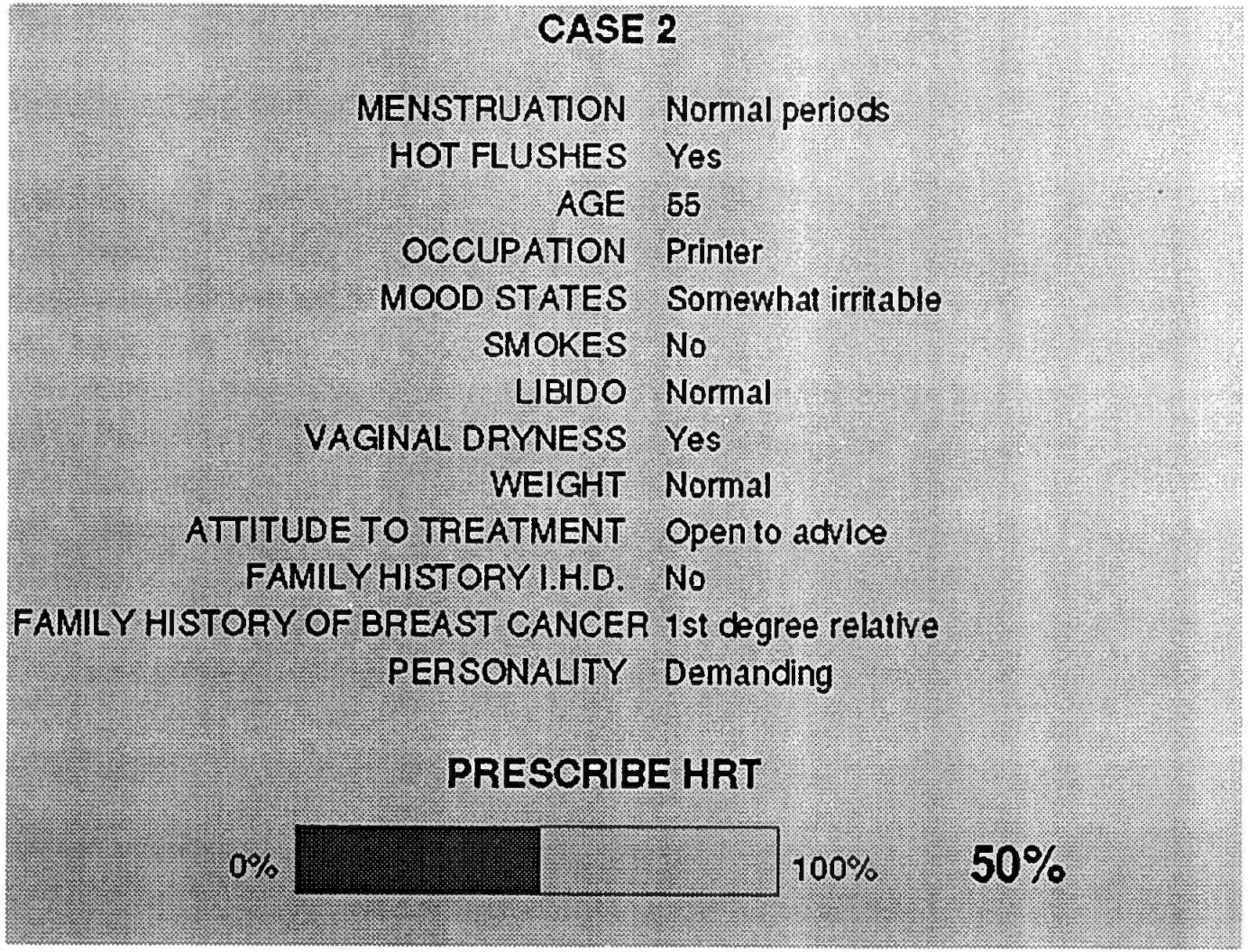


Figure 5.1 (c) Example of a case from the HRT task

Missing data

For two doctors, data was only saved for two tasks even though three were completed. Data for the third task of GP26 failed to be saved because the program was full (an error not allowed to happen again) and data from the first task undertaken by GP27 also failed to save. GP22 ran out of time on the MIGRAINE task and only completed 100 cases. Analysis was done on these. Where a doctor made a mistake or skipped over a case and brought this to the attention of the experimenter the decision and latency data for this case were not included in the analysis. Latencies of cases during which a doctor was interrupted were also excluded from the analysis.

Pilot

The tasks were piloted on both the team GP (John Dean) and another doctor not participating in the study and some adjustment of range was made as a result of this. Adjustment was also made to the instructions for the LIPID task after the first doctor of the sample had completed the task and as a result the data for GP23 for the LIPID task is not included.

Self-insight measure and post task interview

After completing each task the doctor was given a list of the cues that were available on the task and was asked about how they had affected his or her decisions over the 130 cases (see Appendix 9). Firstly an effort was made to quantify the doctor's perception of each cue's relative importance to the decision. They were presented with values at either end of a cue's range. They were asked which of the two would be more likely to make them prescribe, other things being; or if there was no difference between them. They were then asked to give each cue a rating between 0 and 10 to indicate how much of a bearing it had on their decision. The doctor was told that '0' indicated that that dimension had no bearing on the decision, '10' indicated that it had maximum bearing on the decision and that he or she could allocate the same number to more than one dimension. They were also shown an anchored scale indicating these values. Secondly, having rated each dimension, the doctor was given an opportunity to discuss his or her

strategy on the task and to indicate for example whether the affect of one particular dimension was dependent on the value of other dimensions. The doctor was then prompted to talk about their behaviour, with references to cases such as these, in real life. For example, each doctor was asked specifically about how their behaviour on the task compared to that in real life and what sorts of factors would be affecting the decision to prescribe the particular type of drug in real life.

During the post-task interview values and responses were noted down on a standard sheet (Appendix 10). In addition the discussion was tape recorded.

Consistency task

Each consistency task - conducted several months after the original study - consisted of the first 30 cases of the original task in the original order. The on-screen instructions were unchanged except in reference to the number of cases. Responses were entered in the same way as on the original tasks. Cases from the LIPID task were the presented first and were immediately followed by cases from the MIGRAINE task. For doctors who had completed the HRT task in the original study, the consistency test for this followed.

Results and Discussion

There are two main topics to be discussed in this section. Since the three tasks presented predominantly different cues and were asking the judge to make different decisions the cues used will differ between tasks. These differences may be of more interest to medics than to psychologists. A few cues were presented on more than one task and their impact on the different decisions can be compared. However, in all tasks the number of cues presented was the same (13) even though more or less of these may have been relevant, and the doctor was told to do the task in their own time. Similarities and differences in performance and information handling between the tasks can be compared in terms of cognitive control and capacity. The second topic of interest is doctors' ability to state what their policy was.

Definitions

Cognitive Control and consistency:- The judgements or decisions made on cases by each doctor were regressed onto the 13 cues available to give a model of the tacit policy used by each doctor. The total amount of variance in judgement making by the doctor explained by the model can be seen in its linear fit (R^2). If this is low it may be due to two reasons: Firstly, if consistency is low as well, there is no policy to be described in terms of the regression coefficients. It may be that doctors are being inconsistent or that they simply have hardly varied their decision at all (in which case the variance to describe is near zero). Consistency here was measured as the correlation between decisions made on the first thirty cases on their original and subsequent presentation 10 months later. Secondly it may be that the judge is using a non-linear but consistent model - in which case consistency will be high.

Tacit policies:- The judgement analysis is idiographic in that each doctor's policy is formed on an individual basis from the judgements that he or she made on a task. Tacit policies are made up of sets of standardised regression coefficients. Standardised regression coefficients show the relative importance of each cue in this model. A number of other indices could have been used for this, as was discussed in Chapter 4, but these become equivalent when intercue correlations are negligible as here (see method). A standardised regression coefficient indicates the variation in the judgement that would result in a change of one standard deviation of the cue. For example a standardised regression coefficient of 0.8 indicates that for a change in the cue of one standard deviation, the judgements change by 0.8 of their standard deviation. Cues with coefficients close to zero had no significant bearing on the judgement. The definition of cue use throughout depends on the coefficient being significantly different from zero ($p < 0.05$). The sign of the coefficient indicates the way in which the cue influences the judgement. The meaning of positive or negative coefficients can be seen from Table 5.2 in the method. A positive coefficient indicates the doctor was more likely to prescribe for values of that cue towards the right of Table 5.2. A negative coefficient indicates the doctor was more likely to prescribe for values of that cue towards the left.

Explicit policies:- Subjective policies of cue weights comparable to the regression weights

were formed from the subjective ratings of relative importance of the cues given during the post-task interview. These were assigned signs depending on the value of the cue that had been indicated as most likely to be prescribed for and these were comparable to the signs of the regression weights. See Table 5.2 above. Since subjective ratings are on a different scale to standardised regression coefficients, subjective weights were calculated to be directly comparable. Subjective weights are similar to standardised regression coefficients in that when used in a linear equation, with standardised values of cues, a set of decisions with standard deviation 1 and mean 0 is the result. The calculation used for subjective weights can be seen in Appendix 11.

Agreement and Consistency: Comparisons between tasks

Table 5.3 shows the average mean response for doctors on a task and the standard deviation of this. Agreement between doctors in terms of decision response over the 130 cases was calculated for each task in terms of Kendall's W. These are shown in Table 5.3. Greatest agreement was shown on the MIGRAINE task and least on the LIPID task. Kendall's W varies between 0 (no agreement) and 1 (full agreement). Since there is disagreement on the use of lipid lowering therapy in the general medical press this result was expected. However, agreement was moderate even on the MIGRAINE task.

Table 5.3 Concordance between doctors on different tasks.

Task	N	Average response	St. dev. response	Kendall's W*	p
LIPID	33	33.6	23.0	0.25	< 0.01
MIGRAINE	34	41.5	13.9	0.56	< 0.01
HRT	12	55.9	16.6	0.37	< 0.01

* Kendall's W is a measure of concordance ranging from 0 to 1: 1 indicates that there is full agreement, 0 indicates that there is no agreement.

As discussed in Chapter 4, agreement between subjects is affected by both agreement in policies (knowledge) and consistency of utilisation of those policies (cognitive control). Consistency of utilisation of policies can be seen in two ways. One measure is the linear fit of the model of the judgement making. However, linear fit could be low because the subject's behaviour is inconsistent in itself (poor cognitive control) or it

could be low because the subject is being consistent but is using a non-linear policy. In this case another model might be more appropriate. Consistency or reliability measures the subject's ability to give the same judgement to the same case when it is presented more than once: If subjects are inconsistent in their own judgement making, between subject consistency is also likely to be low.

Table 5.4 Mean Consistency (r) and Linear fit (R^2) of models for each task

Task	N	Mean consistency[9]	St. dev. r	Mean R^2	St. dev. R^2
LIPID	33	0.35	0.23	0.42	0.16
MIGRAINE	34	0.71	0.19	0.64	0.14
HRT	12	0.30	0.08	0.47	0.14

The ability of the regression models to explain variance in doctors' judgements varies significantly between tasks, just as agreement does ($F(2,76) = 19.75$, $p < 0.01$). Table 5.4 shows the mean linear fit and mean consistency on the different tasks. The multiple correlation coefficient (transformed to Fisher's z) for doctors was significantly higher on the MIGRAINE than on the LIPID task ($t = 5.45$, $N = 32$, $p < 0.01$). The linear fit and consistency for each doctor on each task are shown in Appendices 12, 13 and 14.

Subjects' consistencies were measured as the correlation between responses on the consistency task and on the original presentation of those cases. The greater agreement between judgements and in linear fit shown on the MIGRAINE task is accompanied by significantly greater consistencies: There was a significant difference between Fisher's z transformations of consistencies on the MIGRAINE and the LIPID task ($t = 6.19$, $N = 29$, $p < 0.01$). Several doctors were inconsistent[10] in making judgements over the first 30 cases of the LIPID task. In contrast on the MIGRAINE task only one doctor was this inconsistent (GP31) and the other 30 had significant and high consistencies. So disagreement or inconsistency between doctors may be related to the degree of cognitive control (as measured by consistency) they show over their own behaviour on the task. Generally, tasks that lead to lower consistencies show poorer linear fits and have less agreement within doctors.

[9] Consistency was the correlation in responses over 30 repeated cases for 30, 31 and 6 doctors on the LIPID, MIGRAINE and HRT tasks respectively.

[10] There was no significant correlation between the judgements made on cases in the original and in consistency tasks.

Within each task there was also a positive correlation between doctors' consistencies and their linear fit. Linear fit can be thought of as a measure of consistency of use of policy. The correlation between Fisher's z transformations of consistency and the multiple correlation coefficient (R) on both the LIPID ($r = 0.403$, $p < 0.05$) and MIGRAINE ($r = 0.732$, $p < 0.05$) tasks was significant. The same correlation over very few doctors completing the consistency judgements on the HRT task was not significant ($r = 0.253$, $p > 0.05$).

Where linear fits are low there are two possible reasons (1) cognitive control or consistency is poor and (2) the subject is using a consistent but non-linear policy. Obviously since here it is a doctor's policy that we wish to capture there is concern as to whether that policy is sufficiently described by a linear model. If any doctor showed good consistency but poor fit of the model that would be reason to suppose that the policy they were implementing was non-linear. No doctor seems to show this pattern and poorer linear fits are associated with worse consistency in decision making. There is one *caveat*. Namely that when a doctor has varied insignificantly in the judgements made across cases, and the standard deviation of responses was low, both the consistency measure and the linear fit will be low. After all there is negligible variance to be explained. This seems to be the case for GP28 on the LIPID task who judged that it was unlikely that he would treat any of the cases with lipid lowering agents[11]. It might be argued that GP28 was consistent in implementing the policy of not prescribing.

Table 5.5 Correlations of R and consistency with standard deviation of responses[12]

Task	N	Correlation of consistency and St. dev.[13]	p
LIPID	30	0.188	> 0.05
MIGRAINE	31	0.189	> 0.05
HRT	6	0.653	> 0.05
	N	Correlation of R and St. dev.[14]	p
LIPID	33	0.475	< 0.05
MIGRAINE	34	0.440	< 0.05
HRT	12	-0.196	> 0.05

[11] His greatest likelihood of prescribing was 35 on the 0 to 100 scale.
[12] Correlation coefficients used as data here are transformed to Fisher's Z.
[13] The standard deviation of judgements on the 30 repeated cases.
[14] The standard deviation of all 130 cases.

The relationships between standard deviation and both linear fit and consistency on the three tasks can be seen in Table 5.5. There was no significant relationship between consistency and standard deviation of the judgements on the repeated cases on any task. However, the correlation between linear fit and standard deviation is significant and positive for both the LIPID and MIGRAINE tasks. In other words where doctors had low standard deviations (*i.e.* varied little in their judgements) they tended to have poor linear fits - after all there was little variation to be explained! However low consistencies were not simply a product of lack of variance on these tasks: Linear fits were related to poor consistency *and* to lack of variance in judgements but these did not amount to the same thing.

Although doctors did differ in consistency within each task, the correlation between consistency shown on the MIGRAINE and LIPID tasks was insignificant ($r = -0.106$, $N = 29$, $p > 0.05$). In other words doctors who were more consistent on the LIPID task were no more likely to be consistent on the MIGRAINE task than other doctors.

Although the first 30 cases were used to measure consistency, there is some evidence to show that doctors were less consistent in decision making over these cases than over the other 100 cases in each task. As Table 5.6 shows, when the first 30 cases were removed from analysis of linear fits (leaving the last 100) on all tasks both linear fits and inter-doctor agreement improved. Most improvement can be seen on the LIPID task. There was negligible change in the policies captured.

Table 5.6 Effects of inconsistency in judgement making on the first 30 cases

	Cases	Kendall's W	Mean R^2
LIPID (N = 33)	130	0.25	0.42
LIPID (N = 33)	last 100	0.28	0.49
Significance of t test of difference			($p < 0.01$)
MIGRAINE (N = 34)	130	0.56	0.64
MIGRAINE (N = 34)	last 100	0.57	0.66
Significance of t test of difference			($p < 0.01$)
HRT (N = 12)	130	0.37	0.47
HRT (N = 12)	last 100	0.38	0.52
Significance of t test of difference			($p = 0.12$)

Complexity of tasks

LIPID task appears to be harder than the other two tasks in that the consistency both within and between doctors is worse on the LIPID task than on the other two tasks. In addition to this, longer was spent on cases in the LIPID decision as can be seen on Table 5.7. A one way ANOVA shows that this between task difference in latencies was significant ($F(2,76) = 11.96$, $p < 0.01$). However, of course the longer latencies could be due to the fact that it was the first task. Similarly, as seen on Table 5.7, there were significant differences between case latencies on all three tasks. This could be because some cases were more difficult and required longer or it could be because doctors spent longer on earlier cases and speeded up as they got used to the task.

Table 5.7 Average Latencies on the LIPID, MIGRAINE and HRT tasks

Task	Mean average latency	St. Dev. of average latencies
LIPID	16.72	6.74
MIGRAINE	10.51	4.69
HRT	10.39	4.29

That the LIPID task is a more complex task also seems to be the perception of the subjects. The LIPID task differed from the MIGRAINE task in that more cues were thought to be clinically relevant to the LIPID decision than the MIGRAINE decision[15]. That this was the perception of the doctors seems to have been confirmed in the way that they have rated the relative importance of cues. Subjective ratings of cues for the three tasks are summarised on Tables 5.8, 5.9, and 5.10. All subjective ratings are shown in Appendices 15, 16, and 17. On all three tasks all cues were rated as having had an effect on decision making in that all cues had at least one non-zero subjective rating. However, on the LIPID task the smallest number of non-zero ratings that any cue had was 10 (**occupation**) whereas on the other two tasks cues had only one non-zero rating (**smoking** on the MIGRAINE task, **occupation** on the HRT task). In addition to this on the LIPID task more cues were getting higher ratings. An ANOVA indicates that there was a

[15] Cholesterol level, hypertension, old age, male gender, evidence of arteriosclerosis, smoking, poor diet, diabetes, being over-weight and having a family history of ischaemic heart disease are all known to increase the risk of ischaemic heart disease. These risks are discussed in several books e.g. Sharp (1994), Heller, Bailey, Gott, and Howes (1987).

significant difference between tasks in the number of cues being rated above 5[16] ($F(2,76) = 9.46$, $p < 0.01$). Doctors rated on average 2 more cues above 5 on the LIPID task than on the MIGRAINE task ($\mu = 1.97$, $t = 4.27$, $p < 0.01$).

Agreement of judgement policy

Agreement between doctors in terms of the cues they felt were important can also be compared. Summary of cue ratings on the three tasks are shown in Tables 5.8, 5.9 and 5.10. Greatest disagreement was again on the LIPID task. This was not just in terms of which cues were said to have influenced their decision making but also in terms of how they were said to have influenced it. On all tasks some cues were seen as important in different ways: some doctors had felt they were more likely to prescribe and had given a positive non-zero rating whilst other doctors had felt the same cue would make them less likely to prescribe and had given it a negative non-zero rating. However, on the LIPID task eight cues had both negative and positive non-zero ratings, compared to four and three on the MIGRAINE and HRT tasks. In the latter cases conflicts were the result of single deviant doctors. There was greater disagreement as to which cues were said to have influenced decision making on the LIPID task than on the MIGRAINE task: On the MIGRAINE task certain cues were said by nearly all doctors to have affected their decision making, whilst the remaining ones were generally said not to have been influential. On the LIPID task a number of cues were said by some doctors not to have had an influence and by some to have been influential.

Actual agreement between doctors' policies is seen in terms of similarity of cue use. Here cues with significant standardised regression coefficients ($p < 0.05$) were defined as used. The LIPID task appears to be a more complex task in that longer was spent on cases, greater inconsistency was seen both between and within doctors judgements making, and more cues were thought to be important. In fact no more cues were actually used on the LIPID task than on either of the other two tasks ($F(2,76) = 1.62$, $p = 0.205$) and the average number used on each task was four. The ranges were similar on each task (LIPID: 2-7; MIGRAINE 1-8; HRT 1-7). It might be argued that the cognitive constraints as to the

[16] Chosen arbitrarily.

amount of information used shown in the task would not apply in real life where information is introduced sequentially and may be reviewed several times during the consultation. Although the time-span of consultations is much longer than time spent on cases in any of these tasks, where latencies differed both between doctors as well as between tasks there was no difference in the number of cues used. The number of significant cues did not correlate significantly with doctors' average latencies on any task (LIPID: $r = 0.169$, $N = 33$, $p < 0.05$; MIGRAINE: $r = 0.265$, $N = 34$, $p < 0.05$; and HRT: $r = 0.068$, $N = 12$, $p < 0.05$). So doctors who spent longer considering cases were not using any more cues than those who spent less time. The number of cues used by each doctor on the tasks is shown along with doctors' policies in terms of cue standardised regression coefficients in Appendices 18, 19, and 20.

The average cue use on each of the tasks are shown in Tables 5.11, 5.12 and 5.13. Some disagreement in cue use policies were seen on all tasks. This would be predictable given that the number of cues used varied. However, it is possible that a core of cues tend to be used by doctors although some doctors use additional cues. This could be said to be the cases on the MIGRAINE task where four cues: **duration, frequency, attitude to treatment** and **response to acute treatment** all appear to be used much more than the others. There are also many significant inter-cue correlations of regression coefficients. These are shown in Appendix 22. Cues either tend to be used instead of each other *e.g.* **occupation** and **misses work** or as well as each other *e.g.* **duration and response to acute treatment**. Significant correlations of cue regression coefficients with the number of significant cues indicate either that as the number of cues increases there are cues that tend to be added (having small weights). Or, as the number of cues increases, the regression coefficients tend towards zero since with the use of more cues each cue has a relatively smaller weight (the correlation coefficient is negative). **Frequency** is the only cue that correlates significantly with the number of significant cues as the number of cues increases the regression coefficient of **frequency** tends to decrease. It doesn't appear to be the case that there is a core of cues being used on the other two tasks. Although it might be argued that Cholesterol, diabetes and attitude are used more than the other cues on the LIPID task. Appendices 21 and 23 show that there were few significant inter-cue correlations of

standardised regression coefficients on the LIPID task and HRT tasks.

Table 5.8 Subjective ratings LIPID task (N = 33)

Cues	+ve	zero	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	33	0	0	8.65	1.40
HYPERTENSION	27	2	4	4.24	4.93
AGE	0	8	25	-5.5	3.70
GENDER	24	9	0	3.99	3.72
OCCUPATION	3	23	7	-0.52	1.54
EVIDENCE OF ARTERIOSCLEROSIS	27	6	0	5.15	3.13
SMOKES	11	4	18	-1.91	6.05
DIABETES	27	3	3	5.05	5.14
COMPLIANCE WITH ADVICE ON DIET	21	11	1	3.68	3.75
WEIGHT	4	8	21	-3.06	4.49
ATTITUDE TO TREATMENT	26	7	7	5.02	3.45
FAMILY HISTORY I.H.D.	30	3	0	6.17	2.92
PERSONALITY	6	20	7	-0.05	2.20

Table 5.9 Subjective ratings MIGRAINE task (N = 34 doctors)

Cues	+ve	zero	-ve	Mean	St. dev.
DURATION OF ATTACK	33	1	0	6.8	2.6
FREQUENCY OF ATTACK	34	0	0	9.2	1.2
AGE	0	32	2	-0.3	1.1
GENDER	1	32	1	0.1	1.1
OCCUPATION	4	24	6	-0.2	2.7
MISSES WORK	0	4	30	-6.2	3
SMOKES	0	33	1	-0.3	1.5
NAUSEA	22	12	0	3.4	3.2
VISUAL DISTURBANCE	22	12	0	3.3	3.1
WEIGHT	0	31	3	-0.6	2.1
ATTITUDE TO TREATMENT	30	4	0	5.9	3.4
RESPONSE TO ACUTE TREATMENT	29	4	1	5.6	3.2
PERSONALITY	8	25	1	0.9	2.7

Table 5.10 Subjective ratings HRT task (N = 12 doctors)

Cues	+ve	zero	-ve	Mean	St. dev.
MENSTRUATION	1	2	9	-5.42	5.23
HOT FLUSHES	12	0	0	8.75	1.14
AGE	4	3	5	-1.33	4.91
OCCUPATION	0	11	1	-0.33	1.16
MOOD STATES	9	3	0	4.08	3.00
SMOKES	4	7	1	0.63	2.82
LIBIDO	11	1	0	5.33	2.90
VAGINAL DRYNESS	12	0	0	7.25	1.71
WEIGHT	0	7	5	-1.42	2.23
ATTITUDE TO TREATMENT	10	2	0	7.08	3.53
FAMILY HISTORY I.H.D.	6	6	0	2.17	2.69
FAMILY HISTORY B.C.	0	8	4	-1.33	2.19
PERSONALITY	3	9	0	0.88	1.86

Table 5.11 Standardised regression coefficients LIPID task (N = 33 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	31	2	0	0.38	0.17
HYPERTENSION	7	25	1	0.06	0.11
AGE	1	23	9	-0.07	0.14
GENDER	5	28	0	0.05	0.12
OCCUPATION	3	30	0	0.03	0.06
EVIDENCE OF ARTERIOSCLEROSIS	6	27	0	0.09	0.19
SMOKES	3	19	11	-0.09	0.18
DIABETES	20	11	2	0.15	0.18
COMPLIANCE WITH ADVICE ON DIET	6	25	2	0.03	0.11
WEIGHT	0	25	8	-0.07	0.13
ATTITUDE TO TREATMENT	16	17	0	0.17	0.20
FAMILY HISTORY I.H.D.	7	26	0	0.06	0.10
PERSONALITY	2	31	0	0.00	0.07

Key:- +ve, -ve = significantly positive, negative ($p < 0.05$), n.s. = not significant

Table 5.12 Standardised regression Coefficients MIGRAINE task (N = 34 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
DURATION OF ATTACK	15	19	0	0.1	0.08
FREQUENCY OF ATTACK	33	1	0	0.67	0.23
AGE	1	32	1	-0.03	0.06
GENDER	4	30	0	0.02	0.06
OCCUPATION	0	31	3	-0.02	0.06
MISSES WORK	0	25	9	-0.06	0.10
SMOKES	1	33	0	0.04	0.05
NAUSEA	4	28	2	0.03	0.09
VISUAL DISTURBANCE	6	28	0	0.06	0.09
WEIGHT	0	33	1	-0.01	0.04
ATTITUDE TO TREATMENT	18	16	0	0.24	0.21
RESPONSE TO ACUTE TREATMENT	20	14	0	0.14	0.11
PERSONALITY	0	32	2	-0.03	0.04

Key:- +ve, -ve = significantly positive, negative ($p < 0.05$), n.s. = not significant

Table 5.13 Standardised regression Coefficients HRT task (N = 12 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
MENSTRUATION	0	9	3	-0.10	0.17
HOT FLUSHES	10	2	0	0.35	0.19
AGE	5	7	0	0.11	0.19
OCCUPATION	0	10	2	-0.02	0.09
MOOD STATES	5	7	0	0.12	0.13
SMOKES	0	12	0	0.02	0.04
LIBIDO	5	7	0	0.12	0.13
VAGINAL DRYNESS	5	7	0	0.13	0.17
WEIGHT	0	10	2	-0.06	0.09
ATTITUDE TO TREATMENT	9	3	0	0.35	0.22
FAMILY HISTORY I.H.D.	0	12	0	0.00	0.05
FAMILY HISTORY B.C.	0	12	0	-0.05	0.06
PERSONALITY	0	12	0	0.01	0.04

Key:- +ve, -ve = significantly positive, negative ($p < 0.05$), n.s. = not significant

Another measure of agreement in terms of cue use, apart from simply which cues are significant is how they are significant. In this respect, greatest disagreement can seen on the LIPID task. Five cues **Hypertension, age, smoking, diabetes,** and **compliance with advice on diet** acted both as indicators and contra-indicators to prescribe. In non-medical terms, some doctors were significantly more likely to prescribe and some doctors were significantly less likely to prescribe given a cue. This is comparable to the positive and negative non-zero ratings that some cues were given. This disagreement was greatest on smoking where, although 14 doctors did not use the cue, three were more likely to prescribe for heavy smokers and 11 were more likely to prescribe for non-smokers. Again two cues on the MIGRAINE task but none on the HRT task had both positive and negative significant standardised regression coefficients.

This differing level of agreement in policy on the three tasks can also be seen when agreement between judgements predicted from tacit policies (standardised regression coefficients) is measured. Kendall's W was calculated for judgements predicted from subjects' tacit policies on the LIPID task (W = 0.46), the MIGRAINE task (W = 0.76 and the HRT task (W = 0.53). The greater concordance between judgements predicted from policies compared to that of actual judgements (Table 5.3) can be put down to the degree of inconsistency in doctors' behaviour. The level of agreement is even greater between judgements predicted from doctors' explicit policies but again there are differences between tasks: agreement is greatest on the MIGRAINE task (W = 0.62, 0.75 and 0.63 on the LIPID, MIGRAINE and HRT tasks respectively).

Of the cues that were available on all three tasks (age, occupation, smoking, attitude to treatment and personality) Attitude to treatment was the only one used by a considerable number of doctors on all three tasks. Regression coefficients for attitude correlated significantly on the LIPID and MIGRAINE tasks (r = 0.560): doctors who were less likely to prescribe for those opposed to treatment on one task were also less likely to do so on the other. Given the very different medical nature of these two tasks this correlation strongly suggests a general personality characteristic which inclines some doctors to be generally more sensitive to the wishes of patients than others. Although weight was only used by one doctor on the MIGRAINE task, on all three tasks there was a

trend for doctors to be less likely to prescribe for patients who were overweight. When gender was significant, on both the LIPID and MIGRAINE tasks, doctors were more likely to prescribe for males. Appendix 24 shows the correlations between indices of relative importance on the different tasks.

Self-insight

Doctors' insight into the effect of cues on their decision making was measured in three different ways. The explicit stating of relative importance can be compared to the measured relative importance of cues (standardised regression coefficients) for each doctor. However, this can only be a global comparison since the two measurements are on completely different scales. Consequently, subjective weights, directly comparable to standardised regression coefficients were calculated from subjective ratings. These were used in one measure of self-insight when predictions of decisions on cases when subjective weights have been used as elements in a regression equation were compared with actual decisions on those cases. Finally, tacit and explicit policies (sets of subjective weights) were compared: the stated importance of cues was compared to the relative importance of cues. Doctors' subjective weights on the three tasks are shown in Appendices 25, 26 and 27.

Subjects were asked for relative importance ratings in terms of the bearing particular cues had had on their decisions. If subjects were able to describe cues' relative importances in this way then where a cue of great importance had a high rating there should be a strong likelihood of prescription for that patient in comparison to other patients. If the doctor's decision making can be described in linear terms at all then there should be a strong correlation between his or her actual decision values and values predicted using his or her subjective weights as elements in a regression equation. Of course this will always be less than the amount explained by the line of best fit (the tacit policy) measured by R.

Using predictions from subjective weights a linear model was formed that could be compared with actual decisions. The ability of this model to describe the decision making behaviour is measured by R_s^2 which is the square of the correlation between the subjective

linear model predictions and the actual decisions. As Table 5.14 shows, there was considerable variation in R_s^2 within tasks. There were significant differences in R_s between tasks too ($F(2,76) = 8.93$, $p < 0.01$). Doctors' R_s on the MIGRAINE were significantly greater than their R_s on the LIPID tasks ($m = 0.14$, $t = 6.14$, $p < 0.01$). R_s are shown in Appendices 25, 26 and 27.

Table 5.14 Square of correlation of subjective predictions and actual decision values (R_s^2)

Task	Mean R_s^2	St. Dev. R_s^2	Mean R_s
LIPID	0.21	0.13	0.42
MIGRAINE	0.36	0.13	0.64
HRT	0.27	0.15	0.47

This measure of insight seems to be doctor related as well as task related: doctors showing greater insight than their colleagues on the LIPID task also did on the MIGRAINE task ($r = 0.515$, $p < 0.05$). Although differences between tasks and doctors were significant when taken over all doctors, when a two way ANOVA was done on the small group of 10 doctors who completed all three tasks, there were neither significant differences between doctors nor between tasks ($F(9,18) = 1.134$ and $F(2,18) = 2.875$ respectively). However, this general inter-task and inter-doctor variation may be a reflection of the difference in ability to be described linearly at all, *i.e.* differences in cognitive control or linearity of behaviour rather than self-insight. On all three tasks there was a significant correlation between R_s and R ($r = 0.54$, 0.57 and 0.65 on the LIPID, MIGRAINE and HRT tasks respectively ($p < 0.05$)). In other words, where doctors' decisions can be modelled well linearly at all, predictions made from fitting subjective weights into a linear equation fit the data better than where decisions are less easily modelled linearly. However, there is a significant correlation between poor linear fit and inconsistency. In this case it would not be surprising if the subject had poor insight. On the HRT task, where doctors had spent longer on cases they tended to also have a better R_s ($r = 0.60$, $p < 0.05$). This correlation was not significant on the LIPID or MIGRAINE tasks ($r = 0.07$ and $r = 0.16$, $p > 0.05$).

In accordance with the suboptimal linear model formed, the correlations between

subjective weights[17] and standardised regression coefficients for each doctor were moderately good on all three tasks as shown in Table 5.15. The correlation for each doctor on each task is shown in Appendices 25, 26 and 27. In stark contrast to the significant differences between tasks when R_S was measured, there was no significant difference between the correlations seen on any of the tasks ($F(2,76) = 0.40$, $p = 0.671$[18]). This is interesting since it contradicts two prior expectations. Firstly, one of the hypotheses in the introduction was that self-insight would be worse on the LIPID task. Many cues would be rated as important but only the same number of cues would be used as on the MIGRAINE task. Both of these are the case but the subjective-objective weight correlation on the LIPID task is not significantly different from the MIGRAINE task. Secondly, considering that consistency within a subjects' judgements was generally worse on the LIPID task anyway subjects would be expected to be less good at explicitly capturing their policy, since they don't have such a clear policy.

Table 5.15 Correlation of regression coefficients with subjective ratings of relative importance

Task	Mean correlation	Standard deviation
LIPID	0.66	0.16
MIGRAINE	0.69	0.14
HRT	0.67	0.21

What is interesting is that the discrepancies between subjective and objective relative importances seem to be quite specific. Plots of subjective ratings against standardised regression coefficients for all doctors are shown in Figure 5.2. A certain clear triangular pattern can be seen: Although cues that are rated low tend to have little impact on the decision making (low standardised regression coefficients). When a cue is rated highly it sometimes does and sometimes does not have a strong impact on the decision making.

[17] These of course would be the same as correlations of subjective ratings here since for each doctor they are perfectly correlated.

[18] Fisher's z transformations of correlations were used for this comparison.

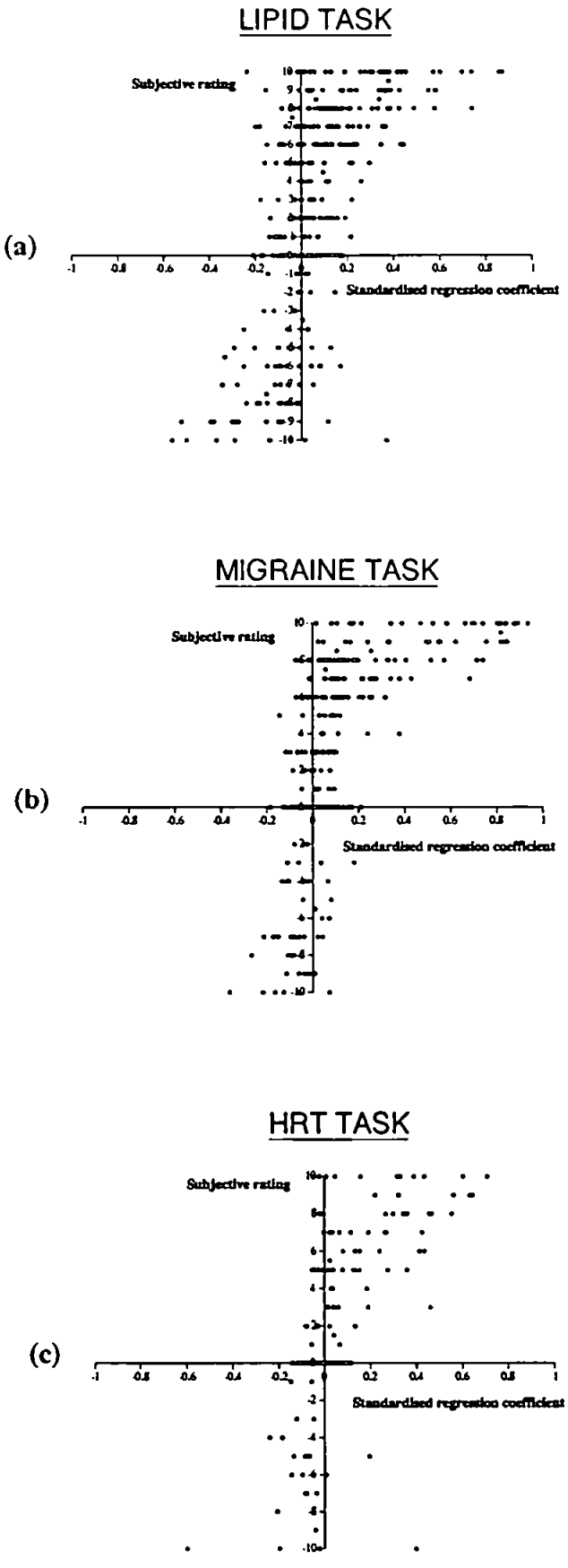


Figure 5.2. Plot of Subjective ratings against standardised regression coefficients for (a) the LIPID task (N = 33), (b) the MIGRAINE task (N = 34), and (c) the HRT task (N = 12).

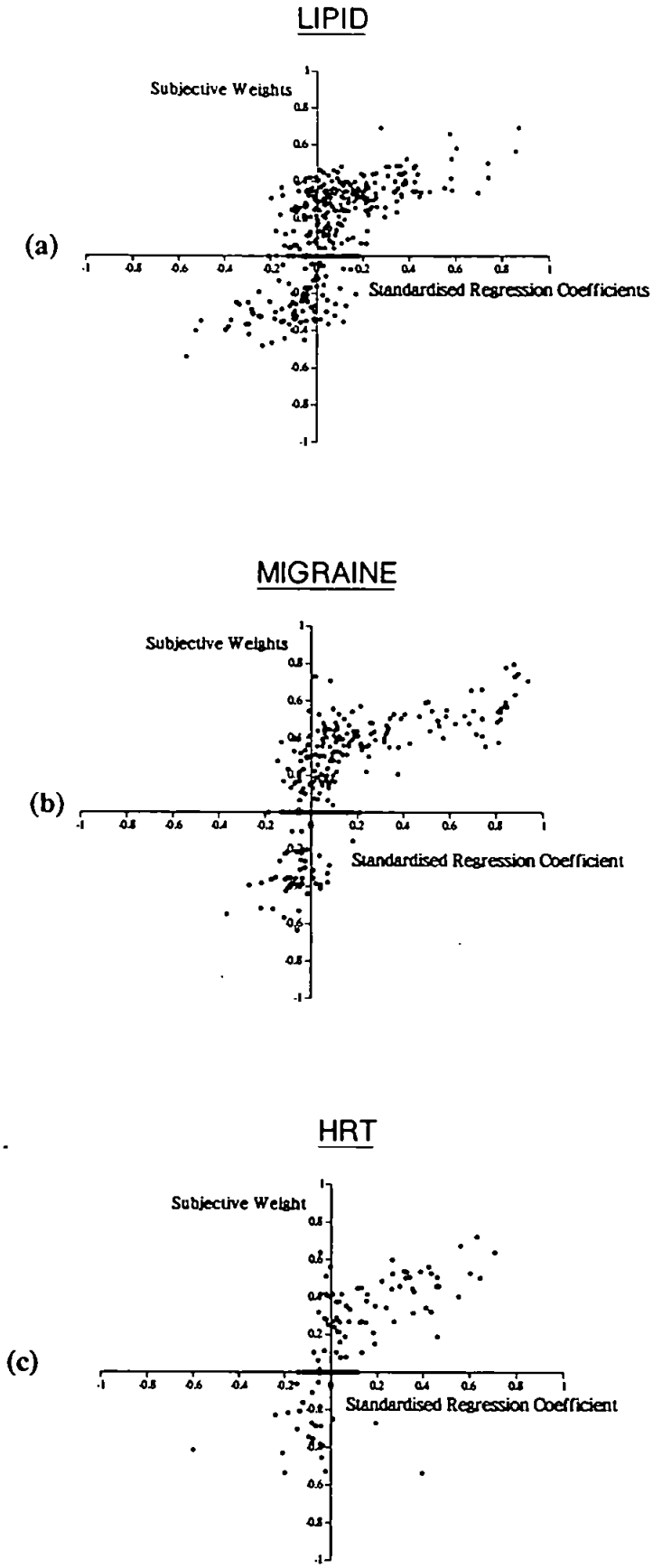


Figure 5.3. Plot of Subjective weights against Standardised Regression Coefficients for all doctors on the (a) LIPID task (N = 33), (b) MIGRAINE task (N = 34), and (c) the HRT task (N = 12)

It might be argued that this triangular pattern could have arisen as a result of different uses of the subjective rating scale: some doctors might rate their cue with highest impact at six and others at ten when they both might have high standardised regression coefficients. However, the triangular pattern can still be seen when subjective weights are plotted against regression coefficients. These triangular plots are shown in Figure 5.3.

Similarly the triangular pattern can be measured statistically in two different ways. Firstly a comparison was made between the number of cues rated above five by a doctor and the number of cues that were significant for him or her. If the doctor was using the rating scale as specified in the instructions then only those cues rated zero would have been thought to have no effect on decisions. Any rating above five could be interpreted as indicating the doctor thought that had a fairly strong bearing on the decision. The number of cues rated above five should have been less than or equal to the number of significant cues if doctors were showing good insight, since certainly all those cues rated above five, and possibly those rated between one and five should have been significant ones. In fact, as Table 5.16 shows, the number of cues a doctor rated above five was *greater than* the number of significant cues for him or her on all three tasks and on the LIPID and MIGRAINE tasks this difference was significant. So doctors were giving more cues high ratings than were even significant.

Table 5.16 Differences between the number of cues doctors rate above five and their number of significant cues

Task	Mean difference between number of cues rated above 5 and number of significant cues	t	p
LIPID	2.455	6.21	< 0.001
MIGRAINE	1.235	3.63	0.001
HRT	0.917	1.42	0.180

Secondly, a comparison was made between the range of regression coefficients for cues given subjective ratings below the doctor's mean subjective rating (A) with the range of regression coefficients for cues given subjective ratings above the doctor's mean subjective rating (B). As would be expected, the mean regression coefficient of each group with subjective ratings above the mean rating was higher than the mean regression

coefficient of the group with subjective ratings below the mean rating. So doctors were showing some degree of insight again: the group of cues that were rated as more important than average were, on average, more important than average. However what is interesting to note is that the standard deviation of regression coefficients of this group of higher rated cues was significantly larger than the standard deviation of the regression coefficients of lower rated cues. The higher rated cues had regression coefficients ranging from high to low whereas regression coefficients for the lower rated cues all tended to be low.

Table 5.17 Differences in means and standard deviations of regression coefficients of cues dichotomized in terms of high and low subjective ratings

Task	Average difference in means of groups A and B	St. dev.	p	Average difference in standard deviations of groups A and B	St. dev.	p
LIPID	0.123	0.072	< 0.01	0.127	0.089	< 0.01
MIGRAINE	0.202	0.080	< 0.01	0.249	0.120	< 0.01
HRT	0.144	0.056	< 0.01	0.144	0.080	< 0.01

Group A = regression coefficients of cues with corresponding subjective ratings below a doctor's mean subjective rating

Group B = regression coefficients of cues with corresponding subjective ratings above a doctor's mean subjective rating.

When actual values of subjective weights and regression coefficients are compared, some degree of insight could be said to be shown in that on average where a cue has been used in a certain way it has on average been said to be important in that way. This can be seen in Figures 5.4, 5.5, and 5.6. **Age** on the HRT task, **personality** on the MIGRAINE task and **occupation** on the LIPID task are exceptions to this. However, it appears that, in accordance with the over-rating seen above, mean subjective weights are nearly always greater than mean regression coefficients. The most obvious exception to this is **frequency** on the MIGRAINE task. On the LIPID and MIGRAINE tasks, although both subjective and objective measures indicate **cholesterol** and **frequency** were the most important cues, there is slight disagreement as to the relative importance of the other cues. On the HRT task there is similar disagreement. The fact that these subjective-objective differences in mean weight vary from cue to cue seems to indicate that the overestimation of relative importance is cue related.

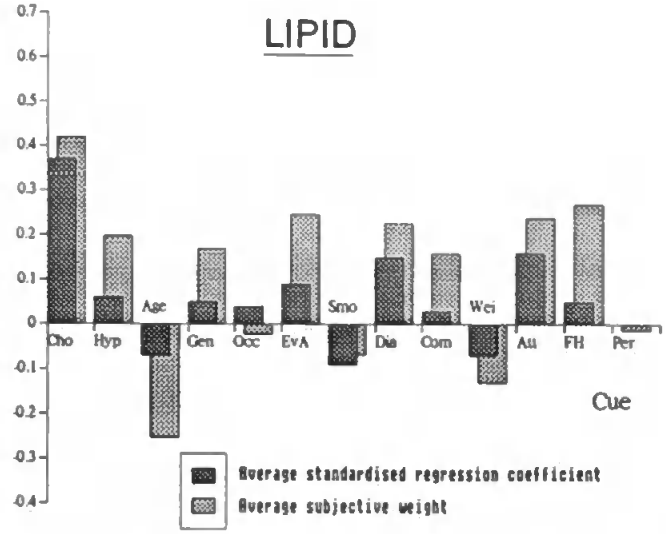


Figure 5.4. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the LIPID task (N = 33).

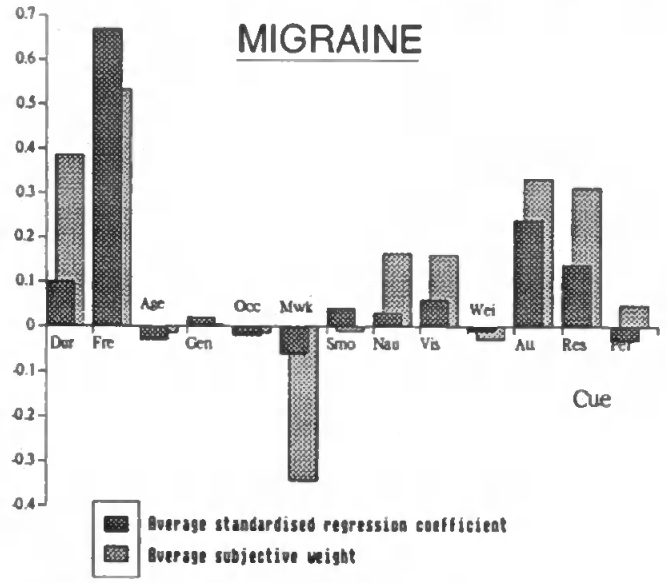


Figure 5.5. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the MIGRAINE task (N = 34).

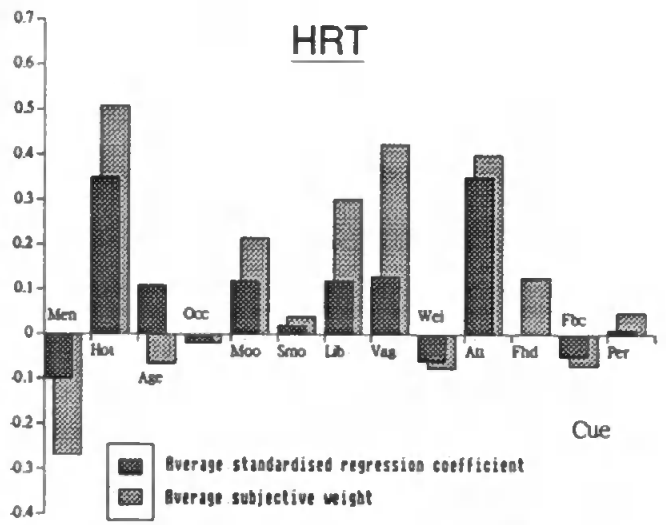


Figure 5.6. Self-insight as shown by the comparison of average subjective weights and average standardised regression coefficients for all doctors on the HRT task (N = 12).

The fact that it is cue related, and not doctor related can be seen even better another way. The results of a two way ANOVA on the absolute value of the differences between doctors' subjective and objective weights for each task are shown in Table 5.18. Although there was no significant difference between doctors in their tendency to over-rate cues, cues did differ significantly in their propensity to be over-rated.

Table 5.18 Two way analysis of variance of subjective weight-regression coefficient differences

Task	F statistic, between doctors	p	F statistic, between cues	p
LIPID	F(32,384) = 1.048	0.4	F(12,384) = 5.532	< 0.01
MIGRAINE	F(33,396) = 1.457	0.05	F(12,396) = 20.49	< 0.01
HRT	F(11,132) = 0.966	0.48	F(12,132) = 4.268	< 0.01

In accordance with this over-rating of the importance of certain cues, there were significant negative correlations between the number of cues rated above five[19] and the self-insight correlation (R_S) on both the LIPID and MIGRAINE tasks (r = -0.4 and -0.48, p < 0.05). This correlation was not significant on the HRT task, in which the sample size was small (r = -0.03, p > 0.05, N = 12). However, the correlation between the self-insight correlation and the number of significant cues for doctors was significant on this task (r = 0.66, p < 0.05). But it was not significant on either the LIPID or MIGRAINE tasks (r = 0.07 and 0.28, p > 0.05). Thus on the HRT task when doctors had used more cues they tended to show better insight. Where doctors had rated several cues highly on the LIPID and MIGRAINE tasks, predictions of decisions were less good.

If overall doctors were to show reasonable insight in terms of matching their subjective weights of relative importance to objective ones, then mean subjective weights of cues should approximate to objective ones. If there is little matching in terms of averages this may either be due to few doctors showing large misjudgements or due to standard discrepancies between objective and subjective weights across all doctors. This could either be because doctors all show poor insight or because of the way subjective weights are calculated or ratings measured. It has been shown above that the importance of cues was sometimes over-rated but rarely under-rated: some cues were more likely to be

[19] Five was chosen arbitrarily.

over-rated than others, but no doctors were significantly more likely to over-rate than other doctors. Perhaps understandably therefore, there was an association between the predictive ability of a linear model of subjective weights and the number of cues that were being rated highly on the LIPID and MIGRAINE tasks. It may be that where more cues were rated highly there was a greater tendency to over-rate. This measure of insight was also associated with the number of cues that were significant: on HRT when more cues were significant for a doctor prediction tended to be better. There is perhaps less chance of over-rating where more cues are important. Although inter-doctor and intertask differences in prediction from linear models of subjective weights were significant, these were related to differences in ability to describe that set of decisions linearly anyway - reflected in lower coefficients of multiple determination or linear fits (R).

This first study seems to indicate that doctors can't explicitly state their policy of cue use very accurately. It might be argued that the pattern of self-insight shown is because we haven't modelled their decision making correctly. Although possible, this seems unlikely because whether doctors have good or poor linear fits they still over-rate cues. It might also be argued that doctors can say what affects their decision making but just can't quantify it. Perhaps there is a scaling problem in the estimation of relative importance as there is in the estimation of correlation in "theory free" data (see Chapter 4). If that were the case, then when subjective and objective measures of relative importance were compared by rank correlations should be good. In fact Spearman's rank correlations of subjective and objective cue weights were quite good but again far from ideal (e.g. mean r = 0.59 on the LIPID task; mean r = 0.64 on the MIGRAINE task).

Conclusions

Generally, GPs' decision making policies on the three tasks can be described using judgement analysis techniques and these capture policies better than explicitly stated relative weights of cues[i]. The LIPID task appeared to be a more difficult task than the other two: longer was spent on cases in the LIPID task, more cues were given high ratings in explicit policies than on the other tasks, less consistency and lower linear fits were seen and there was less agreement between subjects. Between task comparisons may of course

be subject to order effects and the increased latencies and worse consistencies on the LIPID task may have been due to unfamiliarity with the computer presented task.

The actual number of cues used in tacit policies did not differ between tasks despite differences in latencies. Doctors might have been expected to use more cues in tasks on which they spent longer since reduced time pressure might reduce cognitive load. One interpretation is that the limit to information use is an artefact of judging a large number of computer presented cases in one session, with all information being presented at once. In real life where information is gathered sequentially it may be that more information can be taken into account when making the decision. This seems unlikely since limits to cognitive capacity, whether expert or novice, have been demonstrated using different tasks, including those not presented on a computer, using far fewer cases, and not presenting all the information in one go (see Shanteau, 1992; Elstein *et al*, 1990). However, the amount of information used here can be compared with that of Study 2 (Chapter 6) where information is gathered sequentially.

The linear predictions of judgements from subjective weights on the MIGRAINE task were better than that of the other two tasks. This is most likely to be due to better linear describability rather than greater self-insight. There was no significant difference between tasks when tacit and explicit policies were compared directly: Subjective and objective weights were found to correlate moderately well on all tasks. More cues were rated as important than were actually important on all tasks and on all tasks this over-rating of cues tended to be cue but not doctor related. This similarity in self-insight is surprising given that more cues were rated as important on the LIPID task but the same number was actually used. The effects predicted may have cancelled themselves out. The pattern of insight seen was a triangular one whereby when subjects indicated that they did not use a cue they did not. When they indicated that a cue had been used it may or may not have been used.

There could be a number of explanations for this pattern of self-insight. Measurement of self-knowledge is a comparison between an objective measure of a subject's behaviour and a subjective measure of the subject's behaviour as the diagram in Figure 5.7 demonstrates. There are two types of explanation. One suggests that subjects'

self-insight appears to be limited because the method used to elicit either tacit (objective) or explicit (subjective) policies was inappropriate. The other gives explanations for the pattern seen in terms of the information subjects might be giving us. [ii]

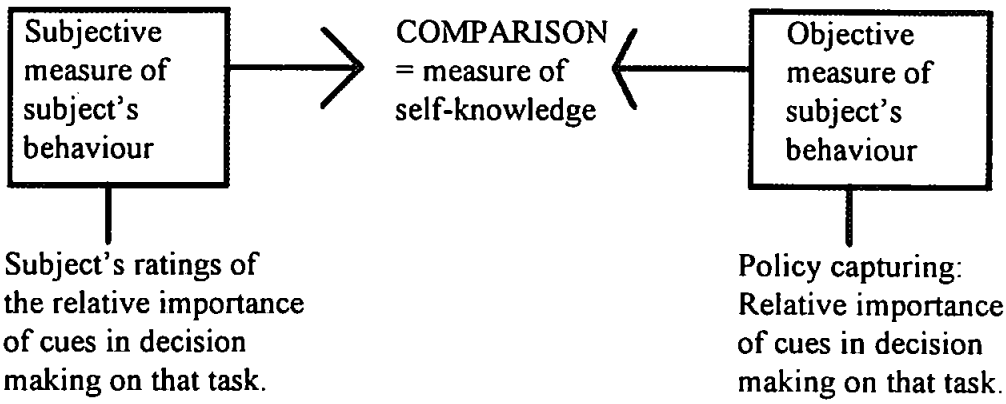


Figure 5.7. Measurement of self-knowledge

Evidence for the inappropriateness of using subjects' stated reports (their explicit policies) as a measure of what they know is examined, using an alternative measurement of subjects' self-insight, in Chapter 7. Another argument might be made that measurement of subjects' tacit (and explicit) policies in terms of a static, linear model is inappropriate. Subjects' ratings of the relative importance of cues might be based on a non-linear model. This Confounded Rating Hypothesis is introduced in Chapter 6 and further evidence for it is examined in Chapter 9 when alternative models are fitted to judgement data.

Another possibility is that doctors are stating some sort of relative importance other than the actual influence cues had on decision making. Often cues are important in clinical terms when their 'presence' increases the patient's morbidity or risk of death. A doctor may have indicated that a cue was very important if it had a strong bearing on the need to do something, even if that something was not the sort of decision making that we were asking about. This hypothesis is discussed further in Chapter 8. However, if this is the case it does not necessarily suggest that the measurement of self-knowledge is flawed since the subjective measure has not been captured. Why subjects were giving this interpretation to the influence of cues would have to be questioned and it might be said to fit the second type of explanation.

Assuming that both explicit policies and tacit policies have been elicited appropriately, to obtain the moderate degree of self-insight consistently seen in judgement analyses, explicit statements about causal knowledge must be based on something. One possibility is that stated above - that they are based on the clinical meaning of importance. Another possibility (the Attention Hypothesis) is that they are based on subjects' attendance to cues which is generally agreed to be both accessible and statable (see Chapter 4). The pattern of self-insight seen involves the overestimation of the importance of some cues. Cues that are looked at more often or are more salient would therefore be rated as being more important than cues that are looked at less often. The relationships between cue attendance, stated cue use and actual cue use are examined in Chapter 6.

Even if there is good evidence that what subjects say correlates well with their attendance to cues this is of course not evidence that one is based on the other. Indeed it begs the question as to on what cue attendance is based. Both cue attendance and stated cue use could be based on something else. For example both could be based on something doctors were taught explicitly - some ideal model, or some a priori hypothesis as suggested by Nisbett and Ross (1980). Both this and the Attention Hypothesis would fit with the type of self-knowledge seen in other studies (discussed in Chapter 4). This possibility is discussed in Chapter 8.

---

[i] The maximum possible linear fit would have been 1, in which case the subject would be consistently implementing a linear additive policy. However, the minimum possible linear fit would not have been zero. Howell (1982, p.498) has noted that an R of 0.1 ($R^2 = 0.01$) would be expected if random data were regressed onto 5 predictors over 50 cases.

[ii] It has been suggested that subjects may have anchored their responses around the 50% mark since at the start of each case the marker was positioned there. Although subjects may have already made their decision before looking at the response bar this is a possibility. Insufficient adjustment from the 50% mark would have led to less consistency and a reduced linear fit. It is unlikely that this could have accounted for the pattern of self-insight seen. If all cues had the same percentage drop in the variance they were accounting for, cues with greater impact might have dropped in importance more than cues with less impact. But if a subject has rated two cues equally, they should have the same reduced impact, given this lack of variance in response. Over-rating of the importance of cues cannot be accounted for by the possible effect of anchoring and adjustment although this would have led to a reduced linear fit.

Chapter Six
Study Two: Information Selection Study

Introduction

Chapter 5 showed the pattern of discrepancy between how GPs said cues affected their decision making (their explicit knowledge) and how they actually used cues (their implicit knowledge). To recap, the number of cues influencing a subject's decision making was small. However, the number of cues given high subjective ratings of importance was greater than this: Some cues were being over-rated. Given the literature discussed in Chapters 3 and 4 it is unsurprising that subjects' estimates of cue use are poor. Firstly, the sort of information requested is essentially causal or correlational knowledge. The regression coefficients used as indices of importance in the tacit policies are a measure of how variance of the cue corresponds to variance in judgements. But causal (or correlational) knowledge is not directly experienced. It has been argued that processes or causal relations are theoretical constructs (White, 1988, p.15-16). In order to get to the relative importance of cues subjects would have to make inferences from their direct experiences over the set of cases. This in itself is not grounds to infer poor self-insight: Knowing which pieces of information they attended to subjects are probably in a better position to state cues' relative importance than others would be.

Secondly, however, policies were modelled on experts' behaviour. Subjects were well practiced at making these sort of decisions.[1] Along with recognition and insight (aha) experiences automatic behaviour has been identified as an area in which verbal reports of behaviour are not useful (Ericsson and Simon, 1980, 1984). Skills become automatic with practice. Indeed the automatised diagnostic skills shown by doctors have been likened to pattern recognition because of the importance of the knowledge base (see Patel and Groen, 1986). Thus subjects may have had little self-insight into processes that had become automatic. Thirdly however, poor self-insight was expected in any case purely because previous judgement analyses have shown that subjects' stated descriptions of causes of behaviour are poor.

[1] Judgement analysis of novices' judgements is of dubious value since they have not developed policies to capture. In this case both linear fits and consistency will be low.

Poor self-insight was expected because in situations like this, where the behaviour is well practised and it is causal knowledge that is requested, subjects do show limited self-insight. A number of possible explanations for this were introduced at the end of Chapter 5. In this chapter the hypothesis to be explored is the Attention Hypothesis. This assumes that the apparent poor performance of subjects' ratings of importance is genuine and that these ratings are based on experiential knowledge or attendance to cues. One of the central themes of metacognition is that humans' knowledge of and ability to state how they do what they do (explicit knowledge) and their knowledge in doing it (implicit knowledge) are two fundamentally different things. Humans have direct awareness (at best) of phenomenal experience. Access to this sort of knowledge, knowledge of the content of short term memory (Ericsson and Simon, 1980; 1984) is feasible and can be verbalised. Access to any other sort of knowledge, such as processes combining the perceived information are only hypotheses. But hypotheses must be based on something. One possibility is that there are commonly held and explicit explanations for behaviour. Another possibility is that they are based on the phenomenal knowledge of which subjects are aware. Here this basic view of metacognition is adhered to. The only phenomenal knowledge subjects would have about the cues would be their attendance to them. Therefore it was hypothesized that subjects' ratings of importance would be based on phenomenal knowledge of attention to cues.

There are a few ways to ascertain which cues a subject is attending to. Firstly, they can talk aloud during the task. In the course of this they should refer to those cues that they heed. However, they may of course attend to cues that they do not mention and where judgement making is virtually automatic, talking aloud may change the nature of the behaviour (Wilson and Schooler, 1991; see also White, 1988). Secondly, eye movements could be traced and the relative amount of time spent on different pieces of information could be used as a measure of the amount it was attended to. However, subjects in these studies were visited in their places of work where it would have been impractical to take the expensive and bulky equipment, and which may have again interfered with the judgement making process. Thirdly, one can adopt the method used in the following study, where subjects' selection of cues can be used as a measure of which cues are attended to.

Here information is only revealed when the subject requests it. Obviously the cues not attended to cannot influence judgement making. The cues influencing judgement making will be the cues selected or some subset of them.

In the mechanical information boards used in some information processing studies all cases (alternatives), and their component cues, are presented simultaneously for inspection. According to Billings and Marcus (1983) subjects' estimates of covariation have been found to be better under simultaneous presentation conditions. Thus self-insight might have improved if this was the presentation format. However, cases in the surgery would not be seen simultaneously and this mode of presentation may reduce the validity of the task. In this study cases were presented sequentially so that subjects were forced to employ an interdimension information search strategy on each case (see Billings and Marcus, 1983).

Where the attendance to cues over the set of cases is compared to their relative influence on judgement making and their stated relative influence on judgement making, the Attention Hypothesis predicts that stated cue use would resemble attendance (or selection) of cues more than it resembles cue use. In this case the correlation between subjective ratings (stated cue use) and cue selection should be high and higher than that between subjective ratings and standardised regression coefficients (actual cue use).

If patterns of subjective ratings bear greater resemblance to regression coefficients than do patterns of cue selection, this would suggest that subjects have better self-insight than they would if they were just relying on information about their selection of cues. If there was no significant difference between these two sets of correlations then the implications are that subjects show no better insight than if they had based their decisions on cue selection.

Another topic to be addressed in this study is the relationship between process tracing and policy capturing analyses of judgement making. A distinction can been made between the information aquisition and information combination phases of judgement or decision making (see *e.g.* Einhorn, 1972). Verbal protocol analysis has given evidence for non-compensatory followed by compensatory processes when subjects are considering a

large number of alternatives (Olshavsky, 1979; cited by Billings and Marcus, 1983)[2]. But the two judgement stages do not necessarily refer to the aquisition and combination phases. Billings and Marcus (1983, p.377) note that process tracing and policy capturing focus on two different aspects of the judgement making process (information gathering and evaluation respectively) and found compensatory and non-compensatory policies used in both. Einhorn, Kleinmuntz and Kleinmuntz (1979) suggested that process tracing implicitly captures the information combination phase of judgement making as well as explicitly capturing the information gathering phase. They argue that the two methods capture the same process at different levels of generality and demonstrated that both could be used to capture the same judgement making. In line with this, in certain studies patterns of cue selection have been taken to be representative of aspects of the underlying judgement policy. For example, a linear compensatory model is inferred where the subject has selected the same information about each case (see Billings and Marcus, 1983).

Einhorn, Kleinmuntz and Kleinmuntz (1979) analysed two different types of judgement using both process tracing and policy capturing. In their first study on judgements of adjustment on MMPI profiles policy capturing was considerably better at predicting the subject's cross validated judgements than was the process tracing model based on an analysis of verbal protocols. Einhorn, Kleinmuntz and Kleinmuntz gave two possible explanations for the poor performance by the process tracing model. Firstly, subjects may be able to say little about their judgement making in the first well practised scenario (MMPI judgements). Secondly, it is incredibly difficult for an experimenter to form sets of rules describing the judgement making process from verbal protocols. In their second study on judgements of cereals the subject's[3] judgements on a cross validation sample were predicted about equally well by both methods. The essential argument - that both types of analysis are useful and complementary is certainly true. They were both useful descriptively at one level. However, to say that both describe the same process but at different levels of generality is misleading. They only describe the same process in as

[2] To recap, compensatory models (*e.g.* linear) imply that the judgement is affected by the values on all the cues. Non-compensatory models (*e.g.* Elimination by aspects) suggest that cases having for example a low value on one cue cannot be compensated for by a high value on another cue (see Chapter 3).

[3] There was only one subject.

much as the experimenter infers the information combination pattern on the basis of the information selected.

Billings and Marcus (1983) too looked for evidence of a convergence of results using different methods. Subjects' judgements on the same set of apartments were analysed using process tracing techniques in one task and using policy capturing in a second task. In both tasks compensatory and non-compensatory policies were compared. In the process tracing analysis these were measured by examining the information search pattern (constant amount per alternative as opposed to a varied amount per alternative). In the policy capturing analysis linear, log models and ANOVAs were used as indices of linear and non-linear behaviour. Although, as predicted by Billings and Marcus, on both tasks there was an increase in compensatory models under increased information load, there was no convergence in the two measures of behaviour. Subjects measured as using a compensatory policy on one task (with one analysis) were equally likely to be using a non-compensatory policy as a compensatory one on the other task (with the other analysis).

The two methods did not capture the same underlying behaviour. Cue selection and cue combination are two quite separate parts of judgement making. One interpretation is that Billings and Marcus (1983) found no relationship between configurality of cue selection and interaction of cue use or curvilinear cue use. The idea that both process tracing and policy capturing capture the same process is misleading. However, although Billings and Marcus presented the same alternatives, cue selection and policy were measured in two separate tasks that differed in some fundamental ways. These presented the alternatives sequentially or simultaneously leading to judgement making or choice respectively and varied different aspects of information load (number of alternatives presented at one time or time pressure). These differences could have lead to differences in behaviour and contributed to a lack of correlation.

In this study information selection and combination were measured on the same task. Where information aquisition forms part of the judgement process non-linearity of information selection might contribute to non-linearity of judgement making. It might be argued that selective revelation of cues would be odd if they were not selectively influential on the judgement. Non-linearity or configurality of information selection would

be seen in terms of selection of differing numbers of pieces of information on different cases. The standard deviation of the number of cues a doctor selected on cases could be used as an index of this. Configurality of cue use could be seen in terms of the amount of systematic variance explained by a linear model. Sets of judgements can be seen as consisting of three components: inconsistent behaviour and consistent behaviour which itself consists of systematic linear behaviour and systematic non-linear behaviour. The linear fit gives a measure of how much systematic linear behaviour there is, but this is limited by both the non-linearity and the consistency of behaviour. Consistency can be used to measure the total amount of systematic behaviour. Of consistent behaviour, the less systematic linear behaviour is, the more configural the behaviour, or the more systematic non-linear behaviour there is.

Where there is more variance in the number of cues selected on different cases, indicating configural cue selection, less of the systematic variance would be explicable with a linear model, indicating configural cue use. In other words when consistency has been partialled out, if the two methods (information selection and combination) are really part of the same phenomenon, a negative correlation would be expected between linear fit and the variance in cue selection. In this study, where both cue selection and cue use are measured this can be tested. However, cue use could still be configural even if the same cues were selected on every case. Linear or non-linear strategies may be used in either of these phases. Cue selection and cue combination are two quite separate parts of judgement making and it is expected that there will be no relationship between configurality of the two.

The second issue raised in Study 1 was that of the limitations to information processing. The apparent limits to information processing capacity were seen in the limits to the number of cues that doctors were taking into account on tasks. In a demonstration of poor self-insight, this was less than the number of cues doctors indicated were of importance. The limits to human information processing capacity lie in working memory. In previous studies, limits to information processing have been seen in terms of the number of hypotheses considered (Elstein *et al*, 1990), limits to memory, number of diagnostic statements (see Shanteau, 1992) and also in the number of cues used in judgement making

(see Chapter 3). These different measures tap different parts of information processing. This study will address whether information aquisition as well as evaluation is limited.

Method

Subjects and logistics

This Study was run during the same session as both the consistency task (Study 1) and the Policy Recognition task of Study 3 (described in Chapter 7). The average time elapsed since subjects were seen for the primary study was ten months (range 8 to 13 months).

Thirty doctors who had participated in Study 1, described in Chapter 5, participated in this study (Study 2).[4] All of these had completed the LIPID task of Study 1. They completed the Consistency task of Study 1 immediately prior to this Study. Four subjects were female.

The Information Selection task

The Information Selection task will be referred to as the IS task. The LIPID task always refers to the LIPID task of Study 1 (described in Chapter 5). Subjects were presented with a series of cases on which they had to make decisions about their likelihood of prescription of a lipid lowering agent. This was the same decision as subjects had made on the LIPID task of Study 1. On each case in the IS task subjects chose the information they considered relevant to a case before making a judgement about their likelihood of prescription of a lipid lowering drug. The cases used were the same as the last 100 cases on the LIPID task of Study 1 (Chapter 5). They were presented in the same order and carried the same case number as the LIPID task. The cues presented and their ranges were shown in Table 5.2 (Chapter 5). As in the LIPID task, in the IS task the likelihood of prescription of a lipid lowering drug was entered using the computer mouse. Prior to this information was selected and revealed using a concept keyboard. Only the information selected about each case by subjects was revealed to them.

[4] These were doctors who wished to participate in a second study and for whom the LIPID task data had been saved. All of these had done the consistency task described in Study 1 (Chapter 5).

The original 130 cases on the LIPID task had been generated to give a maximum inter-cue correlation of 0.2 over 130 cases. The greatest inter-cue correlation on the 100 cases presented in the IS task was 0.26 (between Occupation and Compliance). All inter-cue correlations on the IS task are shown in Appendix 29.

Information Selection

Information revelation was independent for each case. The screen seen at the start of a case is shown in Figure 6.1. Apart from the case number, initially only the *type* of information available was displayed. This was on the left hand side of the screen.

Subjects revealed the information they wanted by pressing the relevant area of a specially adapted A4 sized *concept keyboard*. This contained horizontal, clearly labelled bars which corresponded to the types of information available on the screen and were in the same relative positions. As bars on the keyboard were pressed the information appeared on the right hand side of the screen opposite the corresponding information label as can be seen in Figure 6.2. For example, having pressed 'Gender' on the concept keyboard the word 'Male' or 'Female' would appear opposite the word 'Gender' on the screen.

The bottom section of the keyboard was labelled 'Ready to make decision'. When this was pressed the response bar appeared at the bottom of the screen as shown in Figure 6.3. The mouse then became active and no more information could be revealed about that case. Pressing the mouse button to register a judgement caused the next case to appear on the screen.

Instructions

The instructions for the information seeking task were presented on a laminated sheet due to space restrictions on the Acorn A4 portable computer. These are shown in Appendix 28. Subjects read them in their own time immediately prior to the task. The assumptions about each case were the same as those in the LIPID task of Study 1. In addition to that subjects were given specific instructions on how to reveal information they wanted about each case prior to making a decision.

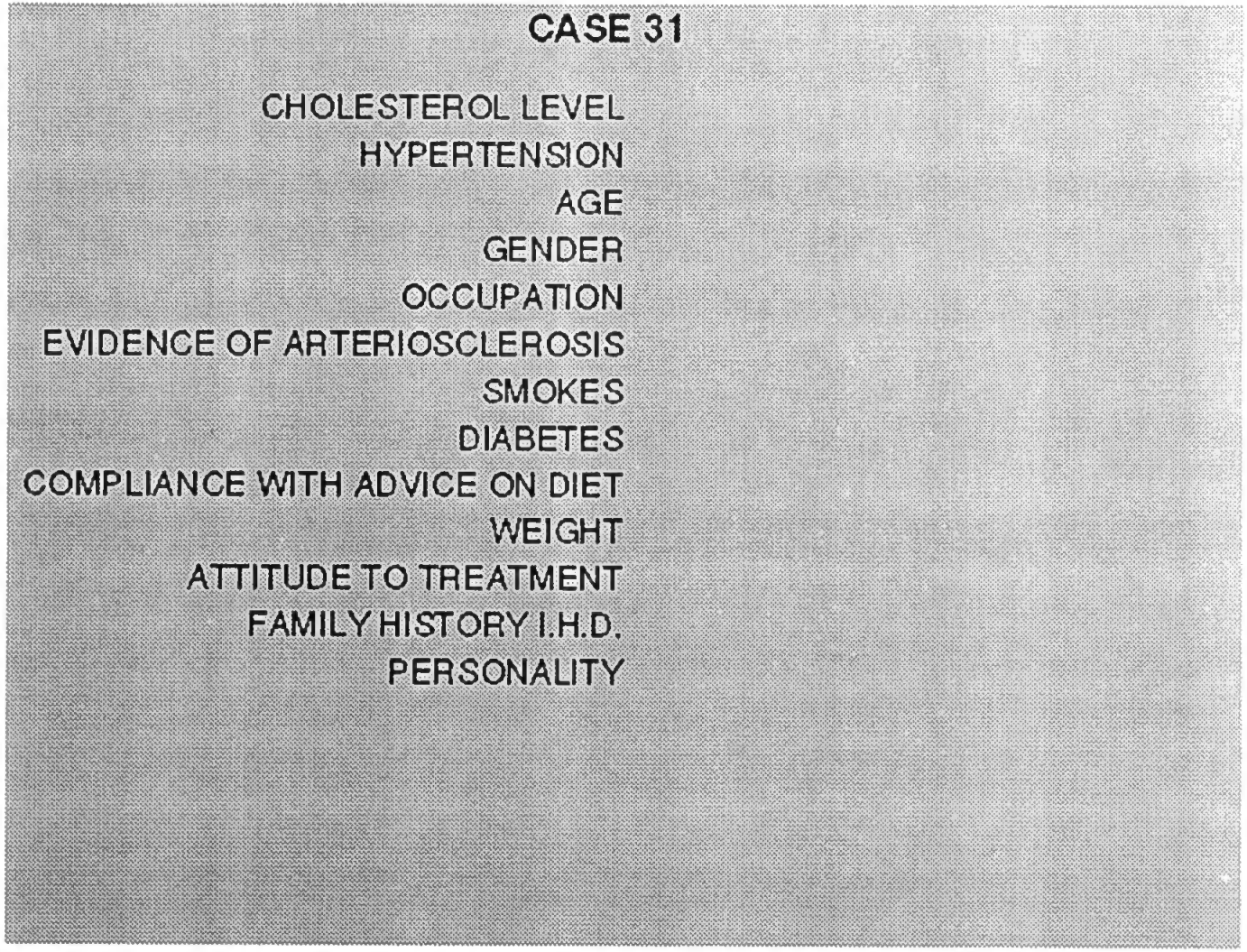


Figure 6.1 The screen seen at the start of a case in the Information Selection task.

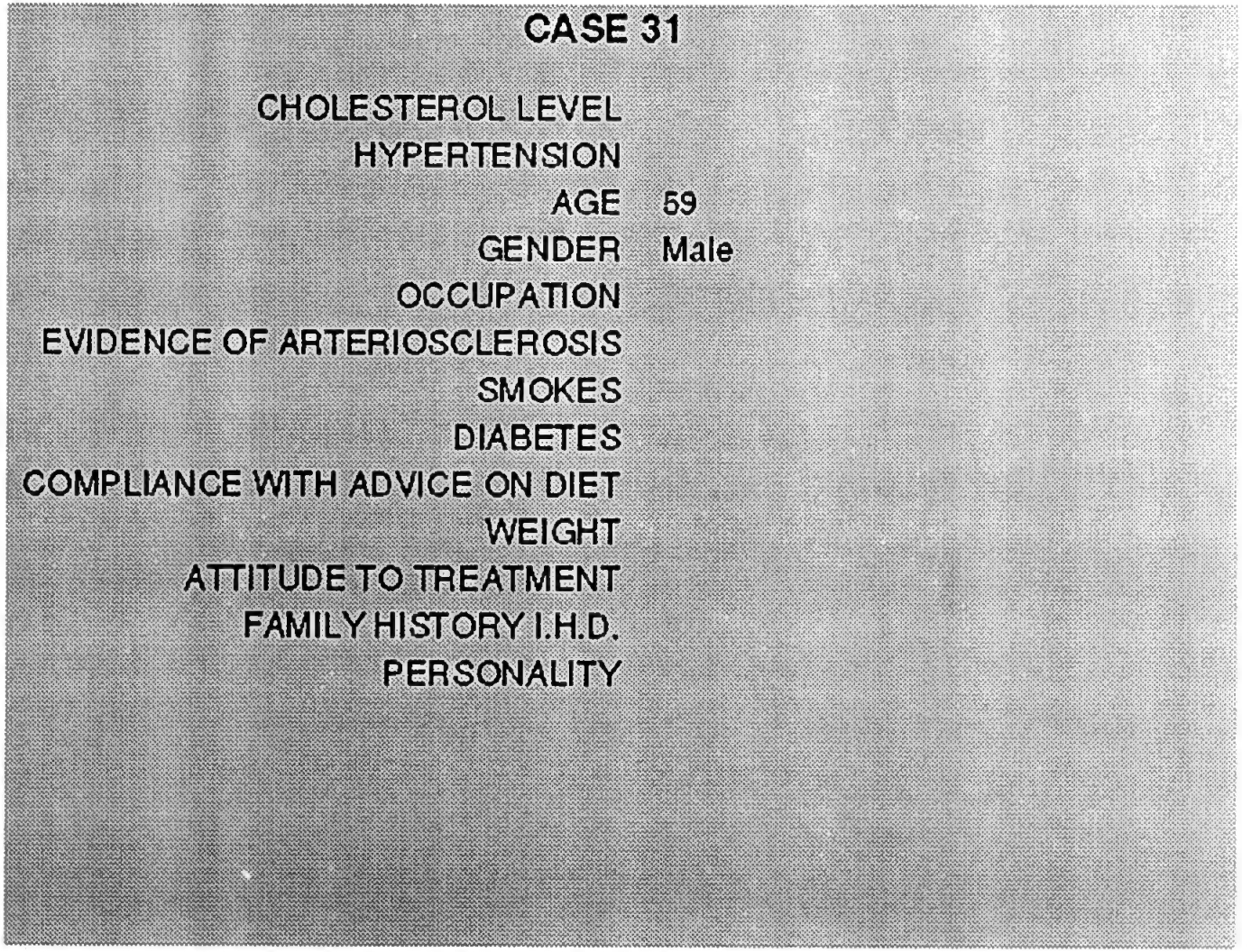


Figure 6.2 The screen seen during information selection on a case in the Information Selection task.

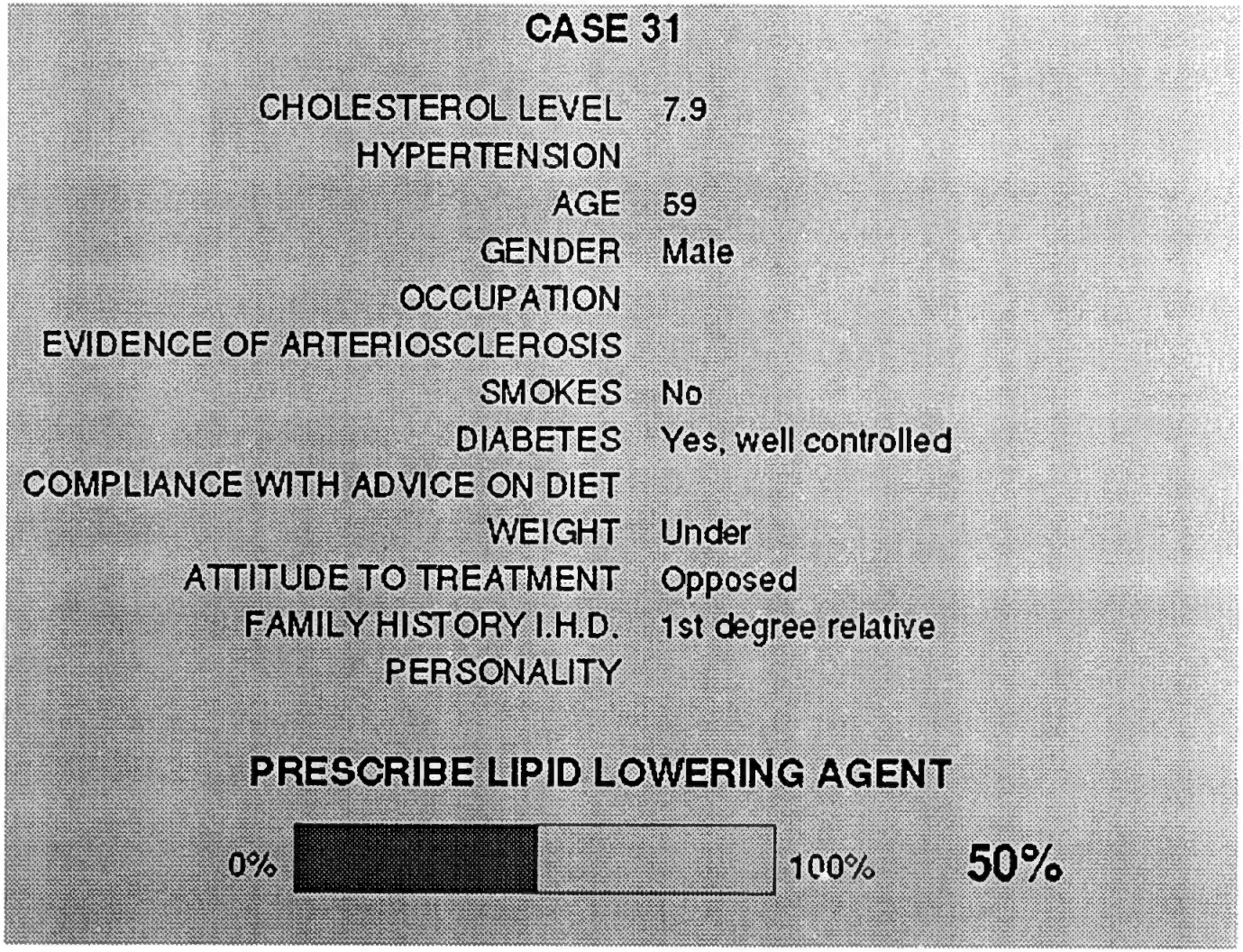


Figure 6.3 The screen seen after information selection and during decision making on a case in the Information Selection task.

Results and Discussion

The two main topics of interest in Study 1, information processing capacity and self-insight are both explored here. The finding of limited use of cues of Study 1 will be looked at in relation to the selection and use of cues in this study. The Attention Hypothesis will be tested as an explanation for the pattern of self-insight seen in Study 1. In addition, cue selection behaviour will be analysed and discussed with reference to its use in process tracing techniques. Firstly however, subjects' behaviour on this information selection (IS) task will be compared with that on the LIPID task of Study 1. Change in behaviour between the LIPID and IS tasks can be looked at in in terms of how similar judgements were, how similar policies were and how much consistency there was again within and between doctors' judgement making on the two tasks. This involves comparison of both judgements made and the tacit policies leading to them on LIPID and IS tasks.

Change and consistency of judgements

Since the cases on the IS task were the same as the last 100 in the LIPID task of Study 1 judgements on the two tasks can be compared. Similarity can be measured in two ways: Responses on cases be directly compared or they can be correlated. To test the significance of the change in judgement making between studies for each doctor, a paired t test was done over the two sets of judgements on the 100 cases. There was no trend here. The IS task does not make doctors any more or less likely to prescribe. Ten doctors were significantly more likely to prescribe on the IS task, twelve were significantly less likely to prescribe and for 8 there was no significant change in likelihood of prescription. Table 6.1 shows that the average mean judgement for both tasks was the same.

Table 6.1 Comparison of indices on the IS and LIPID tasks

Task	N	Average Mean judgement	St. Dev. of mean	Mean R^2	St. Dev. R^2	Kendall's W[5]	d.f.
LIPID (100)	30	33.0	23.7	0.49	0.16	0.28	99
IS	30	33.0	20.1	0.51	0.16	0.35	99

[5] Kendall's W is a measure of concordance ranging from 0 to 1. 1 indicates that there is full agreement, 0 indicates that there is no agreement.

Consistency is a measure of agreement on two presentations of the same case in terms of the correlation. It measures whether doctors were still relatively more likely to prescribe for cases for which they were previously relatively more likely to prescribe. The correlation between judgements made on different presentations of these 100 cases (the inter task consistency) was calculated for each doctor (see Appendix 30). These correlations were reasonable (mean r = 0.48). If judgements were different, it could be due to a change of policy over time, due to inconsistency in judgement making on either task (or both), or due to change of judgement making behaviour with a change in task format. The indication here was that what little change there was was due to inconsistency or change in policy over time rather than change in behaviour induced by the task. General change of decision making over time and inconsistency was seen in performance on the Consistency task of Study 1 which was conducted during the same session as the IS task. In the Consistency task the first 30 cases of the LIPID task were presented in the original format for a second set of judgements. A few doctors had been seen to be inconsistent in judgement making on the first thirty cases of the LIPID task in that the correlation between the two sets of judgements they had made was not even significant. However, over the next 100 cases, with the change in task, only one doctor was not significantly consistent in his decision making behaviour. Same format consistency (mean r = 0.35) was significantly worse than the inter task consistency ($t = 2.61$, $p = 0.01$, $N = 30$)[6] . However, this comparison needs to be viewed with some caution since judgement making on the first thirty cases had been shown to be less consistent within the LIPID task (see Chapter 5). Any change in behaviour is general change in behaviour over time or inconsistency rather than due to the change in task format.

The average linear fits (R^2) of policies on the LIPID and on the IS tasks are shown in Table 6.1. After both were transformed in to Fisher's z there was no significant difference between the multiple correlation R for doctors on the IS task and on the last 100 cases of the LIPID task ($t = 0.76$, $p = 0.45$) and the correlation between Fisher's z transformations of the two was significant ($r = 0.48$, $N = 30$, $p < 0.01$). So linear fits were not significantly different from those on the LIPID task and doctors showing consistency

[6] This is the t test of the difference in Fisher's z transformations of the two sets of consistencies.

in policy use on the LIPID task also tended to on the IS task.

Doctors who had good linear fits on the IS task tended to be those who made the same judgements as on the LIPID task. After both were converted to Fisher's z there was a significant correlation between inter task consistency and R on the IS task ($r = 0.396$, $N = 30$). This can be compared to the significant correlation between consistency and R on the LIPID task in Study 1. Generally doctors whose behaviour was less well captured by a linear model tended to be less consistent over time. However, not all doctors fit this pattern. Doctors having poor between task correlations but good linear fits on the IS task may have changed policy (such as GP21 and GP36) or lacked a policy on the LIPID task (such as GP13). GP14 and GP15 however show both poor between task consistency of judgements and very low linear fits on the IS task. However, GP13, GP14 and GP15 have in common that on one of the tasks their decision making hardly varied. GP15 almost never prescribed on the IS task. GP14, consistent with his behaviour on the last 100 cases of the LIPID task, hardly deviated from the midpoint of the response scale. Similarly GP13, nearly always prescribed for the last 100 cases of the LIPID task.

When judgements vary little, consistency of judgement making will appear low since correlation is used as the index. On the LIPID and MIGRAINE tasks of Study 1 there was a significant correlation between standard deviation of judgements and the multiple correlation coefficient of cues onto those judgements (R). This correlation was also significant on the IS task: $r = 0.45$, $N = 30$, $p < 0.05$. In other words when on the IS task a subject's judgements varied less, less variance in them could be explained by variance of the cues. This would appear to be the case for GP14 and GP15. Appendix 30 shows standard deviations and average judgements for all GPs on the IS task and over the same cases on the LIPID task. However, the inter task consistency in judgement making was not significantly correlated with either the standard deviation of the LIPID task judgements ($r = 0.202$, $p > 0.05$) or of the IS task judgements ($r = 0.196$, $P > 0.05$). In other words, rather than lacking variance in their judgements, some GPs must have been genuinely inconsistent in judgement making or in use of a judgement making policy.

Despite general similarity in the judgements being made, a change of behaviour can be seen on the IS task: Between subject agreement was slightly improved. Kendall's W are

shown for the two tasks in Table 6.1. This improved agreement could be due to increased consistency by doctors or it could be due to greater similarity of policies. Linear fits of models, which have been seen to correlate with consistency, were the same on both tasks. Change in policies will be examined next.

Change in cue use

Consistency of policy is measured in terms of change in the cues used. As before, tacit policies were calculated using standardised regression coefficients as a measure of cues' relative importance (see Chapter 5). Sets of standardised regression coefficients were calculated for each doctor on judgements made on the last 100 cases of the LIPID task (shown in Appendix 31) and on the IS task (shown in Appendix 32). General patterns of cue use on the two tasks were much the same as can be seen when Tables 6.2 and 6.3 are compared. Cholesterol level, Diabetes and Attitude to treatment are the most positively used cues on both tasks. Age, Smokes and Weight are the most negatively used cues on both tasks.

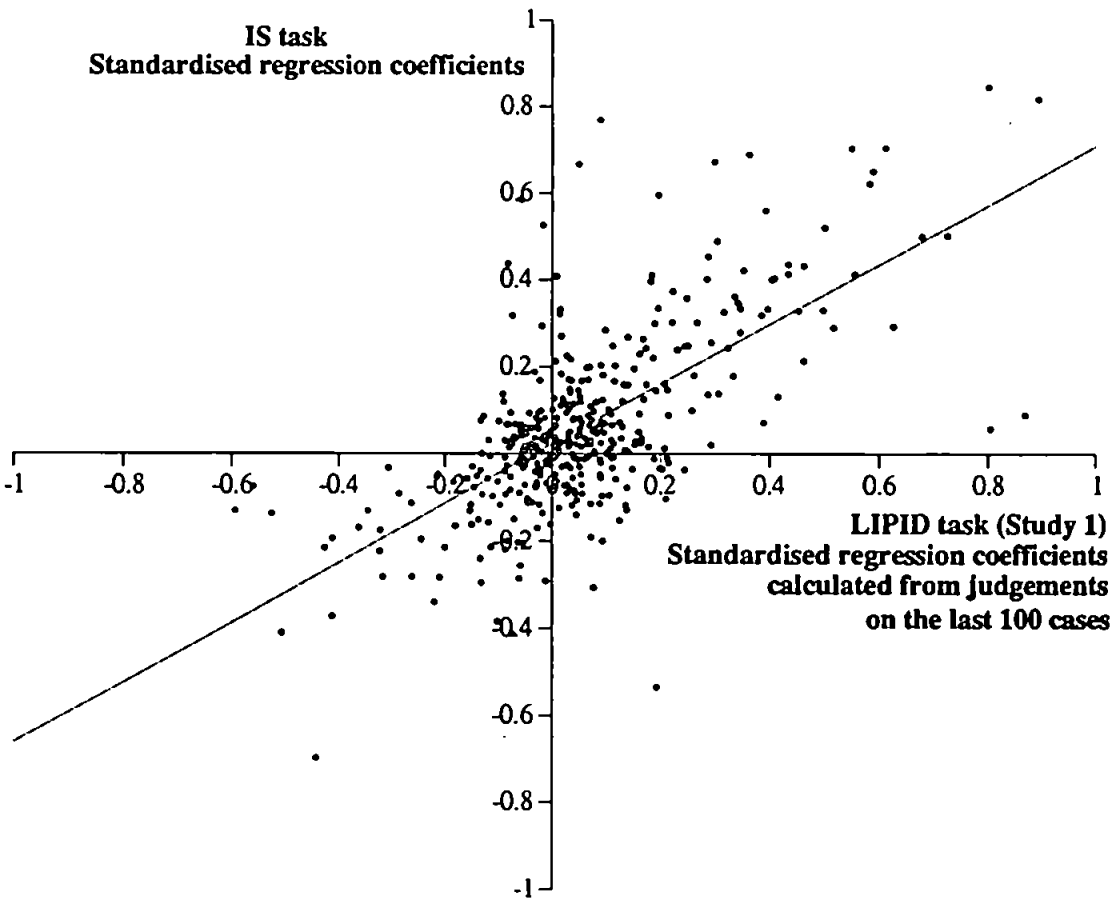


Figure 6.4 Cue Standardised Regression Coefficients for all doctors calculated on the last 100 cases of the LIPID task (Study 1) and the IS task.

Table 6.2 Standardised regression coefficients last 100 cases of LIPID task (N = 30 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	29	1	0	0.41	0.18
HYPERTENSION	6	23	1	0.05	0.11
AGE	1	19	10	-0.09	0.15
GENDER	3	27	0	0.04	0.12
OCCUPATION	3	27	0	0.05	0.07
EVIDENCE OF ARTERIOSCLEROSIS	6	24	0	0.11	0.19
SMOKES	2	19	9	-0.10	0.19
DIABETES	14	14	2	0.15	0.20
COMPLIANCE WITH ADVICE ON DIET	3	26	1	0.02	0.09
WEIGHT	1	24	5	-0.07	0.16
ATTITUDE TO TREATMENT	13	17	0	0.17	0.22
FAMILY HISTORY I.H.D.	4	26	0	0.05	0.10
PERSONALITY	2	28	0	-0.02	0.08

Key:- +ve, -ve = significantly positive or negative regression coefficient; n.s. = not significant.

Table 6.3 Standardised regression coefficients IS task (N = 30 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	26	4	0	0.41	0.21
HYPERTENSION	4	22	4	-0.01	0.12
AGE	0	19	11	-0.12	0.12
GENDER	6	24	0	0.08	0.11
OCCUPATION	1	28	1	0.03	0.06
EVIDENCE OF ARTERIOSCLEROSIS	9	21	0	0.09	0.15
SMOKES	2	18	10	-0.13	0.21
DIABETES	15	15	0	0.17	0.18
COMPLIANCE WITH ADVICE ON DIET	4	26	0	0.04	0.09
WEIGHT	0	24	6	-0.06	0.09
ATTITUDE TO TREATMENT	18	12	0	0.26	0.23
FAMILY HISTORY I.H.D.	7	21	2	0.05	0.13
PERSONALITY	1	28	1	-0.01	0.09

Key:- +ve, -ve = significantly positive or negative regression coefficient; n.s. = not significant.

The relationship between cue use by doctors on the IS and LIPID tasks is plotted in Figure 6.4. The regression coefficients on each task for all doctors are included in this graph. Only one cue appears to have changed in the way it was used but that was of little import (positively) on the LIPID study. Indeed the similarities between policies on the two tasks can again be seen in Figure 6.5 where the mean standardised regression coefficient for each cue on the tasks are compared.

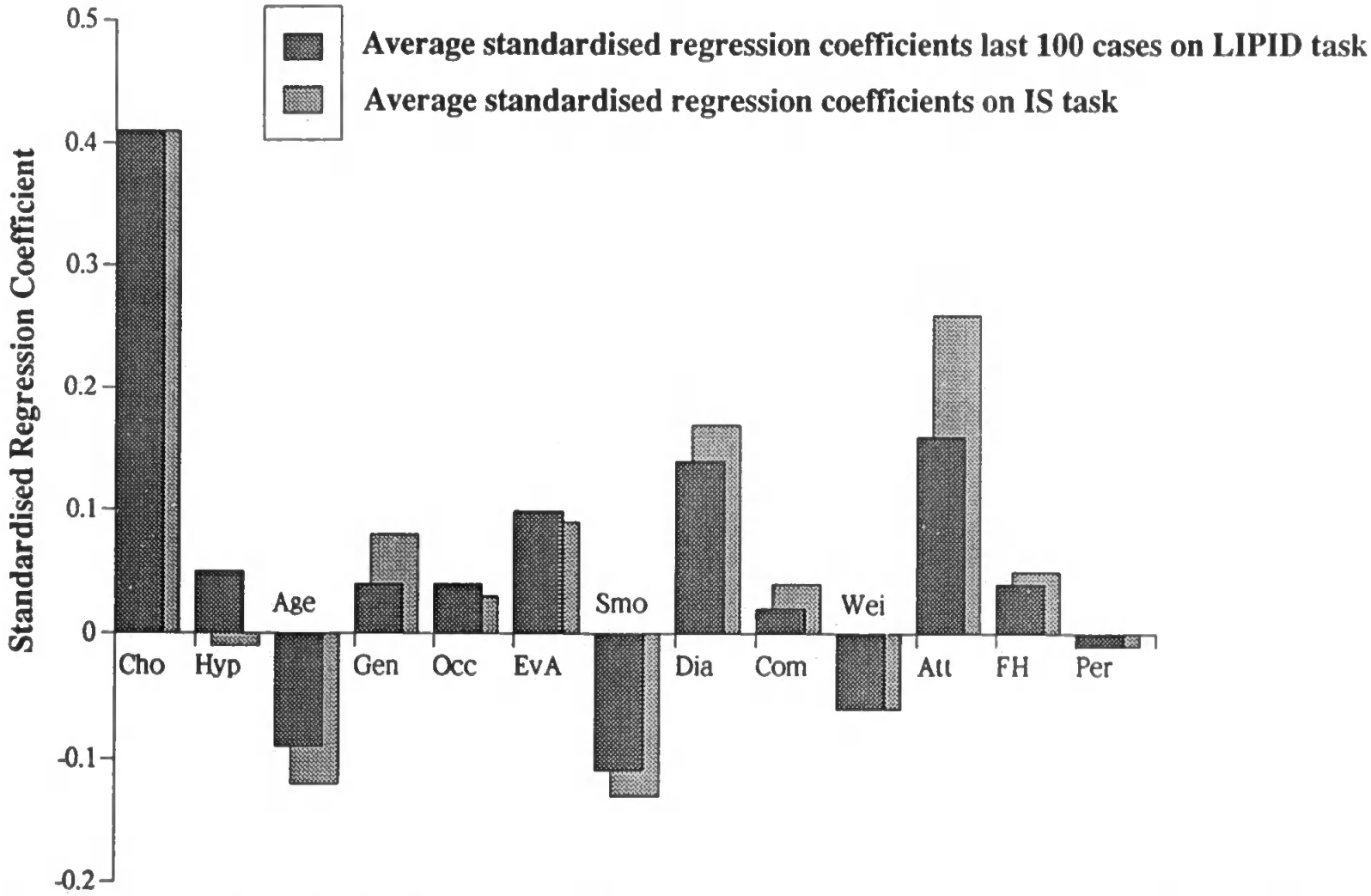


Figure 6.5 Average Standardised Regression Coefficients calculated from judgements on the IS task and on the LIPID task of Study 1 (last 100), N = 30.

A good correlation of a doctor's standardised regression coefficients on the LIPID and the IS task would show consistency of policy for an individual doctor. The correlations between these standardised regression coefficients for each doctor are also shown in Appendix 30. The average correlation between regression coefficients on the two tasks was good (average $r_b = 0.58$, $\sigma = 0.3$). However, the correlation was not significant for ten doctors who may have changed policy between studies. Of these ten, three may have poorly captured policies on one of the tasks due to lack of variance; three doctors who had been significantly less likely to prescribe for smokers on the LIPID task were not influenced by the cue on the IS task; Four doctors were significantly influenced by attitude on the IS task when they had not been on the LIPID task. Three of the doctors were significantly influenced by Evidence of arteriosclerosis on the IS task when they had not been on the LIPID task; one GP, who relied primarily on Evidence of arteriosclerosis on the LIPID task, was not influenced by it at all in the IS task.

Where cues are orthogonal as here, change in policy would lead to a change in set of decisions. Therefore it is unsurprising that the doctors that had higher correlations between standardised regression coefficients of cues on the LIPID and the IS tasks (r_b), showing better consistency in the policy they were using, also tended to show more consistency in decisions made between these two tasks[7] ($r = 0.792$, $N = 30$, $p < 0.05$).

In general doctors are consistent in both policy and judgement making between tasks. However, consistency of policy obviously affects consistency of judgement making. Where doctors were inconsistent in policy use (as measured by linear fit) on the LIPID task they were also on the IS task. Where subjects were inconsistent in policy use between the two studies (perhaps because of inconsistency on the LIPID task) they tended to be less consistent in policy use on the IS task. The slight improvement in agreement on the tasks may be due to the slight change in policy by a few doctors. Two doctors had policies on the IS task that couldn't really be explained in terms of the cues available because they had not varied judgements made on the task.

[7] Fisher's z transformations of r_b and r were used.

Similarity of latency effects

As Table 6.4 shows, the number of cues subjects were selecting on a case is unsurprisingly correlated significantly and positively with the latency on that case for almost all doctors (it does take longer to press more buttons). However, latencies were also still significantly positively correlated with latencies on the LIPID task for a large proportion of doctors. In other words for several doctors there were cases they chose to think about for longer on both the LIPID and IS tasks.

Table 6.4 Summary of within-doctor correlations over the 100 cases on the IS task

	Mean r	St. dev	t	p	No. significant
IS latencies with LIPID task latencies	0.16	0.16	5.40	< 0.01	13/30
IS latencies with number of cues selected	0.48	0.27	9.39	< 0.01	27/29[8]
IS latencies with judgement	0.26	0.25	5.73	< 0.01	19/30
Number of cues selected with judgement	0.31	0.38	4.33	< 0.01	23/29[9]

As in Study 1, several doctors were more likely to prescribe for cases on which they had spent longer. Also, for the majority of doctors, there was a significant positive correlation between the number of cues they looked at on a case and their likelihood of prescribing for that case (mean within-doctor correlation = 0.31, standard deviation = 0.38, N = 29). For three doctors this correlation was significantly negative. One doctor always selected the same five cues and so the calculation was not possible for him. Two different types of policy then can be seen: Twenty doctors, with a significant positive correlation, could be said to be looking for reasons not to prescribe. Cases they spent less time on they were less likely to prescribe for. As soon as they'd found reason not to prescribe they made that judgement. They needed more evidence that they should prescribe. In collecting that evidence, and being more likely to prescribe, they spent more time on the case. Three doctors were looking for evidence to prescribe, and as soon as they'd found it they made that judgement. If they carried on looking at more information it was less likely that they were going to prescribe.

[8] One doctor (GP19) selected the same number of cues on every case and the correlation was not possible.

[9] The correlation was significantly negative for three doctors. GP19 again not included (see footnote 1).

Models of information processing and measurement of capacity

As stated previously, cue selection has been used in information processing studies as a measure of the subject's policy on the task. Einhorn, Kleinmuntz and Kleinmuntz (1979) suggested that the sort of information gleaned in process tracing models captured the same underlying process as policy capturing does but at a different level of generality. They felt that the information combination and use of feedback in judgement making are implicit in process tracing models, which mainly focus on information search. This section will show that the analysis of judgements in terms of information selection is not capturing the same underlying process that policy capturing does.

Doctors varied in the number of cues selected on each case. All but one doctor (who selected all 13 cues) had an average cue selection per case of between 3 and 10 (median = 6.43, N = 30). The mean of the average number of cues doctors selected on cases is notably higher than the number of cues that were significant for doctors (average = 4, range 2 to 8). The cues affecting decision making should be some degree of subset of the cues selected. But doctors selected more cues than affected their decision making. Standard deviations of the number of cues doctors were selecting on cases ranged from 0 to 4.32 (median = 1.7, N = 30). In terms of information processing capacity, this finding is interesting. If limits to information processing capacity lie in working memory then not only should the information affecting decision making be limited but so too should the information attended to. For the majority of doctors this is the case - the information selected, used here as a measure of attention, is less than that available. However, it is more than that actually used. Information is selected sequentially, remains on the computer screen and need not be held in working memory. Thus more information can be selected than is actually used in the combination stage of judgement making.

As indicated in the introduction, linear compensatory rules have been inferred where subjects select the same cues on each case. Configurality of cue use - the use of non-linear, non additive (non-compensatory) policies - might be indicated by selection of different cues on cases. The standard deviation is one measure of the amount a subject's cue selection changes over cases. So this could be taken as an indication of the degree of configurality of cue use. Since doctors had different standard deviations, it might be

thought that they were showing different degrees of configural cue use. However, cue selection and cue use are not the same thing and configurality of cue selection and configurality of cue use are not the same thing. Billings and Marcus (1983) showed the discrepancy between measures of behaviour based on cue selection and those based on cue use policy capturing. However, their measures were on two separate tasks. Here both policy capturing and cue selection measures were carried out on the same task.

In order to test the effect of irregularity of cue selection on configurality of judgement making, the standard deviation of the number of cues a doctor selected on cases was correlated with their multiple correlation coefficient, partialing out consistency. Configurality of cue use would be measured by the paucity of the linear fit (R2) or multiple correlation coefficient (R) in explaining the systematic variance in the subject's behaviour. The linear fit is influenced not only by the linearity of judgement behaviour but also by its consistency. Configurality of cue selection is indicated by the standard deviation of the number of cues a doctor selects. If configurality or irregularity of cue selection has an effect on configurality of judgement making this correlation should be significant and negative. In fact, using Fisher's z transformations of the multiple correlation and consistencies, the partial correlation was insignificant ($r = -0.121$, $p > 0.05$, $N = 30$). In other words those doctors exhibiting irregular cue selection were no more likely to exhibit non-linear cue use than doctors more uniform in their cue selection. Configurality of cue selection cannot be taken as an indicator of configural judgement making. This has interesting implications for the literature discussed in the introduction. It emphasises the difference between the somewhat accessible information gathering stage and the implicit information combination phase of information processing and judgement making. Doctors may show configural cue use regardless of their configurality of selection.

Cue selection Policies

Static models of cue selection, which are comparable to other static measures of cue use such as policy capturing, can be measured in two ways. One index is the number of times a doctor selected a cue at all during the 100 cases or its percentage selection. [These are shown in Appendix 33]. The average percentage of cases on which cues were selected by doctors is shown in Figure 6.6. Another index is the average position at which it was selected. When a cue was selected first it scored a selection rating of 13, when it was selected second it scored 12, third 11 *etc.* When a cue was not selected at all it scored 0 on that case. The average selection position of cues could be calculated for each doctor, showing a selection policy. [These are shown in Appendix 34.] The average of these is shown in Figure 6.7.

Figures 6.6 and 6.7 are obviously extremely similar. The two measures of selection were found to correlate highly with each other over the 13 cues for every doctor (average correlation = 0.92, $\sigma = 0.15$, N = 30). Position of cue selection may have been affected by the order of presentation of cues on the screen. Cues could not be revealed simultaneously and apparent differences in preference for cues might simply be the result of behaviour that has been forced to be sequential. The number of selections out of 100 was used in all the following calculations and this will be referred to as the selection frequency. However, the high correlation between the two indices of selection is important. Firstly it indicates relative consistency of cue selection - an important feature if consistency of cue use is to be obtained. Secondly, it may indicate that those cues selected later whilst looking at the case, were not selected so often. This is important when considering the possible configural use of cues. It indicates that some cues only become relevant or important (and are therefore selected) when other cues have been found to have certain values. However, this correlation may be exaggerated by the proportion of cues never selected and therefore scoring zero in both indices.

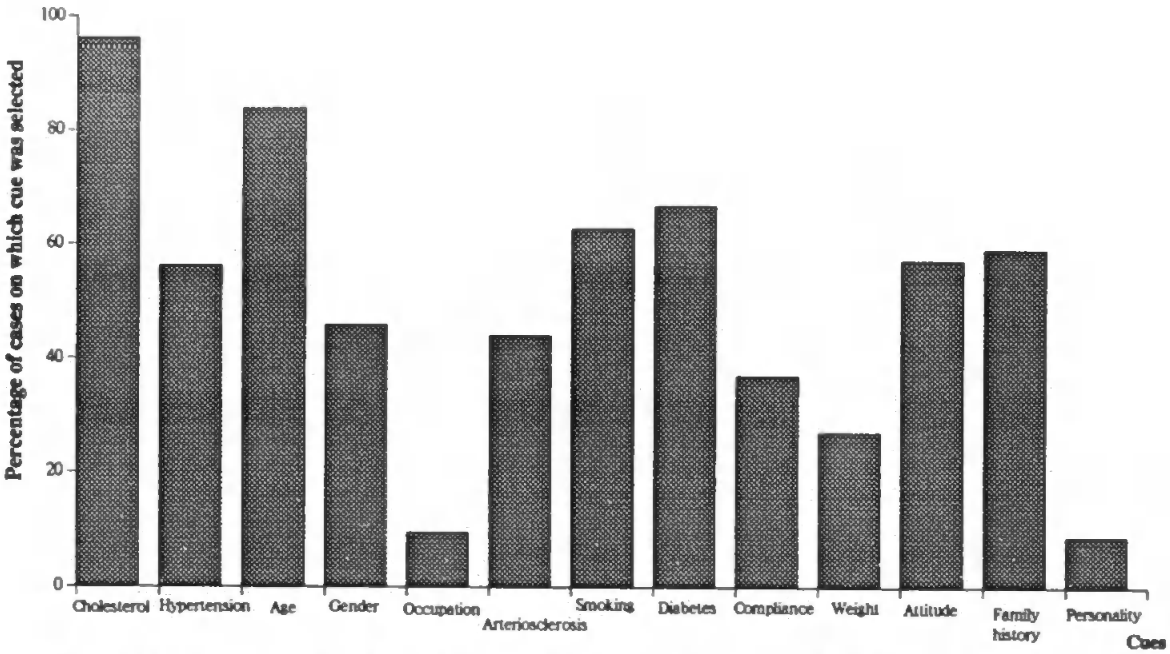


Figure 6.6 Average percentage of cue selection for each cue over all doctors on the Information Selection task.

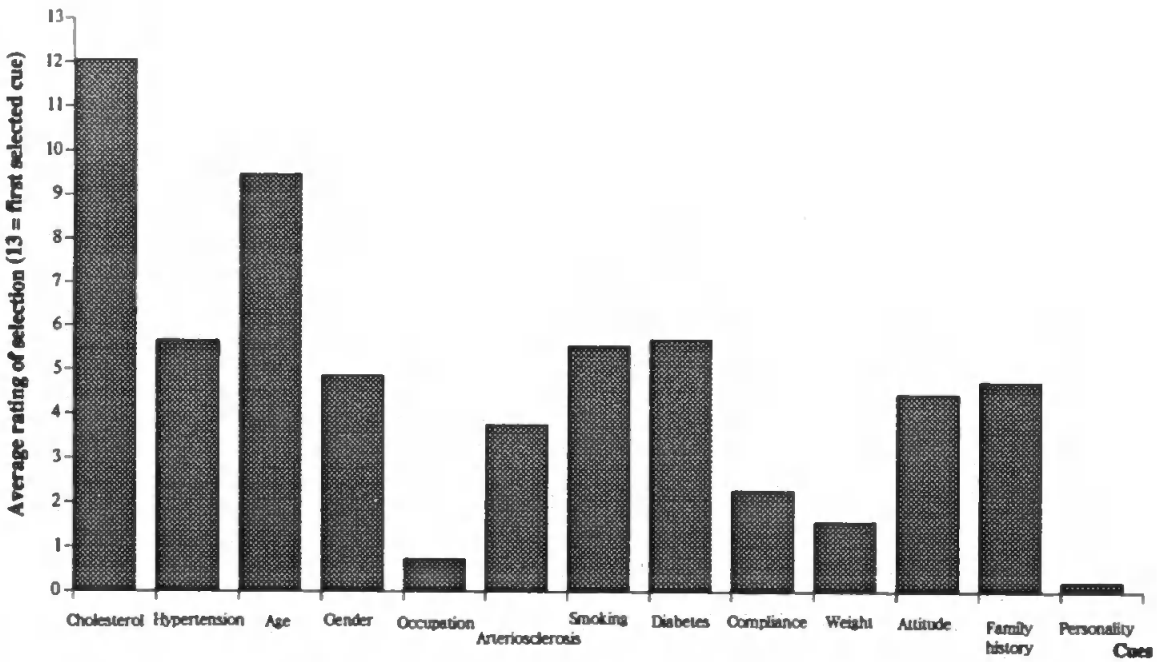


Figure 6.7 Average cue selection rating for doctors on the Information Selection task (13 = first selected cue).

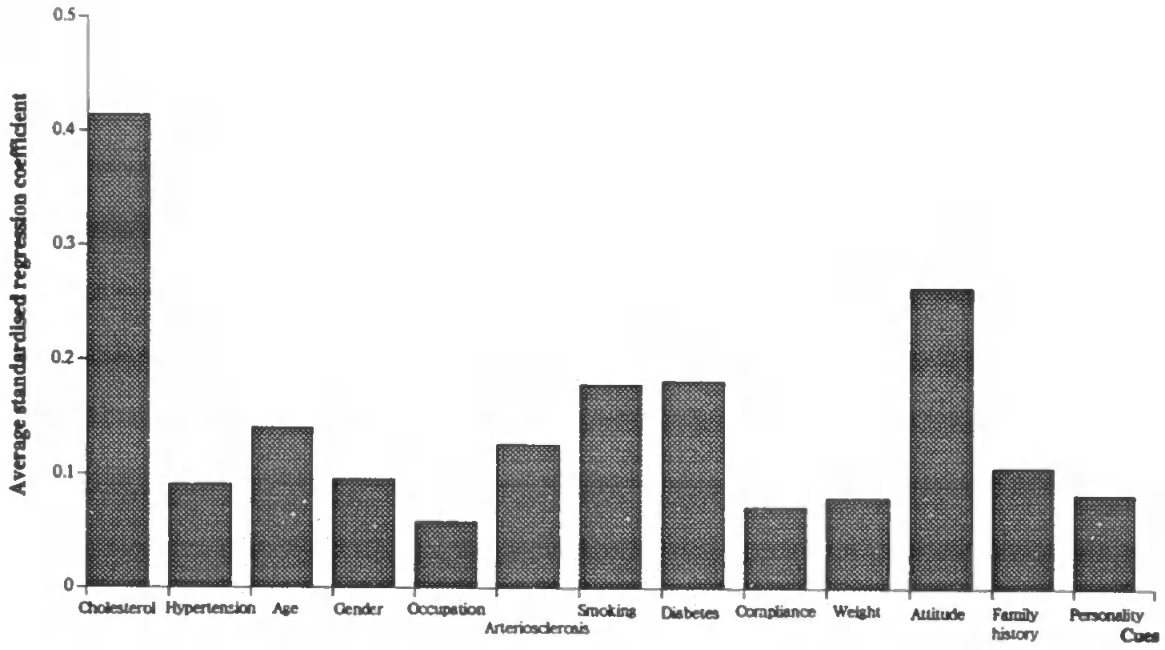


Figure 6.8 Average absolute standardised regression coefficients on the Information Selection task

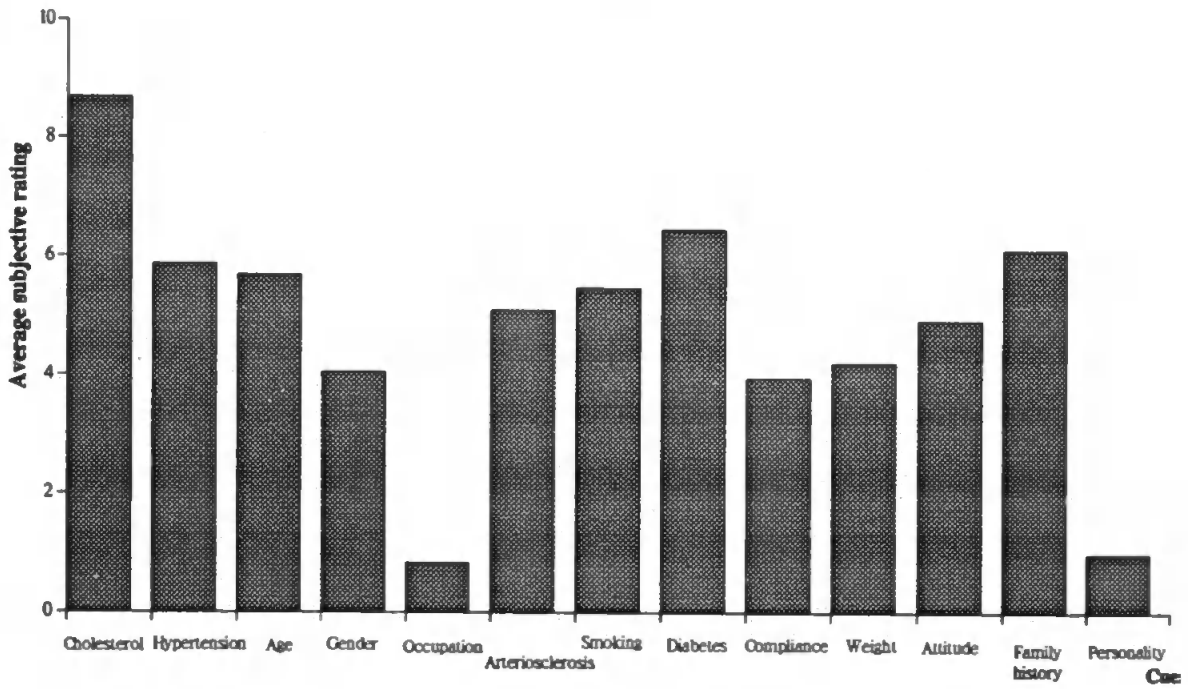


Figure 6.9 Average absolute subjective rating of cue importance (from the LIPID task, Study 1)

Self-insight

Having established that, for the majority of doctors, tacit policies varied little from those on the LIPID task in Study 1, we can compare the subjective ratings of importance obtained on that task with measures made on the IS task such as the attention paid to each cue. Cue selection was used as a measure of attention. One theory of metacognition is that we only have insight into experience and are no better than others at describing the causes of our behaviour. If this is the case in this task subjects should have good knowledge as to the cues they attended to (experiential knowledge) but not the causal knowledge of the relationship between cues and judgements requested. According to the Attention Hypothesis the ratings of importance subjects are giving cues should relate better to this experiential knowledge than to the actual bearing cues had on decision making. Cue selection was used here as a rough measure of attention.

Cue selection can be compared directly to the two other cue related measures - standardised regression coefficients and subjective ratings. Since cue selection is always positive, absolute values of standardised regression coefficients and subjective ratings were calculated for each doctor. The averages of these are shown in Figures 6.8 and 6.9.

To test the Attention Hypothesis, the correlations between subjective ratings and standardised regression coefficients for each doctor were compared with the correlations between subjective ratings and cue selection for that doctor. In all these analyses Fisher's z transforms of correlations were used in comparisons. Only 19 doctors were used for this. The ten doctors who had apparently changed policy between tasks were not included[10] , nor was GP14 whose linear model was a poor description of his policy on the IS task ($R2 = 0.13$). However, for the remaining 19 doctors the correlation between subjective ratings on the LIPID task and cue selection on the IS task (average $r = 0.65$, $s = 0.15$) was significantly greater than the correlation of the subjective ratings with standardised regression coefficients of the IS task for that doctor (average $r = 0.46$, $s = 0.16$) ($t = 4.88$[11] , $p < 0.01$). This indicates that doctors' ratings of importance bore more resemblance to the

[10] The correlation between standardised regression coefficients on the LIPID task and on the IS task were not significant for GP1, GP2, GP13, GP15, G17, GP20, GP21, GP24, GP35, GP36.

[11] This is a t test of the difference in Fisher's z values of the two sets of correlations.

way they looked at cues than to the way they used them, as predicted by the Attention Hypothesis.

If stated cue use is based on patterns of cue selection, predictions of actual cue use (regression coefficients) on the basis of selection should be no worse than those based on subjects' ratings. Again the ten doctors who had changed policies and GP14, whose policy could not be captured on the IS task, were not included in analyses. For the remaining doctors the average correlation between cue selection frequencies and standardised regression coefficients on the IS task was 0.59 (standard deviation = 0.17). As stated earlier, the average correlation between subjective ratings and the standardised regression coefficients (on the IS task) is 0.46 (standard deviation = 0.16). There was a significant difference between Fisher's z transformations of these two sets of correlations: $t = 3.3$, $p < 0.01$). So here cue selection is in fact a better description of cue use than stated cue use is. However, there is a confounding factor in this comparison in that one of these correlations is comparing doctors' behaviours across studies whereas the other compares behaviours within a study and therefore might expect to be better. Both standardised regression coefficients and cue selection were measured on the IS task. Subjective ratings were obtained on the LIPID task.

This can be altered so that both comparisons are within study. Subjective ratings from the LIPID task are correlated with standardised regression coefficients from the (last 100 cases of the) LIPID task (average $r = 0.55$, s.d. = 0.19). On the IS task cue selection is again compared to standardised regression coefficients on the IS task (average $r = 0.59$, s.d. = 0.17). The difference between these for the 19 doctors who did not change behaviour was not significant ($t = 0.83$, $p = 0.42$). If all doctors who did both tasks are included in the comparison there is again no significant difference between the correlations between cue selection and regression coefficients on the IS task (average $r = 0.59$, $N = 30$) and between subjective ratings and regression coefficients on the LIPID task (average $r = 0.51$, $N = 33$): $F(1,61) = 2.6$, $p = 0.112$. Selection of cues was no more similar to the pattern of cue use than stated policies were. In fact there was no difference in prediction of cue use from knowledge of cue selection and prediction of cue use from knowledge of subjects' stated policies.

A note on the Confounded Rating Hypothesis

Grounds for another possible explanation for the pattern of self-insight seen in Study 1 can also be explored here. The Confounded Rating Hypothesis predicts that the apparent over-rating of the importance of certain cues is due to their non-linear use. This questions the validity of both tacit and explicit *linear* models of decision making. The assumption is that subjects do have good self-insight, that subjective ratings are actually reflective of the influence a cue has on some cases, but that this influence is non-linear. Either the importance is dependent on the value of other cues or there is a non-linear relationship between cue and judgements. Although the evidence for the Confounded Rating Hypothesis is tested in Chapter 8, here is a preliminary discussion.

One example of how the Confounded Rating Hypothesis might occur is if a doctor, who would potentially prescribe for a case on the basis of its clinical information, is strongly influenced one way or the other by the patient's attitude. The same doctor would not even bother finding out the patient's attitude to treatment if there was no clinical necessity[12] for the treatment. If the doctor was basing rating of importance on the potential effect of the cue then the doctor might rate attitude highly. However, its regression coefficient would depend on the number of cases for which treatment was considered clinically relevant and therefore on which attitude was selected. Imagine the situation wherein the majority of cases clinically merited prescription: Attitude was selected relatively often and had a high regression coefficient. Here the cue may not have been over-rated. Alternatively imagine the situation where few cases merited prescription, attitude would be ascertained less often and have a lower regression coefficient. Here the amount of times a cue is selected indicates when it is relevant. Those cues that are over-rated should be ones that are not selected that often. This would lead to a dissimilarity between subjective ratings of cues and their selection: The selection pattern should be more like the pattern of standardised regression coefficients than the subjective ratings are.

If the Confounded Rating Hypothesis is true, the relationship between the selection frequency of cues and their regression coefficients should be better than that between their subjective ratings and their regression coefficients. In other words cue selection would be a

[12] Clinical necessity or merit for prescription is as perceived by the doctor.

significantly better predictor of actual policy than stated policy is. Although this has been demonstrated to be false (correlations between cue selection patterns and regression coefficients were no different to those between subjective ratings and regression coefficients), it does not disprove the Confounded Rating Hypothesis. Configurality of cue selection has been taken to be an indicator of configurality of cue use. They were shown earlier to be uncorrelated.

Configural cue selection was measured in terms of the standard deviation of the number of cues selected on cases. Higher values meant there was more variation in the number of cues that might be selected. Doctors who were uniform in the number selected would have a standard deviation of 0. The correlations between absolute values of subjective ratings and standardised regression coefficients on the LIPID task (last 100 cases) are a measure of self-insight. The correlation between self-insight (Fisher's z) and configurality of cue selection was not significantly different from zero ($r = 0.297$, $N = 30$, $p > 0.05$). In other words, those doctors who were selecting cues configurally, showed no worse self-insight, were no more likely to over-rate cues, than other doctors more uniform in their cue selection. However, again this is not evidence for the Confounded Rating Hypothesis since the doctors who were uniform in cue selection could have been using cues configurally (and over-rating their importance) but not selecting them configurally. Thus confounded rating could be shown because of configurality at a later point in the judgement process.

Evidence for the Confounded Rating Hypothesis must be looked at in terms of configurality of cue use rather than configurality of cue selection. This is done in Chapter 8.

Conclusion

Generally behaviour on the IS task had changed little from that on the LIPID task in Study 1 despite the change in presentation of the cases. Slightly more agreement was seen between doctors' judgements. Agreement is affected by both consistency and agreement in policy. Since linear fits were unchanged consistency of behaviour, which was seen to correlate significantly with linear fits in Study 1, was unlikely to have changed.

Change in agreement was most likely to be due to the slight change in policy by some doctors.

The Confounded Rating Hypothesis could not be proved one way or the other here because of the lack of correlation between configurality of cue selection and configurality of cue use. Configurality of cue selection does not appear to affect the overall configurality of judgement making. This is evidence for the relative independence of behaviour in the different stages of information processing. Einhorn, Kleinmuntz and Kleinmuntz (1979) suggested that process tracing and policy capturing models were looking at these different aspects of the same phenomenon. The interpretation here is that they are looking at different phenomena within the same process. No correlation was seen between behaviour on the information selection and information combination phases of judgement making. Although the information involved in the latter is by necessity a subset of the information involved in the former (and subjects were tending to look at more cues than had a bearing on the decision), patterns of non-linear behaviour on the two were different. Both may be linear or non-linear. Assumptions about the latter made on the basis of the former, as occurs in process tracing, may be subject to problems as has occurred in other studies (see Einhorn, Kleinmuntz and Kleinmuntz, 1979). At the same time extrication of the two phases is useful to characterise the whole decision or judgement making process.

Different amounts of information were used in the different phases of judgement making and both appeared to be limited. The information phase seemed less so than the combination phase. It was hypothesized that the sequential and then maintained revelation of the former may have been of help in overcoming cognitive limitations.

The Attention Hypothesis is still viable. It appears that subjects are stating relative importances that are more akin to the amount they look at cues than the amount they use them. Regression coefficients were as predictable from the pattern of cue selection as they were from the pattern of stated cue use. No better self-insight is shown than if cue selection ratings had been used as a measure of the importance of cues. That subjects are basing ratings of cue importance on their cue selection is one interpretation. However, alternative hypotheses also fit this finding. For example, doctors could be basing their ratings on an ideal model of their behaviour. They may also choose to look at cues because they felt they

should be affecting their decision making. Both could be based on the cues the subject believes to be relevant to the judgement in real life. In this case both cue selection and cue importance ratings are based on some other factor. However it would be hard to identify the source of GPs' ideal model of behaviour. However, this is comparable to the idea that subjects give previously held, often socially agreed, causal theories (see Nisbett and Wilson, 1977; Nisbett and Ross, 1980; White, 1988).

In line with this, another possibility introduced at the end of the last chapter is that doctors have grasped at some other meaning of importance. Which cues are of clinical importance is explicitly discussed by doctors and their ratings of relative importance could be based on these. This hypothesis is explained and tested in Chapter 8. However, the results of this study already indicate that this is unlikely. Although subjects might select the cues that they consider relevant to the case, unless all cues that are selected are of clinical importance there is no reason to suppose that the pattern of selection would resemble the pattern of clinical importance.

Another possibility is that cues could be rated as important on the basis of the distribution that they would take in cases seen by the subject in real life rather than on what was seen during the task. The cues that are relevant to this type of decision or judgement would be selected on the task (just as they would be ascertained in real life). The ranges of cues were known on this IS task since the same cases had been looked at previously. If they could state the relative importance of cues for real life decision making it would be surprising if when asked specifically about behaviour on a task they could not give those relative importances. However, it may be that real life ranges of cues were familiar and more easily accessed. Both of these possibilities reinforce the idea that explicit knowledge of behaviour is something separate from tacit policies of behaviour. This possibility will be referred to again in the final discussion in Chapter 10.

However, the correlation between cue selection and subjective rating of importance was not perfect. This could have been because the confounding effect of change of task or perhaps different doctors showed this pattern of insight for different reasons. Alternatively, stated cue use might be additionally influenced by other factors such as cue (or case) salience. The other possibility, mentioned in the concluding section of Chapter 5, is that

the over-rating seen in this measure of self-insight is just typical of the pattern of estimation of covariation. These are affected by prior beliefs about the relationships between cues as described in Chapter 4. This will be further discussed in Chapter 10.

Chapter Seven
Study Three: Policy Recognition

Introduction

In Chapter 5 measurement of self-knowledge was seen as a comparison between a subject's perception of their own behaviour and an objective measure of it. Verbal elicitations were used to identify subjective knowledge. However, as outlined in the review of the literature on self-insight in Chapter 4, one criticism of much research into metacognition lies in the use of subjects' reports. There are two types of argument against the use of verbal reports. These will be discussed below. This chapter presents an alternative measure of self-knowledge (also used by Reilly and Doherty, 1989; 1992). Here the objective model of behaviour used in Study 1 is maintained. But subjects' knowledge of this is elicited in a forced choice task.

One criticism of the use of subjects' verbal reports points out that they are not measuring self-knowledge but the ability to express this verbally. Examples can be given as to why this might be the case. In discussing knowledge of implicit learning Berry and Dienes (1993) suggest a few factors that may contribute to subjects' failure to state the metaknowledge they have: The amount of knowledge may be great and not enough time may be given to its elicitation; subjects may choose not to give knowledge they have low confidence in; finally they suggest that the problem may lie in the actual expression of knowledge -

> "Subjects may not retrieve relevant knowledge in free recall because they may not know what specific questions to ask themselves to reveal their own knowledge."

Other authors have suggested that process or causal explanations are not immediately accessible (Nisbett and Wilson, 1977); that processes are theoretical constructs in any case (Nisbett and Ross, 1980) and that as a consequence subjects are more likely to give previously held (a priori) theories or construct explanations when asked for causal explanations of their own behaviour (Nisbett and Wilson, 1977). Reilly and Doherty (1989, 1992) criticise both the equation of self-knowledge with verbal reports and also the use of correlations, which are quite open to interpretation, as an index of self-insight. The range

between a statistically significant correlation and perfect correlation is large. Where correlations stop showing poor self-insight and start showing moderate self-insight or stop showing moderate self-insight and start showing good self-insight is open to debate.

The second criticism of the use of verbal reports as a measure of self-knowledge lies in the discrepancy between subjects' stated and actual knowledge that is often given as evidence of poor metacognition. In a discussion of subliminal perception for example, Berry and Dienes (1993) refer to the difference in results obtained when measurement is based on subjective reports compared to objective forced choice tests. Below the subjective threshold but above the objective one, subjects are apparently unaware that they have knowledge and believe they are guessing. But performance is significantly greater than chance and can be influenced by priming effects near the objective threshold. This constitutes what Berry and Dienes refer to as the layperson's definition of subliminal perception: semantic influence can occur below the point at which subjects state that they do not know. However, at this level retrieval of knowledge is also possible. The subjective, objective threshold distinction in the realm of subliminal perception is problematic in that both knowledge as measured above the objective threshold and semantic influence of the cue are measured in terms of subjects' above chance performance. Explicitly stated knowledge tends to be measured in comparison to an ideal description.

The subjective threshold has also been the standard measure for explicit knowledge in the implicit learning literature. Subjects' descriptions of underlying rules or policies are generally worse than their implicit performance. But there is evidence that after many trials explicit knowledge may improve (see Berry and Dienes, 1993). However, Berry and Dienes (1993) argue for use of an objective threshold there too. Subjects' knowledge in implicit learning can be characterised by forced choice categorization rather than through free recall and then knowledge of rules underlying a task appears to be better. Care must be taken as to whether explicit or tacit knowledge is being elicited in these cases. If exemplars are used in the forced choice categorization task it may be tacit knowledge rather than explicit knowledge that is being measured.

Forced choice categorization can be used to measure a subject's self-knowledge. It is one thing to objectively measure what a subject knows or how they behave. It is another

thing to measure their knowledge of what they know of their behaviour (their self-knowledge). If a subject's behaviour conforms to certain rules then it could be argued that they have the knowledge of those rules but they do not know they do. Explicit knowledge of the rules could be demonstrated by verbal elicitation. But metaknowledge can also itself be measured through behaviour such as a choice between rules on a forced choice task, or some other measure of performance. The work of Reilly and Doherty (1989 and 1992) does just that in the context of policy recognition. If subjects are aware of their policy and therefore have good self-insight but are unable to state it this should come out in a recognition task.

In their studies Reilly and Doherty (1989, 1992) both measure self-insight in the usual way and also test subjects' recognition[1] of their own policies in a forced choice test. They found that subjects picked out their own policies at much greater than chance levels. However, although the interpretation of standard methods for measuring self-insight is questionable, the significance of subjects' ability to recognise their own policy also does not prove self-insight. It may be that there are other things affecting recognition. Reilly and Doherty (1989) themselves point out that own policy recognition requires knowledge of features that distinguish it from those of fellow subjects. With the conventional measurement of self-insight, certain subjects would apparently have better insight if the explicit policy they give matches the one they happen to be tacitly using. However, these subjects would not necessarily show any better self-recognition, which would additionally require knowledge about what distinguishes their own policy from those of others.

Referring to classic measurement of self-insight as a stating task and Reilly and Doherty's forced choice method as a recognition task, performance on the recognition task should reflect both self-knowledge and some degree of self-discrimination. If the stating task does measure self-knowledge then performance on the recognition task should be a product of performance on the stating task and the degree of similarity between the policies being assessed. If this is not the case then either some other factor is additionally affecting performance on the recognition task or the stated task failed to measure self-insight.

Reilly and Doherty are obviously aware of the effect of similarity between policies

[1] This is not 'recognition' in the usual sense of the word in that subjects have not previously seen the stimulus.

on own policy recognition but they do not measure it explicitly. They do however, measure two related phenomena. They measure "subjective evaluation" as the similarity between a subject's stated policy (the standard measure of self-knowledge) and all tacit policies. Subjects are scored as having a hit on this score if their tacit policy is either most similar or second most similar to their stated policy. This measure takes into account both the similarity between subjects' explicit and tacit policies (the standard stating task measure) and also the similarity between all tacit policies. Obviously in situations where several policies are similar there is less chance of having a hit. Reilly and Doherty also measured similarity between judgements made by correlating judgements by pairs of subjects. [In Study 1 of this thesis Kendall's W was used to measure agreement between all subjects' judgements on a task.] Least agreement was shown between subjects in the condition where twelve orthogonal cues were presented (mean correlation between pairs of subjects' judgements = 0.41). The standard stating measure of self-insight for subjects in this (twelve orthogonal cues) condition was average: the mean correlation between subjective and objective weights was 0.69. However, this was the only condition in which a significant number of subjects had hits on the subjective evaluation score and policy recognition was best in this condition. In the twelve correlated cues condition, however, extremely good self-insight had been shown (average r = 0.92) but subjective evaluation was not significant and recognition, although good, was worse than on the orthogonal condition. Agreement between subjects' judgements was higher here (mean correlation between pairs of subjects' judgements = 0.58).

The decision making tasks Reilly and Doherty used are not as subject to explicitly taught ideals as others might be: Both attractiveness of job offers and of room mates are personal decisions and thus a variety of policies were used and stated as used. Although the actual weighting of the cues used may differ they still may be subject to generally agreed principles[2] . In the event, individual differences in policy are also (if not more of) a feature of clinical judgements. However, as Chapter 5 showed, different tasks may lead to different amounts of agreement between subjects.

[2] For example, most people would agree that a higher paying job was more desirable than an identical but lower salaried job. Most people would not see having an angry room-mate as a benefit.

In addition, other studies have found greater agreement between explicit policies than tacit ones (Chaput de Saintonge and Hattersley, 1985, p. 210). In Study 1 (Chapter 5) judgements predicted from explicit policies generally showed greater inter-subject agreement than those predicted from tacit policies. If this were the case in Reilly and Doherty's study, recognition of subjective policies might expect to be worse than recognition of tacit policies. In fact, in three of the conditions in Reilly and Doherty's study (1992) subjects were significantly better at selecting their subjective policies than their tacit policies. Only on the twelve orthogonal cues condition was there no difference. There may have been a ceiling effect here since in this condition recognition was very high. Reilly and Doherty bring in various factors to explain subjects' better recognition of subjective policies than tacit ones. For example recognition would be good if subjects remembered the actual number they had assigned to a particular cue. However, another interpretation of this superior performance with explicit policies is that subjects really believed these were their policies.

Study 1 ascertained both tacit and explicit policies in decision making tasks for which there is considerable consensus (prescribing prophylaxis for migraine) and for which there is considerably less consensus (prescribing lipid lowering therapy). That this is the case was seen in the differences in agreement of both policies and judgement making on the LIPID and MIGRAINE tasks (Chapter 5). In this study the same subjects were shown selections of these policies and were asked to select their own.

A subject's ability to pick out his or her own policy is a measure of self-insight that could still be due to the incidental matching of explicit and tacit policies. For example, although suboptimal, there is usually a significant correlation between tacit and explicit policies. This suggests that subjects are aware of some features of their tacit policies. It could be these that they use to pick out their tacit policies and significantly often they hit on the correct policy. In other words although it suggests that subjects have explicit knowledge about some features of their policy that distinguish it from others, it does not suggest any greater self-insight than is already seen by the degree of similarity between tacit and explicit policies. However, if subjects select their explicit policies at greater rates than they select their tacit policies it would suggest that these are a better match for the

subject's explicit knowledge. As mentioned earlier Reilly and Doherty (1992) suggested that recognition of explicit policies might be enhanced by memory for factors unrelated to the shape of the policy.

Two hypotheses to be tested in this chapter are (1) that the degree of similarity between the policies being shown would affect the degree of recognition of both tacit and explicit policies and (2) that where doctors show greater self-insight in terms of similarity between tacit and explicit policy they would show greater self-recognition.

Method

Subjects

32 doctors who had participated in Study 1 participated in this study (see Chapter 5).[3] Four subjects were female. The average time elapsed since subjects were seen for the primary study was 10 months (range 8 to 13 months.) This study was run immediately after Study 2 (Chapter 6) and the consistency task of Study 1 (Chapter 5).

Task and instructions

For each of the original tasks in Study 1 (the LIPID, the MIGRAINE and the HRT tasks), a subject was shown a set of tacit policies from which they were asked to identify their own, then a set of explicit policies from which they were asked to identify their own. Instructions, provided on a laminated card described briefly the form of the policies in terms of relative importance. Subjects were also given a key to aid interpretation of the policies for each task and the meaning of this was discussed with the experimenter in the context of an example policy. The instructions, keys and example policies are shown in Appendices 35 to 39. The same examples, which were not real policies, were used for all the doctors. Doctors completed the task in their own time.

Tacit policies always preceded explicit policies on a task[4] and the same order was used with all doctors: recognition of the LIPID policies then recognition of the

[3] These same doctors did the consistency part of Study 1 (Chapter 5) and a subset of 30 participated in Study 2 (Chapter 6).

[4] Except where the data for a task had not been saved and so tacit policies could not be calculated.

MIGRAINE policies. Then two doctors did the HRT recognition task immediately. Six doctors completed the HRT consistency task prior to the HRT policy recognition task.

Twelve doctors' policies were presented in each set and each policy was displayed as a bar chart of cue weights on a separate laminated card. Each subject picked out three policies that might be their own and ranked these three in order of likelihood that they were. Eleven of the set of 12 policy bar charts presented to the doctor had been chosen at random, using a computer program[5], from the set of all doctors' policies for that task. The other was that doctor's own policy.

Formation of bar charts

The cue weights used were standardised regression coefficients where tacit policies were being identified and subjective ratings where explicit policies were being identified. For the LIPID and HRT tasks policies were formed from the standardised regression coefficients or subjective ratings obtained in Study 1 (Chapter 5). The height of the bar for a cue indicated its relative importance in the decision making. Negative weights were shown as such and fell below the x-axis. However, for the MIGRAINE task the sign of the cue response to acute treatment was changed so that a positive cue weight indicated increasing likelihood of prescribing for a patient with a good response to acute treatment. Consequently many of these became negative weights.

Different tasks and type of policy (tacit or explicit) had different coloured card:

Type of policy	**Colour of card**
Lipid tacit policies	White
Lipid explicit policies	Beige
Migraine tacit policies	Lemon yellow
Migraine explicit policies	Pink
HRT tacit policies	Bright yellow
HRT explicit policies	White

Cards were coded on the back with randomly allocated numbers.

[5] See Author's Declaration.

Feedback

Having selected policies for all tasks, doctors were told which policies they had correctly picked out. This was ascertained by matching up policy bar chart codes and doctor codes from a chart. Doctors were then given paper copies of all their policies to keep.

Experimenter bias

The first eight subjects (group A[6]) had a slightly different procedure from the remaining subjects (group B). This manifested itself in two ways: Firstly for subjects in group A, but not group B, the experimenter, who was present during the task, may have unconsciously been aware of the number code on the back of the correct policy and the pattern the policy took. Secondly, and in connection with this, subjects in group B but not group A received copies of their policies after the recognition task (for the purpose of feedback) in sealed envelopes.[7] To see if subjects in group A had unknowingly been cued by the experimenter in any way the results of group A and group B were compared.

Results and Discussion

Definition and probability of a 'hit'

If a doctor managed to pick out their card at all in the three guesses that they had this was counted as a successful recognition (a hit). Doctors picked three out of twelve cards. Therefore the probability that they would pick out their own policy by chance was 0.25.

[6] Group A consisted of GPs 11 13 14 19 20 28 30 and 37. Group B consisted of GPs 1 2 3 4 5 6 7 9 12 15 17 21 22 23 24 26 27 29 31 32 33 34 35 and 36.

[7] For all doctors the random selections of cards to be shown had been pre-generated. Selections of cards for subjects in group A were assembled some time during the 24 hours preceding the study. For this the experimenter looked up the code of each subject in order to ascertain which selection they should be given. At the same time the experimenter also selected the previously prepared bar charts of policies to be given to the subjects during feedback. Thus the experimenter may have been aware of which policy was the subjects during the study and may have unknowingly cued them in some way. For all subjects in group B the selection of cards to be presented was ascertained well in advance. At this time the policies to be given as feedback were all placed in carefully labelled envelopes which were then sealed up.

Table 7.1 Results for all doctors on the Policy Recognition task

	N	Number of doctors with a 'hit'	Proportion of hits	Significance[1]	1st choice[2]	A	B
LIPID tacit policies	30	17	0.57	$p < 0.001$	7	4/8	13/22
LIPID explicit policies	30	18	0.60	$p < 0.001$	12	3/7	15/23
MIGRAINE tacit policies	30	12	0.40	$p = 0.029$	5	2/8	10/22
MIGRAINE explicit policies	30	12	0.40	$p = 0.029$	6	3/8	9/22
HRT tacit policies	8	1	0.13	$p = 0.267$	0	1/3	0/5
HRT explicit policies	9	4	0.40	$p = 0.117$	2	1/3	3/6

[1] Significance of the number of hits is calculated from a binomial distribution with a probability of 0.25 of success over N (the number of doctors) trials.
[2] These were those doctors who ranked their actual policy first out of the three.

Table 7.1 shows the actual rates of own-policy-recognition as well as the number of doctors who ranked their actual policy first out of the three that they had chosen. As can be seen a significantly greater number of doctors were able to pick out their own policy than would be seen by chance. This is similar to the findings of other studies and indicates a certain level of self-insight (see Chapter 4 and introduction to this chapter). Appendix 40 shows which doctors picked out their own policies.

Experimenter bias

Rates of own policy recognition of doctors in groups A and B were compared in two ways. Neither of these found any experimenter cueing effect. Firstly an independent t test of the proportion of correct hits by doctors in groups A and B showed no significant difference ($t = 2.30$, $p = 0.07$).

Differences in self-recognition between doctors in groups A and B were also tested using Fisher's exact test, and no differences were found in recognition of any type of policy. Tacit and explicit recognition on the different tasks was looked at separately. The null hypothesis H_0 was that there was no significant difference between groups A and B on a type of policy on a type of task. This was never rejected in favour of H_1: recognition was greater in either group A or B. Differences between the two groups were not significant for any type of policy (Fisher's exact probability = 0.49, 0.27, 0.28, 0.60, 0.38, 0.76 for the LIPID tacit and explicit, the MIGRAINE tacit and explicit and the HRT tacit and explicit policies respectively).

Tacit versus Explicit policy recognition

A McNemar test on dependent samples was done on the difference between a doctor's recognition of tacit and explicit policies for each task. There was no significant difference between self-recognition of tacit and explicit policies in any of the tasks (LIPID: $\chi^2 = 0.10$, $p = 0.75$, $N = 29$; MIGRAINE: $\chi^2 = 0.07$, $p = 0.79$, $N = 30$; HRT: $\chi^2 = 1.30$, $p = 0.25$, $N = 8$). Subjects' recognition of their explicit knowledge was no better and no worse than that of their tacit knowledge.

Inter-task differences

Recognition of both tacit and explicit policies on the LIPID task were significantly better than on the MIGRAINE task. Twenty-eight doctors did the recognition task on both the LIPID task and the MIGRAINE task. A McNemar test on dependent samples was done using the results of these doctors: the number of doctors who recognised their tacit policy on the LIPID task but not the MIGRAINE task was significantly greater than the number who recognised their MIGRAINE tacit policy but not their LIPID tacit policy ($\chi^2 = 5.79$, $p < 0.05$). The same result was found when the calculation was done using the number of explicit policies recognised.

Own-policy recognition and self-insight

Self-insight had been measured in Study 1 (Chapter 5) as the correlation between subjective ratings and standardised regression coefficients for a doctor. This measure was available for each doctor for each task. Here the proportion of hits they had was the measure of own-policy recognition. It was hypothesized that the ability to pick out one's own tacit policy would be related to one's ability to state it explicitly. This would be refuted if doctors who had picked themselves out did not have significantly better self-insight than doctors who had not picked themselves out. On the LIPID and MIGRAINE tasks although average subjective-objective weight correlations were greater for those doctors who picked themselves out, the difference in Fisher's z transformations of this self-insight index was not significant. [Mean self-insight $r = 0.68$ for the doctors who picked their LIPID tacit policy out, compared to a mean of 0.63 in those doctors who did not; mean self-insight $r = 0.72$ for the doctors who picked their MIGRAINE tacit policy out, compared to a mean of 0.69 in those doctors who did not.]

Confounding effect of policy similarity

If the ability of doctors to pick out their own policy is not significantly related to self-insight there must be other factors which affect this greater than chance policy recognition. Certainly if a doctor is able to pick out their own policy they must have some insight as to the important features in that policy. However, in addition, they must be aware

of the features that distinguish it from the other policies available. Thus in the policy recognition task the similarity between the policies displayed may act as a confounding variable. This would affect recognition of both tacit and explicit policies on all tasks. Where policy recognition is suboptimal it may be due to forgetting (there was on average ten months' delay in this study), or it could be due to the confounding effect of similarity between policies, or in the case of tacit policies it could be due to limited self-insight.

Although doctors were better at stating their policies on the MIGRAINE task than on the LIPID task, ten months later they were better able to identify their LIPID policies than their MIGRAINE policies. This detriment occurred equally on both tacit and explicit policy recognition. Doctors were no better able to recognise their explicitly stated policies than the tacit ones we had calculated. One explanation is that forgetting occurred equally on tacit and explicit policies but was worse on the MIGRAINE task than the LIPID task. Another explanation is that, regardless of the amount of forgetting that occurred, it was harder for doctors to pick out their policies on the MIGRAINE task than on the LIPID task because of the greater agreement shown on this task. Greater similarity was shown both in terms of the tacit policies and in terms of the explicit policies. The greater similarity between explicit policies quashed the advantage of recognition explicit policies might have otherwise had.

Similarity between policies was measured in terms of the correlation between cue weights: either subjective ratings as in the case of the explicit policies or standardised regression coefficients as in the case of the tacit policies. These were calculated between each doctor's policy and all the others presented with it on the particular recognition task. An analysis of variance was done to compare the similarity ratings of doctors who had picked their policy out as either first choice, second choice, third choice or not at all.

Although there were significant differences between the recognition groups on the LIPID explicit policy recognition task ($F(^{8}2,327) = 11.7$, $p < 0.01$) and on the MIGRAINE tacit ($F(3,315) = 8.47$, $p < 0.01$) and explicit ($F(3,326) = 10.81$, $p < 0.01$) policy recognition tasks, this difference was not always in the way expected. On the explicit policy recognition tasks, those doctors who recognised their own policies actually had

[8] There were no doctors who picked out their policy at the third guess.

policies more similar to the ones they were looking at than the doctors who recognised themselves less accurately or not at all. On the MIGRAINE tacit policy recognition task doctors who picked out their policy in their third choice had significantly less similar policies to the other groups. There was no significant difference between groups on the LIPID tacit policy recognition task ($F(3,326) = 0.45$, $p = 0.72$).

So on the MIGRAINE task selection of tacit policies was somewhat affected by similarity between policies as predicted. However, the effect of policy similarity on recognition of explicit policies is odd. Study 1 showed that agreement between predictions from subjects' tacit policies was worse than agreement between predictions from their explicit policies on the LIPID and HRT tasks. Agreement was the same on the MIGRAINE task. Therefore recognition of explicit policies might be predicted to be worse than that of tacit policies on at least the LIPID and HRT tasks. However, as seen earlier, this was not the case.

It may be that the effect of similarity between explicit policies is itself confounded. There are two possibilities here: Where there is a consensus, similarity between policies may be great and, as predicted, it may be harder to pick out one's own policy. Firstly however, this may be exactly the situation where the subject knows their stated policy. Subjects who are aware of their policies may have more similar ones to each other. However, this would suggest an increase in self-insight with increased recognition that was not apparent earlier. Secondly, subjects may all tend to pick commonly agreed explicit policies on the recognition task. Then those subjects that actually stated those sort of policies in the first place, and showed agreement with a *status quo* would be more likely to pick their own. This does not suggest any greater self-insight on the part of those with better self-recognition.

Conclusion

Just as in the Reilly and Doherty studies (1989, 1992), own policy recognition for both types of policy in all tasks was at considerably greater than chance levels. Reilly and Doherty used four different types of cue design: 6 cues or 12 cues that were negligibly correlated or had correlations similar to how they would appear in real life. Although they

pointed out that recognition of own explicit policies was being affected by factors other than knowledge of the unique features of one's policy, for example memory of specific unusual numbers given, recognition of explicit policies was considerably better than recognition of tacit policies in three of the conditions. The fourth condition was similar to tasks used here in that there were 12 orthogonal cues used. In this condition explicit policy recognition was no different to tacit policy recognition but both were high. In this study where 13 orthogonal cues were presented, there was again no difference between tacit and explicit policy recognition.

Two hypotheses were tested in this chapter. Firstly, it was hypothesized that similarity had a negative affect on self-recognition. In Reilly and Doherty's study (1992) the condition with the least inter-subject agreement was the one with the greatest policy recognition. In this study too worse self-recognition was seen in the task in which there was greater similarity between policies (the MIGRAINE task). There was greater agreement between explicit policies than between tacit policies so worse self-recognition would be predicted if similarity between policies has a confounding effect on policy recognition. The lack of difference can be explained if explicit policy recognition would naturally be better than tacit policy recognition.

However, on both LIPID and MIGRAINE explicit policy recognition tasks, doctors who tended to recognise their own policy had policies more similar to others in the set they chose from than doctors who failed to recognise themselves. One possible explanation is that doctors who had well thought out policies and were more aware of them tended to have more similar policies to each other than those doctors who had thought less about the subject, did not follow the consensus and were less aware of their own policies. This relates to the second hypothesis tested that those who were better at picking out their tacit policies would have had a better correlation between tacit and explicit policies. In other words subjects showing better self-insight would also show better self-recognition. This was also not found to be the case. Alternatively, all subjects might show a tendency to pick out the policy matching some (socially agreed) ideal policy. If, ten months previously, that is what they had tended to state, a subject will have a greater chance of picking their own behaviour. In other words, those tending to give a commonly held explicit policy, and

therefore showing greater agreement with others would be more likely to pick themselves out. In this case, not only do subjects not have to have self-insight to have good policy recognition, they also need not remember what they explicitly said!

The Reilly and Doherty (1989 and 1992) results were replicated in that significant self-recognition was shown. However, although the results do not show any greater levels of metacognition than previously seen in self-insight measures, there is no significant correlation between self-insight and self-recognition.

Greater than chance performance on forced choice tasks however does not differentiate between identification of policies similar to one's own (well known one) and poor identification of one's own policy. It may be that the tacit policy describing one's own behaviour is the one on the forced choice task most similar to one's explicitly held policy. This would be selected, apparently showing perfect self-knowledge.

Chapter Eight
Study Four: Risk

Introduction

In this study doctors are asked, in two separate tasks, to make judgements about first their likelihood of prescribing lipid lowering therapy for a patient and second the patient's risk of coronary heart disease (CHD). Judgement analysis is carried out on both of these tasks. The purposes of this chapter are threefold. The first purpose it serves is to examine the relationship between judgements of risk and prescribing. Secondly, and related to this, two theories about the nature of self-insight shown in Study 1 are explored. Thirdly, the level of achievement seen in assessments of risk is measured in the framework of a lens model.

It has been hypothesized that inter-individual differences in patient management may be partly due to differences in perceived risk or probability of a disease (Poses, Cebul and Wigton, 1995). Lipid lowering drugs should, according to the pharmaceutical and general practice advisory literature, be prescribed in the presence of multiple risk factors for, or evidence of, coronary heart disease (CHD) (e.g. British National Formulary 28, September 1994; GP Pocket Guide to Cholesterol, 1994; Grant, 1992, p. 136). Several risk factors have been identified in the medical literature (Heller, Bailey, Gott and Howes, 1987) and studies such as the Framingham Heart study have demonstrated their relative effects on mortality (see Anderson, Wilson, Odell and Kannel, 1991). The relationship between risk of coronary heart disease and prescription of lipid lowering therapy is similar to that between probability of any disease and an appropriate treatment for it in that an increase in perceived probability or risk would lead to an increase in likelihood of treatment. Doctors may differ in treatment because they differ in likely diagnosis. Doctors may also have different probability thresholds at which they would treat. Study 1 (Chapter 5) showed that there are differences in the policies used to prescribe lipid lowering therapy. In this chapter use of lipid lowering therapy and the factors affecting it will be compared to risk judgements and the factors affecting those. The relationship between prescribing and perceived risk will be examined in terms of whether cases at higher risk are more simply

likely to be prescribed for or whether other factors affect the decision. Behaviour on Study 1 suggests that the latter is the case: Subjects were influenced by non-clinical cues and some cues were influential in more than one way.

Assume here that, as the ideal practice described in the literature suggests, doctors' decisions to prescribe are influenced in some way by their judgement of the patient's risk. Doctors may disagree about prescription because they disagree about a patient's degree of risk; or they may disagree about prescription because they have because they have different risk thresholds for prescription; or they may have different ideas about what to do with that high risk person. If doctors have different risk judgement policies then, where cues are orthogonal as in these studies, they will have different prescription decision policies. On the other hand doctors may actually agree about the factors influencing a person's risk but disagree about how to treat them. In this study the degree of individual variation will be compared for the prescription and risk judgement tasks. If agreement between doctors on the risk judgement task is the same as or less than that on the prescription task it indicates that individual differences on the risk task may contribute to individual differences in prescription. If agreement on the risk judgement task is however better, disagreement between doctors must lie at the level of patient management. There are two reasons for expecting this latter result. Firstly, there is more of an explicit criterion in the medical literature as to what are risk factors than as to what should influence prescription. Secondly, the doctors' explicitly stated policies on the LIPID task suggested that many aspects of the decision were being driven by total patient management and use of factors not related to risk such as patient attitude rather than considerations of risk of CHD. An example of this is the explicitly stated, manipulative decision not to prescribe to smokers in order to persuade them to give up. However, this last interpretation of the results of Study 1 will only be borne out in a direct comparison of risk judgements and prescription decision policies.

Risk of CHD is expressed as a continuous variable - as is probability - and thus it is suitable for judgement analysis. In this study subjects were asked to estimate risk along a visual analogue scale anchored by low risk and high risk. There are two reasons for this. Firstly, in real consultations, although general practitioners do make judgements about the

severity or probability of disease this is not a numerical assessment but is more a pragmatic judgement. Secondly, as is shown below, one off numerical values given by subjects in experiments do not reflect their underlying perception of risk or probability.

Studies looking at physicians' estimates of risk and probability almost invariably show there is a tendency to overestimate (e.g. Christensen-Szalanski and Bushyhead, 1981; Tape and Wigton, 1989; Poses, Cebul, Collins, and Fager, 1985; Poses, Cebul, and Wigton, 1986; Poses, Cebul, and Wigton, 1995; Tape, Kripal, and Wigton, 1992). Since probabilities and risks in these studies tend to be small this is in accordance with the general finding that subjects tend to overestimate small probabilities and underestimate large probabilities (Baron, 1988, pp. 198-204). An exception is the accurate estimation of risk of atherosclerotic heart disease (ASHD) by subjects who had a family history of the disease (Tape and Wigton, 1989). This may be because subjects are aware of actual probability values. Indeed calibration can be improved using outcome feedback (e.g. Tape, Kripal, and Wigton, 1992; Poses, Wigton and Cebul, 1995). However, the numerical estimates of probability or risk may not get to the underlying psychological representation of likelihood. It could be argued that a subject showed a good understanding of the situation if he or she rated less likely cases of a disease as less likely, more likely cases of a disease as more likely and used the information that was relevant to the diagnosis in doing so. Christensen-Szalanski and Bushyhead (1981) presented results in which, although physicians overestimated the probability of disease (numerically), they were influenced by a symptom's predictive value of the disease in their judgements: more likely cases were assigned higher probabilities. Measures of accuracy of probability estimates that only look at the nearness of exact probability values (looking at calibration) do not take this into account. Similarly, two individuals could be said to show agreement if they rated one case as higher risk and another as lower risk even if they gave different numerical responses.

There is evidence that prescribing is related to probability estimates and that overestimation of probability coincides with unnecessary prescribing (see Poses, Wigton and Cebul, 1995). However, several other factors could influence the decision to prescribe and lead to unnecessary prescription. Using cognitive feedback and feedback of monthly prevalence data, Poses, Wigton and Cebul (1995) found that despite improved calibration

through a reduction in the experimental group's estimations of the probability of streptococcus pharyngitis in patients with sore throats, the aggregate level of antibiotic prescribing actually increased significantly. For these doctors the relationship between clinical variables and prescribing seemed fairly consistent, This was despite changes in accuracy, and possible changes in discriminating ability, of physicians' probability judgements. Whereas it may be the case that inter-individual differences in probability estimates may reflect differences in perceived probability of a case having a disease, it appears that recalibration of subjects' estimates does not reflect an underlying change in the perception of probability of the disease and the need to do something about it. In this study then the pattern of interest is whether a doctor is more likely to prescribe for cases they perceive to be at high risk; not whether doctors who perceive patients to be at higher risk were more likely to prescribe.

In the above discussion, agreement between subjects refers to the agreement between the policies they actually use. Agreement between explicitly stated policies is something else. In Study 1, the importance of a number of risk factors on the LIPID task, such as Hypertension, Evidence of arteriosclerosis, Gender and Family history of ischaemic heart disease, were over-rated (see Figure 5.4, Chapter 5). Although, in Chapter 6, GPs' ratings of cues importance was found to correlate well with the way they selected cues, this is of course no proof that what they said was based on what they had been looking at. Two possibilities will be discussed here. One is that the rating of importance related to clinical rather than relative importance. In the LIPID task for example, cues can be very important in clinical terms in that they increase risk of CHD and they would be selected in order to find out the risk of CHD. But in terms of influence on the decision making the same cue might be much less important than rated. Doctors could have rated these cues as important because risk is increased and it is important to do something. But in this set of cases these cues' importance was less when considering the decision whether or not to use a lipid lowering drug. Similarly on the MIGRAINE task, over-rating tended to occur on the cues that indicated the effect that the migraine was having on the patient's life. For example, Misses work and Nausea or Visual disturbance were all over-rated (see Figure 5.5). Importance here could have been in terms of severity: It was important to do

something about this person's problem though this was not necessarily prophylaxis. Subjects' self-insight on the risk task should throw light on whether this is the case. If subjects are giving assessments of the clinical importance of cues, they should not show the pattern of over-rating where clinical severity judgements were made.

Although the above suggestion is that subjects reinterpreted the notion of importance, why they would do that has to be questioned. The possibility, introduced in Chapter 6, is that subjects are giving some explicitly held knowledge about judgement policy. This is similar to an ideal policy and relates to Nisbett and Wilson's (1977) suggestion that subjects' reports are based on a priori theories and Nisbett and Ross's (1980) suggestion that subjects' self-knowledge is to a degree theorization about their own behaviour. If subjects' policies are based on theorizations about their behaviour they may be describing how they think the task should be done. A number of risk factors were being over-rated in Study 1. Some risk factors, although stated as being of relevance, were in fact being neglected. The hypothesis here is that they were neglected because of limits to information processing capacity.

Since some criterion measure of actual risk is calculable from epidemiological data it is also suitable for a lens model analysis. Lens model analysis, described in Chapter 3, examines a subject's ability to state a correct judgement (their achievement) in terms of the consistency of implementation and appropriateness of their judgement policy. Predictions from models derived from epidemiological data form the environmental side of the lens model. The equations involved in calculation of the criterion are however, not linear. Tape, Kripal and Wigton (1992) compared subjects' numerically stated judgements of the risk of death from heart disease in the next ten years with those calculated from a model derived from Framingham data. Both matching (correlation between predictions from the model of the criterion and predictions from the model of the judgements) and achievement (usually correlation between criterion values and judgements) were measured using predictions from this model on the criterion side. Initial achievement levels were around $r = 0.4$ or 0.5. Matching was around $r = 0.6$ and 0.7. The coefficient of multiple correlation (R) was about 0.8. However, if linear equations can be used to model risk judgements then they could be used to model risk calculations and estimates of the relative importance of cues in subject

and criterion linear models can be compared. This is done here.

So, perhaps it is the case that doctors are looking at the factors that will tell them how serious the case is but these do not have equal bearing on the decision whether or not to prescribe the particular drug. Here this was tested by comparing individual doctors' ratings of relative importance for a judgement of risk of CHD and their self-insight into the factors affecting their judgements of risk with their ratings of relative importance of cues in prescription decisions. Self-insight is looked at on a risk assessment task where importance in clinical terms would relate well to importance in terms of impact on the decision.

Method

Subject recruitment

A mailshot, including a freepost return envelope and reply slip, was sent to 127 GPs with Cornish "PL" postcodes. Forty-one of these (32.3%) initially responded positively. Thirty-six (28.3%) were able to participate for 2 hours, during the data collection period (January and February 1995). Payment was at the rate of £25 per hour for their time and all participants were promised and given feedback[1].

Subjects

36 General Practitioners from a variety of settings in Cornwall participated in this study. Practices included several health centres and one single-handed practice in town centres as well as rural locations. No doctors had participated in previous studies in this project or were members of practices who had. Three participants were female.

Environment

Most doctors were visited at their practice and carried out the tasks in their consulting rooms. Two doctors were visited at home.

[1] This was sent by post at the end of March 1995 and consisted of barcharts showing that doctor's tacit and explicit policies for each task as well as average tacit and explicit policies for each task.

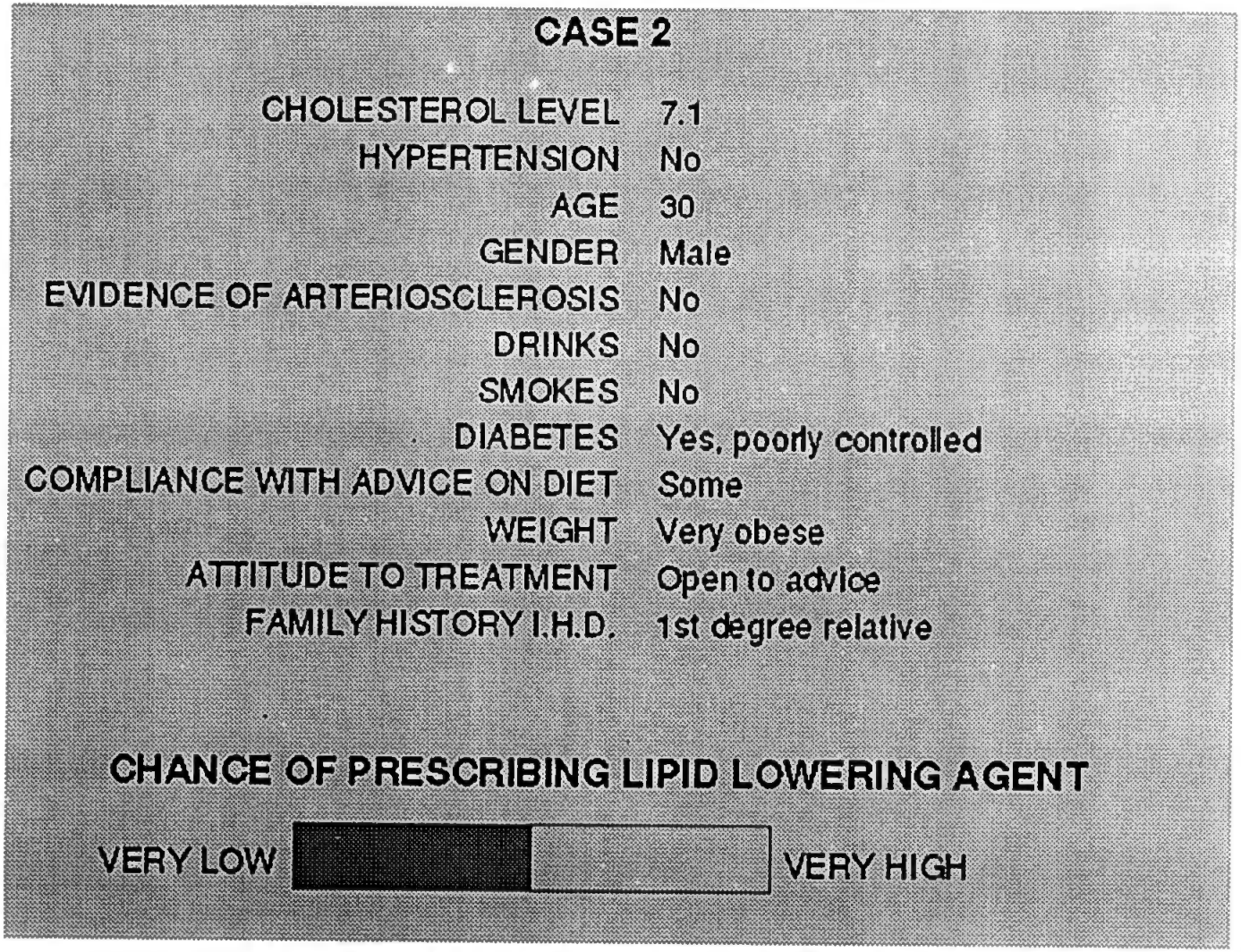


Figure 8.1 An example of a case from the PRESCRIBE task

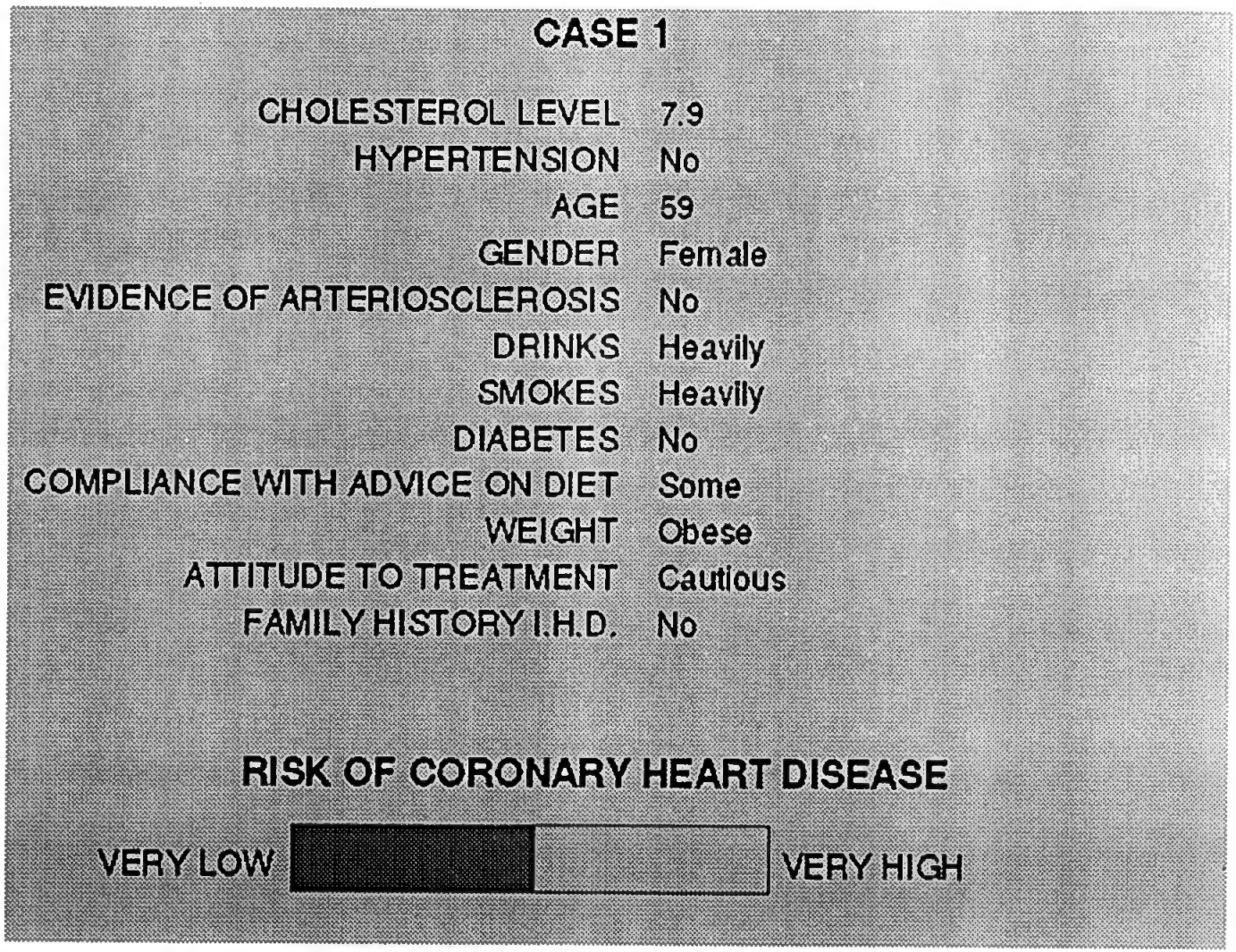


Figure 8.2 An example of a case from the RISK task.

Tasks

Two computer presented tasks were used in this study. Each was followed by a tape recorded discussion as to the factors affecting decision making on the task. The majority of doctors completed both tasks, presented in the same sequence, in one session. One doctor did the second task three weeks later.

The first task (PRESCRIBE) presented a sheet of instructions followed by a series of hypothetical cases described along 12 dimensions (cues). The doctor made a judgement about the likelihood of prescribing a lipid lowering drug for each case. This task was similar to the original LIPID task of Study 1 (see Chapter 5). Again LIPID always refers to the lipid lowering decision task of Study 1. However, the cues and instructions used were slightly different and the bar used to indicate the doctor's chance of prescribing was anchored by "Very low" and "Very high". An example of a case is shown in Figure 8.1.

The second task (RISK) presented the same cases in the same order. However, this time the doctor judged the 'patient's' risk of coronary heart disease and the response bar labelled "Risk of Coronary Heart Disease" was anchored with "Very low" and "Very high". An example of a case from the RISK task is shown in Figure 8.2.

The discussion after each task was designed to identify which cues the doctor felt had been affecting their decisions or judgements and how their behaviour related to their judgement and decision making behaviour in real life.

Case generation and instructions

100 cases were generated for the tasks and these were described along the twelve dimensions shown on the left hand side of Table 8.1. The greatest intercue correlation for these 100 cases was between Age and Diabetes (-0.165). All inter-cue correlations are shown in Appendix 41.

Eleven of these cues and their ranges were identical to those in the LIPID task (see Chapter 5). Two of the LIPID task cues (personality and occupation) were dropped because of their minimal use. One new cue, describing the patient's drinking behaviour, was added in their place.

Table 8.1 Cues and their ranges on the PRESCRIBE and RISK tasks

Cue	CUE WEIGHT: NEGATIVE ← → POSITIVE Range
CHOLESTEROL LEVEL	*6.5-8*
HYPERTENSION	*No/Yes, well controlled/Yes, poorly controlled*
AGE	*30-60*
GENDER	*Female/Male*
EVIDENCE OF ARTERIOSCLEROSIS	*No/Yes*
DRINKS	*No/Occasionally/Regularly/Heavily*
SMOKES	*No/Occasionally/Regularly/Heavily*
DIABETES	*No/Yes, well controlled/Yes, poorly controlled*
COMPLIANCE WITH ADVICE ON DIET	*No/Some/Yes*
WEIGHT	*Under/Normal/Over/Obese/Very obese*
ATTITUDE TO TREATMENT	*Opposed/Cautious/Open to advice/Requesting treatment*
FAMILY HISTORY I.H.D.	*No/2nd degree relative/1st degree relative*

A few doctors in the LIPID task had mentioned that in real life they might want to know the patient's exercise level and alcohol consumption in addition to the information they had been presented on the task. In the instructions for the PRESCRIBE and RISK tasks the exercise level of all patients was stated to be appropriate to their age and general health. The instructions stated that the triglyceride levels of the patient's blood sample were normal. This information was added because in the original study the few doctors that expressed a wish to know the patient's alcohol consumption had done so because triglyceride levels may be raised in heavy drinkers and this could be the cause of a raised cholesterol level. Instructions included additional assumptions about the patient that had been included in the LIPID task. Full written instructions for the prescribing and risk tasks are shown in Appendices 42 and 43.

Consistency

The first thirty cases were repeated at the end of all 100 cases as a measure of consistency. They were renumbered on their second presentation (Case 101- Case 130).

Post task interview

The format of this was similar to the post task interview of Study 1 (Chapter 5). Again firstly an effort was made to quantify the doctor's perception of each cue's relative importance to the decision. He or she was asked which of the values at either end of a cue's range would be more likely to make them prescribe, other things being equal, or if there was no difference between the two. Then the doctor was asked to give a rating between 0 and 10 to each cue to indicate how much of a bearing it had on their decision. The doctor was told that '0' indicated that the dimension had no bearing on the decision, '10' indicated that it had maximum bearing on the decision and that he or she could allocate the same number to more than one dimension. Doctors were able to see the ratings they had given and so could compare these. The self-insight sheet seen by the GPs is shown in Appendix 45. The verbal instructions given at this stage are shown in Appendix 44.

Having rated each dimension the doctor was given an opportunity to discuss his or her strategy on the task and to indicate for example whether the effect of one particular dimension was dependent on the value of other dimensions.

They were also prompted to talk about their behaviour, with references to cases such as these, in real life. They were asked if they did consider prescribing lipid lowering drugs, or make judgements about risk in real life. Each doctor was asked what sorts of factors would be affecting the decision to prescribe lipid lowering drugs in real life or what sorts of factors would be affecting judgements about the degree of risk they felt a real patient had.

Pilot

The study was piloted on two doctors. One of these was a GP in the region and, since nothing was changed, his data was included in the general analysis.[2]

[2] The other was Dr. Mark Harries.

Results and Discussion

The main judgement analysis was carried out using judgements on the last 100 of the 130 cases presented. In this way, where cases had been repeated the second judgement on them was used. A correlation between original responses (on the first 30 cases) and second responses (on the last 30 cases) was used as a measure of consistency. Tacit and explicit policies were calculated as in Study 1 with tacit polices as a set of standardised regression coefficients and explicit policies as a set of subjective ratings or weights. The signs of these were assigned as indicated in Table 8.1. Four types of analysis were carried out. Firstly, the factors affecting decisions on the PRESCRIBE task were compared with those on the original LIPID task of Study 1 and the IS task of Study 2 (Chapters 5 and 6). Secondly, factors affecting judgements of risk were compared to those affecting prescription on the same set of cases. Thirdly, the degree of self-insight shown on the RISK task was compared with that shown on the PRESCRIBE task. Fourthly, a Lens model analysis of subjects answers on the RISK task was done using data from the Framingham study as a criterion.

Agreement

Agreement between doctors over the 100 cases was slightly worse on the PRESCRIBE task with judgements about prescription (Kendall's W = 0.36) than that on the RISK task with judgements about risk (W = 0.39). Whether this is a significant difference cannot be calculated (Kendall 1975, 6.14, p. 102). Kendall's W measurement of concordance can be converted into an average Spearman's rank correlation coefficient for all pairs of doctors (Howell, 1982, p. 229). The corresponding average Spearman's correlations are 0.34 for the PRESCRIBE task and 0.37 for the RISK task. Agreement about prescribing was similar to that seen on the IS task of Study 2 where a slightly different format of presentation was used (W = 0.35). But both were better than that on the judgements on 100 cases in the LIPID task of Study 1 which had the same format (W = 0.28). As discussed in the introduction, better agreement was expected on risk judgements than on prescription judgements if the latter are affected by agreement between subjects as to risk assessment as well as differences in management policy. However, the fact that the

difference is only marginal (if at all) is also of interest. A considerable amount of disagreement was shown in judgements about risk as well as in decisions about prescription.

Kendall's W (concordance) was also calculated across judgements predicted from standardised regression coefficients (doctors' tacit policies) and from subjective weights (doctors' explicit policies). Kendall's W for the predicted judgements were greater on the RISK task (0.62 and 0.82 for tacit and explicit policies respectively) than on the PRESCRIBE task (0.54 and 0.65[3] for tacit and explicit policies respectively) as predicted. There is greater agreement in the medical literature about how to judge risk of coronary heart disease than when (and whether) to treat with lipid lowering drugs. Correspondingly there was greater agreement between policies. Values of concordance are higher than the Kendall's W measures between doctors' actual judgements given earlier. Doctors' actual judgements may be more diverse because they are subject to the doctor's level of inconsistency. Judgements predicted from a model did not have this element of random error. This greater agreement between subjects on the RISK was not large in subjects' actual judgements but these are affected by both consistency and policy and the greater agreement in policy is perhaps offset by a reduction in consistency on this task. Other analyses will further unpack the relationship between the two judgement tasks.

Comparison between PRESCRIBE, LIPID and IS policies

Despite a complete change of subjects and a slight change in the cues presented, where cues were the same, their use was similar on the PRESCRIBE task to that on both the LIPID and IS tasks. The average regression coefficients of the PRESCRIBE task, shown in Figure 8.3 appear similar to those of the IS task seen in Figure 6.5 (Chapter 6). Table 8.2 summarises cue use by doctors on the PRESCRIBE task in a similar manner to the LIPID cue use summary in Table 5.11 and the IS cue use summary in Table 6.3. Age, Smoking and Diabetes again act as both indicators and contra-indicators to prescribe. Differences are that Hypertension and Compliance with advice on Diet did not act as both

[3] Agreement here is between the 30 doctors whose policy could be expressed in linear terms. Otherwise in all cases agreement is calculated on predictions of judgements on 100 cases.

indicators and contra-indicators in the PRESCRIBE task; fewer doctors took account of diabetes, whilst more took account of Evidence of Arteriosclerosis, Attitude to treatment and Family History of ischaemic heart disease.

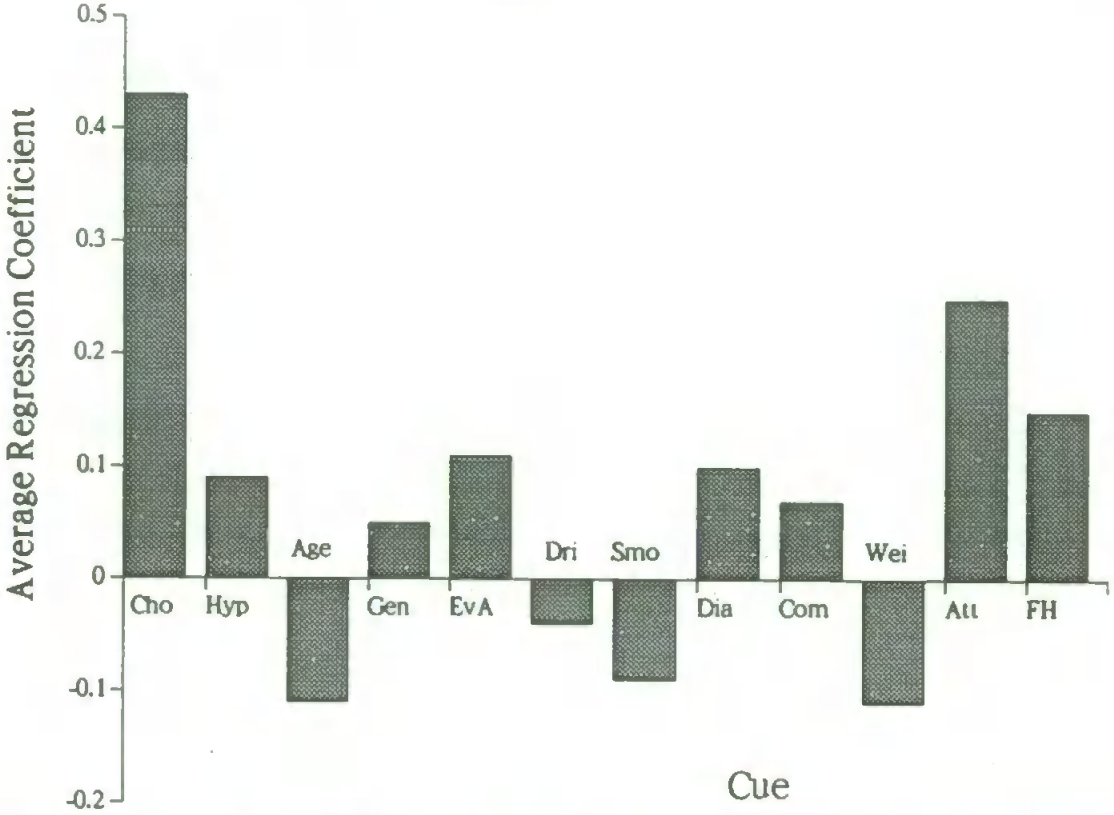


Figure 8.3 Average standardised regression coefficients for doctors on the PRESCRIBE task (N = 36).

Explicit policies can also be compared on these different tasks. However, not all doctors were able to express their PRESCRIBE explicit policies in linear rating terms. For six of the doctors the ratings they attached to cues on the PRESCRIBE task could not all be allocated linear signs. Thus the doctors do not have linear explicit PRESCRIBE policies. There was either an interaction as in the case of Attitude and Smoking for GP40 or a simple non-linear function as Weight took for GP41. The remaining doctors expressed policies which were again remarkably similar to those expressed by subjects in the LIPID task of Study 1. Average subjective ratings of the two tasks in Tables 5.8 and 8.3 are similar both in size and in sign and there are similar levels of dissension as to the effects of Smokes, Diabetes, Compliance with advice on diet and Weight. However, Hypertension in the PRESCRIBE task is only ever rated as an indicator to prescribe. Subjective ratings for all GPs are in Appendices 51 and 52. Subjective weights are in Appendices 53 and 54.

Table 8.2 Standardised regression coefficients PRESCRIBE task (N = 36 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	34	2	0	0.43	0.20
HYPERTENSION	9	27	0	0.09	0.09
AGE	1	23	12	-0.11	0.14
GENDER	5	31	0	0.05	0.12
EVIDENCE OF ARTERIOSCLEROSIS	11	25	0	0.11	0.12
DRINKS	3	28	5	-0.04	0.14
SMOKES	3	22	11	-0.09	0.19
DIABETES	11	23	2	0.10	0.14
COMPLIANCE WITH ADVICE ON DIET	6	30	0	0.07	0.10
WEIGHT	0	23	13	-0.11	0.12
ATTITUDE TO TREATMENT	24	12	0	0.25	0.18
FAMILY HISTORY I.H.D.	15	20	1	0.15	0.16

Key:- +ve, -ve = significantly positive or negative regression coefficients, n.s. = not significant.

Table 8.3 Subjective ratings PRESCRIBE task (N = 36)

Cues	+ve	zero	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	36	0	0	8.53	1.30
HYPERTENSION*	31	4	0	5.67	2.33
AGE	1	4	31	-5.74	3.12
GENDER	24	12	0	3.13	2.87
EVIDENCE OF ARTERIOSCLEROSIS	33	3	0	6.07	2.78
DRINKS	6	17	13	-1.42	4.09
SMOKES*	15	3	17	-0.59	6.65
DIABETES*	30	0	3	5.76	4.09
COMPLIANCE WITH ADVICE ON DIET	25	8	3	3.64	4.49
WEIGHT*	4	9	22	-3.00	3.99
ATTITUDE TO TREATMENT*	32	3	0	5.21	3.29
FAMILY HISTORY I.H.D.	36	0	0	6.97	1.68

* Some ratings could not be assigned signs since non-linear influences were described. This data is omitted.

Linear modelling and consistency

There was no significant difference between linear fits of tacit policies calculated from judgements on the last 100 cases of the PRESCRIBE task and the RISK task. The averages for both are seen in Table 8.4. The average difference of Fisher's z transformations of R were 0.017 ($t = 0.4$, $p = 0.69$, $N = 36$). Just as in Study 1, on both tasks there was a significant correlation between Fisher's z transformations of the multiple correlation coefficient (R) and consistency of judgement making ($r = 0.373$ and $r = 0.651$ on the PRESCRIBE and RISK tasks respectively).

Table 8.4 Mean Consistency and Linear fit of models for each task

Task	N	Mean consistency[4]	Mean R^2	St. dev. R^2
PRESCRIBE	35	0.52	0.52	0.13
RISK	35	0.62	0.53	0.10

However, as can be seen from the averages in Table 8.4, consistencies were significantly better on the RISK than on the PRESCRIBE task ($t = 2.1$, $p = 0.04$, $N = 36$). Individual subjects' consistencies and linear fits are shown in Appendices 46 and 47. The degree of agreement between subjects on tasks, described earlier, is affected by both agreement in policies and the subjects' consistencies. Where a subject shows less consistency (agreement with themselves) they will show less agreement with others. This again suggests that the greater agreement in judgements on the RISK task reflects a greater agreement in policy that is dampened by the level of consistency.

Both tasks showed higher consistencies than the LIPID task and lower consistencies than the MIGRAINE task of Study 1 where there was a delay of 10 months before measurement of consistency. Less consistency might have been expected on these tasks where there was this delay. It may have been the case that where both consistency and linear fit were low it was because there was little variance in the subjects' responses. However, the correlation between Fisher's z transformations of the multiple correlation coefficient and standard deviation was not significant for either task ($r = 0.27$ and $r = 0.09$

[4] Consistency was the correlation in responses over 30 repeated cases.

for the PRESCRIBE and RISK tasks respectively). Nor was that between standard deviation and Fisher's z transformations of consistency (r = 0.238 and r = 0.156 for the PRESCRIBE and RISK tasks respectively). In other words when consistency was low linear fits tended to be low too and it was not because there was little variance in response.

Risk judgements and prescription of lipid lowering drugs

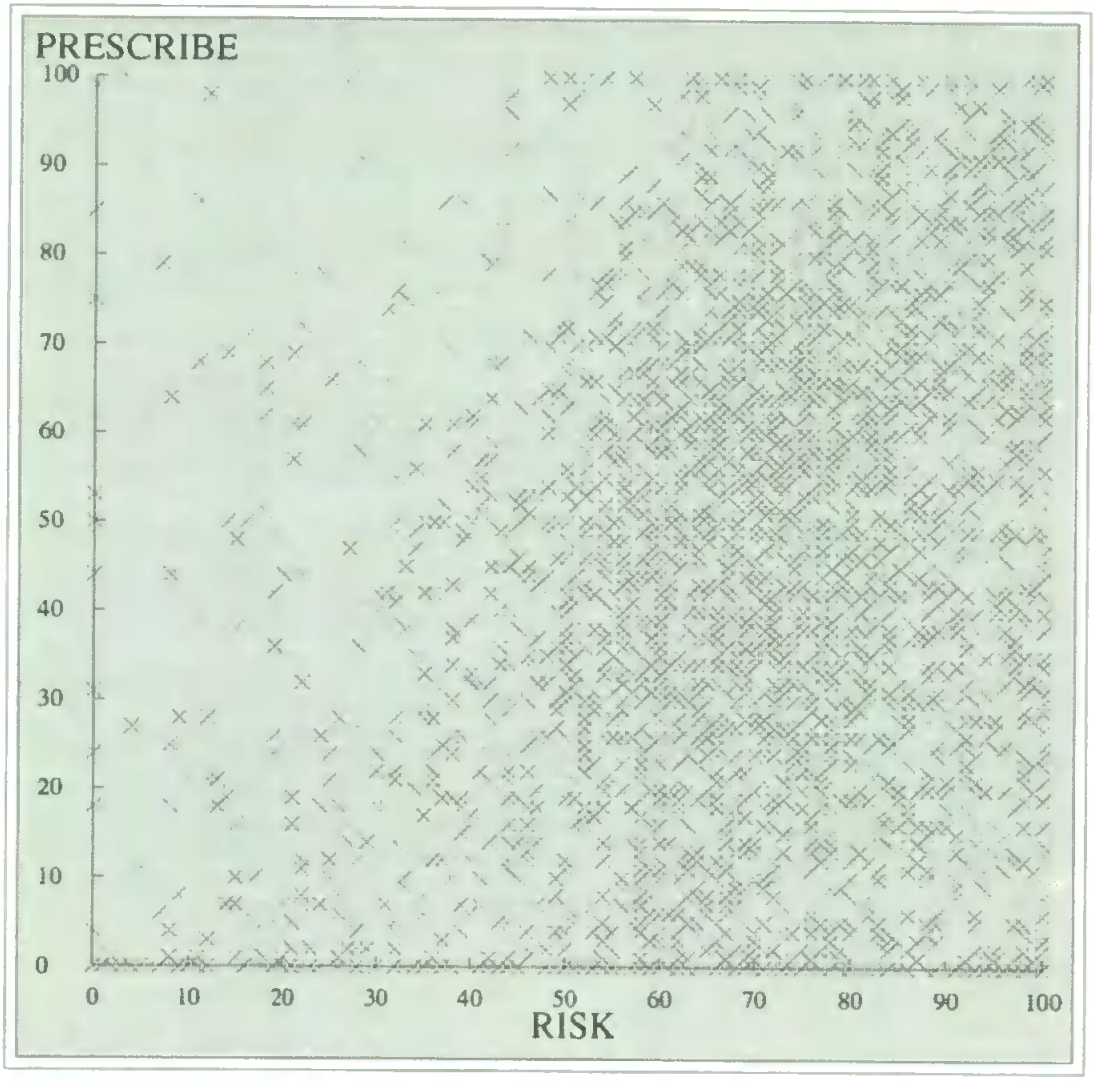


Figure 8.4 Scatter plot showing the relationship between 36 doctors' judgements of risk of Coronary Heart Disease (RISK) and their likelihood of prescribing a lipid lowering drug (PRESCRIBE) for 100 hypothetical patients.

Correlations between doctors' judgements, latencies, regression coefficients and subjective ratings on the PRESCRIBE and RISK tasks are shown in Appendix 55. The correlation between doctors' judgements of risk and their likelihood of prescribing for the last 100 cases is on average significantly different from zero (mean r = 0.22, st.dev. = 0.24,

[5] $t = 5.23$, $p < 0.01$). Figure 8.4 shows this relationship which seems conditional: if the case was prescribed for it will have been rated as a high risk. Prescription rarely occurred when the patient was perceived as having a low risk of coronary heart disease (CHD) but patients rated at high risk of CHD may or may not have been prescribed for. This group of untreated high risk patients might include those who were, for example, at high risk but only have moderate cholesterol levels, are opposed to treatment, who smoke, are overweight or who drink heavily. To test this likelihood of prescription for cases with risk judgements above the doctor's median were correlated with the relevant cues values. These correlations are shown in Appendix 50.

Twenty-nine out of the 36 doctors had a significant positive correlation between cholesterol level and prescription for these cases. For the majority of doctors then, cases at high risk but not being prescribed for had relatively low cholesterol levels. Not to prescribe a lipid lowering drug for someone with relatively low cholesterol levels seems quite rational behaviour. However, the range was such that the lower limit of the cholesterol level cue was 6.5 mmol/l. The range was set thus so as not to be too low for treatment. Given the assessment of high risk doctors might have been expected to prescribe for patients with this level of cholesterol. Fourteen out of 36 doctors had significant positive correlations between attitude and prescribe over these higher risk cases. So some doctors would be less likely to prescribe lipid lowering treatment for high risk patients who were opposed to treatment. For seventeen doctors the correlation between smoking and prescribe was significant and negative. In other words, although these patients were still perceived as at high risk, those who smoked were less likely to be prescribed for by these doctors. The correlation between weight and prescribing on these higher risk patients was significantly positive for five doctors and significantly negative for one. So five doctors were less likely to prescribe for overweight people even though they perceived them as being at high risk. One doctor was more likely to prescribe for overweight people at high risk. Finally, drinking was significant and negative for two doctors. These were less likely to prescribe for heavy drinkers, even though they were at high risk. For two doctors the correlation was significant and positive. However, over all cases, drinks was not a significant factor for

[5] t-test of Fisher's z transformations.

these doctors. In other words two doctors were more likely to prescribe for drinkers if they were at high risk but were not influenced by drinking behaviour generally. This indicates non-linear cue use. In subjective ratings of relative importance, one indicated no influence of drinking behaviour overall and the other indicated being somewhat but negatively influenced (a rating of -4). The relationship between non-linear behaviour and self-insight is discussed in Chapter 9. These correlations indicate that doctors differ in terms of patient management policy even for high risk patients. These differences in policy are not due to differences in perceived risk of the patient.

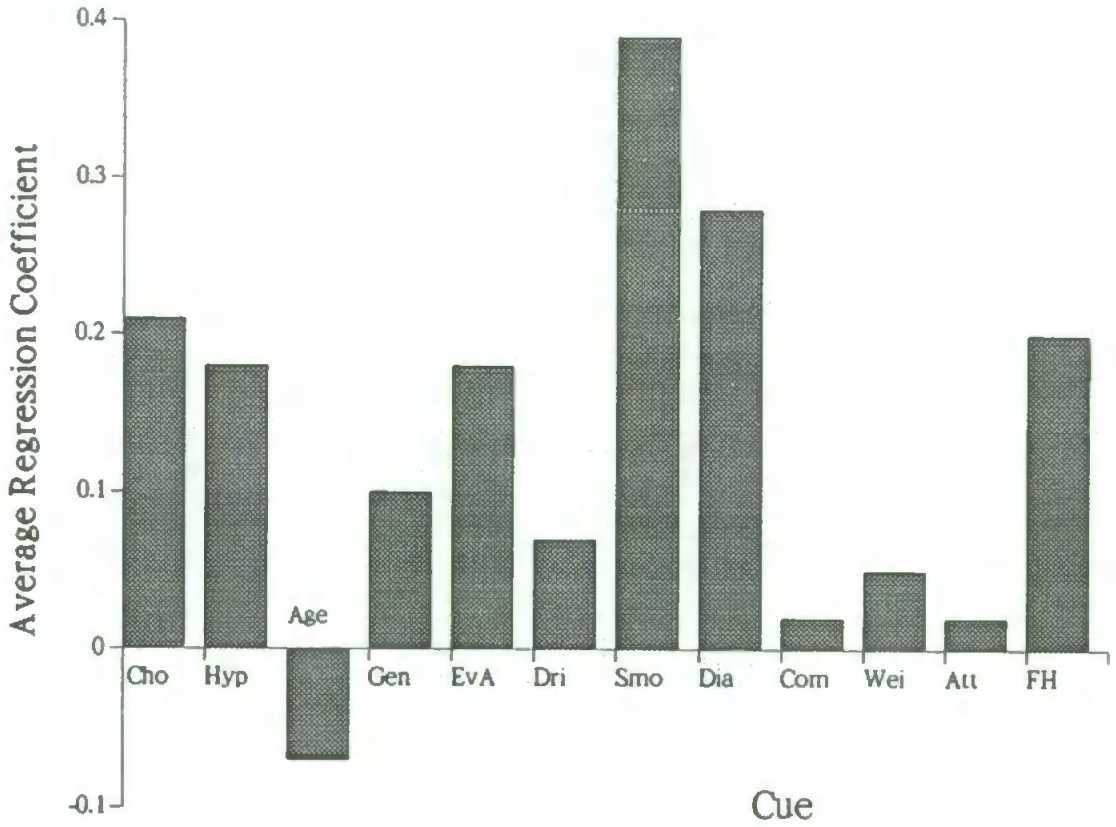


Figure 8.5 Average standardised regression coefficients for doctors on the RISK task (N = 36).

With knowledge of the guidelines on prescription of lipid lowering drugs, outlined in the introduction, and having seen the relationship between judgements made on the two tasks (Figure 8.4) it might be assumed that decisions to prescribe encompass some sort of risk assessment. In this case cues relevant to risk should form a subset of those affecting a doctor's judgements on prescription. There is some evidence for a similarity when cues' regression coefficients on the two tasks are correlated (mean = 0.21, st.dev. = 0.39, t =

2.99, $p < 0.01$). However, the average pattern of cue use on the RISK task (Figure 8.5) is different from that on the PRESCRIBE task (Figure 8.3).

Although an average of four cues were significant for doctors on the PRESCRIBE task five were significant on the RISK task. This is similar to Study 1 and Study 2 where on average four cues were used on tasks. This is again suggestive of limits to information processing capacity as discussed in the previous chapters. However, this fact alone questions the possibility of cues influencing risk judgements being a subset of those influencing prescribe judgements.

Table 8.5 Standardised regression coefficients RISK task (N = 36 doctors)

Cues	+ve	n.s.	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	26	10	0	0.21	0.15
HYPERTENSION	24	12	0	0.18	0.11
AGE	1	28	7	-0.07	0.13
GENDER	10	26	0	0.10	0.14
EVIDENCE OF ARTERIOSCLEROSIS	19	17	0	0.18	0.22
DRINKS	5	31	0	0.07	0.09
SMOKES	32	4	0	0.39	0.17
DIABETES	28	8	0	0.28	0.14
COMPLIANCE WITH ADVICE ON DIET	0	0	0	0.02	0.06
WEIGHT	8	27	1	0.05	0.14
ATTITUDE TO TREATMENT	0	0	0	0.02	0.05
FAMILY HISTORY I.H.D.	20	15	1	0.20	0.17

Key;- +ve, -ve = significantly positive, negative regression coefficients, n.s. = not signficiant.

Table 8.6 Subjective ratings RISK task (N = 36)

Cues	+ve	zero	-ve	Mean	St. dev.
CHOLESTEROL LEVEL	36	0	0	5.83	1.96
HYPERTENSION	35	1	0	6.64	1.69
AGE	11	8	17	-1.58	4.88
GENDER	28	6	2	3.72	3.08
EVIDENCE OF ARTERIOSCLEROSIS	33	3	0	6.46	3.27
DRINKS	29	6	1	3.38	2.88
SMOKES	36	0	0	8.61	1.22
DIABETES	36	0	0	7.81	1.31
COMPLIANCE WITH ADVICE ON DIET	1	21	14	-1.36	2.27
WEIGHT	31	5	0	3.93	2.17
ATTITUDE TO TREATMENT	1	33	2	-0.33	1.67
FAMILY HISTORY I.H.D.	36	0	0	6.82	2.27

Several cues are on average used in the same way on the PRESCRIBE and RISK tasks. Cholesterol has a much greater impact on prescription decisions than risk judgements and attitude to lipid lowering treatment and compliance with advice on diet did not influence any doctor's judgements of risk. The opposing use of cues by different doctors in the PRESCRIBE task can also confuse comparisons of averages. For example, on the PRESCRIBE task the regression coefficients' averages for smoking, drinking and weight were negative, some doctors having significant positive some significant negative weights. On the RISK task however, the average standard regression coefficient for all three of these cues is positive. Smoking has considerable impact on risk.

All GPs' explicit policies on the RISK task could be put in linear terms. The correlations between subjective ratings on the two tasks were significantly different from zero (mean = 0.33, st.dev. = 0.31, $t = 6.06$, $p < 0.01$), and were significantly greater than correlations between regression coefficients ($t = 2.29$, $p = 0.03$). The average subjective ratings on the RISK task (Table 8.6) are again different from those of the PRESCRIBE task (Table 8.3). Perhaps the major difference between ratings is due to differences in the number of cues that are rated positively and negatively by different doctors for example in the case of Age, Drinks, Smokes, Diabetes, Compliance with advice on diet and Attitude to treatment. Attitude to treatment was generally not thought to affect a person's degree of RISK. Whereas in the PRESCRIBE task Smoking, Drinking and Compliance with advice on diet had been used in a slightly moralistic way previously with an argument along the lines of "why help someone who won't help themselves" in the RISK task these were generally seen as increasing a person's risk. Whereas three doctors had felt they would be less likely to prescribe a lipid lowering drug to people with poorly controlled diabetes, all doctors felt people with poorly controlled diabetes would have been at greater risk. Finally although the majority of people had felt they would be more likely to prescribe for younger patients, the perceived influence of Age on risk varied with doctor.

The way Age has affected judgements of risk and was explicitly thought to affect judgements of risk is interesting. Older patients are statistically more at risk from Coronary Heart Disease than younger patients in that for example they are more likely to have a heart attack in the next 5 or 10 years (see the section on lens model analysis below). This is

well known. However, several doctors explicitly and tacitly rated younger people as at greater risk. Doctors may have interpreted the question differently with some interpreting the question as lifetime risk in which case an older person with the same symptoms as a younger person had shown themselves to be a survivor. Some doctors could be expressing a pragmatic interpretation of risk in terms of their anxiety to do something for the patient. Doctors could have been influenced by their behaviour on the previous task where many had explicitly and tacitly been more likely to treat younger patients. The judgements may also reflect a lack of biomedical knowledge.

When a cue's regression coefficients on both tasks are correlated across all doctors the correlations for Age, Gender, Evidence of Arteriosclerosis, Diabetes, Compliance with advice on diet and Family History are all significant (Table 8.7). All of these are known risk factors. Where a doctor took any of these risk factors into account on the risk task it also influenced their prescription. The significance of Compliance with advice on diet is surprising since the cue is not a significant one for any doctor on the RISK task. There are other known risk factors that did not have significant correlations. However, when absolute values of coefficients are correlated the pattern changes slightly (see Table 8.7 column 2). The correlations for Drinks and Weight also become significant and that of smokes has increased. Compliance with advice on diet ceases to be significant. Smoking, drinking and weight were used differently by different doctors on the PRESCRIBE task. Although many of them felt that the cues were relevant they were being used in different ways by different doctors. However, when they were significant on the RISK task they tended to be used in one way. Thus it does seem that doctors who were influenced by drinks and weight on the PRESCRIBE task also saw them as determinants of RISK. It would be expected that the same would be true of smoking. Smoking is interesting in that although it played a large part in risk assessment for all but four doctors, it was only used by fourteen doctors on the PRESCRIBE task.

Table 8.7
Correlations between Cue Weights on the PRESCRIBE and RISK tasks

Cues	Column 1 Signed standardised regression coefficients	Column 2 Absolute standardised regression coefficients	Column 3 Signed subjective weights	Column 4 Absolute subjective weights
CHOLESTEROL LEVEL	0.269	0.248	-0.150	-0.150
HYPERTENSION	0.306	0.242	0.233	0.233
AGE	**0.457**	**0.360**	0.143	**0.551**
GENDER	**0.593**	**0.634**	**0.400**	**0.398**
EVIDENCE OF ARTERIOSCLEROSIS	**0.462**	**0.472**	0.286	0.286
DRINKS	-0.109	**0.552**	0.174	0.090
SMOKES	-0.013	0.269	0.140	0.115
DIABETES	**0.414**	**0.444**	**0.345**	**0.400**
COMPLIANCE WITH ADVICE ON DIET	**-0.350**	-0.149	-0.115	0.119
WEIGHT	0.007	**0.365**	-0.013	0.183
ATTITUDE TO TREATMENT	-0.099	-0.188	0.020	-0.008
FAMILY HISTORY I.H.D.	**0.672**	**0.662**	**0.417**	**0.417**

When signed subjective weights are correlated between tasks (column 3 in Table 8.7) the correlation is only significant for gender, diabetes and family history. Age is also significant when absolute values are used (column 4). In other words those doctors who thought gender was important for prescribing also thought it was an important risk factor, those who thought diabetes was important in prescribing also thought it was an important risk factor and those who thought family history was important in prescribing also thought it was an important risk factor. Those who were more likely to prescribe for either young or old people were more likely to see age as a risk factor too.

The immediate question that needs asking however is why all the risk factor weights don't correlate significantly with prescribing weights. If a factor influences or is believed by the doctor to influence risk judgements then why does that not have an affect on prescribing. According to the medical literature it should: if the patient has risk factors they should be more likely to get treatment (see introduction). One interfering factor may be the lack of variance over some cues. For example cholesterol was important for all but two doctors on the PRESCRIBE task and was also rated as being important by almost all doctors on the PRESCRIBE task. What is odd is that eight doctors who had previously used it as a basis for prescribing did not allow cholesterol to affect judgements of risk. Even if they did not perceive a difference in risk between those with cholesterol levels of 6.5 mmol/l and 8 mmol/l it is still odd that they were more likely to prescribe for the higher cholesterol levels. The odd use of cholesterol is also seen in the fact that doctors who were more influenced on the RISK task by cholesterol were no more likely to prescribe lipid lowering therapy on average than those less influenced ($r = 0.089$, $N = 36$). This may have been influenced by those doctors who were prescribing even though they did not consider cholesterol to be a risk factor or could be influenced by doctors who do feel cholesterol drugs are beneficial despite feeling high cholesterols in the range presented are a problem.

Latencies

Another interesting point to note is that correlations between latencies on the two tasks are significantly different from zero (mean $r = 0.05$, st. dev. $= 0.13$, $t = 2.29$, $p = 0.03$). When doctors spent longer assessing the risk of a patient they also took longer

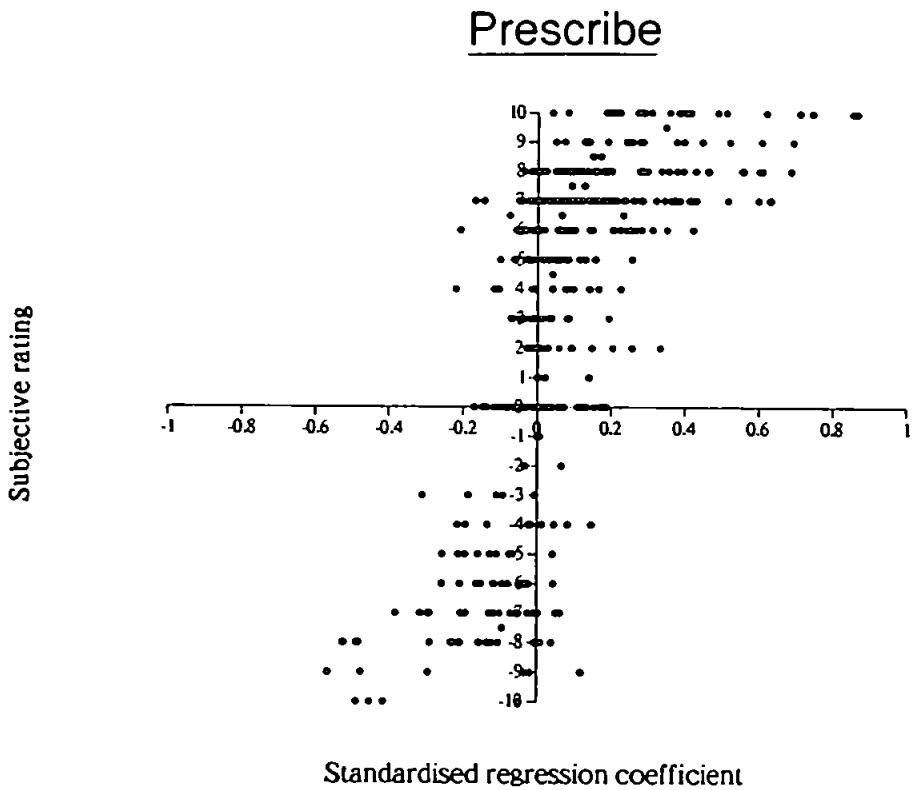
deciding on their treatment. This makes sense if consideration of a person's risk is a prerequisite to prescribing. As in the LIPID task of Study 1 and the IS task, several doctors showed a pattern of decision making whereby they were more likely to prescribe for cases on which they had spent longer on the PRESCRIBE task. On average the correlation was significantly different from zero (mean $r = 0.10$, st. dev. $= 0.20$, $t = 2.91$, $p < 0.01$).The correlation was significantly positive for nine doctors and significantly negative for three (see Appendix 46). However, on the RISK task, nine doctors were significantly less likely to give a high rating of risk to those cases they spent longer on (none had a significant positive correlation). The correlations were significantly different from zero (mean $r = -0.10$, st. dev. $= 0.16$, $t = -3.89$, $p < 0.01$). So on the RISK task doctors tended to give lower ratings of risk the longer they spent on a task. But on the PRESCRIBE task doctors tended to prescribe for cases on which they spent longer.

Self-insight

In this study another of the hypotheses set up in Study 1 (Chapter 5) to explain the triangular pattern of insight will be tested. To recap, the over-rating of importance was found to be cue related. It appeared from Study 1 that the cues that were being over-rated were those that had an impact on the severity of the case and were important in that sense but were less important in terms of prescriptive decision making. In the LIPID task in Study 1 the degree of risk the patient was at was a measure of severity. Here two tests were used. Firstly, over-rating on both the PRESCRIBE and RISK tasks were looked at. However, secondly the cues that were important in terms of risk were measured for the doctor and compared with those cues that were being over-rated on the PRESCRIBE task.

The average correlation between tacit (standardised regression coefficients) and explicit (subjective ratings) policies on the PRESCRIBE task was 0.69, (st.dev. = 0.14). On the RISK task the average correlation was 0.68 (st. dev. = 0.21). These correlations are shown in Appendices 53 and 54. There was no significant difference between these two and both were similar to the average correlation of 0.66 seen on the LIPID task of Study 1.

(a)



(b)

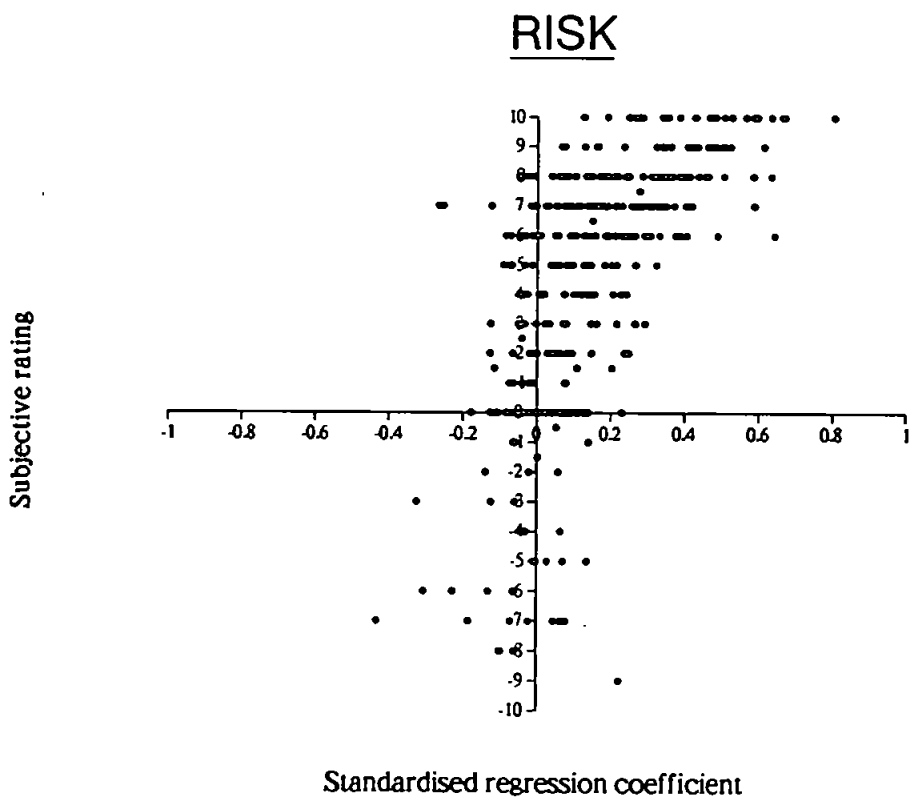
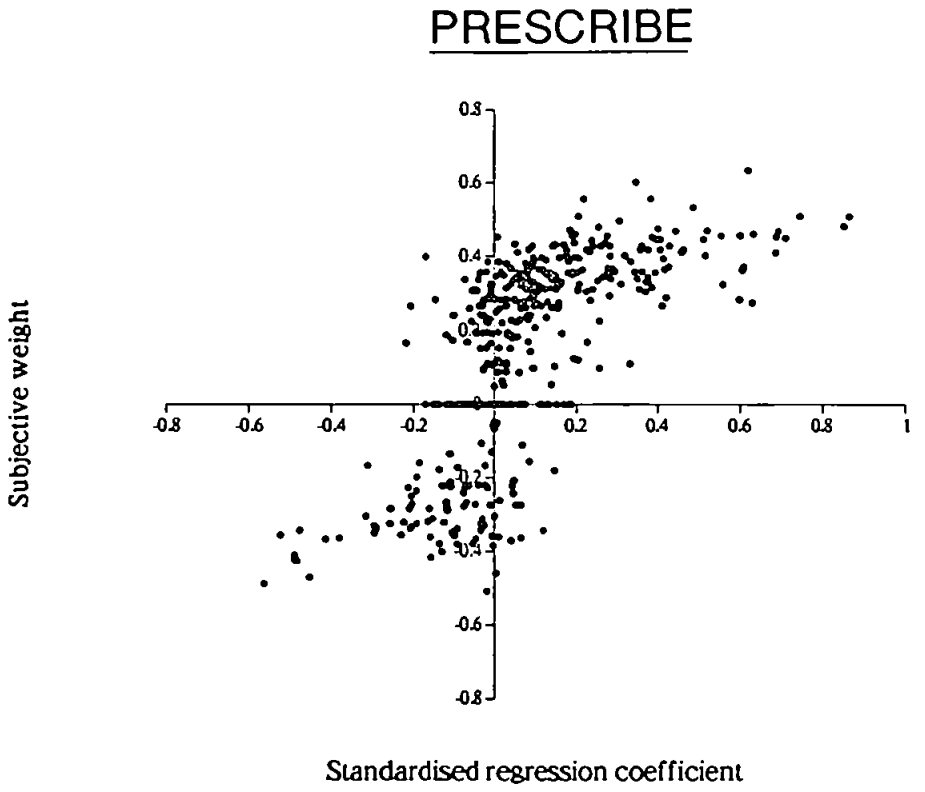


Figure 8.6 Plots of subjective ratings against standardised regression coefficients for doctors on (a) the PRESCRIBE task (N = 30) and (b) the RISK task (N = 36)

(a)



(b)

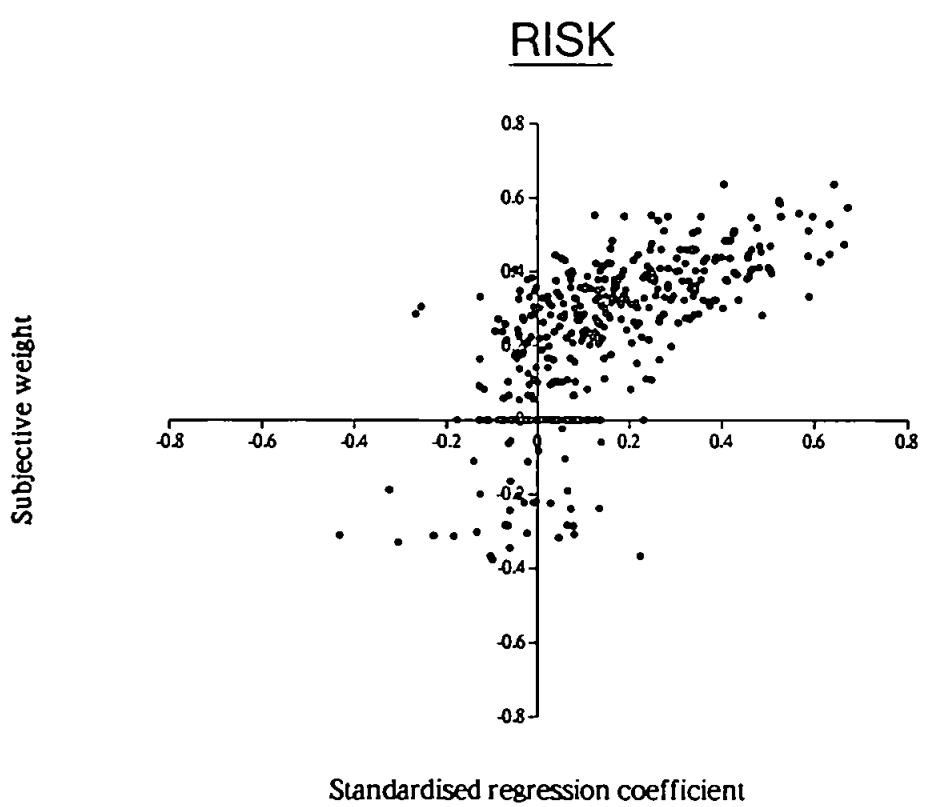


Figure 8.7 Subjective weights plotted against standardised regression coefficients for doctors on the (a) PRESCRIBE (N = 30) and (b) RISK tasks (N = 36).

When explicit policies and tacit policies are plotted against each other a triangular pattern similar to that seen in Figure 5.2 in Study 1 is observed. When cues are rated as not important they are not. But some cues that are being rated as important are not. This over-rating occurs on both tasks as can be seen in Figure 8.6. Again GP40, GP41, GP44, GP59, GP71 and GP74 are not included in the calculations or figures for the PRESCRIBE task. The over-rating pattern is again still there when subjective ratings are changed to subjective weights as in Figure 8.7.

Over-rating on both PRESCRIBE and RISK is cue-related but not doctor-related, just as it was found to be in Study 1. A two way ANOVA of the absolute difference of subjective weights and regression coefficients was significant for cues but not for doctors on both tasks as Table 8.8 shows. Which cues are being over-rated is shown in Figures 8.8 and 8.9.

Table 8.8 Two way analysis of variance of subjective weight-regression coefficient differences

Task	F statistic, between doctors	p	F statistic, between cues	p
PRESCRIBE[6]	$F(29,319) = 1.39$	0.47	$F(11,319) = 3.35$	< 0.01
RISK	$F(35,385) = 1.58$	0.15	$F(11,385) = 4.56$	< 0.01

One hypothesis set out at the beginning of this chapter was that the cues that were being over-rated on the LIPID task were those that the doctor felt were important in terms of the risk of the case. Important was being interpreted in different ways. However, here over-rating occurred on the RISK task as well as the PRESCRIBE task. On the RISK task a clinical interpretation of importance should not have lead to over-rating since if a cue was important in that it increased risk it should have been important in that it increased the judgement of risk! Therefore this different interpretation of "importance" can be rejected.

Although doctors are not re-interpreting the instructions, it does seem to be the case that some doctors were more likely to over rate those cues on the PRESCRIBE task that they explicitly stated were more important in terms of risk. For some doctors there was a significant correlation between their pattern of over rating of cues and their stated pattern

[6] The subjective policies for GP40, GP41, GP44, GP59, GP71 and GP74 were not included since non-linear policies were described on this task.

of cue use on the RISK task. The difference between absolute values of subjective weights and absolute values of standardised regression coefficients (the latter subtracted from the former) was used as the measure of over-rating for each doctor. Absolute values of subjective ratings on the RISK task were used as a measure of the doctors' perception of the relative importance of cues to risk. The correlation was significant for six doctors as can be seen in Appendix 55 but for one of these the correlation was significantly negative: he was less likely to over rate the importance of cues important in risk terms. The average correlation was 0.22 and Fisher's z transformations proved to be significantly different from zero ($p < 0.002$).

There is a second interpretation of the pattern of self-insight. This is that cue over-rating is based on a model the subject feels they ought to be using. Subjective accounts are based on some explicitly held and socially agreed causal explanation. Subjects cannot use the model they profess to be using because it involves the use of more cues than they can take into account. The amount of information affecting decision making has been limited on every task. A number of risk factors were rated as important on the LIPID task but not used to that extent. Statements about the factors influencing decision making on the PRESCRIBE task were more similar to statements about factors influencing judgements on the RISK task than actual policies were similar to each other. On both the PRESCRIBE tasks and the RISK tasks cues stated as being of importance were not so. Another crucial piece of evidence supporting this theory lies in the level of agreement between subjects about judgement policy. Judgements for the 100 cases in the tasks were predicted from policy weights and were used to measure agreement in policy as discussed in the section on agreement earlier. This was done with both subjective weights and regression coefficients and for both the PRESCRIBE and the RISK task. There is greater agreement between predictions from subjective weights ($W = 0.65$ and $W = 0.82$ on the PRESCRIBE and RISK tasks respectively) than from regression coefficients ($W = 0.54$ and $W = 0.62$ on the PRESCRIBE and RISK tasks respectively). This supports the hypothesis that subjects' relative importance ratings are based on socially influenced causal theories.

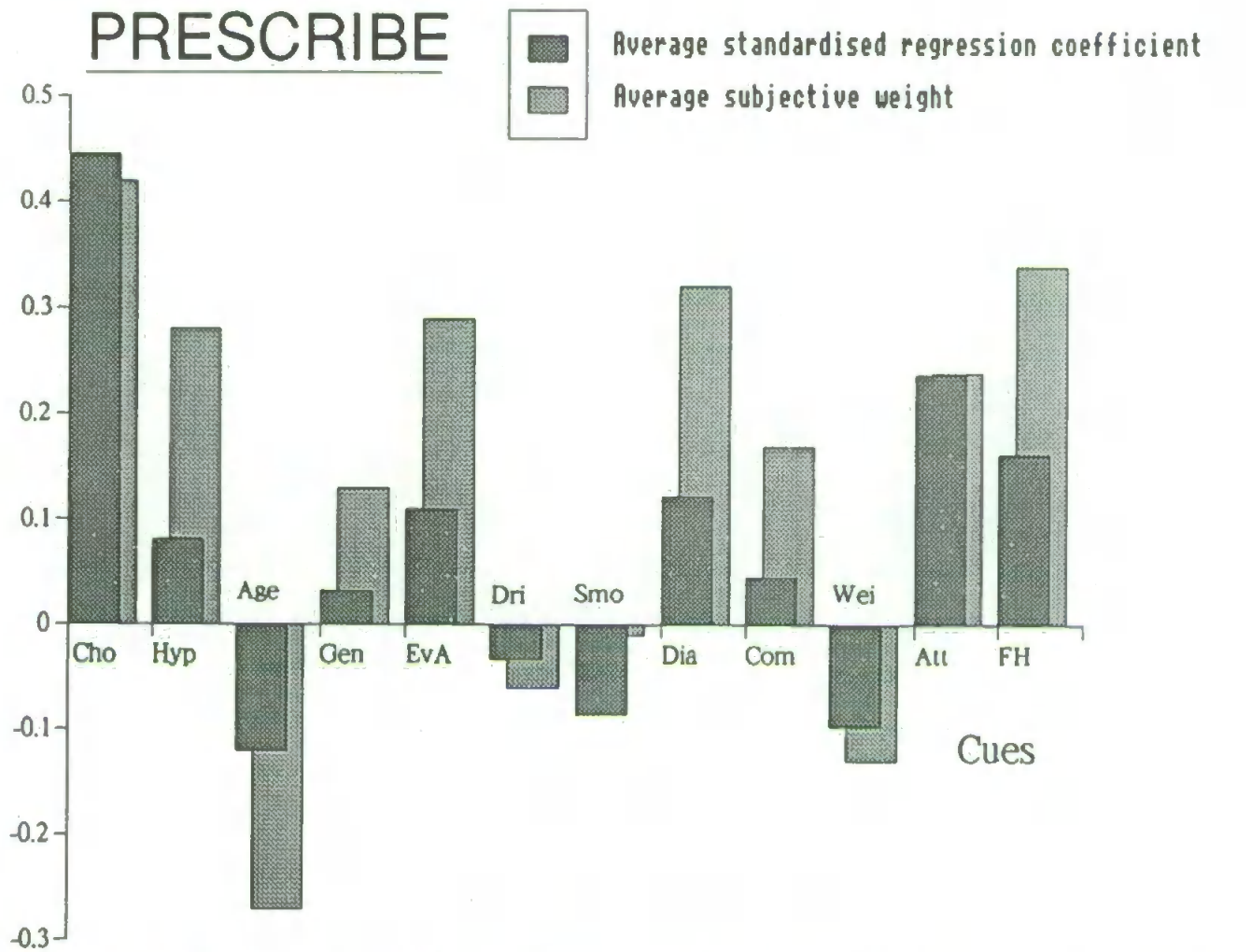


Figure 8.8 Average standardised regression coefficients and average subjective weights for each cue on the PRESCRIBE task (N = 30).

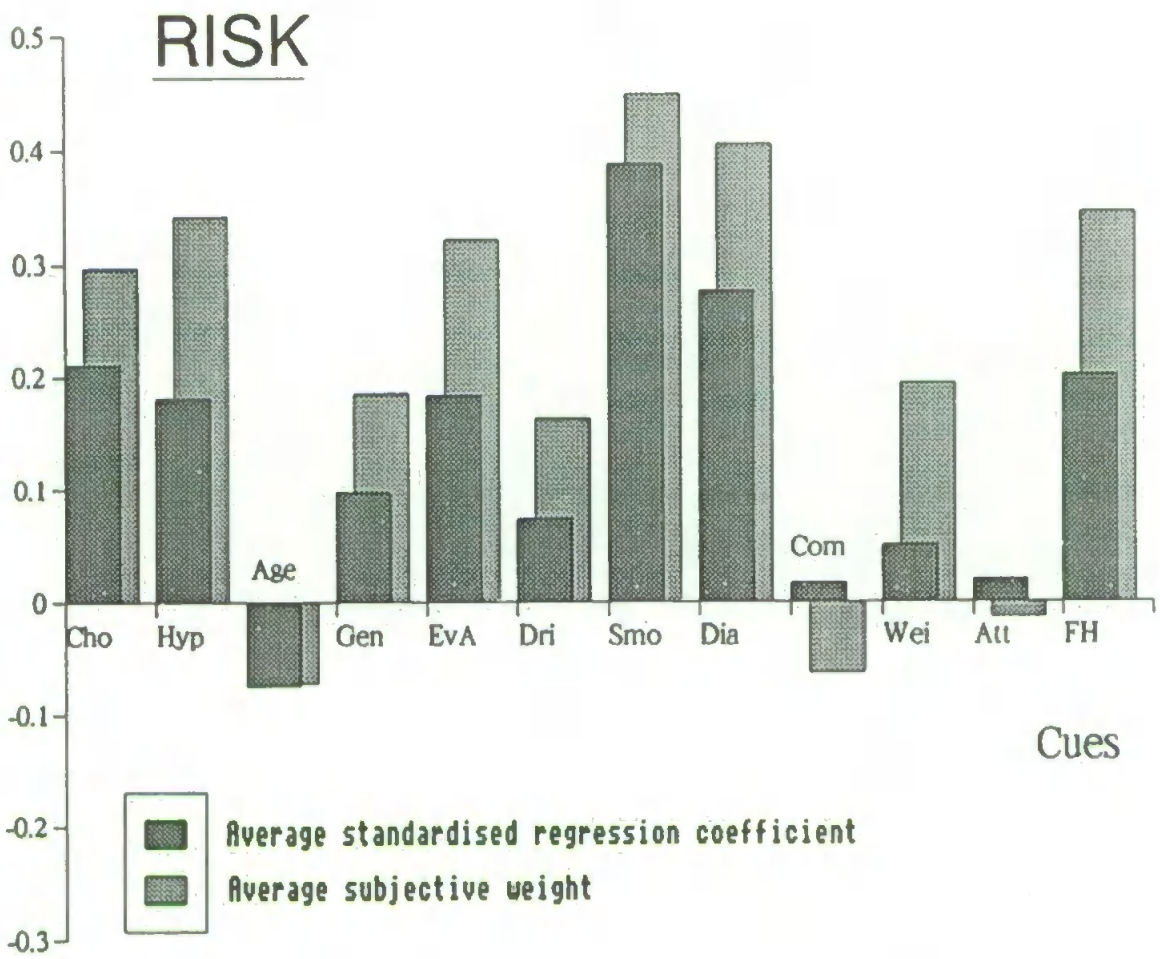


Figure 8.9 Average standardised regression coefficient and average subjective weight for cues on the RISK task (N = 36).

Lens model analysis of the RISK task

The risk task is unique in this set of experiments in that a number of tools exist that purport to give a patient's risk of coronary heart disease. In post-task interviews fourteen doctors explicitly stated that they had some sort of risk assessment tool[7] and eight stated that they did not. Of the former, one doctor used the tool, five used it in some way, such as with new patients or "seldomly", and eight admitted not using it at all. The remaining doctors did not mention any aid or tool.

Table 8.9 Cues and their ranges used in the Mercke, Sharp and Dohme Coronary Risk Calculator

Sex	Male/Female
Age	In years
Systolic blood pressure	Level in mmHg
Total Cholesterol	Level in mmol/l
HDL Cholesterol*	Level in mmol/l
Smoking	Yes/No
Diabetes	Yes/No
Left ventricular hypertrophy	Yes/No

*"If the HDL value is unknown, the program will assume a 'mean' value of 1.15 mmol/l in men and 1.4 mmol/l in women." (HDL = high density lipoprotein)

Here actual risk of CHD could be calculated for each hypothetical patient using the Merke, Sharp and Dohme Coronary Risk Calculator (1993) which is based on equations formed from Framingham study data (Anderson, Wilson, Odell, and Kannel, 1991). This program collects eight pieces of data about the patient and then calculates the patient's ten year risk of a coronary event. It also calculates the risk for a healthy sex and age matched individual. The data entry requirements of the Risk calculator, which are shown in Table 8.9, did not match the information form of the hypothetical cases. For example, one of the pieces of information required was the HDL cholesterol, a piece of information not given in the original task. Where the HDL is unknown, as it was in our set of hypothetical patients, the program assumes a mean value of 1.15 mmol/l in men and 1.4 mmol/l in

[7] Dundee scores (either as discs or in the computer system) and Shaper scores were explicitly mentioned.

women. Systolic blood pressure relates to the RISK task information about Hypertension. Where the patient was described as having poorly controlled hypertension a value of 170 mmHg was entered otherwise the patient was classified as having a systolic blood pressure of 120 mmHg. Smoking was reclassified as "yes" if the hypothetical patient smoked Occasionally, Regularly, or Heavily. Otherwise they were classified as "no". Diabetes was classified as "yes" if the patient had well or poorly controlled diabetes and as "no" if the patient did not have diabetes. Finally the program asked if the patient had left ventricular hypertrophy and for all patients presented in the task the answer was no. Age and sex were taken as they were presented from the age and gender information.

There was information not represented in the program which had been presented in the task. This was Evidence of Arteriosclerosis, Drinks, Compliance with advice on diet, Weight, Attitude to treatment and Family History of Ischaemic heart disease. Although Compliance with advice on diet and attitude to treatment were not influential on any subject's judgement making, Family History of ischaemic heart disease was significant for 21 subjects, Evidence of arteriosclerosis for 19 subjects and Drinks and Weight for five and nine subjects respectively.

The point of the lens model, first described in Chapter 3, is to examine the different components involved in a person getting the right answers in a probabilistic world. In the lens model equation

$$r_A = R_E \cdot R_S \cdot G + C\sqrt{(1 - R_E^2)(1 - R_S^2)}$$

where

- r_A = achievement = the correlation between judgements and the criterion,
- R_E = the degree of systematic linear variance of the criterion, = the correlation between predictions from the linear model of the criterion and its actual values;
- R_S = ('consistency') the degree of systematic linear variance of the subject, = the correlation between predictions from the linear model of the judgements and the actual judgements.
- G (matching) = the correlation between predictions from the model of the judgements and of the criterion,
- C = the amount the remaining variance in subject's judgements, unaccounted for by the model correlates with that of the criterion.

Here the criterion wasn't strictly a real world environmental criterion. An individual's risk or probability can only be calculated from a larger population. Although the real world is probabilistic in that there is a certain amount of random variation in many phenomena, the criterion value calculated here was not. Measures of the Coronary Risk calculation on repeated cases led to a consistency (r_u) of 1. The equation that had been used to form judgements of risk in the programme is a non-linear one (the program is based on equations given in Anderson *et al*, 1991). However, rather than using this as the model of the environment, its predictions were calculated over the set of cases in the RISK task and the relative importance of cues for the criterion were calculated. These calculated risk values over the 100 cases were well described using a linear model. The correlation between predictions from the risk calculator and from the linear model of these (R_E) = 0.93. So there is a small amount of systematic non-linear variance in the criterion.

Different subjects had different policies and a separate lens model analysis was done for each. Figure 8.10 is a diagrammatic representation of the lens model in this context. Using the judgements the subject made over cases, the risk calculation for those cases and linear models of both of these the value of the different component parts of the equation, described above, were calculated for different subjects. These are shown in Table 8.10.

The first interesting and obvious point is that r_A , the measure of achievement, is low in all doctors and only reaches significance for 14 of them. Achievement here is considerably worse than that of medical student subjects in Tape, Kripal and Wigton's study (1992). However, in Tape *et al's* study the additional varying cues included in cases (triglycerides, uric acid and haemoglobin levels) were unrelated to risk of coronary heart disease. Achievement is affected by differences in linear policy as well as differences in the descriptive ability of those policies. Differences in policy occur if a subject uses the cues available that were not included in the risk calculation; if a subject is influenced by cues in a different way to the way the cue influences the risk calculation. Both of these will result in reduced matching (a low G). A subject may also vary judgements unsystematically. This will lead to poor consistency and will again result in low achievement or accuracy.

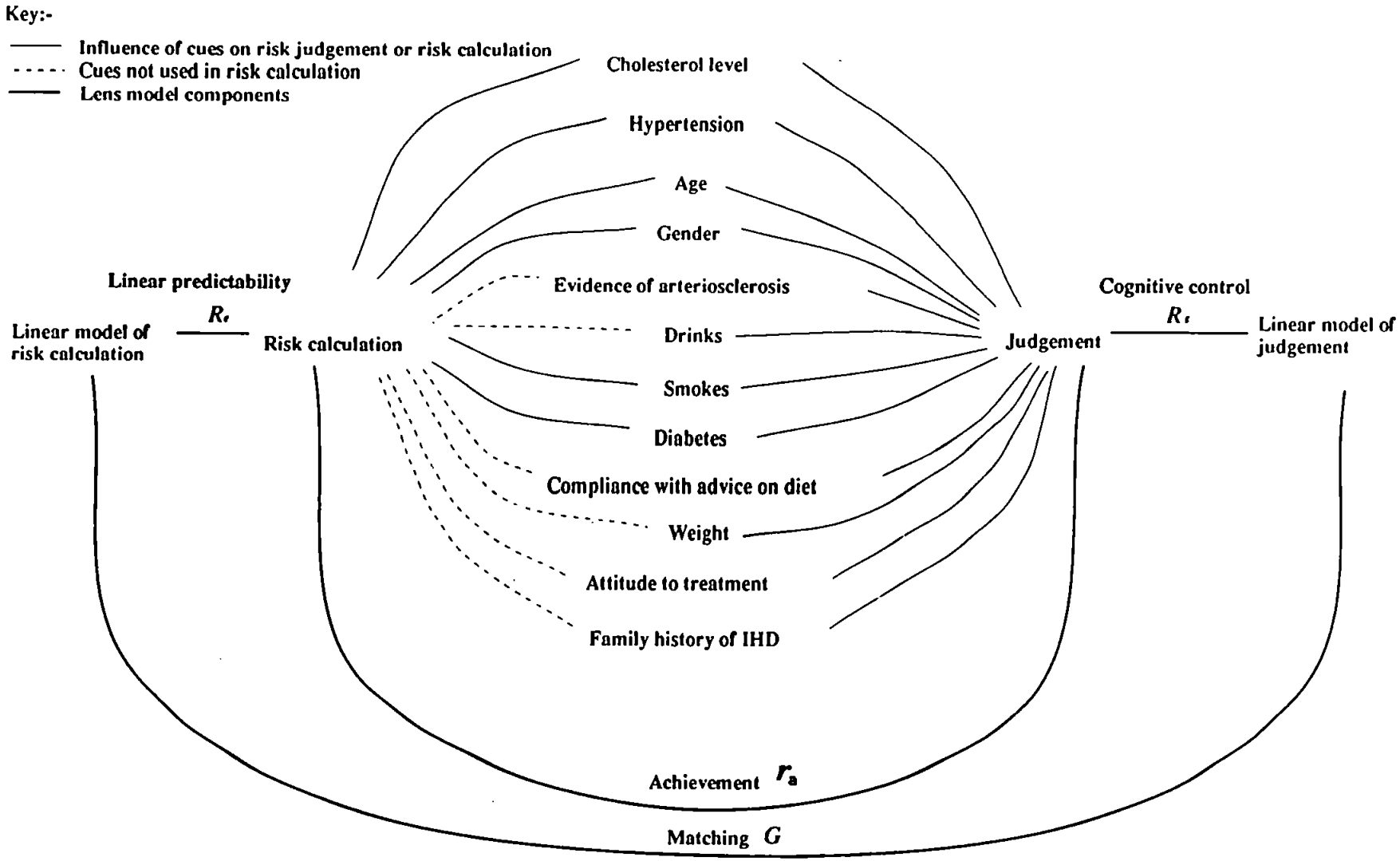


Figure 8.10 Diagrammatic representation of the lens model analysis of risk judgements

Table 8.10
Table of Lens model characters RISK task
$R_E = 0.931$

GP	Non-linear Matching	Cognitive control	Matching	Achievement
	C	R_S	G	r_a
40	0.063	**0.782**	**0.21**	0.167
41	-0.032	**0.713**	**0.411**	**0.265**
42	-0.131	**0.818**	-0.143	-0.136
43	**0.405**	**0.789**	**0.198**	**0.236**
44	0.021	**0.504**	**0.668**	**0.32**
45	0.022	**0.747**	-0.097	-0.062
46	0.034	**0.716**	0.055	0.046
48	0.038	**0.882**	**0.197**	0.168
49	**0.279**	**0.777**	**0.217**	**0.221**
50	0.136	**0.724**	**0.388**	**0.295**
51	-0.07	**0.748**	0.026	0.001
52	-0.034	**0.766**	**0.395**	**0.274**
53	-0.033	**0.763**	0.063	0.037
54	0.062	**0.675**	0.166	0.121
55	-0.127	**0.733**	0.126	0.054
56	0.092	**0.577**	**0.388**	**0.236**
57	-0.136	**0.723**	0.174	0.083
58	0.057	**0.777**	**0.285**	**0.219**
59*	-0.048	**0.686**	0.057	0.024
60	0.075	**0.711**	0.02	0.033
61	-0.095	**0.69**	0.156	0.075
62	-0.001	**0.822**	0.094	0.072
63*	**-0.236**	**0.574**	**0.361**	0.123
64	0.411	**0.734**	0.172	**0.22**
65	0.104	**0.769**	**0.246**	**0.2**
68	0.111	**0.778**	**0.213**	0.18
69	0.114	**0.733**	**0.243**	0.194
71	0.065	**0.739**	**0.285**	**0.212**
72	-0.01	**0.644**	**0.252**	0.149
73	-0.161	**0.742**	**0.306**	0.172
74	-0.199	**0.826**	0.192	0.107
76	0.047	**0.728**	0.029	0.031
77	0.097	**0.726**	0.068	0.07
78	-0.009	**0.712**	**0.361**	**0.237**
79	**0.251**	**0.652**	**0.325**	**0.267**
80	-0.044	**0.674**	**0.331**	**0.196**

* = GPs who used only cues which are used by the coronary risk calculator.

Differences in the actual cues affecting the judgement making can be seen to account for low levels of matching. Only two doctors (GP59 and GP63) were influenced solely by the information that is used in the Risk Calculation. Other doctors might still have reasonable achievement if they were only minimally influenced by information not used by the criterion. This is possibly the case for GP44 who, despite poor consistency (r_u = 0.22) and the additional tendency to rate thin people as at higher risk, was influenced appropriately by the criterion cues and had a high G (0.67) and the highest achievement of all the doctors (r_A = 0.32). Low matching may also have been affected by the way cues influenced judgements. For example coronary risk is greater for older people as can be seen from the importance weights of the risk calculator shown in Figure 8.11. However, for many doctors concern for younger patients was such that they considered them at greater risk. For example, although GP59, used only those cues that were used in the risk calculation he was influenced negatively by Age and G the match between predictions from a linear model of his judgement and those from that of the criterion is poor (0.06).

C, which is a measure of the coinciding systematic non-linear variance, is difficult to interpret if the linear models are very different. Since all but three doctors were significantly influenced by factors not used in the criterion equation it would be surprising if non-linear variance did correspond. For those doctors who did not use other cues, only GP63 has a significant C and this is negative. However, judgement making by GP63 who despite a low linear fit (R = 0.57) was moderately consistent, may not have been well captured by the linear model and he may have had considerable systematic non-linear variance.

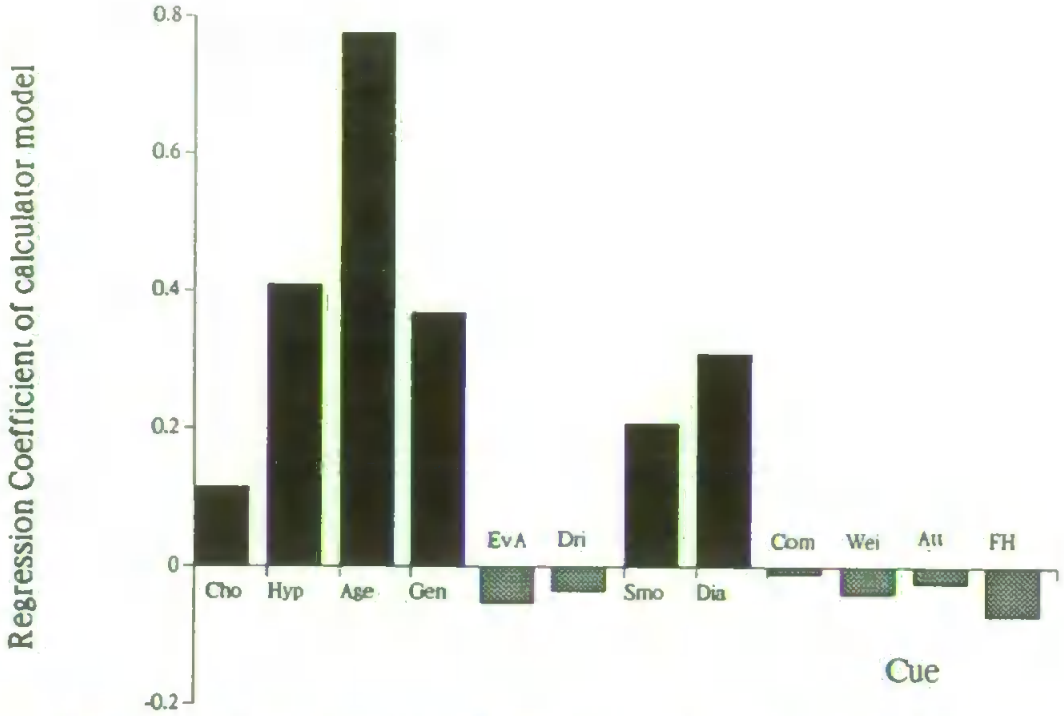


Figure 8.11 Bar graph showing the standardised regression coefficients of the coronary risk calculations over 100 cases.

The limited usefulness of this lens model analysis lies in the inadequacy of the criterion. Despite evidence for the relevance of almost all the task factors in risk of CHD the criterion only takes into account six of them (see Heller, Bailey, Gott, and Howes, 1987). The importance weights calculated from the criterion risk calculations also correlate poorly with the explicitly stated policies of risk judgement. These were shown earlier to reflect a degree of agreement greater than that of prescribing that was thought to relate to the greater agreement in the literature about the factors leading to increased risk. However, the criterion bears no greater resemblance to explicit policies (average correlation = -0.09) than it does to tacit policies (average correlation = -0.09). Plots of the average subjective weights, average regression coefficients and the criterion weights also bear this out. Figure 8.12 shows the weights of the criterion bear poor resemblance to the average subjective weights as they do to the average regression coefficients. Even Tape et al. (1992), who only used five cues in their criterion measure included weight as a predictor of heart disease. When so many subjects were influenced by factors other than those used in the criterion it is not surprising that both matching (G) and achievement (r_A) tend to be low.

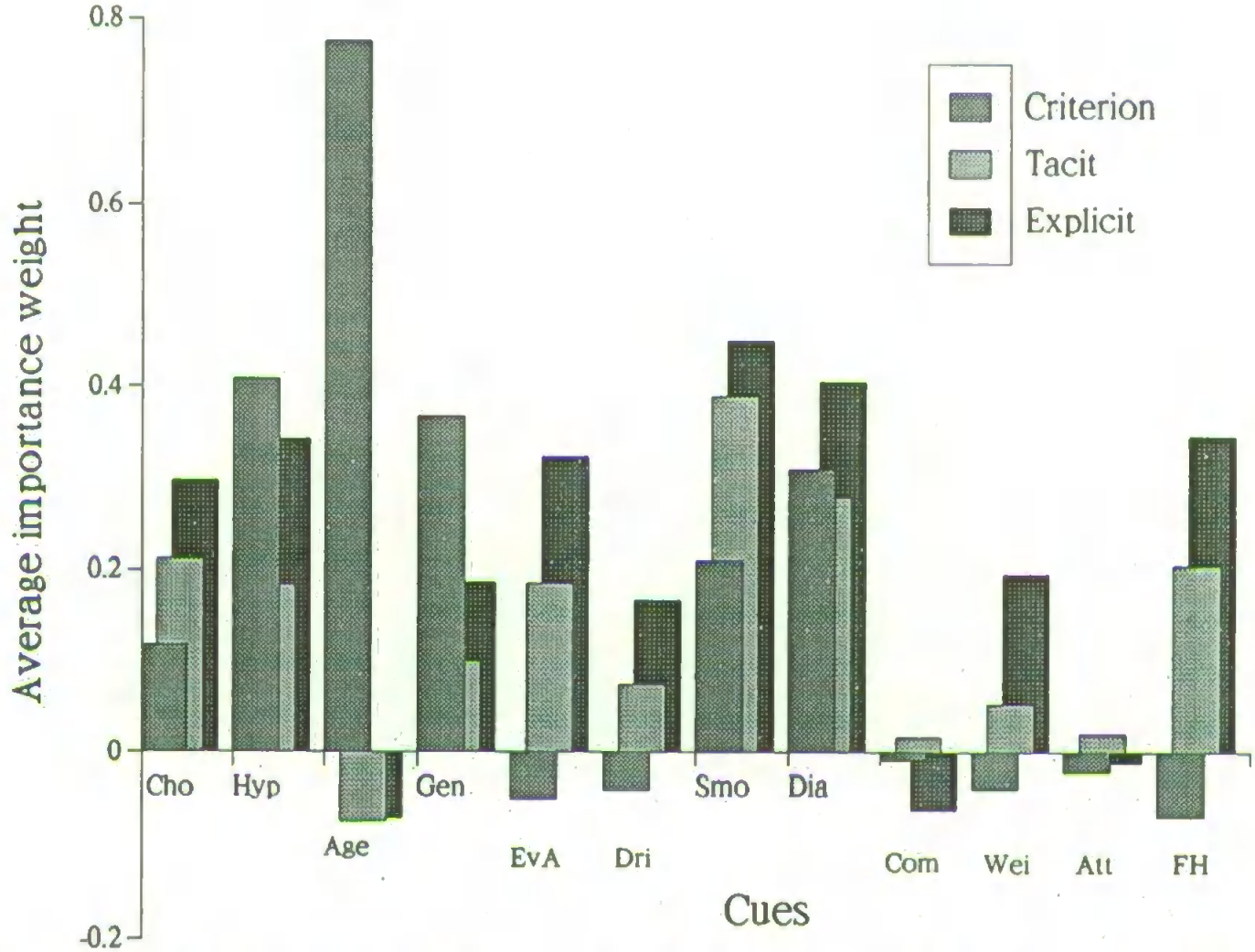


Figure 8.12 Bar graph showing the standardised regression coefficients of the coronary risk calculations over 100 cases.

Conclusion

Judgements of risk of CHD proved to be suitable for judgement analysis and with the uncalibrated scale used gave good linear fits. This study showed that there is a conditional relationship between perceived risk of CHD and prescribing: if subjects were prescribed for they were perceived as being at high risk. However, this does not resemble a threshold model hypothesized in other studies whereby subjects having a probability of a disease above a threshold probability would be prescribed for. Although those who were prescribed for were at high risk and those who were at low risk were not prescribed those at high risk may or may not have been prescribed for. Risk of CHD is multifactorial and several factors could have led to high risk values. The decision to prescribe is clearly based on more than the patient's risk of CHD. When the cues used on each task are compared, although several doctors were influenced by the patient's attitude to treatment on the prescribe task, this was not used to judge risk. However, no more cues were significant for doctors on the PRESCRIBE task than on the RISK task. Average cue use on the tasks was much the same as in previous studies. So although prescription seems in some way dependent on the perceived risk being high, the decision to prescribe seems to use cues very differently to those affecting risk judgements.

Conclusions from the lens model analysis of risk judgements were limited because the criterion model used only a proportion of the cues that were available to and used by doctors. However, there was disagreement even over the use of these. Several doctors either did not use age or rated younger people as at greater risk when in fact the cue was the most important one for the criterion which assessed older people as at far greater risk. Doctors may have been influenced on the RISK task by the PRESCRIBE task they had just completed in which several had been significantly more likely to treat younger patients. On average subjects appear to have under-estimated the importance of hypertension and overestimated the importance of cholesterol and smoking. Hypertension is rated on average as more important than cholesterol for example, and the difference between smoking and the other cues is not so great.

Self-insight on the RISK task was much the same as that on the PRESCRIBE task and on the tasks in Study 1 (Chapter 5). Again all doctors over-rated the importance of

some cues and certain cues had a tendency to be over-rated. However, the importance of cues in the assessment of risk is discussed explicitly in the medical literature. Indeed subjects' tacit and explicit policies on the RISK task showed greater agreement than those on the PRESCRIBE task. In addition stated policies showed greater agreement both within and across tasks than actual (tacit) policies. This explicit agreement about the importance of cues in terms of the severity of the case does not seem to have aided doctors' self-insight into their own judgement and decision making behaviour. A few doctors tended to over-rate those cues on the prescribe task that they thought were important in terms of risk. One interpretation is that subjects' stated policies are partly hypotheses about the cues that should have been affecting their behaviour rather than a mere reflection on those that did.

Chapter Nine
The Confounded Rating Hypothesis Revisited

Introduction

Study 1 had shown that subjects had relatively poor insight into the factors affecting their decision making. They rated more cues as important than were important. Although they only ever tended to use about four cues, other cues had a tendency to be over-rated. This is similar to the pattern of self-insight described in other studies whereby few cues are used and more are rated as important (Slovic and Lichtenstein, 1971, p. 684). Study 2 confirmed this finding and showed that the pattern of explicit policy was more like the pattern of frequency of selecting cues than pattern of tacit policy and that prediction of tacit policies from stated ones was no better than their prediction from patterns of cue selection. This gave evidence for the distinction between explicit and tacit knowledge and supported the Attention Hypothesis. Subjects tended to select more cues than were actually influencing their decision making. Configurality of cue selection, whereby the amount of information collected varied from case to case, did not appear to influence the configurality or non-linearity of judgement making or influence self-insight. Study 3 showed that self-recognition, although saying something slightly different about subjects' self-knowledge, did not suggest that subjects had greater self-insight than we had already found. Study 4 showed that even on a task where, coincidentally, importance in terms of severity was what was being measured and therefore should have matched importance in terms of influence on the decision, cues were still over-rated. A similar number of cues were being used on these tasks.

It looks as if subjects are showing limited self-knowledge. They are not giving a clinical interpretation of importance. It looks as if they are aware of how they looked at cues but not how they were influenced by them. This chapter will examine whether the pattern of self-knowledge seen so far could be due to nonlinear or non-additive use of cues. Evidence for the Confounded Rating Hypothesis, described in Chapter 6, is examined with reference to both previous literature and the studies described so far. The results of Study 4 (Chapter 8) are also analysed further.

To recap, the Confounded Rating Hypothesis, first introduced properly in Chapter 6, states that subjects' ratings of importance might well correspond to the influence that cues could have on cases. But their influence is affected by their configural or non-linear use. The effect of interactions between cues used would lead to some cues having apparently low regression coefficients when within particular cases they were very influential. In this model subjects have a degree of understanding of the relationship between cues and decision making.

There are actually two parts to this hypothesis. Firstly there is an assumption within this model that the linear model of subjects' tacit policy has not captured all subjects' consistent decision making. Subjects' explicit policies were also inappropriately forced into a linear model. If it is the case that the linear model is the best model of their behaviour then the Confounded Rating Hypothesis must be false. Secondly, the over-rating of cues should correspond to this non-linear cue use: where subjects are using cues consistently but non-linearly, they should show worse self-insight. To test this, the addition of some non-linear components or the use of non-linear models should significantly improve linear fits and the cues with non-linear components added should be those that are being over-rated for those doctors whose linear fits are improved.

Non-linear policies can take a number of forms. Analysis using multiple linear regression assumes that the behaviour is linear and additive (Stewart, 1988). Linearity describes a relationship between cue and judgement: as the value of the cue increases regularly, the value of the judgement also increases regularly. Non-linear policies include those with irregular increases such as step functions, inverted-U or U-shaped functions or S-shaped functions. Being additive means that cues affect the judgement independently of the value of other cues. [Billings and Marcus (1983) refer to this sort of policy as compensatory.] Non-compensatory or non-additive models are such that the effect of a cue on the judgement depends on the value of other cues. These include conjunctive, disjunctive, multiplicative and absolute difference policies, elimination by aspects (Tversky, 1972, described in Wright, 1984, p.105-107) and the selective consideration of cues dependent on the value of other cues. These models are termed non-linear because the cues have non-linear relationships with the judgement or because they interact in their effects on the judgement.

Evidence in the literature for non-linear policy use

The literature to be discussed here is that introduced in Chapters 3 and 4, on clinical judgement analysis and on self-insight. The use of multiple linear regression assumes that cues have a roughly linear and additive relationship with judgements or decisions. Linear models are a good fit for the majority of subjects' judgement making in the majority of studies, even if subjects actually have a configural or non-linear pattern of cue use (Brehmer and Brehmer, 1988). The addition of non-linear terms or interactions leads to little if any improvement in fit. The emphasis in policy capturing studies is that what is captured is a paramorphic representation of the subject's behaviour. Subjects' policies in lens model studies are well described by linear models even when the environmental criterion takes a non-linear pattern (Hammond and Summers, 1965). The phenomenon of bootstrapping (see Chapter 3) has also suggested that even if subjects are using a more complex, configural model they are not consistent in applying it (Meehl, 1954; 1965).

However there is some evidence for non-linear behaviour. If subjects are told about a non-linear criterion then they can behave non-linearly (Hammond and Summers, 1965). Bootstrapping could be explained in terms of subjects use of the wrong non-linear model: A linear estimation of this is nearer the criterion than subjects' actual policies. Individual subjects are exceptions in some judgement analyses: For them policy is significantly better captured when non-linear components were added to the linear model. In terms of the Confounded Rating Hypothesis, this significant improvement with non-linear components could make all the difference.

Einhorn (1970) has used parabolic and hyperbolic mathematical formulae to model conjunctive and disjunctive behaviour; Billings and Marcus (1983) have distinguished between conjunctive behaviour and Elimination by aspects (EBA) as non-compensatory models. More recently, Ganzach and Czackes (1995) compared Einhorn's models to two others (the "scatter" and the "true conjunctive-disjunctive" models) in descriptions of theoretically conjunctive and disjunctive behaviour. They found that linear models were better fits than either parabolic or hyperbolic models on all types of task. Only the scatter models were significantly better fitting than linear models and this was only the case in two of the five data sets where non-linear non-compensatory behaviour had been predicted.

However, this model was a better fit than all the other non-linear models on all data sets. The scatter model takes account of the degree to which the case departs form the average in terms of all cues, measuring its "scatter". Other researchers have simply added non-linear or interactive terms to linear models (e.g. Wiggins and Hoffman, 1968; Summers, Talioferro and Fletcher, 1970).

In a few studies non-linear policies have been found to be the best fitting model for a large proportion of subjects. Which non-linear models they use can be seen to be affected by the type of judgement. Einhorn (1972) found the same (conjunctive) model was the best fit for severity of Hodgkin's disease judgements for all three of his pathologist subjects. Einhorn (1971) found that subjects rating jobs tended to use conjunctive models if their behaviour was not linear. However, the conjunctive model was only significantly better for sixteen sets of judgements out of 78, the linear model was significantly better for five. But there was no significant difference between the models for the rest of the judgement sets. On a graduate assessment task linear models were significantly better in eleven out of sixty data sets, non-linear models (disjunctive or conjunctive) were significantly better for twelve. Wiggins and Hoffman (1968) similarly found that although the linear model was best for 12 subjects' judgements of MMPI profiles, sign models were best for 13 and quadratic models were best for three subjects. The sign models used were linear combinations of a number of terms that could be specified from the data. These terms included simple cues and complex combinations of cues. So on a few studies, even when for a proportion of subjects, judgements are better described by a non-linear model, the non-linear model that is best may vary. Different non-linear models seem more likely on different tasks.

In addition to type of policy being related to the type of task, some authors have proposed that it might be related to the information load being presented (Billings and Marcus, 1983; Einhorn, 1971). Information load is changed by changing the time pressure, the number of alternatives being considered or the number of pieces of information about each alternative (the number of cues presented). Einhorn (1971) pointed out that mathematically more complex models are not necessarily cognitively more difficult (and may be simpler) to carry out. He hypothesized that with increasing numbers of cues

subjects would revert to more non-linear, non-compensatory strategies. On both of his tasks fits of all models were better for the two cue condition than for either the four or six cue condition. However, on only one task was there an interaction between the number of cues presented and the type of policy being used. This was a result of very low fits of a disjunctive model when six cues were presented.

There is some evidence that subjects may use non-linear policies, and they may be more inclined to do so where more cues are being presented. Although well described with linear models, judgement making on the tasks presented in Studies 1 to 4 may be non-linear. If subjects are behaving non-linearly their consistency would be high but linear fit limited.

Stated non-linear policy use and self-insight

Subjects' elicited models, whether linear or non-linear have consistently been found to be worse descriptions of behaviour than the objective linear model (Slovic and Lichtenstein, 1971; Reilly and Doherty, 1992; see Chapter 4). If subjects are able to describe their behaviour as being non-linear when it is then that is certainly some degree of self-insight. Cook and Stewart (1975) showed that where non-linear subjective models were formed insight was no better than where linear models were.

Summers, Talioferro and Fletcher (1970) however, describe two findings that suggest another degree of self-insight. Over half their subjects had indicated using a non-linear policy. Summers *et al* found that multiple correlations of cues were worse for those subjects who described using a non-linear policy than for those who did not. This could have been the result of inconsistent rather than non-linear behaviour. When new, non-linear models were formed little improvement in fit was seen. However, those who had described non-linear behaviour had significantly more improvement than those who had not. On one reading of this subjects had some self-insight as to whether their behaviour was non-linear. However, although not all subjects stated non-linearities all showed some improvement when they were added. Summers *et al* do not state whether the subjects' descriptions of the nature of the non-linearity of their own behaviour proved to be correct.

Evidence from Studies 1 to 4

Subjects' judgement and decision making in Studies 1-4 could be described reasonably well with linear models. However, these are only paramorphic representations of subjects' policies. A non-linear model might form a slightly better fit of a subject's behaviour. Patterns of self-insight may be due to non-linear cue use. Where cues have been over-rated they may be used in a non-linear way or have interactions with other cues that have not been captured by the linear model.

Study 3 (Chapter 7) does not help to clarify the situation one way or the other. Regardless of why the discrepancies arose between subjective and objective cue weights, significant correct policy recognition could have been due to the tactics used by doctors to pick out policies. For example, subjects could have picked themselves out on the basis of the cues that felt they didn't use: Any policy with those cues as significant is not their policy. Reilly and Doherty (1992) gave subjects a post-experiment questionnaire in which subjects described the method they used to identify their own policies. Sixteen percent described looking for cues they knew they hadn't used and eliminating policies in which that cue was important. Fifty seven per cent described looking for the cue they felt was most important and eliminating policies where that was not used. If either of these strategies was used where the same pattern of self-insight had been seen as on our task recognition would generally be successful: Cues that subjects stated not using generally were not used and subjects often were able to identify their most important cue. Even if the importance of other cues was confused due to non-linear interactions, if either of these two policies was being used the correct policy would be picked out. The results of Study 3 do not confirm or deny the Confounded Rating Hypothesis.

In Study 2 (Chapter 6) the cues selected were noted. If cues were thought to be important they would be selected regardless of whether they were used linearly or not. Patterns of cue importance seemed to be based on known patterns of selection: stated cue use and cue selection were found to be similar but.both were equally dissimilar from actual linear policy. If it was configural cue use that was leading to the apparent over-rating of cues then where cues were known to be selected configurally less self-insight and less linearity might be expected. Where cue selection is configural it should be more similar to patterns of actual cue use than stated policies are. Subjects' patterns of cue selection were

no more like their pattern of cue use than their subjective ratings were and subjects selecting cues configurally were no more likely to over-rate cues than other subjects. But configurality of cue selection was also seen to be uncorrelated with configurality of cue use.

Although on all tasks described so far (in Chapters 5, 6, and 8) there have been significant correlations between consistency and linear fit, these correlations have not been perfect. Thus, although it can be said that a significant amount of doctors' non-linear behaviour is related to inconsistency, it is not all. There may still be a certain amount of systematic non-linear behaviour. Self-insight was much the same on all tasks: objective-subjective weight correlations were similar and there was over-rating of certain cues. However, the correlation between consistency and linear fit varied between tasks: on the MIGRAINE task it was much higher than on the LIPID task for example. Thus, although the tasks differed in terms of systematic non-linear behaviour, they did not differ in terms of self-insight. This would suggest that the over-rating phenomena being observed are not due to poor capturing of a non-linear policy.

In order to disprove the Confounded Rating Hypothesis subjects would have to be found who had used cues configurally but had not over-rated their relative importance. Over-rating by subjects who had apparently used cues in a linear additive manner would mean that there could be more than one explanation for over-rating. But it would not disprove the Confounded Rating Hypothesis.

Testing the hypothesis

On an individual by individual basis, non-linear behaviour would be indicated by a significant increase in fit with the addition of non-linear components to a policy model. In addition, if over-rating was due to non-linear behaviour, all those cues that were over-rated should also be those involved in a better fitting non-linear model. The obvious way to test the Confounded Rating Hypothesis is to build non-linear models and see if where there is improvement in the model's fit there was over-rating of the importance of a cue. If there was over-rating of a cue but there was no improvement of fit with a non-linear model it may be that the over-rating is still caused by a non-linear pattern but not one that we have

put into the model. However, if there is improvement in a model with the addition of a non-linear component where there was no over-rating it suggests that not capturing a non-linear component does not necessarily lead to over-rating. The problem comes in deciding which non-linear or configural models to try out.

There are a myriad of different types of model. As Stewart (1988) points out there are an infinite number of possible non-additive policies. Obviously it is impractical as well as inefficient to test all of these. General non-linear models of the types used in other studies (described earlier) could be fitted or individual non-linear components could be added to existing linear models to test out specific interactions. Something has to be used as the basis on which to add non-linear components to an existing model or to create a completely different type of non-linear model. Apart from the actual judgements made only two other types of information were collected in the studies described so far: cue selection and the content of the post task interviews. In previous studies non-linearity of actual judgement making policy has not been predictable from subjects' descriptions of their policies. Even when complex descriptions of policy are given often linear models have been found to be the best fit. But these descriptions can be used to ascertain the types of non-linear behaviour that subjects might exhibit.

Study 4 (Chapter 8) reanalysed

The first difficulty in testing the Confounded Rating Hypothesis is to decide which interactions or non-linear patterns to start including in the model. In Study 4 (Chapter 8) subjects were explicitly asked if they were happy with the ratings they had given and if more important cues were getting higher ratings. They were asked if they wished to describe their policy in any other terms. They were asked if there were any interactions between cues whereby the importance of one cue was dependent on the value of another. Not all doctors described non-linear cue use or cue interactions. The comments made are presented in Appendices 56 and 57.

A few doctors were identified on the PRESCRIBE task for whom cues could not be given signed ratings. These cues had either been described as being involved in interactions in some convoluted way, or had been described as affecting decision making

in some other non-linear way. Other interactions were described between cues but these did not prevent assignment of signed ratings. Not all doctors described any interactions or non-linear behaviour on that task, but all doctors tended to over-rate cues. On the RISK task descriptions of interactions were often general. For example : -

"..I think the combination of family history in a first degree relative and .. er .. established pathology such as arteriosclerosis, hypertension; that plus any one of the others, let alone combination with any of the others, was er critical in risk assessment, yeah, I think." (GP41)

"..Females over fifty I would put at higher risk..

"Higher risk than females under fifty yup and I mean again I'd be looking for a combination of risk factors so ..somebody who had a lot of risk factors was more at risk than somebody who only had one or two" (GP50)

"Well once again you're making a judgement - a real judgement on a combination of factors aren't you.. and if you've got somebody with poorly controlled diabetes and poorly controlled hypertension obviously that is going to magnify the risk more than if they had poorly controlled hypertension and were say overweight. It is a multifactorial decision...

... Well I think that smoking interacts with everything - well all. Put it this way, all the factors interact. But I'd have thought that smoking was probably the most significant multiplier" (GP46)

" Um, well I suppose smoking and diabetes were - increased my likelihood of assessing them at high risk. That was probably my main one actually" (GP51)

Even if non-linear patterns of cue use are found to account for the discrepancies between objective and subjective accounts of cue importance, subjects still show lack of self-insight in that these were not always described. Subjects' accounts of their own cue use policy cannot be relied upon to describe or predict their judgement and decision making behaviour.

However, non-linearity or cue interaction could still have been the cause of apparent over-rating even where doctors had not explicitly identified these. The discrepancy between a doctor's subjective and objective weights give an indication of which cues are being over-rated. Where restructuring of a cue leads to improvement in a model's fit it would not be surprising if that cue had previously been over-rated. In this chapter a number of analyses are done, each of which reforms subjects' models to include a non-linear analysis or an interactive component. One change in the models is tried out at a time and significant improvements in a model's fit are tested for. Non-linear use of two cues is looked at: Diabetes and Hypertension. Cue-interactions between age and gender, between Cholesterol level and all other cues, attitude to treatment and all other cues and

between smoking and all other cues are also examined.

(1) *Non-linear cue use:* **diabetes and hypertension.**

A couple of doctors described a non-linear use of diabetes: although they were more likely to prescribe for well controlled diabetics, if the diabetes was poorly controlled they were less likely to prescribe. It may have been the case for other doctors that diabetes, whether or not it was well controlled would lead to increased prescribing. The following analysis was carried out on data from all doctors. Diabetes, which had originally been coded as 1, 2 or 3 depending on whether there was no diabetes, well controlled diabetes or poorly controlled diabetes, was put into two columns. In one column only well controlled diabetes was coded as 1 and anything else was coded as 0. In the other column poorly controlled diabetes was coded as 1 and anything else was coded as 0. Judgements in both PRESCRIBE and RISK tasks were regressed onto cases with these two diabetes measures.

When diabetes was recoded into two "cues", the new linear model was just a less constrained version of the original linear model. The significance of any increase in linear fit was therefore calculated for each doctor using an F test: $F(1, 86) = 86 * (R^2_N - R^2_O)/(1-R^2_N)$ where R^2_O = the linear fit of the doctor's original model and R^2_N = the linear fit of the new model (Howell, 1982 p.435). This difference was found to be significant for eight doctors on the PRESCRIBE task and for 13 doctors on the RISK task.

Hypertension was recoded in an identical manner to the recoding of diabetes and new models were calculated where there were two standardised regression coefficients relating to hypertension. Well controlled hypertension was the only non-zero for one hypertension "cue" and poorly controlled hypertension was the only non-zero for the other hypertension"cue". The significance of the difference in linear fits was again calculated. The difference was significant for four doctors on the PRESCRIBE task and six doctors on the RISK task.

The relationship between over-rating of cues and significant increase in linear fit was examined using Fisher's exact statistic. Over-rating of cues was looked at in terms of the difference between subjective and objective weights. Doctors were classified as having over-rated the importance of a cue if and only if the subjective weight was greater than the

objective weight. Otherwise they were classified as not having over-rated it. The significance of this difference could obviously not be measured since subjective weights are not statistics.

When hypertension was recoded as two cues there was no significant relationship between a significant increase in linear fit and tendency to over-rate hypertension on the PRESCRIBE or RISK tasks ($p = 0.313$, $p = 0.569$). When diabetes was recoded there was no significant relationship between significant increase in linear fit and tendency to over-rate diabetes on the PRESCRIBE task ($p = 0.56$). However there was a significant relationship on the RISK task ($p = 0.047$). Four out of the thirteen subjects who had significant improvements in linear fit had not over-rated diabetes whereas all but one of the 23 subjects who did not have significant improvements in linear fit had over-rated the cue. So in this task subjects who had significant increases in linear fit were less likely to have over-rated than subjects who did not have significant increases in linear fit. This is the opposite way round to the result predicted by the Confounded Rating Hypothesis.

The above analyses show that non-linear behaviour does not necessarily lead to over-rating. No general conclusions can be drawn from the large numbers of doctors had over-rated the importance of the cues diabetes and hypertension but did not have improvement in linear fit. Other cue interactions could have caused the over-rating in these instances.

(2) *Interactive cue use:* **age and gender**

Another type of non-linear model could involve interactions between cues. Some doctors had indicated that the importance of age and gender were related. For example GP45 indicated increased likelihood of prescription for young females over young males but that this difference disappeared in old males and females.

An interactive term was added to the linear equation which then included the 12 original cues plus a product of age and gender. With this additional independent variable five doctors showed a significant improvement in linear fit on the PRESCRIBE task and three showed a significant improvement in linear fit on the RISK task.

There was no significant relationship between over-rating of gender and

improvement of linear fit on either the PRESCRIBE ($p = 0.51$) or RISK ($p = 0.58$) task. There was also no significant relationship between over-rating of age and improvement of linear fit on either task ($p = 0.45$, $p = 0.58$).

(3) *Multiplicative models of interaction* of **cholesterol** with all other cues and interaction of **attitude to treatment** and all other cues. Disjunctive model: Interaction of **smoking** with other cues.

Some doctors had described a policy whereby for some of the range of cholesterol level they would be very unlikely to prescribe regardless of other factors. Outside this range other factors would play some part in the decision. To test to see if providing an interaction between cholesterol and all other cues made a difference to the fit of the model to the data, all cues were recoded as the product of the cue with cholesterol. Cholesterol was included as it had originally been. Since the new model differed quite substantially from the original one the difference in linear fits could not be tested for significance. However, the model would also not have improved automatically. Since the number of cues were the same in the original and the new model, linear fits were compared directly and classified as either having improved or not having improved. On the PRESCRIBE task twelve doctors had a better fitting model when an interaction was included between cholesterol and all the other cues. On the RISK task no doctor's model improved.

Just as some doctors had indicated that there was an interaction between cholesterol and other cues, a similar interaction was mentioned for attitude by some doctors. For example, GP56 indicated on the PRESCRIBE task that if subjects indicated opposition to treatment they would not prescribe, regardless of the other factors. A model similar to the one described above was formed for all the doctors on each task. This time attitude to treatment remained the same but all other cues were the product of their original value and attitude to treatment. Again there was no test of significance of any increase of linear fit but subjects were classified as having an improved linear fit or not having an improved linear fit. Five doctors (not including GP56) showed improved linear fit with this new model on the PRESCRIBE task. No doctors showed improvement on the RISK task.

The above two interactions were only described by doctors as occurring on the

PRESCRIBE task and were not mentioned on the RISK task. They were also only found to have an effect on the PRESCRIBE task and not on the RISK task. This again indicates some sort of insight. Some doctors specifically mentioned smoking as one of the factors involved in interactions on the RISK task. Accordingly a new model was formed for all doctors whereby the value of all cues was multiplied by the value of smoking bar itself which was included as was. Again any improvement in linear fit was noted. No improvement was shown by any doctor on the PRESCRIBE task. Two doctors showed improvement on the RISK task.

Since a number of cues had been affected by these multiplicative policies, over-rating of individual cues could not be looked at. However, over-rating of all cues could be seen in the pattern of self-insight or the correlation between objective and subjective policies. To test the effect of this type of non-linear cue use on self-insight, the self-insight of doctors with and without significant improvement in linear fits on the PRESCRIBE task were compared.

If this analysis was done for each policy separately, differences in self-insight between subjects with significant improvements in fit and those without may not show up. The group without significant linear fits could include subjects exhibiting a different sort of non-linear behaviour (who thus have poorer self-insight). As a result of this subjects with significant improvements from any of the non-linear policies tried here were grouped into the 'non-linear' group. The rest were grouped into the 'linear' group. The (Fisher's z) self-insight shown by subjects in these two groups was compared.

Twenty-one doctors had a significant improvement in linear fit on the PRESCRIBE task with one of the non-linear or interactive adjustments described above. These will be referred to as having non-linear PRESCRIBE policies. Fifteen doctors showed no such improvement. There was no significant difference in self-insight between these two groups ($F(1,34) = 2.65$, $p = 0.113$). On the RISK task nineteen doctors had a significant improvement and were classified as having non-linear RISK policies. Seventeen doctors showed no such improvement. Again there was no significant difference in the groups' self-insight ($F(1,34) = 1.03$, $p = 0.316$).

Conclusions

Evidence for the Confounded Rating Hypothesis has been looked for. Comparison between self-insight by those who had non-linear policies and those who had linear ones showed no difference between the two groups for the majority of non-linear adjustments tried. When subjects' policies involved particular cue interactions they were no more likely to over-rate those cues than other subjects and in one case they were less likely to do so. It could be argued that the subjects without a significant pattern of cue interaction or non-linearity could still be using those cues in some non-linear way which would lead to apparent over-rating and no significant differences between non-linear and linear groups. But analyses in this chapter have failed to find any evidence to confirm the Confounded Rating Hypothesis. Again, the interactions and non-linear uses tested for were the ones that were mentioned by subjects. This would suggest that even if over-rating does relate to non-linear cue use, subjects are not stating that use.

Chapter Ten
General Discussion

Introduction

There have been three major findings in the studies presented. Firstly, the use of multiple regression based policy capturing techniques has proved useful in analysing different types of decisions and judgements made by GPs. Systematic differences, which may be more of clinical than psychological interest, have been observed between GPs in terms of the information that they both actually use and think that they use. There was greater inter-subject agreement in stated than tacit policies. Secondly, and of wholly psychological interest, doctors' ability to use the information was limited. Thirdly, the pattern of self-insight shown has been looked at in greater detail than in previous studies. Possible accounts for the pattern were tested and results indicate that self-hypothesising by GPs as to the factors affecting their decisions is no better than an account based solely on the information selected.

On the capturing of GPs' policies

The nature of general practice was introduced in the first chapter of this thesis. The fundamental characteristics are that consultations are short: Decisions are made under time pressure. But the GP's knowledge of the patient is built up over a long period of time and over several (albeit brief) consultations. GPs see a wide range of problems of varying degrees of severity. Brooke and Sheldon (1985) also add that a large number of psychosocial problems are presented. General practice is more oriented to data collection for patient management rather than following the taught pattern of data collection, diagnosis, therapy planning, and management (see Sheldon, Brooke and Rector, 1985). Structures have still been used to analyse the consultation but these tend to be idealistic and simple. However, the judgement or decision process can still be divided into parts that contribute to individual differences in patient management. Different data may be collected. These may be interpreted differently. They may be combined to affect the decision differently.

Individual differences in the first, and perhaps in the second, can be identified through analysis of recorded consultations or verbal protocols of the doctors' retrospective analysis of their consultations. Differences in the third are more difficult to analyse. Through the use of judgement analysis the same data was presented to subjects in the same form. Individual differences in information combination could be seen and subjects' ability to state this was also noted.

At first it looks as if the problems presented differed in their suitability for policy capturing. Linear fits and consistencies varied between the tasks. At the same time decision making by different GPs appeared to be better or worse captured using linear regression. Again some individuals were less consistent than others. Differences in linear fit of policy captured could be due to non-linear behaviour or due to general inconsistency in decision making. The latter was found to correlate significantly with linear fit on all tasks. Although in Chapter 9 some doctors' policies were significantly improved by the addition of non-linear components, a significant amount of variance in decision making was accounted for using linear models. Policy capturing techniques are still of use in identifying the relative importance of cues.

There were several interesting individual differences in policy, some of which were systematic in that they occurred across tasks. Doctors differed as to the amount they were affected by the patient's attitude to treatment. Those that did take it into account on the LIPID task also tended to on the MIGRAINE task (Chapter 5). Two thirds of doctors took it into account on the PRESCRIBE task (Chapter 8). This seems to indicate some differences in sensitivity to the patients' wishes. Doctors taking the patient's attitude into account might be said to be patient-centred. McWhinney (1985) describes the differences between patient-centred and doctor-centred (models of) decision making. Doctors may practice either (or both). Being patient-centred is characterised by being aware of the patient's point of view and attitude towards their illness and taking their wishes into account. Doctor-centred approaches focus on the illness itself, its identification, and how best to overcome it. Of course those not doing so here may pay attention to patients' wishes in real life, where the patient is actually present and influencing both the consultation and information available to the doctor.

There were other systematic differences between doctors. For example, a few doctors tended to prescribe less for those who were overweight. There may be quite separate, clinical reasons for favouring treatment for those of normal or below average weight on the different tasks. For example, the policy of some doctors not to prescribe lipid lowering therapy for overweight patients might be justifiable on the grounds that dieting would reduce the lipid levels of the blood, reducing that risk factor. The tendency for some doctors to prescribe more for males however, whilst dubiously clinically justifiable on the grounds of risk of CHD on the LIPID task, is certainly questionable on the MIGRAINE task. Only four doctors showed this behaviour. Even on the LIPID and PRESCRIBE tasks however, this gender related preference should not be great. Being male is a risk factor for CHD. Where a man and a woman both have the same symptoms, the man may be more at risk. However, CHD is one of the biggest killers of women as well as men and gender preference should only occur where the other risk factors independently lead to a moderate level of risk (see Sharp, 1994 for a discussion of this point). Again only five doctors showed this preferential treatment of men and they were not strongly influenced.

On the LIPID (and IS and PRESCRIBE) task certain cues acted as contra-indicators to prescribe for some doctors but as indicators for others. Indeed two of these - weight and smoking - were only ever perceived as increasing risk but still affected doctors' management decision making differently. The relationship between risk and prescribing was examined in Chapter 8. Although those patients prescribed for were a subset of those perceived to be at high risk, the use of cues on the different tasks was very different. If a judgement about risk is made prior to prescribing, the cues used in the risk judgement might be expected to be a subset of those used in the prescribe judgement. This was certainly not the case. Some cues reduced the likelihood of a doctor prescribing despite recognition of a patient's high risk. As mentioned earlier, the avoidance of drug treatment for overweight people may be explicable on the LIPID task. However, the policy of some doctors on this task amounts to avoiding prescribing lipid lowering treatment for smokers despite the recognition that they are at greater risk. These doctors explicitly stated they would try to persuade the person to stop smoking rather than give them drugs. But stopping smoking, whilst reducing overall risk, does nothing for risk caused by having a high

cholesterol level. This reflects a prejudice against smokers on the part of some doctors but others felt the patient was more likely to give up smoking if they were not being given lipid lowering treatment (and a false sense of security about the reduction in their risk).

On limits to information processing capacity

Throughout all the studies presented, subjects have been limited in the number of cues that they use. Although 12 or 13 cues were always presented, subjects never used more than eight and used an average of four or five on all tasks. Individual differences were shown on tasks but did not carry across tasks. In other words, where a doctor used fewer cues on one task, they did not necessarily use few cues on another task. Differences in the number of cues used cannot then be attributed to individual differences in working memory capacity.

In the literature reviewed in Chapter 3, the effects of expertise on working memory were discussed briefly. Although the number of chunks of information that could be used is still limited in experts, experience is such that information can be chunked into units of known recurring patterns. Thus chess masters, for example, can reproduce complex patterns of pieces when they are possible chess situations but are hampered if the arrangement is not feasible (de Groot, 1978; described in Simon, 1979). The importance of chunking can be seen in discussions of medical expertise where pattern matching is seen as an intrinsic part of diagnosis within the domain of expertise. Similarly the use of identification of forceful features in diagnosis as discussed by Grant and Marsden (1987; Gale and Marsden, 1985) would only be successful in the context where other features, characteristic of the disease, are likely to occur alongside that forceful feature. If different doctors were more expert on a particular task it would be expected that they could take more information into account. They would be more aware of patterns of cues. Expertise is certainly domain dependent and so different doctors might take more information into account on different tasks.

Although this was the pattern found over the studies presented here, differences in expertise cannot explain the task specific individual differences in quantity of cues used. Advantages of expertise are lost where cues are orthogonal as they were in the tasks in this

thesis. In real life where cues correlate to a certain degree, recognisable patterns may be learnt and sets of features may be picked out on the basis of one or two forceful ones. Where cues are not intercorrelated the use of one or two cues is just that and not the recognition of several associated cues. So the differences in quantity of cue use between GPs on our task is not due to differences in patterns of expertise. This is confirmed by the general limit to the number of cues used. If differences in expertise really could account for these differences then the number used would be greater. As it is, the number being used is in the range expected if each item of information were taking up one chunk: Cue use is limited for all doctors.

If not domain specific expertise then what can account for task specific individual differences in the number of cues used? Here the differences can only be explained through differences in policy. There are individual differences in the number as well as choice of cues. It might then be argued that differences in the number of cues used is not evidence for limits to information processing capacity and that these differences are simply policy differences. There are two facts that make this unlikely. Firstly, this does not explain why no GP used more than eight cues. Secondly, in statements about cue use GPs consistently said they were using more cues than they actually were. Explicit policies were not limited in cue use in the same way as tacit policies were.

On self-insight

One reason for using judgement analysis is its lack of reliance on subjects' stated policies. It might seem odd to then compare doctors' stated policies with their captured policies. If doctors had shown good self-insight obviously there would be no need for the statistical calculations involved in judgement analysis. But doctors' lack of self-insight into their decision making can be seen elsewhere. For example, another study on the same sample of doctors used in Chapter 5 showed that they either lacked any insight or gave responses in terms of some sort of academic decision theory when asked how they make decisions (DiCaccavo and Reid, 1995). However, the point in these studies was not to justify the use of policy capturing techniques by simply showing that GPs lack self-insight. Here the particular pattern of self-insight has been examined and reasons for it discussed.

The pattern found was a triangular one. As in previous studies subjective policies were flatter than objective ones: More cues were rated as important than actually were (Slovic and Lichtenstein, 1971). Previously, where the pattern has been described, important cues have been said to be under-rated and unimportant ones over-rated (Slovic and Lichtenstein, 1971). Here however, where a GP said a cue was unimportant it was not significant. Cues that GPs indicated were important may or may not have been. Where cues were actually significant GPs tended to say they were. So there was selective over-rating of certain insignificant cues. Some cues had more of a tendency to be over-rated than others but there were no significant differences between doctors in their tendency to fit this pattern. Important cues were not under-rated. This pattern could be seen when subjective ratings were used as stated and also when subjective weights, comparable to regression coefficients were used in the comparison. This pattern occurred on both decision tasks and the risk judgement task and with the two different samples of doctors used.

In Chapter 5, self-knowledge was defined as a measure of the match between a subjective model of behaviour (or knowledge) and an objective model of behaviour (or knowledge) (see Figure 5.7). There are two types of explanation for the pattern of self-insight seen in these studies. Firstly, the pattern could be classed as some artefact of the method: the subjective model is inappropriate or the objective model is inappropriate (or both). Secondly, reasons can be hypothesized for the pattern of self-insight which is assumed to be genuine.

Given that this pattern was found on all tasks, one conjecture then was that the pattern is actually caused by the task or the means of analysis. One thing all the tasks had in common was the use of orthogonal cues. This could conceivably have caused the pattern of self-insight seen, as explained below. Call this the "Artificiality Hypothesis". Another possibility is that the method used to capture policies causes this pattern. Perhaps it is where cues are used non-linearly that there is apparent over-rating. This has been referred to as the "Confounded Rating Hypothesis".

If, as in real life, cues are correlated, then their use may be substituted over a series of cases. For example, if overweight people usually have hypertension then a doctor could

use the fact that a person in front of them was overweight or could measure their blood pressure to make a management decision. These things are not perfectly correlated - some thin people for example have hypertension. However, for the management of the majority of patients it wouldn't matter which cue the doctor used. Suppose a doctor alternated which cue he or she paid attention to. If the doctor was to then rate their importance, he or she might rate them both equally as very important. In the situation where they were both correlated this might be true.

But suppose now that the doctor had been making decisions on a set patients where cues were orthogonal. In this case the doctor might rate both cues as equally very important again, assuming that the use of one really meant the use of the other. However, in this situation the alternating use of cues would mean two things. Firstly, both cues would be equally important (if they had been used an equal amount of times) but only moderately important. As a consequence of this both would appear to be over-rated in comparison to other cues that would not be naturally correlated. Secondly, the doctor's decision making might appear slightly inconsistent and the fit of the model might not be ideal. After all they have been alternating cue use policy.

This, of course, is exactly the pattern that was found - in all studies, on all tasks. However, if the Artificiality Hypothesis is correct then subjects making decisions on realistically correlated sets of cues should show better self-insight. If they do not, the Artificiality Hypothesis can be rejected. Studies with both intercorrelated cues (*e.g.* Kirwan, Chaput de Saintonge, Joyce, Holmes, and Currey, 1986) and orthogonal cues (*e.g.* Fisch, Hammond, Joyce, and O'Reilly, 1981) have shown the same lack of self-insight. Reilly and Doherty (1989) found that subjects' ratings of the relative importance of cues given just before the task showed significantly worse self-insight than those given after the task. In other words subjects had been influenced to some degree by the pattern of cues they had seen. However, in the study by Reilly and Doherty (1992) subjects did show better self-insight where cues were intercorrelated, in a representative design, than when they were orthogonal. But in this measure of self-insight Reilly and Doherty did not compare cues' subjective and objective importance weights because of the difficulty of identifying the separate relative importance of cues when they are correlated. They

grouped together (twelve or six) cues into five factors for which they calculated usefulness indices and summed the relative importance ratings for the cues in each factor. This grouping may have led to the apparently excellent self-insight shown by subjects in these conditions. [The mean correlation between objective and subjective factors in the twelve cue condition was 0.92.] This suggests that the artificiality of the task set may contribute to the pattern of over-rating seen. It cannot explain it totally, since self-insight is still lacking where cues are correlated. What this means in terms of GPs' decision making is that in theory policies are not completely consistent, although in practice they may appear to be so.

The Confounded Rating Hypothesis was discussed in Chapter 9. This hypothesized that over-rating of cues might have occurred where cues were used (systematically) non-linearly. Despite being given the opportunity, not all doctors stated non-linear cue use so their self-insight can still be said to be limited. However, their ratings of importance may have been confounded nevertheless, leading to this triangular pattern of self-insight. Where non-linear cue use was demonstrated, doctors were no more likely to over-rate those cues than doctors who had used them linearly. In one case they were actually significantly less likely to do so. This, the reverse result to that predicted, seems enough to reject the hypothesis. However, it can always be argued that those doctors classified as not using those cues non-linearly may have had a different non-linear policy than the one tested. Since all doctors tended to over-rate certain cues then, if the Confounded Rating Hypothesis were correct, all doctors must have tended to use certain cues non-linearly. However, linear fit is also somewhat a measure of non-linearity. Linear fits varied in their ability to describe subjects' behaviour. Although this variation was significantly related to consistency of cue judgement making, that correlation was not perfect and there must also have been differences in systematic non-linear behaviour. However, there were no significant differences between doctors in terms of the pattern of self-insight.

The suggestion from the above two hypotheses is that subjects may have had self-insight but that this was warped by the method used - by the artificiality of the task or by the use of a linear model. In these cases both objective and subjective models were inappropriate. It has similarly been suggested that subjects may have better self-insight

than shown by asking for explicit ratings: the subjective model is inappropriate. Reilly and Doherty (1989, 1992) investigated this by looking at subjects' recognition of their own policies. In Chapter 7 the same was done here and subjects' ability to pick out their own policies was significantly greater than by chance. However, they were no better at picking out their own explicit policy than they were at picking their tacit policy. Performance in picking out subjective policies may have been reduced because of the influence of the similarity between policies or because of forgetting. But it would still be expected to be better than selection of tacit policies if the pattern of difference between them was caused by inaccuracy on the part of the tacit policy.

The results of the studies contained in this thesis then show that subjects cannot state the relative importance of cues - they cannot state their combinatorial policies. Even if the orthogonality of cues may have contributed to the pattern of lack of self-insight, as described in the Artificiality Hypothesis above, it suggests that subjects are insensitive to the covariation of cues and their use. Explicit and implicit knowledge of these policies appear to be separate. However, it appears, in fitting with the findings of other researchers that subjects can state explicitly certain aspects of the decision making process. But that others remain elusive and can only be hypothesized about, just as another person might.

The part of decision making that subjects seem to be able to state is which cues they have paid attention to - that they know phenomenally. When subjects' stated policies were compared with both their tacit policies and their cue selection policies in Study 2 (Chapter 6) tacit policies could be predicted from cue selection policies just as well as from stated policies. Their stated policies were more similar to their cue selection policies that to their tacit policies.

Attention to cues is the focus of process tracing techniques, which often use verbal protocol analysis. Although retrospective verbal protocols have been criticised as being open to memory loss and reconstruction (Ericsson and Simon, 1980), the post task analysis of the relative importance of cues in the studies presented here produced a fairly good likeness to the cues selected. Although not requested, subjects seem able to give this sort of phenomenal self-knowledge retrospectively.

Although cue selection policies correlated well with stated cue use, it could of

course be the case that the two are correlated with something else. One possibility is that both cue selection and stated cue use correlate with cue use in real life decision making, where cues are correlated. This would fit the results shown. But this is the artificiality hypothesis which was earlier rejected as a full explanation of the findings.

Another possibility considered was that the type of importance was misinterpreted and that subjects were giving some measure of clinical importance rather than the bearing cues had had on the decision making. The same cues would be looked at in that they had a bearing on risk, and would be rated highly in terms of clinical risk. However, they need not affect decision making quite this strongly. That doctors were not giving just a clinical interpretation of importance was seen in Chapter 8 when the fact that the pattern of self-insight was still apparent in statements of policy on risk assessments.

Another possibility is that subjects could be stating the cues they thought they ought to have attended to and also selected these. Explicit policies are hypotheses based on taught knowledge. It is unlikely they would have done this deliberately. They were asked about their decision making policies specifically on the task they had just completed and participation was anonymous. However, two lines of evidence support this. Firstly, risk factors, or factors contributing to the severity of the case, tended to be the ones over-rated. These are the cues that doctors will have been taught are of import. Secondly predictions of judgements from explicit policies showed greater agreement than those from tacit policies. This was greater on the RISK task than on the PRESCRIBE task. Again doctors will have been explicitly taught the factors of relevance to a risk calculation. A variation of this is that knowledge of the clinical importance of cues comes into some ideal model of the decision policy. If this were the case it would again be evidence that the sort of knowledge that GPs can give is based on that which they have been taught explicitly, have constructed from explicitly held knowledge or are phenomenally aware of. There was greater agreement between subjective models than objective models. Even when the socially agreed ideal combination of cues is known, this is not what doctors are actually doing.

That much of self-knowledge is hypothesizing, influenced by generally agreed theories was put forward by both Nisbett and Wilson (1977) and Nisbett and Ross (1980 chapter 9). Relative importance is a measure of influence of pieces of information in

covariant terms and the pattern of self-insight here seems to match general patterns of estimation of correlation. Estimates of correlation have been found to be subject to attentional bias whereby evidence positively confirming the presumed relationship is disproportionately attended to (Baron, 1988 chapter 14). This would mean that a decision to prescribe for a hypertensive patient might be seen as evidence for the strong effect of hypertension. No account is taken of prescription decisions when hypertension was not present, or of the value of hypertension when the value of other cues lead to non-prescription. This pattern of assessment has been seen both when subjects estimate covariation or correlation from summary tables and from real life phenomena (see Nisbett and Ross, 1980 chapter 5). The attentional bias is related to the other bias mentioned by Baron - the effect of prior beliefs on estimations of covariation and correlation. Subjects (including clinicians) have been found to perceive correlations consistent with their hypotheses and to miss actual correlations in the data that they would not have predicted (Nisbett and Ross, 1980 chapter 5). In keeping with an attentional bias subjects may only attend to data that fits their hypothesis. The attentional bias is also however, exhibited in correlation estimations on "theory free data".

A comparison can be made with confirmation bias (see Evans, 1989, chapter 3) whereby subjects only seek and notice evidence consistent with their theory or hypothesis. Obviously it is unlikely that such a policy would lead to a correct assessment of correlation and it is easy to see how it leads to overestimation of correlations. However, only the importance of cues attended to are likely to be overestimated. If this were the case, those cues not attended to would be correctly identified as uninfluential. Thus the pattern of similarity between cue selection and stated cue use is not surprising. Of those cues that are attended to there is no way of predicting which would be over-rated and which would not. It appears that subjects are selecting cues and rating them as important on the basis of hypotheses about their own behaviour.

So either way, the suggestion is that doctors lacked self-insight in terms of how information was combined. At best they were aware of their relative attention to different pieces of information, which seems to be based on factors they believed were relevant to themselves for a particular type of case. The results of these studies have shown that

explicit knowledge and implicit knowledge are again separate phenomena.

Implications for communication

It has been suggested that the methods of process tracing and policy capturing are complementary and that both look at different aspects of the same process (Einhorn, Kleinmuntz and Kleinmuntz, 1979). The former identifies the information attended to and its order of aquisition. This is then formed into a series of rules by the analyst. Policy capturing in contrast looks at how much of a bearing each piece of information has on the decision. As was seen in Chapter 6, the relative importance of cues has a something of a relationship with cue attendance (or selection in this case). Subjects certainly have some insight into the information they have attended to as can be seen from the success of certain reported protocol analyses (*e.g.* Boreham, 1989) and this may be what they are stating here.

General Practitioners were the subjects used in all studies presented. The tasks used were ones they encounter regularly and were familiar with. One argument might be that GPs may appear to lack self-insight more because of the way they aquire their expertise. However, all subjects lack self-insight in terms of cue combination. Although they can state phenomenal knowledge or knowledge they were taught explicitly, where self-insight is shown it is because this happens to coincide with how cues are actually used. The point of this section is to show that this limit to self-insight does not have bad consequences for teaching of or communication of skills.

GPs' behaviour is learnt through experience, and is very different to the practice that is taught explicitly during clinical training. This can be seen even in different models of the consultation. Even in hospital medicine, procedure and knowledge bases change substantially as expertise develops (Schmidt and Boshuizen, 1993). Some behaviour is taught explicitly, some is learnt implicitly through interaction with the patient. Whereas, in other areas of expertise, where some processes are taught explicitly and can be explicitly stated when asked for, in general practice that explicit knowledge of the process was not taught. It would be unfeasible to teach trainee GPs how to combine information to make patient management decisions for every type of case that could come up. During clinical

training they were guided towards establishment and judgment of diagnosis on the basis of information presented. As stated earlier, it may be because much behaviour is learnt intuitively, through interaction with patients that general practice is seen by some as an art. GPs are aware of the factors of relevance to disease patterns and can state and seek out these.

In situations where procedural knowledge is learnt (taught) explicitly subjects may appear to have self-knowledge. By recalling this they may describe something that fits their pattern of behaviour and communicate to others so that they can exhibit similar behaviour. Indeed some of the literature discussed in Chapter 4 showed this to be the case. Where procedural knowledge is not available only phenomenal, experiential knowledge may be elicited. Here all GPs, who don't have this explicit procedural knowledge to call upon, showed the same lack of self-insight.

However, even where experts are explicitly taught processes, it is unlikely that they will have been taught them in terms of the relative importance of cues. In accordance with this it is rare for subjects to show good self-insight in terms of being able to state the relative importance of cues. However, they may be able to state the cues they attended to. One argument is that whenever procedural information is being communicated it is the information about cue attendance rather than combination that is transferred.

Imagine being taught a skill, such as driving. You can be instructed to feel the biting point, instructed verbally on hand eye co-ordination in steering. You can be taught the different parts of the engine and how they work together. This latter information is not necessary for the driving procedure. But the procedural instruction must be given. It would take a person of exception to be able to work out how to drive solely intuitively. However, nobody can learn to drive solely on the basis of verbal instruction about how to proceed. They must actually do it. Driving ability is not instant but is something that is built up till it becomes automatic. This isn't just a physical skill. Considerable judgement and decision making is involved until eventually an experienced driver can take familiar routes without recalling the journey. When less familiar routes are taken more concentration must be given. However, if asked how they drove the driver would not describe the automatic process that he or she regularly exhibits but would describe the process in terms of its

procedural components. If asked how they drove along a particular route the driver would give information about turnings, traffic lights and pubs - in terms of the things that had featured in their attention during the journey.

In a similar way, a GP asked about the process employed in patient management might talk about data collection and diagnosis, or problem identification, maybe hypothesis testing. Asked about management of a particular patient, they might talk about the features of that patient that they had particularly attended to. However, they would not include a description of how the features are incorporated into a judgement or into a decision.

The studies presented here have shown that when asked about the bearing cues have on decision making, GPs may give information about cue aquisition in the process but not its relative influence. Indeed they were able to do this retrospectively. Again this is the sort of information given in verbal protocols. The construction of actual cue use policies from information about which cues were attended to is obviously a skilled procedure. However, if, as in this study, subjects are only actually influenced by a subset of cues they selected, analysis may be misled by sole reliance on verbal protocols. On the other hand, analysis of verbal protocols can be useful in identifying which cues might (in some cases) influence behaviour. The reliance on this sort of information with respect to skill learning is not surprising. It is also not surprising that subjects can communicate their policies satisfactorily to one another (Brehmer and Brehmer, 1988 cite a paper by Ekegren, 1983). They can describe their policies on one level, but this is not the level of cue combination. Skills can be learnt through knowledge of cues to attend to and aquisition of factual knowledge of the situation.

The pattern of learning from feedback (see Chapter 4) also provides evidence for this two tier knowledge system. The information relayed is what the subject is or should be attending to. By changing the focus of attention or data collection, cue use policies can change considerably. In the phenomena of feedback only cognitive feedback has been found to be of benefit: information about policies is useful, whereas simply giving the correct response (outcome feedback) is not. However, not all cognitive feedback is of benefit, although task information (the ideal policy) leads to improvements in decision or judgement making, cognitive information (information about one's own policy) makes no

difference. At first this might suggest that subjects do not then lack self-insight but lack knowledge as to what they should be doing. However, evidence presented both in this work and in other studies suggests that subjects do lack self-insight. Another interpretation is that the information subjects glean from these policies is information about the cues they should be attending to. Feedback of cognitive information is redundant not because subjects know the relative importance of cues but because they already know which cues they have been attending to.

Policies can be communicated and changed through description at the level of cue attention. The lack of self-insight about cue combination is not a problem.

On analysis of judgement and decision making

It appears that there are two tiers to knowledge in judgement and decision making. The first is on the relevant facts for the situation. Subjects may have self-insight into this as can be seen in verbal protocols. It is this that is given explicitly in teaching and in conveying the essentials of a practice. The second is data combination, into which subjects do not apparently have self-insight and which is measured in policy capturing. The combination of data, the use of selective features to trigger pattern recognition is something that develops only with experience.

With the policy capturing techniques described here the relative bearing information has on the decision making can be seen without interruption of the process and without requesting self-insight from the GP. However, there is one caveat. This relates to the model fitting cue combination. It has been pointed out that mathematically simple models such as linear regression are not necessarily cognitively simple ones. The apparent ease of analysis using linear regression may shadow the ease of use of other models. The policies captured in these studies are paramorphic ones: They are useful ways of comparing decision making by different subjects but probably do not represent their actual method of cue combination. The capturing of relative importance of cues can be compared to the description of a painting in terms of its colours. Not in terms of which colours lie where but in terms of the percentages of green and of yellow *etc.* in the picture. Subjects can say which colours are in the paint box. Policy capturing in terms of relative

importances identifies which colours are in the painting and how much of each colour is there. But whether these are combined in large areas, in a pointillistic style, in stripes, which colours are next to which *etc.* is not defined by identification of relative percentages. Relative importances are not enough to identify how cues are combined. Other sorts of rules and mathematical models can do that. They can predict which colours go next to which - which cues lead to the use of other cues *etc.* But the basic identification of relative importance certainly tells us more than a subject can.

Clinical judgement analysis of GP decision making is advantageous in that these selective features can be identified and possible reasons for interindividual differences can be highlighted. The use of process tracing techniques is possible but verbal protocol analysis may interrupt the semi-intuitive process of patient management under a time pressured consultation. Where the process is akin to pattern recognition or is tripped by the recognition of forceful features as has been suggested, verbal protocol analysis may be unsuitable. It is precisely in recognition processes or automatised processes that Ericsson and Simon (1980) suggested verbal protocol analysis may be less useful. However, it appears that even retrospectively in these instances subjects may show insight into the information that they attended to.

Future Research

This chapter has outlined the major findings of this study and its practical implications. What is left is to outline areas that would be useful to explore in the future. Research developments related to the work discussed in this thesis are of medical or psychological interest or both.

On the medical side, judgement analysis of other types of decision or judgement might be of interest or the development of a multifactorial model might be considered. Where several management options are possible on cases, policies for each could be captured and compared just as risk judgements and prescribing policies on the same cases were compared. A second issue for medics might rest in the validity of the studies: the relation of behaviour on these tasks to everyday practice. Although other tasks have shown paper presented cases elicited much the same responses as real life ones there may be

doubts about any study that has not actually tested this. To test validity, responses on real and paper versions of the same cases could be compared directly or policies on hypothetical cases with realistic cue ranges and combinations could be compared to policies on real life patients. In either of these cases the doctors' behaviour on real life cases might well be changed by having to note information proforma (as would be necessary). There are many other factors that additionally might affect doctors' behaviour in real life. For example, time pressure has been given as a reason for prescribing (to end the consultation); the information available is affected by what the patient chooses to disclose; the patient may bring several problems to the doctor at the same time and may dictate the focus of consultation. Generalisations can not be made about doctors policies other than on the range of cases on which they have made decisions where other factors do not vary. However, by reducing the variance of other factors (such as time pressure) systematic differences between doctors can be seen.

Another possible practical follow-up of interest to the medic is the application of lens model analysis and cognitive feedback to everyday judgements such as risk. The majority of doctors have computers on their desks; many of them have risk assessment packages on them (many of them may do but don't know about it). As was seen in Chapter 8, risk assessment is a complex multifactorial judgement. Tacit policies have been found to be limited in cue use. Even explicit policies included inappropriate relative importances of cues. For example many doctors rated younger people as at greater risk when statistically risk increases with age. However, doctors could be trained to make better judgements of patients' risk using cognitive feedback. Currently they may receive outcome feedback through use of computer risk calculators on a patient's data. Outcome feedback, as was discussed in Chapter 4 is of no benefit in changing behaviour. To give doctors versions of the mathematical models from which risk assessments are calculated would also be of no use in that they would be meaningless to the majority of doctors. However, feedback of task information (and possibly cognitive information) would indicate to doctors the relative importance they should give each piece of information (and how they should be influenced by it). The weighting would be specific to the range of cases that the doctor saw and so would be more meaningful to him or her. Of course, the possibility of such a model would

rely on a good criterion measure being found rather than one that omits to take into account important features as was seen in Chapter 8.

The effect of feedback on self-insight might be of interest to psychologists. Although when task information has been fed back subjects' self-insight has not improved (Luckett and Hirst, 1989), it might be the case that subjects could be trained to perceive the relative importance of cues with repeated feedback of cognitive information. In this case subjects would become sensitive to the cues that they attend to whose variance affects their decision making little (and which have been overrated in these studies).

Although the case mix was unrealistic it is of interest to psychologists that even on a relatively simplified set of cases subjects could not state their own policies accurately. However, the account that has been left somewhat hanging in this thesis is is the effect of a real life case mix on both cue use and self-insight. Studies using realistic cases have reported poor self-insight too but using realistic sets of cases seems to have improved self-insight in at least one study (Reilly and Doherty, 1992). The effect of realistic case sets on cue use seems to depend on the measure. If the relative importance of a cue is measure in terms of its regression coefficient then those cues accounting for most of the variance get high regression coefficients. Cues that correlate somewhat with these may appear insignificant. In this case, if subjects use several correlated cues, the number of cues they use may be under-estimated. If the relative importance of cues is measured otherwise, such as through analysis of variance or correlation coefficients, the number of cues subjects use may be over-estimated because of the inter-cue correlation. For example cues that correlate with cues that the subject has used may appear to have affected their decision making even if they have not attended to them. So although it is of interest to investigate the effects of inter-cue correlations on self-insight; obtaining an objective measure of the subject's policy in these circumstances is difficult.

One way to proceed might be to use information from more than one type of analysis to obtain an objective measure of behaviour. Information about the cues doctors selected to attend to could be used to eliminate the importance of those unselected but correlating cues. However, there is still the question of the relative importance of the information that they did attend to, which will be intercorrelated. As was shown in the last

paragraph, no individual cue index of relative importance is useful under these circumstances. The only alternative might be to group the cues into psychologically meaningful factors as Reilly and Doherty (1992) did. However, in this case subjects' explicit understanding of the factorial grouping of cues might also be of interest.

Although there are a couple of studies that directly compare process tracing and policy capturing methods of describing behaviour (*e.g.* Einhorn, Kleinmuntz and Kleinmuntz, 1979; Billings and Marcus, 1983), there is still need of a thorough investigation of the relationship between process tracing and policy capturing, particularly in relation to the issue of self-insight. Neither Einhorn, Kleinmuntz and Kleinmuntz (1979), who used subjects' verbal protocols as the basis for their process tracing models, nor Billings and Marcus (1983), who formed process tracing models and policy capturing ones on behaviour on separate tasks, measured subjects' self-insight. Both process tracing techniques (other than those based on verbal protocols) and policy capturing are ways of objectively measuring behaviour, against which subjective measures might be compared. However, as shown earlier, they measure different aspects of the decision making process. That subjects have self-knowledge of one aspect of their behaviour but not the other might be clearly demonstrated in a task in which both measures were used. The selection of cues in Study 2 might have been used as the basis for a process tracing model but time limitations meant that that was not possible. In addition to clarifying the nature of self-insight, the use of process tracing models could complement judgement analysis: if a cue is not always selected, it cannot always be used and models of non-linear aspects of behaviour can be guided (but not dictated) by non-linear patterns of cue selection. If the same behaviour is modelled using both techniques they might also be combined: rather than just having two separate (complementary) descriptions of aspects of the process one is obtained.

A number of possible follow up studies have been suggested, some of which will be more complex than others and some of which are of more interest to medics rather than psychologists. But the work in this thesis has pointed out the limits to processing capacity and limits to self-insight that can only be identified through the use of orthogonal cues. Pragmatically speaking, some of the time neither of these matter: few cues can be used in

real life as a basis for decisions because cues are correlated with each other, doctors implicitly learn to recognise frequently occurring patterns of cues through use of key features to make decisions. This works well when the pattern of cues seen in a patient fits the standard pattern of cue. However, where cases are atypical (as in fact many are) decision making based on this sort of automatic behaviour may be inappropriate and policy differences between doctors may show up. Whether or not doctors are using an appropriate few cues in their automatic behaviour can best be identified through judgement analysis and lens model analysis. Behaviour can perhaps be modified with cognitive feedback. The limits to self-insight are also inconsequential in a pragmatic sense since subjects do show some self-insight and are able to communicate their policies to other practising individuals. But it takes time for those individuals to learn behaviour and patterns of cue combinations. Subjects' explicit directions indicate which cues might be relevant but again not how or which might be the key cues to attend to. Judgement analysis gives a useful description of judgement behaviour over cases. It is important to model the task environment as well as the relationships between cues and judgements to really identify subjects' key cues.

References

Adelman, L., Sticha, P.J., & Donnell, M.L. (1984). The Role of Task Properties in Determining the Relative Effectiveness of Multi-attribute Weighting Techniques. *Organizational Behavior & Human Performance 33*, 243-262.

Allais, M. (1953). Le comportement de l'Homme Rationnel devant le Risque, Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica 21*, 503-546.

Anderson, K.M., Wilson, P.W.F., Odell, P.M., & Kannel, W.B. (1991). An Updated Coronary Risk Profile: A statement for health professionals. (AHA statement). *Circulation 83 (1)*, 356- 362.

Arkes, H.R., & Hammond, K.R. (Eds.). (1986). *Judgment and Decision Making: An interdisciplinary reader*. Cambridge University Press.

Audit Commission (1994). *A Prescription for Improvement: Towards More Rational Prescribing in General Practice*. London: H.M.S.O..

Balzer, W.K., Doherty, M.E., & O'Connor, R. (1989). Effects of Cognitive Feedback on Performance. *Psychological Bulletin 106 (3)*, 410-433.

Balzer, W.K., Sulsky, L.M., Hammer, L.B., & Sumner, K.E. (1992). Task Information, Cognitive Information, or Functional Validity Information: Which Components of Cognitive Feedback Affect Performance. *Organizational Behavior & Human Decision Processes 53*, 35-54.

Baron, J. (1988). *Thinking and Deciding*. Cambridge University Press.

Barrett, B.J., Parfrey, P.S., Foley, R.N., & Detsky, A.S. (1994). An economic analysis of strategies for the use of contrast media for diagnostic cardiac catheterization. *Medical Decision Making 14 (4)*, 325-335.

Bech, P., Haaber, A., Joyce, C.R.B. & the Danish University Anti-depressant Group (1986). Experiments on clinical observation and judgment in the assessment of depression: profiled videotapes and judgment analysis. *Psychological Medicine 16*, 873-883.

Berry, D.C. (1994). Implicit Learning: Twenty-five Years on: A Tutorial. In C. Umilta & M. Moscovitch (Eds.), *Attention and Performance XV*, (pp. 755-782). Cambridge Mass.: MIT Press.

Berry, D.C., & Dienes, Z. (1993). *Implicit Learning: Theoretical and Empirical Issues*. Lawrence Erlbaum Associates.

Billings, R.S., & Marcus, S.A. (1983). Measures of Compensatory and Noncompensatory Models of Decision Behavior: Process Tracing versus Policy Capturing. *Organizational Behavior and Human Performance 31*, 331-352.

Blood, M.R. (1971). The Validity of Importance. *Journal of Applied Psychology 55 (5)*, 487-488.

Boreham, N.C. (1989). Modelling Medical Decision-Making under uncertainty. *British Journal of Educational Psychology 59*, 187-199.

Brehmer, B. (1987). Note on Subjects' Hypotheses in Multiple-Cue Probability Learning. *Organizational Behavior & Human Decision Processes 40*, 323-329.

Brehmer, B. (1988). The Development of Social Judgment Theory. In B. Brehmer & C.R.B. Joyce (Eds.), *Human Judgment: The SJT view*. Elsevier Science Publishers B.V. (North Holland).

Brehmer, A., & Brehmer, B. (1988). What have we learnt about human judgment from thirty years of policy capturing? In B. Brehmer & C.R.B. Joyce (Eds.), *Human Judgment: The SJT view*. Elsevier Science Publishers B.V. (North Holland).

Brehmer, B., & Joyce, C.R.B. (Eds.). (1988). *Human Judgment: The SJT View*. Elsevier Science Publishers B.V. (North Holland).

Brooke, J.B., & Sheldon, M.G. (1985). Clinical Decision = Patient with Problem + Doctor

with Problem. In M. Sheldon, J. Brooke, & A. Rector (Eds.), *Decision- Making in General Practice*. The Macmillan Press Ltd.

Brunswik, E. (1952). *The Conceptual Framework of Psychology*. International Encyclopedia of Unified Science: Foundations of the unity of Science 1(10). University of Chicago Press.

Centor, R., Witherspoon, J., Dalton,H., Brody, C., & Link, K. (1981). The Diagnosis of Strep Throat in Adults in the Emergency Room. *Medical Decision Making 1 (3)*, 239-246.

Chaput de Saintonge, D.M., Crane, G., Rust, N., Karadia, S., & Whittam L. (1988). Modelling Determinants of Expected Rewards in Healthy Volunteers. *Pharmaceutical Medicine 3*, 45-54.

Chaput de Saintonge, D.M., & Hathaway, N.R. (1981). Antibiotic Use in Otitis Media: Patient Simulations as an Aid to Audit. *British Medical Journal 283*, 883-884.

Chaput de Saintonge, D.M., & Hattersley, L.A. (1985). Antibiotics for Otitis Media: Can We help Doctors Agree? *Family Practice 2 (4)*, 205-212.

Chaput de Saintonge, D.M., Kirwan, J.R., Evans, S.J.W., & Crane, G.J. (1988). How can we design trials to detect clinically important changes in disease severity? *British Journal of Clinical Pharmacology 26*, 355-362.

Cheng, P. (1985). Restructuring versus automaticity: Alternative accounts of skill aquisition. *Psychological Review 92*, 414-423.

Christensen-Szalanski, J.J.J., & Bushyhead, J.B. (1981). Physicians' use of Probabilistic Information in a real Clinical Setting. *Journal of Experimental Psychology 7 (4)*, 928-935.

Clancey, W.J. (1984). Methodology for building an intelligent tutoring system. In W. Kintsch, J. Miller, & P. Polson (Eds.). *Methods and Tactics in Cognitive Science*, pp.55-84. Erlbaum Hillsdale, New Jersey.

Cook, R.L., & Stewart, T.R. (1975). A Comparison of Seven Methods for Obtaining Subjective Descriptions of Judgmental Policy. *Organizational Behavior and Human Performance 13*, 31-45.

Cooksey, R.W., & Freebody, P. (1985). Generalised multivariate Lens model analysis for complex Human Inference tasks. *Organizational Behavior & Human Decision Processes 35*, 46-72.

Cooksey, R.W. (In Preparation). *Judgment Analysis: Theory, Methods and Applications.*

Coombs, C. H. (1975). Portfolio Theory and the measurement of risk. In M. Kaplan & S Schwartz (Eds.). *Human Judgment and Decision Processes*. Academic Press Inc.

Darlington, R.B. (1968). Multiple Regression in Psychological Research and Practice. *Psychological Bulletin 69 (3)*, 161-182.

Dawes, R.M. (1968). Algebraic models of cognition. In C.A.J. Vlek (Ed.). *Algebraic models in psychology: Proceedings of the NUFFIC international summer session in science*. Netherlands University foundation for International cooperation.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist 34 (7)*, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin 81 (2)*, 95-106.

de Groot, A.D. (1978). *Thought and Choice in chess*. The Hague: Mouton. 2nd Edition.

Dennett, D. (1982). How to study consciousness empirically or nothing comes to mind. *Synthese 53 (2)*, 159-180.

Di Caccavo, A., & Reid, F. (1995). Decisional Conflict in General Practice: Strategies of Patient Management. *Social Science & Medicine 41(3)*, 347-353.

Doherty, M.E., &Balzer, W.K. (1988). Cognitive Feedback. In B. Brehmer & C.R.B. Joyce (Eds.). *Human Judgment: An SJT Viewpoint*. Elsevier Science Publishers B.V.

(North-Holland).

Doubilet, P., & McNeil, B.J. (1985). Clinical Decisionmaking. Reprinted in J. Dowie & A.S. Elstein (Eds.). *Professional judgment. A reader in clinical decision making.* Cambridge University Press.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* The Free Press (New York).

Edland, A., & Svenson, O. (1993). Judgment and Decision Making Under Time Pressure: Studies and Findings. In A. J. Maule & O. Svenson (Eds.). *Time Pressure & Stress in Human Judgment and Decision Making.* New York:Plenum.

Edwards, W., & Newman, J.R. (1982). Multiattribute evaluation. Reprinted in H.R. Arkes & K.R. Hammond (Eds.). *Judgment and Decision Making: An interdisciplinary Reader.* Cambridge University Press.

Einhorn, H.J. (1970). The Use of Nonlinear, Noncompensatory Models in Decision Making. *Psychological Bulletin 73,* 221-230.

Einhorn, H.J. (1971). Use of Nonlinear, Noncompensatory Models as a Function of Task and Amount of Information. *Organizational Behavior and Human Performance 6,* 1-27.

Einhorn, H.J. (1972). Expert Measurement and Mechanical Combination. *Organizational Behavior and Human Performance 7,* 86-106.

Einhorn, H.J. (1974). Expert Judgment: Some necessary conditions and an example. *Journal of Applied Psychology 59 (5),* 562-571.

Einhorn, H.J., Kleinmuntz, D.N., & Kleinmuntz, B. (1979). Linear Regression and Process Tracing Models of Judgment. *Psychological Review 86 (5),* 465-485.

Ekegren, G. (1983). *Verbal reports about strategies in probabilistic inference tasks.* Upsala: Acta Universitatis Upsaliensis, Studia Psychologica Upsaliensa No. 8.

Elstein, A.S., Holzman, G.B., Belzer, L.J., & Ellis, R.D. (1992). Hormone Replacement Therapy: Analysis of Clinical Strategies Used by Residents. *Medical Decision Making 12 (4),* 265-273.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1978). *Medical Problem Solving: An analysis of Clinical Reasoning.* Harvard University Press

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1990). Medical Problem Solving: A Ten-Year Retrospective. *Evaluation & The Health Professions 13 (1),* 5-36.

Engelhart, M.D. (1936). The technique of path coefficients. *Psychometrika 1,* 287-293.

Engel, J.D., Wigton, R., LaDuca, A., & Blacklow, R.S. (1990). A social judgment theory perspective on clinical problem solving. *Evaluation & The Health Professions 13,* 63-78.

Ericsson, K.A., & Simon, H.A. (1980). Verbal Reports as Data. *Psychological Review 87(3),* 215-251.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol Analysis: Verbal reports as data.* MIT Press.

Essex, B.J. (1985). Decision Analysis in General Practice. In M. Sheldon, J. Brookes & A. Rector (Eds.). *Decision Making in General Practice.* The Macmillan Press Ltd.

Evans, J.St.B.T. (1984). Heuristic and Analytic Processes in reasoning. *British Journal of Psychology 75,* 451-468.

Evans, J.St.B.T. (1989). *Bias In Human Reasoning: Causes and Consequences.* Lawrence Erlbaum Associates.

Evans, J.St.B.T. (1992). Bias in Thinking and Judgement. In M.T. Keane & K.J. Gilhooly (Eds.). *Advances in the psychology of thinking (Vol.1).* Harvester-Wheatsheaf.

Evans, J.St.B.T., Harries, C., Dennis, I., & Dean, J. (1995). Tacit and Stated Policies in the Prescription of Lipid Lowering Agents. *British Journal of General Practice 45,* 15-18.

Fasoli, A., Lucchelli, S., Blasi, F., Tosi, C., & Colombini, W. (1992). Judgment Analysis of physicians facing the problem of provisional diagnosis of coronary heart disease. *Medical Decision Making 12(4)*, p.335 (abstract).

Fenichel, G.S., Murphy, J.G., Wigton, R.S., Schwartz, J.S. (1984). Results of Physician Decision Making using Results of Conjoint Analysis compared to Physician Decision Making in Actual Practice. *Medical Decision Making 4(4)*, 528 (abstract).

Fisch, H.-U., Gillis, J.S., & Daguet, R. (1982). A Cross-National Study of Drug Treatment Decisions in Psychiatry. *Medical Decision Making 2 (2)*, 167-177.

Fisch, H.-U., Hammond, K.R., Joyce, C.R.B., & O'Reilly, M. (1981). An Experimental Study of the Clinical Judgment of General Physicians in evaluating and prescribing for Depression. *British Journal of Psychiatry 138*, 138-109.

Fishburn, P.C. (1988). Normative Theories of Decision Making Under Risk and Under Uncertainty. In D.E. Bell, H. Raiffa, & A. Tversky (Eds.). *Decision Making: Descriptive, normative & prescriptive interactions*. Cambridge University Press.

Flanagan, O. (1991). Cognitive Science and Artificial Intelligence: Philosophical Assumptions and Implications. In O. Flanagan (Ed.). *The Science of the Mind*. MIT Press.

Ford J.L. (1987). *Economic choice under uncertainty. A perspective theory approach.* Edward Elgar Publishing Ltd.

Fox, J. (1984). Formal and Knowledge-based methods in Decision Technology. *Acta Psychologica 56*, 303-331. Also in J. Dowie & A.S. Elstein (Eds.). *Professional Judgment: A reader in clinical decision making*. Cambridge University Press.

Fox, J. (1985). Knowledge and Judgement in Decision-making. In M. Sheldon, J. Brooke & A. Rector (Eds.). *Decision-Making in General Practice*. The Macmillan Press Ltd.

Frisse, M.E. (1992). Medical Informatics in Academic Health Science Centers. *Academic Medicine, 67 (4)*, 238-241.

Gale, J., & Marsden, P. (1985). Diagnosis: Process not Product. In M. Sheldon, J. Brooke, and A. Rector (Eds.). *Decision-Making in General Practice*. Macmillan Press Ltd.

Ganzach, Y., & Czackes, B. (1995). On Detecting Nonlinear Noncompensatory Judgment Strategies: Comparison of Alternative Regression Models. *Organizational Behavior & Human Decision Processes 61 (2)*, 168-176.

Gilhooly, K.J. (1990). Cognitive psychology and medical diagnosis. *Applied Cognitive Psychology, 4*, 261-272.

Gilhooly, K.J., & Simpson, S. (1992). Deep Knowledge in Human Medical Expertise. In E. Keravnou (Ed.). *Deep Models for Medical Knowledge Engineering*. Elsevier Amsterdam.

Goldberg, L.R. (1968). Simple Models Or Simple Processes?: Some research on clinical judgments. *American Psychologist 23*, 483-496.

Goldstein, H.C., & Mitzel, W.M. (1992). The relative importance of relative importance: Inferring other peoples' preferences from relative importance ratings and previous decisions. *Organizational Behavior & Human Decision Processes 51*, 382-415.

Goran, M.J., Williamson, J.W., & Gonnella, J.S.A. (1973). The validity of patient management problems. *Journal of Medical Education 48*, 171-177.

GP Pocket Guide to Cholesterol (1994). Haymarket Publishing Services Ltd.

Grant, J., & Marsden, P. (1987). The structure of memorised knowledge in students and clinicians. *Medical Education 21*, 92-98.

Grant, R. (1992). *Which? Medicine*. Which? Books Consumers' Association and Hodder and Stoughton.

Griffin J.P. (1990). Factors affecting medicine usage and rates of consultation. In Jane Griffin (Ed.). *Factors Influencing Clinical Decisions in General Practice*. Office of Health Economics.

Groen, G.J., & Patel, V.L. (1985). Medical Problem Solving: Some questionable assumptions. *Medical Education 19*, 95-100.

Hammond, K.R., McClelland, G.H., & Mumpower, J. (1980). *Human Judgment and Decision Making: Theories, Methods and Procedures.* Hemisphere Publishing Corporation.

Hammond, K.R., Stewart, T.R., Brehmer, B., & Steinmann, D.O. (1975). Social-Judgment Theory. In M. Kaplan & S. Schwartz (Eds.). *Human Judgment & Decision Processes.* Academic Press Inc.

Hammond, K.R., & Summers, D.A. (1965). Cognitive Dependence on Linear and Non-linear cues. *Psychological Review 72 (3)*, 215-224.

Hammond, K.R., & Summers, D.A. (1972). Cognitive Control. *Psychological Review 79 (1)*, 58-67.

Hamm, R.M. (1988). Clinical Intuition and Clinical Analysis: Expertise and the Cognitive Continuum. In J. Dowie and A.S. Elstein (Eds.). *Professional Judgment: A reader in clinical decision making.* Cambridge University Press.

Harries, C., Evans, J.St.B.T., Dennis, I.., & Dean, J. (Accepted). A Clinical Judgment Analysis of Prescribing Decisions In General Practice. *Le Travail Human.*

Heller, R.F., Saltzstein, H.D., & Caspe, W.B. (1992). Heuristics in Medical and Non-medical Decision Making. *The Quarterly Journal of Experimental Psychology 44A(2)*, 211-235.

Heller, T. (1987) The pathology of coronary heart disease. In T. Heller, L. Bailey, M. Gott, and M. Howes (Eds.), *Coronary Heart Disease: Reducing the risk.* Chichester, U.K. John Wiley & sons.

Heller, T., Bailey, L., Gott, M., & Howes, M. (Eds.). (1987). *Coronary Heart Disease: Reducing the Risk.* Chichester Wiley.

Hixon, J.G., & Swann, W.B. (1993). When Does Introspection Bear Fruit? Self-Reflection, Self-Insight, and Interpersonal Choices. *Journal of Personality and Social Psychology 64(1)*, 35-43.

Hoffman, P.J. (1960). The Paramorphic Representation of Clinical Judgment. *Psychological Bulletin 57(2)*, 116-131.

Hoffman, P.J., & Blanchard, W.A. (1961). *A study of the effects of varying amounts of predictor information on judgments.* Oregon Research Institute Research Bulletin, 1961.

Hoffman, P.J., Slovic, P., & Rorer, L.G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin 69 (5)*, 338-349. (Abbreviated version in Arkes & Hammond, 1986, p.568).

Holmes, M.M., Rovner, D.R., Rothert M.L. Schmitt, N., Given, C.W., & Ialongo, N.S. (1989). Methods of Analyzing Physician Practice Patterns in Hypertension. *Medical Care 27(1)*, 59-68.

Holzemer W.L., Schleutermann, J., Farrand, L.L., & Miller, A.G. (1981). A Validation Study: Simulations as a Measure of Nurse Practitioners' Problem Solving Skills. *Nursing Research 30 (3)*, 139-144.

Holzman, G., Ravitch, M., Metheny, W. Rothert, M., Holmes, M., & Hoppe, R. (1984). Physicians' Judgments About Estrogen Replacement Therapy for Menopausal Women. *Obstetrics & Gynecology 63 (3)*, 303-311.

Howell, D.C. (1982). *Statistical Methods for Psychology.* Duxbury Press, Boston, Massachusetts.

Hursch, C.J., Hammond, K.R., & Hursch, J.L. (1964). Some Methodological considerations in multiple-cue probability studies. *Psychological Review 71 (1)*, 42-60.

Hynes, L.M., Levine, A.S., Littenberg, B., & Nease, R.F. (1994). Development and comparison of two utility based measures of regret. Presented at 16th Annual

Meeting of Society for Medical Decision Making. *Medical Decision Making 14 (4)*, 433 (abstract).

Janis, I.L. & Mann, L. (1977). *Decision Making: A psychological analysis of conflict, choice and commitment.* New York: Free Press.

Kahneman, D. & Tversky, A. (1979). Prospect Theory: An analysis of Decision under risk. *Econometrica 47 (2).* 263-291.

Kaplan, M.F. (1975). Information Integration in Social Judgment: Interaction of Judge and Informational components. In M. Kaplan & S. Schwartz (Eds.). *Human Judgment & Decision Processes.* Academic Press Inc.

Kassirer, J.P., Kuipers, B.J., & Gorry, A. (1982). Toward a theory of clinical expertise. *American Journal of Medicine 73, 251-9.* Also in J. Dowie and A.S. Elstein (Eds.) *Professional Judgment: A reader in clinical decision making.* Cambridge University Press.

Kelley, H.H. (1972). Causal schemata and the attribution process. In E.E. Jones, D.E. Kanouse, H.H.Kelley, R.E. Nisbett, S. Valins, & B. Weiner (Eds.). *Attribution: Perceiving the causes of behavior.* General Learning Press.

Kellogg, R.T. (1982). When can we introspect accurately about mental processes? *Memory and Cognition 10 (2),* 141-144.

Kirwan, J.R., Chaput de Saintonge, D.M., Joyce, C.R.B., Currey, H.L.F. (1983a). Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Diseases 42,* 644-647.

Kirwan, J.R., Chaput de Saintonge, D.M., Joyce, C.R.B., & Currey, H.L.F. (1983b) Clinical Judgment in rheumatoid arthritis II. Judging 'current disease activity' in clinical practice. *Annals of the Rheumatic Diseases 42, 648-651.* (Also in Arkes & Hammond, 1986).

Kirwan, J.R., Chaput de Saintonge, D.M., Joyce, C.R.B., & Currey, H. (1983c). Clinical Judgement Analysis - Practical Application in Rheumatoid Arthritis. *British Journal of Rheumatology 22 (supplement),* 18-23.

Kirwan, J.R., Chaput de Saintonge, D.M., Joyce, C.R.B., Holmes, J., & Currey, H. (1986). Inability of rheumatologists to describe their true policies for assessing rheumatoid arthritis. *Annals of the Rheumatic Diseases 45,* 156-161.

Kirwan, J.R., Chaput de Saintonge, D.M., Joyce, C.R.B. (1990). Clinical Judgment Analysis. *Quarterly Journal of Medicine. New Series 76 (281),* 935-949.

LaDuca, A., Engel, J.D., & Chovan, J.D. (1988). An Exploratory Study of Physicians' Clinical Judgment: An application of Social Judgment Theory. *Evaluation & The Health Professions 11 (2),* 178-200.

Laker, M., Reckless, J., Durrington, P., Miller, P., Nicholls, P., Sheperd, J., & Thompson, G. (1991). Facilities for the management of patients with lipid disorders in the UK: Results of the British Hyperlipidaemia Association Survey. *Health Trends 23 (4),* 147-149.

Lee, W. (1971). *Decision Theory and Human Behavior.* John Wiley & Sons, Inc.

Lipe, M.G. (1990). A Lens model analysis of covariation research. *Journal of Behavioral Decision Making 3,* 47-59.

Loomes, G., & Sugden, R. (1982). Regret Theory: An alternative theory of rational choice under uncertainty. *The Economic Journal 92,* 805-824.

Luckett, P.F., & Hirst, M.K. (1989). The impact of feedback on inter-rater agreement and self insight in performance evaluation decisions. *Accounting Organizations and Society 14 (5/6),* 379-387.

Lundsgaarde, H.P. (1987) Evaluating Medical Expert Systems. *Social Science & Medicine 24(10),* 805-819.

MacGregor, E.A. (1993). Prescribing for Migraine. *Prescribers' Journal 33 (2),* 50-58.

Marmot, M. (1994). The cholesterol papers. *British Medical Journal 308*, 351-352.

Martin, H.T. Jnr. (1957). *The Nature of Clinical Judgment*. Unpublished Doctoral Dissertation, Washington State College, 1957.

Mathews R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J-R., & Druhan, B. (1989). The role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory and Cognition 15*, 1083-1100.

Maule, A.J., & Svenson, O. (1993) Theoretical and Empirical Approaches to Behavioral Decision Making and Their Relation to Time Constraints. In A. John Maule & Ola Svenson (Eds.). *Time Pressure & Stress in Human Judgment & Decision Making*. New York: Plenum.

McGeorge, P., & Burton, M. (1989). The effects of concurrent verbalisation on performance in a dynamic systems task. *British Journal of Psychology 80*, 455- 465.

McNeil, B.J., Pauker, S.G., Sox, H.C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine 306*, 1259-1262.

McNeil, B.J., Pauker, S.G., & Tversky, A. (1988). On the Framing of Medical Decisions. In D.E. Bell, H. Raiffa & A. Tversky (Eds.). *Decision Making: Descriptive, normative and prescriptive interactions*. Cambridge University Press.

McWhinney, I.R. (1985). Patient-centred and Doctor-centred models of Clinical Decision Making. In M. Sheldon, J. Brooke and A. Rector (Eds.). *Decision-Making in General Practice*. MacMillan Press Ltd.

Mear, R., & Firth, M. (1987). *Cue Usage and Self-Insight of Financial Analysts*. The Accounting Review LXII (1). 176-182.

Meehl, P.E. (1954). *Clinical Versus Statistical Prediction*. University of Minnesota Press.

Meehl, P.E. (1965). Seer over sign:the first good example. *Journal of Experimental Research in Personality 1*, 27-32.

Merikle, P.M. (1992). Perception Without Awareness: Critical Issues. *American Psychologist 47*, 792-795.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 81-97.

Morrell, D.C., Evans, M.E., Morris, R.W., & Roland, M.O. (1986). The "five minute" consultation: effect of time constraint on clinical content and patient satisfaction. *British Medical Journal 292*, 870-873.

Morrell, D.C., &.Roland, M.O. (1990). Analysis of referral behaviour: responses to simulated case histories may not reflect real clinical behaviour. *British Journal of General Practice 40*, 182-185

Newell, A., & Simon, H.A. (1972). *Human Problem Solving*. Prentice-Hall.

Nisbett, R.E., & Ross, L. (1980). *Human inference: strategies & shortcomings of social judgment*. Prentice Hall.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we know: Verbal Reports on Mental Processes. *Psychological Review 84 (3)*, 231-258.

Norman, G.R., & Feightner, J.W. (1981). A comparison of behaviour on simulated patients and patient management problems *Medical Education 15*, 26-32.

Oliver, M.F. (1987). Strategies for preventing and screening for coronary heart disease. In T. Heller, L. Bailey, M. Gott & M. Howes (Eds.). *Coronary Heart Disease: A Reader*. Chichester John Wiley & sons.

Olshavsky, R.W. (1979). Task complexity and contingent processing in decision making: A replication and extension. *Organizational Behavior and Human Performance 24*, 300-316.

Page, G.G., & Fielding, G.W. (1980). Performance on PMPs and performance in

practice: are they related? *Journal of Medical Education 55*, 529-537.

Patel, V.L. & Groen, G.J. (1986). Knowledge Based Solution Strategies in Medical Reasoning. *Cognitive Science 10*, 91-116.

Patel V.L., Groen, G.J., & Arocha, J.F. (1990). Medical expertise as a function of task difficulty. *Memory & Cognition 18(4)*. 394-406.

Phelps, R.H., & Shanteau, J. (1978). Livestock Judges: How Much Information Can an Expert Use? *Organizational Behavior and Human Performance 21*, 209- 219.

Politser, P.E. (1991). Do medical decision analyses largest gains grow from the smallest trees? *Journal of Behavioural Decision Making 4*, 121-138.

Poses, R.M., Cebul, R., Collins, M., & Fager, S. (1985). The Accuracy of Experienced Physicians' Probability Estimates for Patients With Sore Throats: Implications for Decision Making. *Journal of the American Medical Association 254(7)*, 925-929.

Poses, R.M., Cebul, R.D., & Wigton, R.S. (1986). Feedback on Simulated Cases Improves Doctor's Probability Estimates. *Clinical Research 34 (2)*. 832A (abstract).

Poses, R.M., Cebul, R.D., & Wigton, R.S. (1995). You Can Lead a Horse to Water-Improving Physicians' knowledge of Probabilities May Not Affect Their Decisions. *Medical Decision Making 15 (1)*, 65-75.

Poses, R.M., Cebul, R.D., Wigton R.S., Collins, M. (1986). Feedback on Simulated Cases to improve Clinical Judgment. *Medical Decision Making* p. 274 (abstract).

Reber, A.S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology (General). 118 (3)*, 219-235.

Redelmeier, D.A., Koehler, D.J., Liberman, V., & Tversky, A. (1995). Probability Judgment in Medicine: Discounting Unspecified Possibilities. *Medical Decision Making 15*, 227-230.

Reilly, B.A., & Doherty, M.E. (1989). A note on the Assessment of Self-Insightin Judgment Research. *Organizational Behavior and Human Decision Processes 44*, 123-131.

Reilly, B.A., & Doherty, M.E. (1992). The Assessment of Self-Insight in Judgment Policies. *Organizational Behavior & Human Decision Processes 53*, 285-309.

Richardson, D.K., Gabbe, S.G., & Wind, Y. (1984). Decision Analysis of High-Risk Patient Referral. *Obstetrics and Gynecology 63 (4)*, 496-501.

Rich, M.C. (1979). Verbal Reports on Mental Processes: Issues of Accuracy and Awareness. *Journal for the Theory of Social Behaviour 9*, 29-37.

Roose, J.E., & Doherty, M.E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior & Human Performance 16*, 231-249.

Roose, J.E., & Doherty, M.E. (1978). A Social Judgment Theoretic Approach to Sex Discrimination in Faculty Salaries. *Organizational Behaviour and Human Performance 22*, 193-215.

Rorer, L.G. (1971). A circuitous route to bootstrapping. In H.B. Haley, A.G. D'Costa, A.M. Schafer (Eds.). *Conference on personality measurement in medical education.* Washington DC Association of American Medical Colleges.

Rothert, M.L. (1982). Physicians' and Patients' Judgments of Compliance with a Hypertensive Regimen. *Medical Decision Making 2 (2)*, 179-195.

Rothert, M., Rovner, D., Elstein A., Holzman, G., Holmes, M., & Ravitch, M. (1984). Differences in Medical Referral Decisions for Obesity Among Family Practitioners General Internists and Gynecologists. *Medical Care 22 (1)*, 42-55.

Rovner, D.R., Rothert, M., Holmes, M., Ravitch, M., Holzman, G., & Elstein, A. (1985) Rationale for Physicians' Decisions to Refer Obese Patients. *Medical Decision Making 5 (3)*, 279-292.

Rovner, D,R., Rothert, M.L., Holmes, M.M., Given, C.W., & Ialongo, N.S. (1986). Validation of case vignettes with clinical practice: the case of UTI. *Medical Decision*

Making 6 (4), 272 (abstract).
Sabini, J. & Silver, M. (1981). Introspection and Causal Accounts. *Journal of Personality and Social Psychology 40*, 171-179.
Schmidt, H.G., & Boshuizen, H.P.A. (1993). On Aquiring Expertise in Medicine. Unpublished manuscript.
Schmitt, N., Gogate, J., Rothert, M., Rovner, D., Holmes, M., Talarcyzk, G., Given, B., & Kroll, J. (1991). Capturing and Clustering Women's Judgment Policies: The Case of Hormonal Therapy for Menopause. Journal of Gerontology: *Psychological Sciences 46 (3)*, 92-101.
Schmitt N., & Levine R. L. (1977). Statistical and Subjective weights: some problems and proposals. *Organizational Behavior and Human Performance 20*, 15-30.
Schneider, W., & Shiffrin, R. (1985). Categorisation (restructuring). and automatisation: Two separable factors. *Psychological Review 92*, 424-428.
Schooler, J.W., Ohlsson, S., & Brooks, K. (1993). Thoughts Beyond Words: When language Overshadows Insight. *Journal of Experimental Psychology (General). 122(2)*, 166-183.
Shackle, G.L.S. (1952). *Expectation in Economics.* Cambridge University Press.
Shackle, G.L.S. (1961). *Decision, Order and Time in Human Affairs.* Cambridge University Press.
Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica 81*, 75- 86.
Sharp I. (Ed.). (1994). *Coronary heart disease: Are women special?* National Forum for coronary heart disease prevention.
Sheldon, M., Brooke, J., & Rector, A. (Eds.). (1985). *Decision-making in General Practice.* The Macmillan Press Ltd.
Simon, H.A. (1979). How big is a chunk? 1974 in H.A. Simon (1979) *Models of thought.* Yale University Press.
Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral Decision Theory. *Annual Review of Psychology 28*, 1-39.
Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgement. *Organizational Behavior & Human Performance 6*, 649-744.
Slovic, P., Rorer, L.G., & Hoffman, P.J. (1971). Analyzing the use of diagnostic signs. *Investigative Radiology 6*, 18-26.
Smith, D.G., & Wigton, R.S. (1983). Use of conjoint analysis to determine how physicians weight ethical considerations in making clinical judgments. *Medical Decision Making 3*, 376 (abstract).
Speroff, T., Connors, A.F., & Dawson, N.V. (1989). Lens Model Analysis of Hemodynamic Status in the Critically Ill. *Medical Decision Making 9 (4)*, 243-252.
Steinmann, D.O. (1974). Transfer of Lens Model Training. *Organizational Behavior & Human Performance 12*, 1-16.
Stewart, T.R. (1988). Judgment Analysis: Procedures. In B. Brehmer & C.R.B. Joyce (Eds). *Human Judgment: The SJT View.* Elsevier Science Publishers B.V. (North-Holland).
Stewart, T.R., & Joyce, C.R.B. (1988). Increasing the Power of Clinical Trials Through Judgment Analysis. *Medical Decision Making 8 (1)*, 33-38.
Summers, D.A., Talioferro, J.D., & Fletcher, D.J. (1970). Subjective vs. objective description of judgment policy. *Psychonomic Science 18 (4)*, 249-250.
Tape, T., Heckerling, P., Ornato, J., & Wigton, R. (1991). Use of Clinical Judgment Analysis to Explain Regional Variations in Physicians' Accuracies in Diagnosing Pneumonia. *Medical Decision Making 11 (3)*, 189-197.

Tape, T.G., Kripal, J., & Wigton, R.S. (1989). Learning to estimate cardiac risk from computer generated cases: The type of feedback is critical to success. *Medical Decision Making* p.323 (abstract).

Tape, T.G., Kripal, J., & Wigton, R.S. (1992). Comparing Methods of Learning Clinical Prediction from Case Simulations. *Medical Decision Making 12 (3)*, 213-221.

Tape, T.G., & Wigton, R.S. (1989) Medical Students' and Residents' Estimates of Cardiac Risk. *Medical Decision Making 9 (3)*, 170-175.

Tolman, E.C. (1948). Cognitive Maps in Rats and Men. *Psychological Review 55*, 189-208.

Tucker, L.R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch and by Hammond, Hursch and Todd. *Psychological Review 71*, 528-530.

Tversky, A. (1972). Elimination by Aspects: A theory of choice. *Psychological Review, 79*, 281-299.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Reprinted in D. Kahneman, P. Slovic & A. Tversky (Eds.). *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press.

Tversky, A., & Kahneman, D. (1988). Rational Choice and the Framing of Decisions. In D.E. Bell, H. Raiffa and A. Tversky (Eds.), *Decision Making: Descriptive, normative and prescriptive interactions.* Cambridge University Press.

Tversky, A., & Koehler, D.J. (1994). Support Theory: a non-extensional representation of subjective probability *Psychological Review 101*, 547-67.

Ullman, D.G., & Doherty, M.E. (1984). Two determinants of the diagnosis of hyperactivity: The child & the clinician. *Advances in Developmental and Behavioral Pediatrics 5*, 167-219.

Ullman, D., Egan, D., Fiedler, N., Jurenec, G., Pliske, R., Thompson, P., & Doherty, M. (1981). The Many Faces of Hyperactivity: Similarities and Differences in Diagnostic Policies. *Journal of Consulting and Clinical Psychology 49 (5)*, 694-704.

Verhoef, L.C.G., de Haan, A.F.J., & van Daal, W.A.J. (1994). Risk attitude in Gambles with Years of Life: Empirical Support for Prospect Theory. *Medical Decision Making 14 (2)*, 194-200.

von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior.* Princeton: Princeton University Press. (3rd Edn., 1953)

von Winterfeldt, D., & Edwards, W. (1986) *Decision Analysis and Behavioural Research.* Cambridge: Cambridge University Press.

White, P.A. (1988). Knowing more about what we can tell: 'Introspective access' and causal report accuracy 10 years later. *British Journal of Psychology 79*, 13-45.

Wiggins, R., & Hoffman, P.J. (1968). Three models of Clinical Judgment. *Journal of Abnormal Psychology 73 (1)*, 70-77.

Wigton, R.S. (1987). The use of computer simulation in teaching clinical diagnosis. *Computer Methods and Programs in Biomedicine 25*, 111-114.

Wigton, R.S. (1988a). Use of Linear Models to Analyze Physicians' Decisions. *Medical Decision Making 8 (4)*, 241-252.

Wigton, R.S. (1988b). Applications of Judgment analysis and cognitive feedback to medicine. In B. Brehmer & C.R.B.Joyce (Eds.). *Human Judgment: The SJT View.* Elsevier Science Publishers B.V. (North-Holland).

Wigton, R.S., Hoellerich, V.L., & Patil, K.D. (1986). How Physicians Use Clinical Information In Diagnosing Pulmonary Embolism: An application of conjoint analysis. *Medical Decision Making 6 (1)*, 2-11.

Wigton, R.S., Patil, K.D., & Hoellerich, V.L. (1986). The Effect of Feedback In Learning Clinical Diagnosis. *Journal of Medical Education 61*, 816-822.

Wigton, R.S., Poses, R.M., Collins, M., & Cebul, R.D. (1990). Teaching Old Dogs New Tricks: Using Cognitive Feedback to Improve Physicians' Diagnostic Judgments on Simulated Cases. *Academic Medicine 65(9)*, s5-s6.

Wilson, T.D., & Schooler, J.W. (1991). Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions. *Journal of Personality and Social Psychology 60 (2)*, 181-192.

Wright, G. (1984). *Behavioural Decision Theory. (An Introduction)*. Penguin Books.

Young, M.J., Woodiscroft, J.O., & Holloway, J.J. (1986). Determining the policies of a residency selection committee. *Journal of Medical Education 61*, 835-837.

Appendix 1
Examples of JA Studies

This appendix contains details from 43 judgement analysis studies presented in two sets. Studies 1 to 20 are presented on pages 252 to 255, studies 21 to 43 are presented on pages 256 to 259.

Key:-
Name and date = reference of study
Cues = number of cues presented
Cases = number of cases presented
Repeats = number of cases repeated for a consistency check
Design = case and cue design of the study:
- FF = Full factorial design
- Frac F = fractional factorial design
- Real = real life judgements
- Real paper = details of real cases presented in paper form
- Nomothetic = analysis done over more than one subject's responses
- Linear or non-linear describes the relationship of cues to criterion
- Orth. = orthogonal cues
- Random = randomly generated cases
- Real distribution = cases designed with the range and mix of cues resent in the real world
- Paper = paper cases used
- MCPL = multiple cue probability learning task

Subjects = type of subject used
Judgement = judgement or decision being made

Crit = linear fit of the criterion
Obj = linear fit of the objective model
subj = linear fit of the subjective model
Equal = linear fit of an equal weights model
Rand = linear fit of a random weights model
cons = consistency
r(s-i) = self-insight measure: correlation of subjective and objective weights
r_A = achievement or agreement

Notes = notable aspects of the study
(N) = see notes.

See also Table 4.1 and text.

Appendix 1 continued.

	Name	Date	Cues	Cases	Repeats	Design
1	Balzer, Sulsky, Hammer, & Sumner	1992	5	50		Real paper
2	Billings & Marcus	1983	4	27		FF
3	Blood	1971	5	-	-	Nomothetic
4	Brehmer	1987	2/4	25+25		linear fn
						non-linear
5	Chaput de Saintonge & Hathaway	1981	Proforma	48	41 +photo	Real(N)
6	Chaput de Saintonge & Hattersley	1985	24	50	19	Real paper
7	Chaput de Saintonge *et al*	1988	10	30	20	Real paper
8	Cook & Stewart	1975	3	50		Orth.
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
			3	50		
			7	50		
9	Dawes & Corrigan	1974	11	?1200		Real
			10	90		Real
			3	?		Real
			3	?		?
10	Einhorn	1974	9	193	26	Real slides
11	Einhorn, Kleinmuntz & Kleinmuntz	1979	10	96		Real
			4			
			11	67	20	Real
12	Elstein *et al*	1992	2	6	6	FF
13	Fisch, Hammond, Joyce & O'Reilly	1981	8	41	39	Random
			8			
14	Fisch, Gillis & Daguet	1982	8	30	10	Random
15	Holmes *et al*	1989	4	16	4	Real distn.
16	Holzman *et al*	1984	4	24	6	FF
17	Kirwan, Chaput de Saintonge, Joyce & Currey.	1983a	15	Total 70	70	Real, paper
18	Kirwan *et al* (as above).	1983b	5	Total 19	19	Real, paper
19	Kirwan *et al* (as above).	1983c	20	50	20	Paper
20	Kirwan *et al* (as above + Holmes)	1986	10	30	20	Paper

Appendix 1 continued.

	Subjects	Judgement
1	133 students	Baseball team wins
		+ second session
2	48 students -Time pressure	Apartment desirability
	-No time pressure	
3	380 clerical workers	Job satisfaction
4	32 Students	MCPL task
5	7 GPs	Probability of Otitis Media,
		other, & treatment given
6	5GPs 1, trainee GP	" " " "
7	48 rheumatologists (15)	Degree of change
8	21 students & staff	Financial Assessment (FA)
	21 students & staff	Grade Assessment (GA)
	19 students & staff	FA
	19 students & staff	GA
	17 students & staff	FA
	17 students & staff	GA
	29 students & staff	FA
	29 students & staff	GA
	21 students & staff	FA
	21 students & staff	GA
	16 students & staff	FA
	16 students & staff	GA
	20 students & staff	FA
	20 students & staff	GA
9	29 clinical psychologists	MMPI, neurotic/psychotic
	80 students	Grade point averages
	Faculty	Admissions rating
	subjects	"Values" of elipses
10	3 pathologists	Features of a biopsy
		+ disease severity
11	MMPI users (1 selected)	Degree of adjustment
	1 subject	Nutrition of cereals by category
12	21 residents (3rd yr)	Certainty of prescribing HRT,
		later categorised.
		Knowledge base & practice consistent for 13/21
13	15 GPs (Swiss)	Severity of depression &
	14 GPs	any drug prescribed
14	24 swiss,70 US Psychiatrists	Treatment of psychiatric patients
15	98 physicians	Number of tests ordered
16	25 GPs, 25 gynecologists	Probability of HRT
17	9 rheumatologists	Severity of RA
18	2 rheumatologists	Severity of RA
19	48 rheumatologists	Change in severity of RA
20	89 rheumatologists	Change in severity of RA
	4 rheumatologists	

Appendix 1 continued.

	Crit	Obj	Subj	Equal	Rand	cons.	r(s-i)	r_A
1	0.89	0.74						0.34
		0.79						0.56
		0.83						0.43
		0.77						0.54
		0.79						0.48
2		0.67						
		0.78						
3		R = 0.43						
4	0.96							0.85, 0.89
								0.16, 0.29
5						p < 0.05		
6		80-100%				79-100%	(N)	
7		>0.8				> 0.8		k = 0.3
8		0.79	0.75					
		0.73	0.58					
		0.82	0.76					
		0.69	0.49					
		0.82	0.72					
		0.71	0.49					
		0.81	0.74					
		0.69	0.54					
		0.82	0.75					
		0.76	0.55					
		0.81	0.72					
		0.67	0.5					
		0.83	0.75					
		0.7	0.52					
9								0.28
								0.33
								0.19
								0.84
10		0.46, 0.27,				0.19-0.93		0.33-0.80
		0.49 (N)				0.63(N)		0.27(N)
11		R = 0.88						
		0.87						
						3 errors		
12						6/21 consi		
13		R = 0.77				0.69		
		R = 0.86	0.49	0.25	0.22	0.77		
14						93%,83%(N)		
15		0.56				78%		
16		0.69						
17						0.90, 0.97		
18		0.95, 0.94						
19		2/3s > 0.75				c0.75		0.63,0.64
								0.76
20		0.73	0.39	0.41				
		0.88	0.34					

Appendix 1 continured.

	Notes
1	Components of Cognitive Feedback: no feedback, TI, CI, TI + CI, TI + CI + FVI (see Chapter 3 & 4).
2	Comparison with process tracing
3	Cues classified in terms of their importance used in regression. Weights not corresponding.
4	No effect of no. of cues. S concerned with cue combinations rather than ind. rns.
5	Real life, written & photo presentation compared. Validity test.
6	Agreement compared over time. Logical models not statistical ones. Expressed policies compared with fitted ones: worse fit.
7	Results reported for 15 subjects. Little agreement between them.
8	Comparison of different measures of self-insight. 100 Rate (1-100) Rate (1-100) Paired comparison Paired comparison Ratio Ratio No. of times influenced No. of times influenced ICR I + 100 ICR I + 100 ICR II ICR II
9	Summarised others studies: (Goldberg, 1965) (Wiggins & Cohen, 1971) (Dawes, 1971) (Yntema & Togerson, 1961)
10	R^2 of 3 subjects. Judgements made of signs and also global judgements
11	12 step forced normal distn. (Q sort) Regression model good compared to Process Tracing model Cues amalgamated: 3 in model
12	(1) Think aloud during decision making (2) Interview re probabilities & utilities
13	Based on HDRS GP chose own variables
14	Comparison between Groups, agreement on drug class.
15	Validity: written cases comp. chart audit. Standard algorithm vs consistent cue use.
17	Validity of paper cases, proforma used. All real presented as paper, some paper repeats.
18	Real, paper & paper repetitions in model. Flat distribution of subjective weights shown.
19	CJA, before & after info as cues in model. Average of 6 cues used per doc. Agreement between 2 s: no improvement with discussion but improvement with discussion over CJA models.
20	Equal > subjective model for 49/89 s. 4 subjects interviewed re. subjective policies to obtain considered policy (no better).

Appendix 1 continued.

	Name	Date	Cues	Cases	Repeats	Design
21	LaDuca, Engel & Chovan	1988	5	30		Realisitic
22	Lipe	1990	4	34		(N)
23	Mear & Firth	1987	10	30		Orthogonal
24	Poses *et al*	1985	27	308		Real
25	Reilly & Doherty	1989	19	115 + 45		Paper
26	Reilly & Doherty	1992	12	100		Orth.
			12	100		Real paper
			6	100		Orth.
			6	100		Real paper
27	Richardson, Gabbe & Wind	1984	7	18		Paper
28	Roose & Doherty	1976	64(9)	200 + 160		Real
29	Roose & Doherty	1978	*28	175	25	Real
30	Rothert	1982	4	27		Frac. F
31	Rothert *et al.*	1984	4	24	-	FF
32	Rovner *et al*	1985	4	24	6	FF
33	Russell	1985	?18	c200 each		Real
34	Schmitt *et al*	1991	3	8	8	FF
35	Sperroff, Connors & Dawson	1989	32	500		Real
36	Steinmann	1974	3	100+100		MCPLtask
37	Summers, Talioferro & Fletcher	1970	4	175		
38	Tape *et al*	1991	12 clusters			Real
39	Ullman *et al*	1981	19	80	15	Random*
40	Ullman & Doherty	1984	7	52		Real, paper
			19	80	15	Random*
41	Wigton, Hoellerich & Patil	1986	8	27		Frac. F
42	Wigton, Patil & Hoellerich	1986	7	18 (X3)		Frac. F
43	Wigton, Poses, Collins & Cebul	1990	7	12(X4)		Frac. F

Appendix 1 continued.

Ref	Subjects	Judgement
21	13 physicians	Severity of Congestive Heart Failure
22	Mainly students	Judgements of covariation
23	38 financial analysts etc.	Portfolio risk & return
24	10 physicians	Probability of strep throat & treatment (antibiotics).
25	40 faculty students	Desirability of job offers.
26	19 students	Room-mate desirability
	18 students	Room-mate desirability
	19 students	Room-mate desirability
	21students	Room-mate desirability
27	211 Obs & Gyne.	Likelihood of referral
28	16 Agency managers	Sucess/failure + confidence.
29	42 faculty judges	Salary
30	30 patients	Likelihood of compliance
	30 physicians	of hypothetical patients.
31	45 primary care physicians	Endocrine Disorder
32	Total of 45 physicians	Likelihood of referral. (A)
	15 Family practitioners	Morbidity due to weight. (B)
	15 Gynecologists	Endocrine disorder. (C)
	15 Internal Medicine	
33	10 Assessors	Overall Assessment Rating.
34	265 perimenopausal women	Likelihood of taking HRT
35	123 physicians	Hemodynamic dysfunction
36	8 students: LMF (A)	
	4 students: FF (B)	
	4 students: OF (C)	
37	131 students	Level of S.E. development
38	Illinois Physicians	Est chest radiograph would show pneumonia
	Nebraska Physicians	
	Virginia Physicians	
39	16 clinical psychologists	Hyperactive plus confidence
40	11 professionals	Hyperactive plus confidence
	74 professionals	Hyperactive plus confidence
41	19 Medical students (group)	Likelihood of Pulmonary Embolism
	First year HOs (grouped)	? treat.
	Third year HOs (grouped)	
	13 Faculty members (group)	
	All grouped	
42	11 medical students (CF)	Likelihood of +ve UTI
	16 Medical students (OF)	
43	11 student physicians	Probability of strep throat
	12 medical students (2nd yr)	

Appendix 1 continued.

Ref	Crit	Obj	Subj	Equal	Rand	cons	r(s-i)	r_A
21	0.93	0.70-0.91						0.03-0.99
22	0.89	0.74						0.91
23		0.73,0.74	0.34,0.33	0.09,0.23(N)			0.30, 0.32	
25		0.70*	0.34*					
26		0.73					0.69	
		0.81					0.92	
		0.73					0.43	
		0.84					0.66	
28	0.43	0.75						0.13
29	0.83	0.87				0.91		
30		0.53	0.41					
		0.68	0.37					
31		0.68						
		0.82						
		0.75						
32		0.76						0.86
33		0.69-0.82						
35	0.45	0.59						0.42
36	R = 0.73	0.31, 0.72						0.18, 0.45
	R = 0.73	0.75, 0.81						0.37, 0.53
	R = 0.75	0.74, 0.85						0.51, 0.57
	R = 0.78	0.84, 0.87						0.58, 0.65
	R = 0.74	0.91, 0.90						0.55, 0.59
		0.56, 0.74						0.31, 0.52
		0.69, 0.85						0.44, 0.54
		0.69, 0.82						0.36, 0.56
		0.77, 0.85						0.51, 0.58
		0.66, 0.73						0.43, 0.49
		0.78, 0.80						0.55, 0.51
		0.86, 0.89						0.55, 0.50
		0.72, 0.79						0.43, 0.50
		0.86, 0.86						0.56, 0.59
		0.85, 0.79						0.55, 0.44
37		R = 0.75	R = 0.60					
38	0.39	0.59						0.41
	0.64	0.7						0.66
	0.59	0.63						0.55
39		0.75				67%		
40		(?)0.67						
		0.72				69: >70%		
41		0.41						
		0.34						
		0.57						
		0.43						
	0.2	0.4						
42								.55,.73,.79
								.55,.62,.64
43		R = .81,.81						0.4, 0.69
		R = .90,.98						0.34, 0.93

Appendix 1 continued.

Ref	Notes
21	Lens model analysis
22	Meta-analysis of several studies: different cells emphasised.
23	Results for risk & return judgements. Signed equal weights: 0.31, 0.25.
24	Overestimation for 81% Discussion of R-ship between probability & treatment.
25	* Correlations on holdout sample. Significant self-recognition. 1-10 cues used.
26	Representative data: cues grouped into five factors for analysis. Cross validation: beta-weights better than subjective weights for all but 2 subjects. Significant self-recognition. 1-10 cues used.
27	Conjoint analysis. Diabetes pregnancy.
28	Unit weights good for bootstrapping. Non-zero insight for 3 subjects.
29	* not including gender: Sex bias.
31	Av. judged typicality of cases = 67%. r(A,C) = 0.42, r(A,C) = 0.28. Two articles re. same study.
32	+ questionnaire. 7 point scale: >6 = referral.
33	3 or 4 factor solutions.
34	Oestrogen only vs. plus progestin. Discounted subjects whose $R^2 < 0.6$ (14). Clustering. Regression against subject variables.
36	Plus 20 practice trials per session Lens Model Feedback, Feedforward (no UI) or Outcome Feedback. (8, 5 & 5 tasks) Differences between groups & sessions. Same 5 tasks. First then second session. N.B. R not R^2 given.
37	Significantly smaller R^2 where non-linear models described, increase in fit with non-linear greater for those than for subjects describing linear models.
39	* ranges as in reality.
41	Conjoint analysis. Average individual $R^2 = 0.7$
42	Feedback, OF vs.CF
43	Lecture + 3 feedback sessions. All 4 in one go.

Appendix 2
Instructions for the LIPID task (Study 1, Chapter 5)

You will be presented with a series of 130 hypothetical cases in which you have to decide whether or not the individual should be prescribed a lipid lowering agent. Assume that the option to refer is not available. In each case you will be supplied with a number of details about the patient and their medical history.

You can assume that in each case you originally tested the patient's blood cholesterol level at least six months previously and have offered the usual advice on alterations to diet and have recommended, where appropriate, that the patient should give up smoking cigarettes. The blood cholesterol level given in the problem is the current one and reflects any changes in the patients life-style that he or she has made or is likely to make.

Given all the information, you have to decide whether or not you would now prescribe a lipid lowering drug. We realise that the cases may not be clear cut and that in practice you might seek further information. However, you can express your preference for prescribing the drug in percentage terms. If you are certain that you would prescribe then give a response of 100% and if you arecertain you would not give a response of 0%. Where you are not certain then please given any intermediate value you wish to convey the likelihood that you would prescribe.

In each case you can set the percentage by pulling the mouse to the left or right. When the reading is correct then please press any button on the mouse to proceed to the next problem.

PRESS ANY MOUSE BUTTON TO START

Appendix 3
Instructions for the MIGRAINE task (Study 1, Chapter 5)

The following are a series of hypothetical cases in which all the patients presented have been having some symptoms of migraine headache. You must decide whether or not to prescribe prophylactic treatment. None of the patients has had any symptoms to suggest that there is any more ominous cause and you have a high degree of confidence that this is purely a migraine headache.

You have seen the patient previously and in each case the patienthas modified their lifestyle to avoid trigger factors such as certain foods and flashing lights and to reduce stress.

Given the information in each case you have to decide whether or not you would treat the patient with prophylaxis. We realize that the cases may not be clear cut and that in practice you might seek further information. However, you can express your preference for prescribing a prophylactic in percentage terms. If you are certain that you would prescribe then give a response of 100% and if you are certain you would not give a response of 0%. Where you are not certain then please given any intermediate value you wish to convey the likelihood that you would prescribe.

In each case you can set the percentage by pulling the mouse to the left or right. When the reading is correct then please pressany button on the mouse to proceed to the next problem.

PRESS ANY MOUSE BUTTON TO START

Appendix 4
Instructions for the HRT task (Study 1, Chapter 5)

You will be presented with a series of 130 hypothetical cases in which you have to decide whether or not to treat the patient by prescribing some Hormone Replacement Therapy. In each case you will be supplied with certain pieces of information about the patient. The option to refer is not available.

You can assume that in each case there is no reason why the patient would be already receiving hormone therapy. They have a normal pelvic examination and there is no relevant medical history, other than that displayed in each case, which would affect your decision. Where appropriate you have given advice about smoking and diet.

Given all the information, you have to decide whether or not you would now treat the patient with some sort of hormone replacement therapy. We realize that the cases may not be clear cut and that in practice you might seek further information, and use all the information to make a choice between types of hormone replacement therapy. However, you can express your preference for prescribing some sort of HRT in percentage terms. If you are certain that you would prescribe then give a response of 100% and if you are certain you would not give a response of 0%. Where you are not certain then please given any intermediate value you wish to convey the likelihood that you would prescribe.

In each case you can set the percentage by pulling the mouse to the left or right. When the reading is correct then please press any button on the mouse to proceed to the next problem.

PRESS ANY MOUSE BUTTON TO START

Appendix 5
Inter-cue correlations on the LIPID task

	Cho	Hyp	Age	Gen	Occ	Eva	Smo	Dia	Com	Wei	Att	FH	Per
Cho	1	0.1	-0.03	0.05	0.13	-0.07	0.13	0.04	0.08	0.07	-0.05	-0.08	-0.17
Hyp	0.10	1	0.05	0.17	0.02	-0.01	-0.04	0.02	0.09	-0.03	-0.06	0.04	0.05
Age	-0.03	0.05	1	-0.02	0.07	-0.05	-0.11	0.03	-0.09	-0.09	0.01	0.03	-0.08
Gen	0.05	0.17	-0.02	1	-0.08	0.19	-0.01	-0.08	-0.06	0.02	-0.11	-0.05	-0.13
Occ	0.13	0.02	0.07	-0.08	1	-0.09	0.03	0.02	0.19	0.13	0.05	0.11	0.09
Eva	-0.07	-0.01	-0.05	0.19	-0.09	1	-0.1	-0.07	-0.02	0.09	-0.11	0.07	-0.14
Smo	0.13	-0.04	-0.11	-0.01	0.03	-0.10	1	0.01	-0.05	-0.11	0.00	-0.15	-0.14
Dia	0.04	0.02	0.03	-0.08	0.02	-0.07	0.01	1	0.12	-0.04	-0.09	0.09	-0.05
Com	0.08	0.09	-0.09	-0.06	0.19	-0.02	-0.05	0.12	1	0.1	-0.06	0.13	-0.07
Wei	0.07	-0.03	-0.09	0.02	0.13	0.09	-0.11	-0.04	0.1	1	-0.03	-0.04	0.07
Att	-0.05	-0.06	0.01	-0.11	0.05	-0.11	0.00	-0.09	-0.06	-0.03	1	0.15	-0.05
FH	-0.08	0.04	0.03	-0.05	0.11	0.07	-0.15	0.09	0.13	-0.04	0.15	1	0.08
Per	-0.17	0.05	-0.08	-0.13	0.09	-0.14	-0.14	-0.05	-0.07	0.07	-0.05	0.08	1

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality.

Appendix 6
Inter-cue correlations on the MIGRAINE task (Study1)

	Dur	Fre	Age	Gen	Occ	Mwk	Smo	Nau	Vis	Wei	Att	Res	Per
Dur	1	-0.06	-0.01	0.12	0.09	-0.03	0.04	0.08	-0.02	-0.09	0.02	0.06	0.05
Fre	-0.06	1	-0.03	-0.09	-0.09	-0.02	-0.04	-0.06	-0.06	-0.02	-0.01	0.01	-0.04
Age	-0.01	-0.03	1	-0.06	0.08	0	-0.07	-0.04	-0.01	0.01	0.01	-0.19	0.05
Gen	0.12	-0.09	-0.06	1	0.03	-0.06	-0.05	-0.05	-0.08	-0.08	-0.03	0.02	-0.01
Occ	0.09	-0.09	0.08	0.03	1	0.02	-0.13	0.05	0.09	0.02	0.14	0.02	0.08
Mwk	-0.03	-0.02	0.00	-0.06	0.02	1	0.07	0.03	-0.11	0.07	0.10	-0.13	0.19
Smo	0.04	-0.04	-0.07	-0.05	-0.13	0.07	1	0.07	0.09	-0.02	-0.12	0.04	0.11
Nau	0.08	-0.06	-0.04	-0.05	0.05	0.03	0.07	1	0.12	0.02	-0.19	-0.06	0.03
Vis	-0.02	-0.06	-0.01	-0.08	0.09	-0.11	0.09	0.12	1	0.10	-0.07	-0.06	0.02
Wei	-0.09	-0.02	0.01	-0.08	0.02	0.07	-0.02	0.02	0.1	1	-0.07	0.16	0.11
Att	0.02	-0.01	0.01	-0.03	0.14	0.10	-0.12	-0.19	-0.07	-0.07	1	-0.13	0.01
Res	0.06	0.01	-0.19	0.02	0.02	-0.13	0.04	-0.06	-0.06	0.16	-0.13	1	-0.04
Per	0.05	-0.04	0.05	-0.01	0.08	0.19	0.11	0.03	0.02	0.11	0.01	-0.04	1

Key:- DUR = duration, FRE = frequency, AGE = age, GEN = gender, OCC = occupation, MWK = misses work, SMO = smoking, NAU = nausea, VIS = visual disturbance, WEI = weight, ATT = attitude, RES = response to acute treatment, PER = personality.

Appendix 7
Inter-cue correlations on the HRT task (Study 1)

	Men	Hot	Age	Occ	Moo	Smo	Lib	Vag	Wei	Att	FHD	FBC	Per
Men	1	-0.03	0.02	0.04	0.01	-0.04	-0.05	-0.05	0.02	-0.05	0.09	0.14	0.07
Hot	-0.03	1	0.00	-0.04	0.04	0.08	-0.13	-0.14	-0.08	-0.04	-0.17	0.04	0.14
Age	0.02	0.00	1	0.05	-0.01	0.18	-0.16	0.01	0.10	0.12	0.19	-0.07	0.19
Occ	0.04	-0.04	0.05	1	-0.08	-0.08	-0.07	-0.14	0.18	-0.04	-0.05	0.05	-0.01
Moo	0.01	0.04	-0.01	-0.08	1	0.06	-0.11	-0.05	0.05	-0.07	0.00	0.06	-0.13
Smo	-0.04	0.08	0.18	-0.08	0.06	1	-0.08	-0.05	-0.04	-0.03	-0.03	0.05	0.08
Lib	-0.05	-0.13	-0.16	-0.07	-0.11	-0.08	1	0.06	0.12	-0.13	-0.11	-0.09	0.11
Vag	-0.05	-0.14	0.01	-0.14	-0.05	-0.05	0.06	1	0.11	-0.08	0.06	0.01	-0.05
Wei	0.02	-0.08	0.10	0.18	0.05	-0.04	0.12	0.11	1	-0.16	0.06	-0.01	0.03
Att	-0.05	-0.04	0.12	-0.04	-0.07	-0.03	-0.13	-0.08	-0.16	1	-0.05	0.01	-0.11
FHD	0.09	-0.17	0.19	-0.05	0.00	-0.03	-0.11	0.06	0.06	-0.05	1	0.00	-0.05
FBC	0.14	0.04	-0.07	0.05	0.06	0.05	-0.09	0.01	-0.01	0.01	0.00	1	0.03
Per	0.07	0.14	0.19	-0.01	-0.13	0.08	0.11	-0.05	0.03	-0.11	-0.05	0.03	1

Key:- MEN = menstruation, HOT = hot flushes, AGE = age, OCC = occupation, MOO = mood states, SMO = smoking behaviour, LIB = libido, VAG = vaginal dryness, WEI = weight, ATT = attitude, FHD = family history of ischaemic heart disease, FBC = family history of breast cancer, PER = personality.

Appendix 8

Social Class groupings and job titles for Occupation cue used in the LIPID, MIGRAINE and HRT tasks of Study 1, Chapter 5

I (Social class one)

"Barrister","Solicitor","Chartered Accountant","Industrial Statistician","University Lecturer","Management Consultant","Civil Engineer","Architect","Company Director","Head Teacher"

II (Social class two)

"Claims Assessor","Systems Analyst","Building Inspector","Secondary School Teacher","Journalist","Laboratory Technician","Air Traffic Controller","Surveyor","Social Worker","FE Lecturer"

III NM (Social class three, non-manual)

"Manager's PA","Driving Instructor","Camera Operator","Receptionist","Shop Assistant","Police Constable","Clerk","Cashier","Caravan Site Manager","Office Machine Operator"

III M (Social class three, manual)

"Chef","Hairdresser","Shoe Repairer","Butcher","Fishmonger","Printer","Carpet Fitter","Painter and Decorator","Electrician","Travel Steward"

IV (Social class four)

"Bus Conductor","Packer","Production Worker","Spray Painter","Laundry Manager","Post Office Worker","Gardener","Bar Staff","Horticultural Worker","Market Trader"

V (Social class five)

"Refuse Worker","Office Cleaner","Kitchen Hand","Sewage Plant Attendant","Lavatory Cleaner","Window Cleaner","Road Sweeper","Janitor"

Selected from OPCS Classification of Occupation, 1980, Appendix B.1. (page lxxxiii-civ).

Appendix 9

Sample sheet to remind doctors of the cues available on the LIPID task whilst rating the cues' relative importance (Study 1, Chapter 5)

EXERCISE ONE

CHOLESTEROL LEVEL
HYPERTENSION
AGE
GENDER
OCCUPATION
EVIDENCE OF ARTERIOSCLEROSIS
SMOKES
DIABETES
COMPLIANCE WITH ADVICE ON DIET
WEIGHT
ATTITUDE TO TREATMENT
FAMILY HISTORY OF I.H.D.
PERSONALITY

IMPORTANCE RATING SCALE

0	1	2	3	4	5	6	7	8	9	10
No Attention paid										Of Maximum importance

Appendix 10

Standard sheet used by the experimenter to note down subjective ratings in Study 1, Chapter 5

INSIGHT: EXERCISE ONE

I would like you to be as honest as possible in trying to estimate what affects your decisions. All the following questions are trying to ascertain what you feel did, not what you feel should have, affected your behaviour.

1) DIRECTIONALITY:

Firstly, for each item on the list I would like to ascertain what values or category would be more likely to make you prescribe. For example, other things being equal, do you feel you would be more likely to prescribe for a sixty year old or for a thirty year old or would age make no difference?

Go through each item asking if A or B would make you more likely to prescribe.

Things to note: A or B or neither (in this case should score 0 on second half of insight assessment).

Attribute	**A**	**B**
Cholesterol level	8	6.5
Hypertension	No	Poorly controlled
Age	30	60
Gender	Male	Female
Occupation	Architect	Window cleaner
Evidence of arteriosclerosis	Yes	No
Smokes	Heavily	No
Diabetes	No	Poorly controlled
Compliance with advice on diet	Yes	No
Weight	Very obese	Under
Attitude to treatment	Requesting	Opposed
Family history of IHD	First degree rel.	No
Personality	Co operative	Demanding

2) WEIGHTING:

Now I would like to get an idea of how much you feel that each piece of information has a bearing on your decision. I would like you to assign each information type a number between 0 and 10 to indicate how much you feel you take note of this attribute . If you feel that you pay no attention at all to a piece of information then you should assign it 0. If you feel that it is the most important factor then you should assign it 10 points. Other intermediate values should be assigned as appropriate to attributes which you feel have some affect on your decision.

Note value assigned (0-10) for each cue.

3) REPRESENTATIVENESS:

Have you got anything to add about how you might do this task in the real world?

Record answer, note any other varibles that might be taken into account, any variables showing interactions, any indication of non-linearity of variable.

Would you normally refer patients whom you thought might need lipid lowering drugs?

Appendix 11

Mathematical explanation for the conversion of subjective ratings to subjective weights (ψ_i)[1]

Let ψ_i = subjective weights ($1 \leqslant i \leqslant 13$)

Let S_i = original subjective ratings

Let Y_S = the set of 130 decisions predicted by treating S_i as regression coefficients in a linear equation and then inserting X values from the standardised X matrix into this equation.

Let Y_ψ = the set of 130 decisions predicted by treating ψ_i as regression coefficients in a linear equation and then inserting X values from the standardised X matrix into this equation.

μ_s = the mean of $Y_s = 0$ since all predictor cues are standardised

σ_s = the standard deviation of Y_s

σ_ψ = the standard deviation of Y_ψ

ψ_i are calculated for direct comparison with standardised regression coefficients (β_i). These are the multiple linear regression coefficients which when multiplied by the standardised X matrix produce a set of values with mean = 0 and standard deviation = 1. This same property is desired in subjective weights:

$$\psi_i = cS_i : \sigma_\psi = 1$$

$$Y_S = \Sigma S_i X_i$$

$$Y_\psi = \Sigma \psi_i X_i = \Sigma c S_i X_i = c \Sigma S_i X_i = c Y_S$$

$$\sigma_\psi = 1 \therefore \text{Var}(Y_\psi) = 1 \therefore \text{Var}(cY_S) = 1$$

$$c^2 \text{Var}(Y_S) = 1$$

$$c = \frac{1}{\sqrt{Var(Y_S)}} = \frac{1}{\sigma_s}$$

$$\text{Thus } \psi_i = \frac{S_i}{\sigma_s}$$

[1] All credit to Ian Dennis

Appendix 12

Indices on the LIPID task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency.

GP	Consistency	R^2	Mean Judgement	St. dev Judgement	Mean[2] latency	St. dev. latencies	Correlation[3] Judgement & latency
1	0.60	0.43	35.9	27.3	2394.3	968.8	0.080
2	0.29	0.37	11.2	17.9	1224.8	792	0.199
3	0.30	0.50	58.7	15.6	1931.5	947.9	-0.001
4	0.77	0.61	16.6	13.6	1476.8	917.3	-0.090
5	0.44	0.48	24.12	32.0	3024.1	1079.9	0.109
6	0.17	0.36	44.8	17.9	514.8	438.6	0.131
7	0.32	0.36	6.6	15.2	1198.6	660.9	**0.593**
8	-	0.45	47.8	23.6	1692.6	489.7	0.027
9	0.19	0.48	10.8	19.8	1424.6	912.9	**0.418**
10	-	0.38	44.4	36.2	1939.5	1133.9	**0.201**
11	0.15	0.39	17.3	23.7	1646.4	1013.1	0.156
12	0.34	0.11	7.8	10.5	2829.5	966.4	0.088
13	0.46	0.22	92.5	11.7	424.2	512.2	-0.125
14	-0.12	0.25	49.2	5.5	1408.8	649.9	0.028
15	-0.12	0.29	8.3	20.1	2350.4	1302.2	-0.020
17	0.26	0.28	47.3	38.8	1536.1	886	-0.052
19	0.43	0.50	16.3	13.9	1850.8	1253.7	**0.378**
20	0.23	0.40	18.8	24.0	1856.8	1359.1	**0.397**
21	0.36	0.80	54.1	37.6	2415.7	1462.1	-0.036
22	0.10	0.30	32.4	15.6	2020.9	1148.5	**-0.213**
23	-	-	-	-	-	-	-
24	0.32	0.22	18.0	20.1	1834.5	972	0.105
25	-	0.32	38.8	32.6	1900.0	1508.3	-0.123
26	0.72	0.81	55.7	22.2	1821.2	820.6	0.021
27	-	-	-	-	-	-	-
28	0.22	0.61	16.0	8.5	1395.4	733.9	**-0.223**
29	0.44	0.20	19.4	14.5	2528.4	1075.8	-0.066
30	0.49	0.39	13.3	20.5	1134.4	633.7	**0.289**
31	0.68	0.59	75.8	37.9	1124.4	1453.1	-0.120
32	0.64	0.38	20.1	32.3	1151.0	829.8	**0.304**
33	0.37	0.55	15.8	33.0	896.6	821.1	**0.256**
34	0.28	0.29	10.1	21.2	947.7	926.9	0.014
35	0.51	0.38	40.7	30.4	1823.5	1044.7	0.056
36	0.63	0.58	65.3	22.6	538.6	485.5	0.139
37	-0.11	0.53	74.6	31.6	2301.8	912.7	-0.101
Mean	**0.35**	0.42	33.6	22.7	1653	942.8	0.085
St.dev.	**0.23**	0.16	23.0	9.1	640	285.7	0.192

[2] Measured in centiseconds.

[3] Numbers in bold are significant ($p < 0.05$).

Appendix 13

Indices on the MIGRAINE task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency.

GP	Consistency	R^2	Mean Judgement	St. dev. Judgement	Mean[4] Latency	St. dev. latencies	Correlation[5] Judgement & latency
1	0.67	0.55	42.7	45.8	1303.2	739.2	0.174
2	0.79	0.63	39.8	46.0	823.1	807.1	**0.355**
3	0.91	0.80	40.3	31.0	1190	443.9	0.087
4	0.82	0.67	58.7	25.7	1525.7	955.7	**0.222**
5	0.88	0.60	20.1	31.4	1543.7	890.7	**0.408**
6	0.69	0.56	50.3	22.5	527.2	270.6	0.137
7	0.88	0.76	38.0	37.8	726.5	723.1	**0.329**
8	-	0.69	47.4	24.1	1263.2	489.2	0.065
9	0.94	0.76	42.2	36.3	1293	2802	0.056
10	-	0.72	49.2	46.6	576.9	552.2	-0.009
11	0.56	0.72	45.6	41.3	1212	736.5	0.122
12	0.66	0.73	43.9	36.9	736.7	419.7	0.057
13	0.87	0.75	67.0	39.4	258.4	183.5	**-0.222**
14	0.55	0.47	50.0	13.0	583	347.9	**0.206**
15	0.83	0.65	28.1	33.4	1396.7	1510	**0.212**
17	0.39	0.54	55.1	42.8	1209.7	1423.2	**-0.122**
19	0.94	0.86	53.0	38.2	1098.6	906.9	**0.190**
20	0.76	0.62	52.9	31.0	1216.5	1800.8	0.104
21	0.89	0.67	30.7	36.1	1535.6	1331.3	**0.565**
22	0.81	0.62	29.8	25.1	1828	1241.9	0.050
23	0.62	0.50	6.5	11.1	220.9	156.2	**0.427**
24	0.50	0.50	19.5	29.7	1063.1	480.1	**0.307**
25	-	0.62	25.0	28.1	1047.5	453.2	**0.319**
26	0.68	0.71	48.0	25.9	1195.6	591	-0.063
27	0.84	0.77	49.4	38.6	677	492.9	-0.146
28	0.73	0.72	24.7	26.3	749.4	685.5	**0.303**
29	0.71	0.77	47.7	25.6	2472.1	1186.8	0.017
30	0.91	0.79	51.4	46.4	839.9	414.2	0.162
31	0.06	0.12	63.6	17.4	1069.5	916.6	0.102
32	0.52	0.73	35.2	42.3	664.7	634.2	0.129
33	0.75	0.71	34.6	40.5	540.2	425.5	0.002
34	-	-	-	-	-	-	-
35	0.67	0.67	42.9	38.3	1149.9	720.4	0.003
36	0.63	0.43	57.7	34.6	580.1	376.9	**0.133**
37	0.61	0.43	20.2	34.9	1611.6	909.4	**0.427**
Mean	**0.71**	**0.64**	**41.5**	**33.1**	1050.9	794.6	**0.150**
St.dev.	**0.19**	**0.14**	**13.9**	9.2	469.0	527.9	0.177

[4] Measured in centiseconds.

[5] Significant numbers in bold ($p < 0.05$).

Appendix 14

Indices on the HRT task (Study 1) - Consistency, Linear fit, mean judgement and mean latency, correlation between judgement and latency.

GP	Consistency	R^2	Mean Judgement	St.dev.	Mean latency[7]	St.dev.	Correlation[6] Judgement & latency
1	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-
9	0.39	0.61	64.3	25.2	1510	1835	**-0.244**
10	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-
13	0.19	0.28	71.4	39.7	476.0	223.5	**-0.200**
14	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-0.104
19	-	0.31	86.4	19.5	1176.5	931.9	-
20	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-
24	-	0.45	60.9	37.0	1655.2	823.5	-0.163
25	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-
27	0.38	0.33	53.0	15.0	892.2	369.6	0.033
28	-	0.67	43.9	20.7	1229.7	913.3	-0.081
29	-	-	-	-	-	-	-
30	-	0.59	54.8	24.4	1446.2	606.1	-0.032
31	0.25	0.52	51.5	27.1	581.8	480.5	-0.123
32	-	-	-	-	-	-	-
33	0.24	0.54	60.8	26.2	681.7	1095.9	-0.024
34	0.32	0.39	18.4	35.3	658.4	838.5	**0.518**
35	-	0.58	44.4	35.6	1513.6	694.7	-0.176
36	-	0.32	60.8	32.7	647.3	427.9	-0.038
37	-	-	-	-	-	-	-
Mean	**0.3**	**0.47**	**55.9**	**28.2**	**747**	**770**	**-0.05**
St.dev.	**0.08**	**0.14**	**16.6**	**7.8**	**651**	**425**	**0.197**

[6] Significant numbers in bold ($p < 0.05$).

[7] Measured in centiseconds.

Appendix 15
Explicit policies: Subjective ratings on the LIPID task (Study 1)

GP	CHO	HYP	AGE	GEN	OCC	EVA	SMO	DIA	COM	WEI	ATT	FH	PER	No.>5
1	10	7	-5	8	0	0	7	7	0	-5	0	6	0	6
2	9	2	-2	2	0	0	-6	8	6	4	9	1	-2	5
3	9	6	-6	3	1	8	-7	7	6	-6	8	8	3	10
4	9	4	-3	3	0	6	6	9	0	0	6	1	0	5
5	9	9	-9	2	0	0	-9	9	9	-4	6	7	0	8
6	9	7	-6	6	-4	8	-8	9	5	-5	4	6	-5	8
7	10	7	-8	8	-2	7	-10	7	9	-9	10	8	-3.5	11
8	9	6	-5	2	0	2	8	4	6	-5	8	5	8	6
9	10	7	0	2	0	7	-2	1	2	0	4	0	0	3
10	10	8	0	8	0	10	3	8	0	-3	3	8	0	6
11	9.5	5	-7.5	8.5	0	0	0	-5.5	4.5	0	7.5	8.5	0	6
12	7	7	-5	0	0	6	-3	6	5	-7	7	8	0	7
13	10	8	-10	0	0	8	0	0	0	0	0	10	0	5
14	7	6	0	6	-5	6	-6	0	5	-5	6	6	0	7
15	8	8	-8	9	-2	7	-9	9	9	-7	1	9	-1	10
17	6	4	-7	9	0	8	-4	10	0	-1	0	6	0	6
19	9	7	-10	0	0	3	4	8	0	-1	0	0	0	4
20	7	6	-6	0	0	7	-9	7	0	-8	10	8	0	9
21	10	1	0	1	0	10	2	2	0	0	5	0	0	2
22	8	-8	-8	6	1	7	-9	-8	6	-7	7	8	2	11
24	6	6	0	8	-5	5	-8	-10	5	-7	9	7	0	8
25	6	-6	-6	0	0	5	-6	8	7	-5	7	7	0	8
26	7	2	-8	0	0	0	2	2	0	2	10	8	-3	4
27*	4	-4	-7	8	0	6	-9	7	6	5	7	8	2	8
28	7	6	-7	8	0	3	6	8	0	0	0	5	0	6
29	7	6	-7	0	0	0	-6	5	6	0	7	7	0	7
30	10	7	0	2	0	3	-4	6	1	-8	0	3	0	4
31	10	9	0	0	0	3	8	8	0	3	0	8	0	5
32	9	7	-9	1	0	7	7	6	5	-9	3	6	1	8
33	10	-5	-10	10	-2	5	-10	10	10	-10	5	10	0	8
34	10	0	-10	10	-1	6	0	0	10	0	5	5	2	5
35	10	-9	-9	1	0	6	6	9	5	-7	9	6	-1	9
36	8	0	0	0	0	8	0	7	5	-1	7	8	-4	5
37	10	10	-10	8	2	9	-6	10	-5	10	2	10	2	9
Mean	8.52	4.00	-5.54	4.10	-0.50	5.18	-2.12	5.10	3.75	-2.82	5.07	6.22	0.02	6.74
St.dev.	1.59	5.05	3.64	3.73	1.52	3.09	6.08	5.08	3.72	4.63	3.41	2.89	2.19	2.23
No. > 5	33	24	21	15	0	19	22	26	12	12	18	25	1	

* = only subjective data available for this subject.

KEY: CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality and No. >5 = Number of ratings greater than 5.

Appendix 16
Explicit policies: Subjective ratings on the MIGRAINE task (Study 1)

GP	DUR	FRE	AGE	GEN	OCC	MWK	SMO	NAU	VIS	WEI	ATT	RES	PER	No. > 5
1	0	10	0	0	0	-10	0	6	2	0	9	9	7	6
2	2	10	0	0	0	-8	0	3	3	0	8	2	0	3
3	3	9	0	0	0	-9	0	0	0	0	6	7	3	4
4	7	9	0	0	0	-10	0	3	4	0	8	6	2	5
5	8	10	0	0	3	-7	0	4	4	0	10	7	0	5
6	8	8	-5	0	-2	-6	0	8	8	0	5	6	0	6
7	7	10	0	0	0	-10	0	0	0	0	0	0	0	3
8	9	7	0	0	0	-7	0	7	7	0	8	6	8	8
9	10	10	0	0	0	-7	0	0	0	0	8	6	0	5
10	10	10	0	0	0	-5	0	3	0	0	3	8	0	3
11	7.5	9.5	0	0	0	-5.5	0	0	0	0	8.5	8.5	0	5
12	9	9	0	0	0	-7	0	3	3	0	8	7	0	5
13	10	10	0	0	0	0	0	0	0	0	0	0	0	2
14	7	7	0	0	0	-6	0	0	0	0	7	6	0	5
15	8	10	0	0	-8	-9	0	6	6	0	3	4	8	7
17	6	9	0	-3	-2	-10	0	7	7	0	4	6	0	6
19	10	10	0	0	0	-3	0	0	0	0	1	0	0	2
20	6	6	0	0	0	0	0	0	0	0	10	9	0	4
21	4	10	0	6	-4	-7	0	0	0	-8	10	3	0	5
22	8	9	0	0	0	-8	0	6	6	0	7	7	3	7
23	5	10	0	0	0	-6	0	6	3	0	8	6	-7	6
24	8	10	0	0	0	-7	0	0	3	0	8	8	0	5
25	9	9	0	0	0	-9	-9	9	9	-9	1	9	0	8
26	6	8	0	0	0	-3	0	8	6	0	8	7	0	6
27	3	10	0	0	0	0	0	0	0	0	0	7	0	2
28	3	10	0	0	0	-3	0	3	3	0	7	0	0	2
29	5	10	0	0	3	-7	0	6	6	-4	7	5	0	5
30	3	10	0	0	0	-4	0	1	2	0	5	6	0	2
31	8	8	0	0	0	-8	0	8	8	0	0	8	0	6
32	7	9	0	0	-4	-4	0	1	1	0	2	8	2	3
33	10	10	0	0	10	-10	0	8	8	0	8	10	0	8
35	7	10	0	0	2	-7	0	6	8	0	10	-4	0	6
36	8	6	0	0	0	0	0	0	0	0	9	8	0	4
37	8	10	-4	0	-4	-7	0	5	5	0	3	4	5	3
Mean	6.8	9.2	-0.3	0.1	-0.2	-6.2	-0.3	3.4	3.3	-0.6	5.9	5.6	0.9	5
St.dev.	2.6	1.2	1.1	1.1	2.7	3	1.5	3.2	3.1	2.1	3.4	3.2	2.7	2
No. > 5	26	34	0	1	2	24	1	13	11	2	21	24	4	

KEY: DUR = duration, FRE = frequency, AGE = age, GEN = gender, OCC = occupation, MWK = misses work, SMO = smoking, NAU = nausea, VIS = visual disturbance, WEI = weight, ATT = attitude, RES = response to acute treatment, PER = personality, No. > 5 = number of ratings greater than 5.

Appendix 17
Explicit policies: Subjective ratings on the HRT task (Study 1)

GP	MEN	HOT	AGE	OCC	MOO	SMO	LIB	VAG	WEI	ATT	FHD	FBC	PER	No.>5
9	-5	8	-5	0	6	2	6	8	-4	8	5	-6	0	6
13	-10	10	-10	0	4	0	2	7	0	0	5	0	0	4
17	-7	9	-3	-4	5	0	10	5	-1	6	0	-2	4	4
24	-8	8	-5	0	5	5	0	7	-5	10	3	0	0	4
26*	-2	6	0	0	6	0	6	5	0	10	6	-9	0	6
27	-9	7	-7	0	3	5	7	8	-6	8	0	0	0	7
28	-7	9	8	0	6	0	5	5	0	9	0	0	0	5
30	-6	8	3	0	6	0	5	7	-1	8	3	-3	1	5
31	7	9	1	0	0	0	3	7	0	0	0	0	0	3
33	0	10	0	0	0	0	5	10	0	7	8	0	0	4
34	-10	10	0	0	10	-6	10	10	0	10	0	-5	0	7
35	-10	10	2	0	4	1.5	5	5	0	10	2	0	5.5	4
36	0	7	0	0	0	0	6	8	0	9	0	0	0	4
Mean	-5.2	8.5	-1.2	-0.3	4.2	0.6	5.4	7.1	-1.3	7.3	2.5	-1.9	0.8	4.8
St. dev.	5.1	1.3	4.7	1.1	2.9	2.7	2.8	1.8	2.2	3.5	2.8	3	1.8	1.3
No. > 5	9	13	3	0	5	1	6	9	1	11	2	2	1	

KEY: MEN = menstruation, HOT = hot flushes, AGE = age, OCC = occupation, MOO = mood states, SMO = smoking behaviour, LIB = libido, VAG = vaginal dryness, WEI = weight, ATT = attitude, FHD = family history of ischaemic heart disease, FBC = family history of breast cancer, PER = personality, No. > 5 = number of ratings greater than 5.
* = Subjective ratings only collected for this subject.

Appendix 18

Tacit policies: Standardised Regression Coefficients (β_i) on the LIPID task (Study 1)

GP	CHO	HYP	AGE	GEN	OCC	EVA	SMO	DIA	COM	WEI	ATT	FH	PER	No.S
1	**0.42**	**0.15**	0.05	**0.15**	0.02	-0.10	0.05	**0.37**	**0.17**	-0.04	0.07	0.14	0.06	5
2	**0.36**	-0.04	0.04	0.03	0.12	-0.01	**-0.25**	**0.31**	-0.15	0.00	**0.43**	-0.14	0.00	4
3	**0.37**	-0.07	-0.09	-0.03	0.01	**0.49**	**-0.18**	-0.01	**0.17**	0.08	**0.33**	0.12	0.00	5
4	**0.58**	0.04	-0.03	0.00	0.02	0.07	-0.08	**0.39**	0.06	-0.06	**0.43**	-0.09	-0.01	3
5	**0.58**	0.04	0.12	-0.01	0.02	0.03	**-0.15**	**0.18**	0.09	0.03	**0.35**	0.04	-0.03	4
6	**0.55**	0.06	-0.06	-0.10	0.00	-0.04	**-0.24**	**0.21**	-0.05	-0.06	0.11	-0.08	-0.10	3
7	**0.45**	0.14	-0.02	0.10	-0.05	0.13	**-0.37**	0.12	0.03	**-0.28**	0.06	0.03	0.01	3
8	**0.24**	**0.35**	**-0.20**	0.00	0.01	**0.16**	**0.21**	0.12	0.00	-0.04	**0.39**	**0.21**	-0.08	7
9	**0.57**	0.07	**0.17**	0.03	0.04	0.01	-0.02	**0.21**	0.01	0.05	**0.26**	0.02	-0.12	4
10	**0.44**	0.08	**-0.17**	0.06	0.09	**0.37**	**-0.18**	**0.18**	-0.04	**-0.16**	**0.22**	0.04	0.00	7
11	**0.38**	0.05	**-0.15**	**0.34**	0.05	-0.05	-0.13	**-0.34**	0.10	-0.10	-0.04	0.06	-0.10	4
12	**0.20**	0.16	-0.10	0.05	0.00	**0.22**	-0.12	**0.22**	-0.11	-0.01	0.00	0.10	0.08	3
13	**0.31**	0.03	**-0.24**	0.06	0.02	0.11	0.14	0.14	0.12	0.14	0.04	0.11	-0.02	2
14	0.02	**0.19**	0.04	-0.01	-0.01	0.11	0.08	**0.17**	-0.05	0.13	**0.44**	**0.20**	0.06	4
15	**0.21**	-0.05	-0.15	**0.17**	0.15	0.12	-0.08	0.04	-0.01	**-0.35**	-0.07	**0.37**	0.01	4
17	0.14	0.00	-0.09	0.05	0.04	**0.37**	**-0.25**	**0.30**	0.14	-0.15	-0.03	0.00	-0.01	3
19	**0.34**	0.11	**-0.57**	0.09	-0.03	0.00	0.02	**0.20**	0.07	0.03	-0.01	-0.02	0.02	3
20	**0.26**	0.06	-0.11	-0.03	0.11	-0.01	**-0.53**	0.12	**-0.21**	-0.07	**0.26**	**0.15**	-0.05	5
21	**0.28**	0.07	0.06	-0.03	0.02	**0.87**	0.05	**0.08**	-0.04	0.05	0.03	0.05	-0.05	3
22	**0.25**	-0.10	**-0.18**	-0.01	0.04	0.01	**-0.38**	-0.01	**0.23**	-0.06	**0.16**	**0.18**	0.05	6
24	**0.23**	0.11	-0.03	0.08	-0.04	-0.03	-0.09	**-0.29**	**0.29**	-0.09	**0.22**	0.04	-0.07	4
25	**0.24**	0.04	0.04	0.09	**0.17**	-0.01	-0.01	**0.42**	0.07	**-0.30**	0.07	-0.02	0.03	4
26	**0.25**	0.04	-0.05	0.04	**0.09**	0.02	0.05	**0.08**	-0.01	-0.03	**0.86**	**0.09**	-0.03	5
28	**0.35**	**0.17**	**-0.12**	0.09	**0.18**	0.04	0.06	**0.58**	0.09	-0.02	-0.04	0.00	-0.02	5
29	**0.22**	**0.22**	0.05	-0.05	-0.02	-0.03	-0.10	**0.30**	0.05	0.05	**0.29**	-0.02	0.04	4
30	**0.60**	0.01	0.08	-0.13	-0.03	0.05	-0.06	0.12	-0.09	**-0.19**	-0.01	0.09	-0.11	2
31	**0.74**	0.03	-0.02	0.09	0.01	0.06	**0.16**	**0.12**	-0.02	0.03	0.08	**0.13**	**0.13**	5
32	**0.34**	**0.20**	-0.10	-0.12	0.03	0.14	**-0.20**	**0.24**	**-0.16**	**-0.39**	-0.10	0.07	-0.11	6
33	**0.42**	0.10	-0.05	**0.34**	0.12	0.00	**-0.50**	**0.13**	**0.13**	**-0.14**	0.10	-0.05	**0.18**	7
34	**0.36**	0.02	-0.14	**0.37**	-0.02	-0.09	-0.10	0.09	**0.19**	-0.02	**0.22**	-0.06	0.13	4
35	**0.38**	**-0.29**	**-0.30**	0.07	-0.05	0.08	0.01	-0.15	0.04	**-0.28**	**0.18**	0.10	-0.02	5
36	**0.74**	**0.16**	**-0.12**	0.00	-0.02	-0.03	0.06	0.01	-0.05	0.02	0.07	-0.01	0.02	3
37	**0.70**	-0.02	0.01	-0.04	0.00	0.03	**0.17**	**0.19**	-0.05	0.00	**0.19**	0.04	0.11	4
Mean	**0.38**	**0.06**	**-0.07**	**0.05**	**0.03**	**0.09**	**-0.09**	**0.15**	0.03	**-0.07**	**0.17**	**0.06**	0.00	4.24
St.dev.	0.17	0.11	0.14	0.12	0.06	0.19	0.18	0.18	0.11	0.13	0.20	0.10	0.07	1.32
No. S	31	8	10	5	4	6	14	22	8	8	16	7	2	

KEY: CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality and No. S = Number of significant regression coefficients ($p < 0.05$).

Appendix 19

Tacit policies: Standardised Regression Coefficients (β_i) on the MIGRAINE task (Study 1)

GP	DUR	FRE	AGE	GEN	OCC	MWK	SMO	NAU	VIS	WEI	ATT	RES	PER	NO.S
1	0.06	**0.66**	-0.04	0.04	0.07	**-0.16**	-0.01	0.02	0.04	-0.01	**0.33**	**0.16**	-0.01	4
2	-0.09	**0.74**	0.08	**0.15**	-0.01	-0.05	0.07	0.07	0.09	-0.02	**0.36**	0.08	-0.04	3
3	0.03	**0.85**	-0.01	0.03	-0.03	**-0.11**	0.02	-0.01	0.06	0.00	**0.32**	**0.10**	0.02	4
4	**0.11**	**0.55**	-0.03	-0.01	0.01	**-0.36**	0.01	0.01	**0.11**	-0.04	**0.52**	**0.11**	-0.04	6
5	0.10	**0.58**	**-0.18**	-0.08	-0.07	**-0.17**	0.00	-0.05	0.04	0.04	**0.47**	-0.01	0.01	4
6	-0.01	**0.74**	-0.04	-0.01	-0.08	-0.05	0.08	**0.12**	**0.19**	0.01	0.03	0.04	0.00	3
7	0.06	**0.88**	-0.04	0.03	0.00	-0.06	0.07	0.02	0.02	-0.01	0.00	0.04	-0.06	1
8	**0.34**	**0.22**	-0.03	0.03	0.02	**-0.15**	0.02	**0.34**	**0.38**	0.04	**0.57**	**0.25**	-0.07	7
9	**0.18**	**0.81**	-0.05	0.01	-0.04	-0.08	0.08	-0.01	0.04	0.00	**0.27**	**0.12**	**-0.12**	5
10	**0.21**	**0.84**	-0.08	**0.10**	-0.04	0.08	-0.03	0.07	0.08	-0.02	0.09	0.06	0.00	3
11	0.06	**0.82**	-0.04	0.09	-0.06	0.01	0.04	0.00	0.05	0.01	**0.26**	**0.11**	-0.06	3
12	**0.14**	**0.82**	-0.09	-0.02	-0.01	-0.06	0.07	0.04	-0.05	0.05	0.08	**0.13**	**-0.13**	4
13	0.02	**0.88**	-0.03	0.03	0.02	-0.05	-0.01	0.03	-0.02	0.02	0.04	0.05	-0.01	1
14	0.05	**0.26**	0.03	-0.04	-0.04	0.04	0.05	0.12	0.02	-0.01	**0.68**	**0.20**	-0.08	3
15	**0.18**	**0.80**	-0.03	0.06	-0.08	-0.02	0.00	-0.02	0.09	-0.01	0.05	0.07	-0.05	2
17	**0.14**	**0.55**	-0.06	**0.18**	-0.04	**-0.22**	0.09	**0.14**	**0.21**	-0.02	**0.38**	**0.16**	-0.06	8
19	**0.08**	**0.94**	0.01	0.03	0.04	-0.06	0.05	0.04	-0.02	-0.01	0.03	0.03	-0.05	2
20	**0.12**	**0.13**	**0.13**	**0.13**	-0.04	0.07	0.00	0.05	0.05	0.06	**0.69**	**0.50**	-0.03	6
21	0.05	**0.74**	-0.09	0.06	-0.04	-0.10	0.06	-0.03	-0.01	-0.09	**0.34**	0.08	-0.02	2
22	**0.15**	**0.62**	-0.07	0.03	-0.10	-0.11	0.10	-0.04	0.09	-0.10	**0.43**	**0.21**	0.00	4
23	0.12	**0.70**	-0.06	-0.08	0.00	0.07	-0.05	-0.07	0.04	-0.06	0.03	0.10	0.02	1
24	**0.18**	**0.52**	-0.08	-0.09	-0.01	**-0.21**	0.09	0.00	-0.07	0.01	**0.20**	**0.33**	-0.07	5
25	**0.16**	**0.76**	-0.01	0.00	-0.04	-0.06	-0.04	0.05	0.02	0.01	0.10	**0.24**	0.07	3
26	**0.26**	**0.71**	-0.03	0.05	0.06	**-0.11**	0.06	**0.18**	**0.14**	-0.02	**0.33**	**0.25**	-0.06	7
27	0.07	**0.88**	-0.02	-0.02	-0.02	-0.04	0.02	0.04	-0.06	-0.02	0.04	**0.10**	0.01	2
28	0.07	**0.84**	-0.02	-0.05	-0.01	0.04	0.02	**-0.10**	-0.03	-0.01	-0.01	0.05	0.02	2
29	0.08	**0.82**	0.05	0.02	**-0.12**	0.04	0.07	-0.02	**0.22**	**-0.11**	**0.28**	**0.10**	-0.05	6
30	0.01	**0.89**	-0.03	0.02	0.01	-0.03	0.04	0.01	-0.01	-0.03	0.05	**0.11**	-0.02	2
31	-0.07	0.05	0.02	0.05	0.17	**-0.27**	**0.21**	0.11	0.14	0.05	0.09	0.13	-0.05	2
32	**0.11**	**0.84**	0.01	0.00	**-0.13**	-0.02	-0.05	-0.05	0.07	-0.05	-0.03	0.03	0.01	3
33	**0.16**	**0.81**	-0.09	0.00	**-0.13**	0.07	0.06	-0.02	-0.02	0.01	0.03	**0.11**	0.04	4
35	0.08	**0.74**	-0.05	0.04	-0.02	-0.09	0.04	0.00	0.07	-0.02	**0.39**	-0.02	-0.04	2
36	0.12	**0.32**	0.06	0.03	-0.04	0.11	-0.02	0.03	0.05	0.06	**0.51**	**0.41**	-0.13	3
37	0.06	**0.59**	-0.10	-0.05	-0.04	-0.04	0.10	**-0.14**	0.08	0.01	0.10	**0.24**	-0.04	3
Mean	**0.10**	**0.67**	**-0.03**	**0.02**	**-0.02**	**-0.06**	**0.04**	0.03	**0.06**	-0.01	**0.24**	**0.14**	**-0.03**	3.53
S.d.	0.08	0.23	0.06	0.06	0.06	0.10	0.05	0.09	0.09	0.04	0.21	0.11	0.04	1.81
No. S	15	33	2	4	3	9	1	5	6	1	18	20	2	

KEY: DUR = duration, FRE = frequency, AGE = age, GEN = gender, OCC = occupation, MWK = misses work, SMO = smoking, NAU = nausea, VIS = visual disturbance, WEI = weight, ATT = attitude, RES = response to acute treatment, PER = personality, No. S = number of significant regression coefficients ($p < 0.05$).

Appendix 20

Tacit policies: Standardised Regression Coefficients (β_i) on the HRT task (Study 1)

GP	MEN	HOT	AGE	OCC	MOO	SMO	LIB	VAG	WEI	ATT	FHD	FBC	PER	NO.S
9	-0.04	**0.47**	-0.07	0.01	**0.41**	-0.03	**0.24**	**0.46**	**-0.24**	**0.30**	-0.03	-0.10	0.00	6
13	**-0.20**	**0.33**	**0.40**	0.11	0.04	0.04	0.02	0.02	-0.05	0.01	0.04	-0.07	0.07	3
17	-0.09	**0.22**	-0.12	**-0.19**	0.13	-0.01	**0.32**	0.01	-0.06	**0.43**	0.02	-0.08	0.03	4
24	**-0.21**	**0.36**	**0.19**	-0.12	**0.27**	0.08	0.10	0.03	-0.09	**0.39**	0.04	-0.08	0.08	5
27	-0.04	0.07	-0.08	-0.01	**0.19**	-0.01	**0.19**	0.00	-0.15	**0.55**	-0.03	0.02	-0.07	3
28	-0.03	**0.32**	**0.35**	-0.01	0.08	0.05	0.03	-0.02	-0.01	**0.64**	-0.05	-0.03	0.01	3
30	-0.05	**0.34**	**0.46**	0.05	**0.15**	-0.01	**0.36**	**0.26**	**-0.15**	**0.46**	0.06	-0.05	-0.06	7
31	0.00	**0.63**	0.06	**-0.14**	-0.07	-0.04	0.02	**0.42**	0.07	-0.01	-0.01	0.09	0.02	3
33	0.05	**0.71**	0.10	-0.07	0.08	0.06	-0.05	-0.05	0.00	0.11	-0.02	-0.09	-0.03	1
34	**-0.60**	0.01	0.03	0.08	-0.05	0.01	-0.02	0.05	0.10	**0.16**	-0.01	-0.14	0.05	2
35	-0.02	**0.43**	**0.13**	-0.06	**0.18**	0.04	**0.15**	**0.12**	-0.11	**0.60**	-0.08	-0.02	0.02	6
36	0.00	**0.27**	-0.09	0.10	-0.03	0.03	0.13	**0.27**	-0.02	**0.56**	0.09	-0.03	-0.01	3
Mean	-0.10	**0.35**	0.11	-0.02	**0.12**	0.02	**0.12**	**0.13**	-0.06	**0.35**	0.00	**-0.05**	0.01	3.83
St.dev.	0.17	0.19	0.19	0.09	0.13	0.04	0.13	0.17	0.09	0.22	0.05	0.06	0.04	1.80
No. S	3	10	5	2	5	0	5	5	2	9	0	0	0	

KEY: MEN = menstruation, HOT = hot flushes, AGE = age, OCC = occupation, MOO = mood states, SMO = smoking behaviour, LIB = libido, VAG = vaginal dryness, WEI = weight, ATT = attitude, FHD = family history of ischaemic heart disease, FBC = family history of breast cancer, PER = personality, No. S = number of significant regression coefficients ($p < 0.05$).

Appendix 21
Correlation of cue weights LIPID task, Study 1

Correlation of cue standardised regression coefficients (β_i)

N = 33 therefore significance ($p < 0.05$) when $r > 0.33$.

	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	CMP	WEI	ATT	FHD
HYP	-0.150											
AGE	0.204	0.115										
GEN	-0.100	-0.078	-0.229									
OCC	-0.215	-0.022	0.102	0.217								
ART	-0.222	-0.053	0.029	-0.261	-0.026							
SMO	0.147	0.164	-0.038	-0.114	-0.196	0.042						
DIA	0.025	0.316	0.240	-0.240	**0.403**	-0.025	0.082					
CMP	-0.142	-0.148	-0.131	**0.464**	-0.122	-0.111	0.016	-0.219				
WEI	0.075	0.181	0.103	-0.108	-0.213	0.063	**0.334**	0.066	0.129			
ATT	-0.089	-0.023	0.236	-0.173	0.004	-0.104	0.081	-0.017	-0.035	**0.373**		
FHD	-0.322	-0.021	-0.140	-0.061	0.024	0.159	0.192	-0.318	-0.063	-0.121	-0.070	
PER	0.026	-0.073	-0.048	**0.450**	0.035	-0.128	-0.020	0.154	0.212	0.173	-0.006	-0.033

Correlation of cue subjective weights (ψ_i)

N = 34 therefore correlations are significant ($p < 0.05$) when greater than 0.329

	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	CMP	WEI	ATT	FHD
HYP	0.312											
AGE	**0.360**	0.205										
GEN	-0.288	-0.057	0.002									
OCC	0.230	-0.092	-0.303	-0.318								
ART	0.229	-0.071	**0.350**	-0.016	-0.034							
SMO	**0.498**	0.197	-0.037	-0.141	0.290	-0.017						
DIA	0.030	0.101	-0.043	-0.223	0.323	-0.007	0.194					
CMP	-0.325	**-0.402**	-0.073	0.059	-0.299	-0.231	**-0.436**	-0.197				
WEI	0.137	0.061	-0.069	0.047	0.311	-0.056	0.168	0.109	-0.327			
ATT	-0.097	**-0.385**	0.123	-0.293	-0.081	-0.078	-0.239	-0.274	**0.401**	-0.005		
FHD	**-0.587**	-0.104	-0.219	0.025	-0.010	-0.200	-0.185	-0.187	0.090	-0.076	0.026	
PER	-0.018	0.037	-0.124	0.109	0.237	-0.029	0.188	-0.165	0.079	-0.092	-0.068	-0.053

KEY:- Cho = cholesterol level, Hyp = hypertension, Age = age, Gen = gender, Occ = occupation, Art = evidence of arteriosclerosis, Smo = smoking behaviour, Dia = diabetes, Cmp = compliance with advice given on diet, Wei = weight, Att = attitude to treatment, Fhd = family history of ischaemic heart disease, Per = personality.

Appendix 22
Correlation of cue weights MIGRAINE task, Study 1

Correlation of cue standardised regression coefficients (β_i)

N = 34: significance ($p < 0.05$) where $r > 0.33$.

	DUR	FRE	AGE	GEN	OCC	MWK	SMO	NAU	VIS	WEI	ATT	RES
FRE	-0.076											
AGE	-0.259	-0.308										
GEN	-0.088	-0.082	**0.430**									
OCC	-0.206	-0.310	0.128	0.057								
MWK	0.027	0.232	0.266	0.025	**-0.465**							
SMO	-0.295	-0.261	-0.044	0.107	0.283	**-0.377**						
NAU	0.301	**-0.421**	0.255	**0.370**	**0.375**	-0.239	0.134					
VIS	0.257	**-0.386**	0.140	**0.344**	0.007	-0.241	0.167	**0.581**				
WEI	0.021	**-0.438**	0.092	-0.012	0.299	-0.004	-0.015	0.309	-0.063			
ATT	0.176	**-0.645**	0.250	0.248	-0.005	-0.193	0.016	**0.338**	**0.343**	0.115		
RES	**0.337**	**-0.650**	0.401	0.099	0.082	0.087	0.024	0.247	0.088	**0.364**	**0.500**	
PER	-0.136	0.319	-0.172	-0.164	-0.216	0.096	**-0.365**	-0.286	-0.146	-0.251	**-0.341**	-0.307

Correlation of cue subjective weights (ψ_i)

N = 34 therefore correlations are significant ($p < 0.05$) when greater than 0.329

	DUR	FRE	AGE	GEN	OCC	MWK	SMO	NAU	VIS	WEI	ATT	RES
FRE	-0.066											
AGE	-0.066	0.122										
GEN	-0.151	0.008	0.019									
OCC	-0.018	-0.074	0.285	-0.174								
MWK	**0.374**	0.188	0.001	0.077	0.027							
SMO	0.036	0.263	-0.043	0.013	-0.020	0.014						
NAU	-0.286	**-0.390**	-0.303	-0.270	0.063	-0.326	-0.220					
VIS	-0.208	**-0.383**	-0.313	-0.270	0.076	-0.317	-0.227	**0.905**				
WEI	0.228	0.181	-0.075	**-0.624**	0.074	0.038	**0.595**	-0.079	-0.089			
ATT	-0.321	-0.216	0.154	0.190	0.237	-0.024	0.267	-0.113	-0.088	0.035		
RES	-0.132	-0.303	0.062	-0.134	0.023	-0.174	-0.041	-0.138	-0.228	0.097	0.152	
PER	-0.107	-0.168	-0.139	-0.025	**-0.374**	-0.253	0.059	0.173	0.175	0.103	-0.090	0.043

KEY: Dur = duration of untreated attack, Fre = frequency of attack, Age = age, Gen = gender, Occ = occupation, Mwk = whether they miss work or not, Smo = smoking behaviour, Nau = nausea, Vis = visual disturbance, Wei = weight, Att = attitude to treatment, Res = response to acute treatment, Per = personality.

Appendix 23

Correlation of cue weights HRT task, Study 1

Correlation of cue standardised regression coefficients (β_i)

N = 12: significance ($p < 0.05$) when $r > 0.5324$

	MEN	HOT	AGE	OCC	MOOD	SMO	LIB	VAG	WEI	ATT	FHD	FBC
HOT	**0.609**											
AGE	-0.028	0.179										
OCC	-0.327	-0.395	0.287									
MOOD	0.219	0.073	-0.068	-0.203								
SMO	-0.063	0.107	0.375	0.071	0.001							
LIB	0.278	-0.237	-0.116	-0.109	**0.535**	-0.451						
VAG	0.249	0.337	-0.166	0.061	0.127	**-0.642**	0.297					
WEI	-0.395	-0.014	-0.004	-0.035	**-0.894**	0.147	**-0.712**	-0.237				
ATT	0.300	-0.368	-0.068	-0.006	0.333	0.168	0.502	-0.157	-0.417			
FHD	-0.081	-0.172	0.092	0.309	-0.251	0.018	0.239	0.162	0.077	-0.103		
FBC	**0.543**	0.251	-0.021	-0.264	-0.256	-0.371	0.000	0.332	0.081	0.083	-0.124	
PER	**-0.542**	-0.040	0.139	-0.114	-0.126	0.356	-0.400	-0.149	0.373	-0.351	0.062	-0.288

Correlation of cue subjective weights (ψ_i)

N = 13 therefore correlations are significant (in bold) when > 0.514 ($p < 0.05$)

	MEN	HOT	AGE	OCC	MOO	SMO	LIB	VAG	WEI	ATT	FHD	FBC
HOT	**0.649**											
AGE	0.261	0.274										
OCC	0.118	0.034	0.118									
MOO	**-0.625**	**-0.650**	0.086	-0.086								
SMO	-0.172	-0.226	-0.394	0.083	-0.185							
LIB	0.118	-0.144	0.242	-0.500	0.028	-0.447						
VAG	**0.659**	**0.539**	-0.034	0.348	**-0.675**	-0.103	0.053					
WEI	0.300	0.498	**0.524**	-0.044	-0.132	**-0.796**	0.242	0.053				
ATT	-0.250	-0.499	0.365	0.135	0.238	0.138	0.273	-0.086	-0.142			
FHD	0.033	0.080	-0.206	0.256	-0.013	0.065	-0.244	0.233	0.051	0.080		
FBC	0.006	0.492	-0.007	0.002	**-0.604**	0.234	-0.317	0.219	-0.007	-0.291	-0.367	
PER	-0.319	0.060	0.142	-0.540	0.084	0.017	0.239	**-0.509**	0.186	0.080	-0.204	0.130

KEY: Men = menstruation, Hot = hot flushes, Age = age, Occ = occupation, Mood = mood states, Smo = smoking behaviour, Lib = libido, Vag = vaginal dryness, Wei = weight, att = attitude, Fhd = family history of ischaemic heart disease, Fbc = family history of breast cancer, Per = personality.

Appendix 24
Between task correlation of Relative Importances of cues, Study 1

LIPID and MIGRAINE

Cue	β_i	S_i
	N = 32	N = 33
Age	-0.130	0.168
Gender	0.139	-0.238
Occupation	-0.170	-0.051
Smoking	0.031	0.111
Weight	0.138	-0.050
Attitude	**0.560**	**0.421**
Personality	0.077	**0.503**

LIPID and HRT

Cue	β_i	S_i
	N = 11	N = 12
Age	-0.118	0.049
Occupation	-0.124	-0.128
Smoking	-0.280	-0.379
Weight	0.204	-0.144
Attitude	-0.195	**0.653**
FHD	0.037	0.142
Personality	0.003	-0.064

MIGRAINE and HRT

Cue	β_i	S_i
	N = 11	N = 12
Age	-0.133	*
Occupation	-0.135	0.296
Smoking	-0.465	*
Weight	**0.599**	*
Attitude	0.354	**0.784**
Personality	-0.264	*

Key:- * = All doctors rated the cue on one or other task at 0.
Significant figures are in bold ($p < 0.05$).
β_i = standardised regression coefficients
S_i = subjective ratings

Appendix 25

Subjective Weights (w_i) on the LIPID task, Study 1

GP	r	R_S	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	CMP	WEI	ATT	FHD	PER
1	0.70	0.55	0.48	0.33	-0.24	0.38	0.00	0.00	0.33	0.33	0.00	-0.24	0.00	0.29	0.00
2	0.73	0.47	0.49	0.11	-0.11	0.11	0.00	0.00	-0.32	0.43	0.32	0.22	0.49	0.05	-0.11
3	0.64	0.51	0.39	0.26	-0.26	0.13	0.04	0.34	-0.30	0.30	0.26	-0.26	0.34	0.34	0.13
4	0.73	0.63	0.52	0.23	-0.17	0.17	0.00	0.35	0.35	0.52	0.00	0.00	0.35	0.06	0.00
5	0.52	0.44	0.35	0.35	-0.35	0.08	0.00	0.00	-0.35	0.35	0.35	-0.16	0.24	0.27	0.00
6	0.58	0.33	0.37	0.28	-0.24	0.24	-0.16	0.32	-0.32	0.37	0.20	-0.20	0.16	0.24	-0.20
7	0.79	0.51	0.34	0.24	-0.27	0.27	-0.07	0.24	-0.34	0.24	0.31	-0.31	0.34	0.27	-0.12
8	0.65	0.57	0.44	0.29	-0.24	0.10	0.00	0.10	0.39	0.20	0.29	-0.24	0.39	0.24	0.39
9	0.61	0.51	0.66	0.46	0.00	0.13	0.00	0.46	-0.13	0.07	0.13	0.00	0.26	0.00	0.00
10	0.72	0.51	0.44	0.35	0.00	0.35	0.00	0.44	0.13	0.35	0.00	-0.13	0.13	0.35	0.00
11	0.81	0.51	0.45	0.24	-0.35	0.40	0.00	0.00	0.00	-0.26	0.21	0.00	0.35	0.40	0.00
12	0.62	0.30	0.35	0.35	-0.25	0.00	0.00	0.30	-0.15	0.30	0.25	-0.35	0.35	0.40	0.00
13	0.69	0.41	0.48	0.38	-0.48	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.48	0.00
14	0.21	0.24	0.37	0.32	0.00	0.32	-0.27	0.32	-0.32	0.00	0.27	-0.27	0.32	0.32	0.00
15	0.67	0.42	0.28	0.28	-0.28	0.32	-0.07	0.25	-0.32	0.32	0.32	-0.25	0.04	0.32	-0.04
17	0.74	0.47	0.29	0.19	-0.34	0.43	0.00	0.39	-0.19	0.48	0.00	-0.05	0.00	0.29	0.00
19	0.90	0.67	0.49	0.38	-0.54	0.00	0.00	0.16	0.22	0.43	0.00	-0.05	0.00	0.00	0.00
20	0.80	0.52	0.31	0.27	-0.27	0.00	0.00	0.31	-0.40	0.31	0.00	-0.36	0.44	0.36	0.00
21	0.80	0.74	0.69	0.07	0.00	0.07	0.00	0.69	0.14	0.14	0.00	0.00	0.35	0.00	0.00
22	0.81	0.47	0.34	-0.34	-0.34	0.25	0.04	0.29	-0.38	-0.34	0.25	-0.29	0.29	0.34	0.08
24	0.80	0.42	0.25	0.25	0.00	0.34	-0.21	0.21	-0.34	-0.42	0.21	-0.29	0.38	0.29	0.00
25	0.50	0.30	0.30	-0.30	-0.30	0.00	0.00	0.25	-0.30	0.40	0.35	-0.25	0.35	0.35	0.00
26	0.71	0.71	0.39	0.11	-0.45	0.00	0.00	0.00	0.11	0.11	0.00	0.11	0.56	0.45	-0.17
27*	-	-	0.17	-0.17	-0.31	0.35	0.00	0.26	-0.39	0.31	0.26	0.22	0.31	0.35	0.09
28	0.65	0.44	0.37	0.31	-0.37	0.42	0.00	0.16	0.31	0.42	0.00	0.00	0.00	0.26	0.00
29	0.63	0.32	0.36	0.31	-0.36	0.00	0.00	0.00	-0.31	0.26	0.31	0.00	0.36	0.36	0.00
30	0.74	0.38	0.58	0.41	0.00	0.12	0.00	0.17	-0.23	0.35	0.06	-0.47	0.00	0.17	0.00
31	0.55	0.46	0.50	0.45	0.00	0.00	0.00	0.15	0.40	0.40	0.00	0.15	0.00	0.40	0.00
32	0.66	0.32	0.40	0.31	-0.40	0.04	0.00	0.31	0.31	0.27	0.22	-0.40	0.13	0.27	0.04
33	0.66	0.41	0.35	-0.17	-0.35	0.35	-0.07	0.17	-0.35	0.35	0.35	-0.35	0.17	0.35	0.00
34	0.73	0.24	0.44	0.00	-0.44	0.44	-0.04	0.26	0.00	0.00	0.44	0.00	0.22	0.22	0.09
35	0.81	0.24	0.41	-0.37	-0.37	0.04	0.00	0.24	0.24	0.37	0.20	-0.29	0.37	0.24	-0.04
36	0.30	0.33	0.42	0.00	0.00	0.00	0.00	0.42	0.00	0.37	0.26	-0.05	0.37	0.42	-0.21
37	0.20	0.10	0.34	0.34	-0.34	0.27	0.07	0.31	-0.20	0.34	-0.17	0.34	0.07	0.34	0.07
Mean		0.44	0.41	0.19	-0.25	0.18	-0.02	0.24	-0.08	0.24	0.17	-0.12	0.24	0.28	0.00
S.d.		0.14	0.11	0.23	0.16	0.16	0.07	0.16	0.27	0.23	0.15	0.20	0.17	0.13	0.10

* = only subjective data available for this subject.

KEY: CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, ART = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, CMP = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FHD = family history of ischaemic heart disease, PER = personality, R_S = correlation between values predicted from these subjective weights when used as regression weights and the actual decision values, r = correlation between subjective weights and standardised regression coefficients.

Appendix 26

Subjective Weights (w_i) on the MIGRAINE task, Study 1

GP	r	R_S	DUR	FRE	AGE	GEN	OCC	MWK	SMO	NAU	VIS	WEI	ATT	RES	PER
1	0.72	0.58	0.00	0.52	0.00	0.00	0.00	-0.52	0.00	0.31	0.10	0.00	0.47	0.47	0.36
2	0.78	0.66	0.13	0.66	0.00	0.00	0.00	-0.53	0.00	0.20	0.20	0.00	0.53	0.13	0.00
3	0.73	0.68	0.19	0.57	0.00	0.00	0.00	-0.57	0.00	0.00	0.00	0.00	0.38	0.44	0.19
4	0.88	0.72	0.38	0.49	0.00	0.00	0.00	-0.55	0.00	0.16	0.22	0.00	0.44	0.33	0.11
5	0.76	0.56	0.41	0.52	0.00	0.00	0.16	-0.36	0.00	0.21	0.21	0.00	0.52	0.36	0.00
6	0.54	0.45	0.41	0.41	-0.26	0.00	-0.10	-0.31	0.00	0.41	0.41	0.00	0.26	0.31	0.00
7	0.69	0.62	0.44	0.63	0.00	0.00	0.00	-0.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.75	0.71	0.45	0.35	0.00	0.00	0.00	-0.35	0.00	0.35	0.35	0.00	0.40	0.30	0.40
9	0.76	0.69	0.54	0.54	0.00	0.00	0.00	-0.38	0.00	0.00	0.00	0.00	0.43	0.33	0.00
10	0.64	0.59	0.57	0.57	0.00	0.00	0.00	-0.29	0.00	0.17	0.00	0.00	0.17	0.46	0.00
11	0.65	0.61	0.43	0.54	0.00	0.00	0.00	-0.31	0.00	0.00	0.00	0.00	0.48	0.48	0.00
12	0.61	0.56	0.50	0.50	0.00	0.00	0.00	-0.39	0.00	0.17	0.17	0.00	0.45	0.39	0.00
13	0.68	0.62	0.73	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.63	0.46	0.48	0.48	0.00	0.00	0.00	-0.41	0.00	0.00	0.00	0.00	0.48	0.41	0.00
15	0.53	0.48	0.39	0.49	0.00	0.00	-0.39	-0.44	0.00	0.29	0.29	0.00	0.15	0.20	0.39
17	0.76	0.60	0.31	0.46	0.00	-0.15	-0.10	-0.52	0.00	0.36	0.36	0.00	0.21	0.31	0.00
19	0.73	0.70	0.71	0.71	0.00	0.00	0.00	-0.21	0.00	0.00	0.00	0.00	0.07	0.00	0.00
20	0.86	0.73	0.39	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.59	0.00
21	0.78	0.65	0.20	0.50	0.00	0.30	-0.20	-0.35	0.00	0.00	0.00	-0.40	0.50	0.15	0.00
22	0.70	0.55	0.42	0.48	0.00	0.00	0.00	-0.42	0.00	0.32	0.32	0.00	0.37	0.37	0.16
23	0.49	0.41	0.27	0.55	0.00	0.00	0.00	-0.33	0.00	0.33	0.16	0.00	0.44	0.33	-0.38
24	0.89	0.66	0.44	0.55	0.00	0.00	0.00	-0.38	0.00	0.00	0.16	0.00	0.44	0.44	0.00
25	0.55	0.43	0.36	0.36	0.00	0.00	0.00	-0.36	-0.36	0.36	0.36	-0.36	0.04	0.36	0.00
26	0.81	0.72	0.42	0.42	0.00	0.00	0.00	-0.36	0.00	0.00	0.00	0.00	0.42	0.36	0.00
27	0.85	0.77	0.24	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00
28	0.67	0.60	0.23	0.78	0.00	0.00	0.00	-0.23	0.00	0.23	0.23	0.00	0.55	0.00	0.00
29	0.63	0.57	0.28	0.56	0.00	0.00	0.17	-0.39	0.00	0.33	0.33	-0.22	0.39	0.28	0.00
30	0.78	0.72	0.22	0.75	0.00	0.00	0.00	-0.30	0.00	0.07	0.15	0.00	0.37	0.45	0.00
31	0.46	0.28	0.39	0.39	0.00	0.00	0.00	-0.39	0.00	0.39	0.39	0.00	0.00	0.39	0.00
32	0.67	0.60	0.46	0.59	0.00	0.00	-0.26	-0.26	0.00	0.07	0.07	0.00	0.13	0.53	0.13
33	0.26	0.26	0.38	0.38	0.00	0.00	0.38	-0.38	0.00	0.30	0.30	0.00	0.30	0.38	0.00
35	0.70	0.57	0.35	0.50	0.00	0.00	0.10	-0.35	0.00	0.30	0.40	0.00	0.50	-0.20	0.00
36	0.85	0.62	0.53	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.59	0.53	0.00
37	0.58	0.43	0.44	0.55	-0.22	0.00	-0.22	-0.39	0.00	0.28	0.28	0.00	0.17	0.22	0.28
Mean		0.58	0.39	0.53	-0.01	0.00	-0.01	-0.34	-0.01	0.17	0.16	-0.03	0.33	0.31	0.05
St.dev.		0.12	0.15	0.12	0.06	0.06	0.13	0.16	0.06	0.15	0.15	0.10	0.19	0.19	0.14

KEY: DUR = duration, FRE = frequency, AGE = age, GEN = gender, OCC = occupation, MWK = misses work, SMO = smoking, NAU = nausea, VIS = visual disturbance, WEI = weight, ATT = attitude, RES = response to acute treatment, PER = personality, R_S = correlation between values predicted from weights when used as regression weights and the actual decision values, r = correlation between subjective weights and standardised regression coefficients.

Appendix 27

Subjective Weights (w_i) on the HRT task, Study 1

GP	r	R_S	MEN	HOT	AGE	OCC	MOO	SMO	LIB	VAG	WEI	ATT	FHD	FBC	PER
9	0.85	0.69	-0.29	0.46	-0.29	0.00	0.34	0.11	0.34	0.46	-0.23	0.46	0.29	-0.34	0.00
13	0.13	0.05	-0.54	0.54	-0.54	0.00	0.21	0.00	0.11	0.37	0.00	0.00	0.27	0.00	0.00
17	0.84	0.50	-0.38	0.49	-0.16	-0.22	0.27	0.00	0.54	0.27	-0.05	0.32	0.00	-0.11	0.22
24	0.70	0.53	-0.43	0.43	-0.27	0.00	0.27	0.27	0.00	0.38	-0.27	0.54	0.16	0.00	0.00
26*	-	-	-0.12	0.36	0.00	0.00	0.36	0.00	0.36	0.30	0.00	0.60	0.36	-0.54	0.00
27	0.62	0.38	-0.45	0.35	-0.35	0.00	0.15	0.25	0.35	0.40	-0.30	0.40	0.00	0.00	0.00
28	0.74	0.66	-0.39	0.50	0.45	0.00	0.34	0.00	0.28	0.28	0.00	0.50	0.00	0.00	0.00
30	0.78	0.66	-0.38	0.51	0.19	0.00	0.38	0.00	0.32	0.44	-0.06	0.51	0.19	-0.19	0.06
31	0.78	0.60	0.56	0.72	0.08	0.00	0.00	0.00	0.24	0.56	0.00	0.00	0.00	0.00	0.00
33	0.43	0.29	0.00	0.64	0.00	0.00	0.00	0.00	0.32	0.64	0.00	0.45	0.51	0.00	0.00
34	0.56	0.37	-0.42	0.42	0.00	0.00	0.42	-0.25	0.42	0.42	0.00	0.42	0.00	-0.21	0.00
35	0.73	0.61	-0.53	0.53	0.11	0.00	0.21	0.08	0.26	0.26	0.00	0.53	0.11	0.00	0.29
36	0.89	0.55	0.00	0.52	0.00	0.00	0.00	0.00	0.45	0.60	0.00	0.67	0.00	0.00	0.00
Mean		0.49	-0.26	0.50	-0.06	-0.02	0.23	0.04	0.31	0.41	-0.07	0.42	0.14	-0.11	0.04
St. dev.		0.19	0.30	0.10	0.26	0.06	0.15	0.13	0.14	0.13	0.12	0.20	0.17	0.17	0.10

KEY: MEN = menstruation, HOT = hot flushes, AGE = age, OCC = occupation, MOO = mood states, SMO = smoking behaviour, LIB = libido, VAG = vaginal dryness, WEI = weight, ATT = attitude, FHD = family history of ischaemic heart disease, FBC = family history of breast cancer, PER = personality, R_S = correlation between values predicted when using weights as regression weights and the actual decision values, r = correlation between subjective weights and standardised regression coefficients.

* = Subjective weights only collected for this subject.

Appendix 28
Instructions for the IS task (Study 2, Chapter 6)

INSTRUCTIONS

You will be now be presented with 100 hypothetical cases for which you have to decide whether or not the individual should be prescribed a lipid lowering agent. You should assume that the option to refer is not available. In each case you originally tested the patient's blood cholesterol level at least six months previously. You offered the usual advice on alterations to diet and recommended, where appropriate, that the patient should give up smoking cigarettes. The blood cholesterol level given in the problem is the current one and reflects any changes in the patient's life-style that he or she has made or is likely to make.

A number of pieces of information about the patient and their medical history are available. These will appear on the screen as the relevant bar on the attached keyboard is pressed. Please choose only that information which you consider is necessary to make the decision and in the order in which it seems natural to seek it. When you feel you have uncovered all relevant facts available press the bar at the bottom of the board. This will activate the mouse. We realise that the cases may not be clear cut and that in a real life consultation you might wish to know other information before making a definite decision.

Given all the information you have chosen to look at, you should indicate your preference for prescribing a lipid lowering drug in percentage terms. If you are certain that you would prescribe then give a response of 100% and if you are certain you would not prescribe then give a response of 0%. Where you are not certain then please given any intermediate value you wish to convey the likelihood that you would prescribe.

In each case you can set the percentage by pulling the mouse to the left or right. When the reading is correct then please press any button on the mouse to proceed to the next problem.

PRESS ANY MOUSE BUTTON TO START

Appendix 29
Inter-cue correlations on the IS task (Study2)

	Cho	Hyp	Age	Gen	Occ	Eva	Smo	Dia	Com	Wei	Att	FH	Per
Cho	1	0.09	-0.05	0.00	0.17	-0.11	0.11	0.07	0.15	0.08	-0.01	-0.03	-0.10
Hyp	0.09	1	0.05	0.19	0.03	-0.01	-0.04	0.02	0.09	-0.1	-0.02	-0.02	-0.02
Age	-0.05	0.05	1	0.07	0.01	-0.04	-0.14	0.00	-0.17	-0.03	-0.03	0.02	-0.14
Gen	0.00	0.19	0.07	1	-0.12	0.22	-0.04	-0.07	-0.06	-0.06	-0.10	-0.12	-0.03
Occ	0.17	0.03	0.01	-0.12	1	-0.11	0	0.04	0.26	0.13	0.04	0.06	0.03
Eva	-0.11	-0.01	-0.04	0.22	-0.11	1	-0.17	-0.06	0.03	0.04	-0.15	0.09	-0.19
Smo	0.11	-0.04	-0.14	-0.04	0.00	-0.17	1	-0.01	-0.05	-0.12	0.01	-0.17	-0.12
Dia	0.07	0.02	0.00	-0.07	0.04	-0.06	-0.01	1	0.02	-0.04	0.00	0.05	0.03
Com	0.15	0.09	-0.17	-0.06	0.26	0.03	-0.05	0.02	1	0.16	0.01	0.14	-0.12
Wei	0.08	-0.1	-0.03	-0.06	0.13	0.04	-0.12	-0.04	0.16	1	0.00	-0.05	0.09
Att	-0.01	-0.02	-0.03	-0.10	0.04	-0.15	0.01	0.00	0.01	0.00	1	0.14	-0.12
FH	-0.03	-0.02	0.02	-0.12	0.06	0.09	-0.17	0.05	0.14	-0.05	0.14	1	0.06
Per	-0.10	-0.02	-0.14	-0.03	0.03	-0.19	-0.12	0.03	-0.12	0.09	-0.12	0.06	1

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality.

Appendix 30

Comparison of indices on the LIPID task of Study 1 and the IS task of Study 2

Task GP	LIPID Consistency	IS Consistency	LIPID (100) Mean judgement	IS Mean judgement	LIPID (100) R^2	IS R^2	Consistency of policy
1	0.60	0.48	43.8	42.1	0.55	0.60	0.31
2	0.29	0.51	11.4	23.9	0.53	0.64	0.40
3	0.30	0.65	57.6	47.7	0.57	0.50	0.82
4	0.77	0.55	17.9	12.7	0.66	0.52	0.67
5	0.44	0.50	26.9	59.4	0.51	0.75	0.61
6	0.17	0.54	45.5	36.5	0.33	0.65	0.88
7	0.32	0.66	6.8	17.0	0.42	0.51	0.87
8	-	-	46.2	-	0.60	-	-
9	0.19	0.86	10.9	17.4	0.52	0.58	0.90
10	-	-	38.1	-	0.54	-	-
11	0.15	0.53	20.1	8.7	0.45	0.28	0.75
12	0.34	0.34	6.1	11.4	0.32	0.41	0.72
13	0.46	0.19	92.7	62.6	0.27	0.39	0.28
14	-0.12	0.28	48.6	47.6	0.36	0.13	0.64
15	-0.12	0.33	7.1	0.5	0.32	0.16	-0.01
17	0.26	0.51	54.8	49.4	0.48	0.56	0.37
19	0.43	0.55	16.0	15.4	0.62	0.68	0.56
20	0.23	0.39	16.9	19.4	0.38	0.44	0.03
21	0.36	0.22	50.2	21.9	0.79	0.66	0.12
22	0.10	0.52	36.8	29.3	0.45	0.61	0.78
23	-	-	-	-	-	-	-
24	0.32	0.39	20.6	14.5	0.22	0.40	0.38
25	-	-	43.8	-	0.39	-	-
26	0.72	0.77	56.1	68.0	0.87	0.60	0.94
27	-	-	-	-	-	-	-
28	0.22	0.55	16.9	55.1	0.72	0.73	0.58
29	0.44	0.22	21.6	53.5	0.22	0.47	0.77
30	0.49	0.54	11.7	21.0	0.45	0.70	0.84
31	0.68	0.93	75.0	71.1	0.63	0.71	0.99
32	0.64	0.48	16.3	22.1	0.46	0.54	0.47
33	0.37	0.43	17.8	15.2	0.55	0.24	0.86
34	0.28	0.57	11.0	17.5	0.37	0.37	0.80
35	0.51	0.43	39.0	27.8	0.42	0.45	0.35
36	0.63	0.08	62.8	60.0	0.69	0.51	-0.05
37	-0.11	0.37	72.5	40.4	0.60	0.57	0.74
Mean	0.35	0.48	33.9	32.97	0.49	0.51	0.58
St.dev.	0.23	0.19	22.7	20.1	0.16	0.16	0.30

Key :- r_b = correlation between absolute values of doctor's standardised regression coefficients for cues on information study and on original study (N = 13) [$p < 0.05$ when $r_b > 0.553$ (2 tailed test)]

r = consistency/reliability = correlation between decisions made on information seeking study and same cases in original task (N = 100) [$p < 0.05$ when $r > 0.2$].

R^2 = linear fit.

Appendix 31

Tacit policies: Standardised Regression Coefficients based on decisions made on last 100 cases of the LIPID task of Study 1

GP	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	COM	WEI	ATT	FHD	PER
1	0.50	0.21	-0.01	0.09	0.07	0.01	0.08	0.46	-0.06	0.03	0.05	0.14	-0.02
2	0.43	-0.05	-0.06	0.01	0.21	0.02	-0.29	0.27	-0.22	-0.01	0.46	-0.16	0.01
3	0.35	-0.12	-0.15	0.03	0.00	0.39	-0.21	0.00	0.19	0.03	0.40	0.16	-0.08
4	0.58	0.05	-0.04	0.00	0.07	0.15	-0.03	0.39	-0.03	0.01	0.42	-0.07	-0.05
5	0.56	0.00	0.08	0.01	0.07	0.02	-0.20	0.20	0.02	0.09	0.33	0.03	-0.14
6	0.55	0.09	-0.06	-0.09	0.04	-0.03	-0.15	0.21	-0.08	-0.05	0.08	-0.05	-0.07
7	0.50	0.06	0.01	0.03	0.01	0.10	-0.41	0.14	0.03	-0.32	0.03	-0.05	-0.07
8	0.27	0.26	-0.29	0.03	-0.05	0.12	0.19	0.17	0.08	-0.07	0.46	0.22	-0.13
9	0.59	0.09	0.16	0.01	0.08	0.02	-0.03	0.16	-0.03	0.07	0.29	0.08	-0.13
10	0.58	0.12	-0.08	0.04	0.05	0.40	-0.14	0.20	0.01	-0.14	0.17	0.10	0.04
11	0.45	0.09	-0.21	0.32	0.04	0.03	-0.09	-0.36	-0.03	-0.08	-0.07	0.07	-0.10
12	0.35	0.15	-0.06	0.05	0.08	0.24	-0.10	0.34	-0.02	-0.13	0.05	0.17	-0.02
13	0.25	0.09	-0.22	0.02	-0.04	0.20	0.19	0.17	0.21	0.20	-0.02	0.06	-0.03
14	-0.08	0.04	0.04	0.00	0.00	0.05	0.07	0.13	0.03	0.11	0.63	0.12	0.11
15	0.29	-0.08	-0.01	0.10	0.16	0.14	-0.13	-0.03	0.07	-0.43	-0.05	0.29	0.06
17	0.18	0.03	-0.18	0.01	0.08	0.52	-0.26	0.34	-0.06	-0.06	-0.08	-0.03	-0.09
19	0.36	0.14	-0.59	0.06	0.02	-0.02	0.11	0.22	0.07	0.04	0.00	-0.02	0.03
20	0.29	-0.01	-0.08	-0.05	0.09	-0.02	-0.56	0.16	-0.08	-0.13	0.24	0.03	0.00
21	0.30	0.11	0.05	-0.04	0.01	0.87	0.05	0.07	-0.03	0.09	0.01	0.04	-0.04
22	0.23	-0.10	-0.34	-0.09	0.11	0.14	-0.44	0.01	-0.03	-0.01	0.21	0.22	0.04
24	0.29	0.12	-0.09	0.04	-0.02	-0.03	-0.13	-0.31	0.17	-0.09	0.20	-0.02	-0.15
25	0.31	0.04	0.01	0.03	0.18	0.10	0.06	0.44	-0.05	-0.30	0.12	0.01	-0.06
26	0.22	0.03	-0.12	0.05	0.08	0.05	0.05	0.03	-0.01	-0.01	0.89	0.09	0.03
28	0.32	0.22	-0.14	0.01	0.18	0.09	0.05	0.68	0.04	0.05	-0.06	0.00	0.01
29	0.25	0.26	-0.02	-0.05	-0.02	0.03	-0.11	0.30	-0.03	0.11	0.26	-0.01	-0.03
30	0.61	-0.06	0.03	-0.12	-0.07	-0.04	-0.10	0.12	0.09	-0.32	-0.06	0.05	-0.11
31	0.80	0.00	0.03	0.03	-0.03	0.10	0.13	0.10	0.03	0.00	0.09	0.05	0.14
32	0.39	0.11	-0.08	-0.13	0.08	0.10	-0.26	0.19	-0.05	-0.53	-0.09	0.00	-0.05
33	0.40	0.09	-0.05	0.31	0.15	-0.01	-0.51	0.13	0.12	-0.12	0.07	-0.06	0.17
34	0.35	0.04	-0.18	0.43	-0.04	-0.15	-0.15	0.05	0.19	0.03	0.18	-0.08	0.05
35	0.41	-0.32	-0.41	0.13	0.00	-0.02	-0.07	-0.13	0.02	-0.25	0.11	0.11	-0.06
36	0.80	0.13	-0.08	0.00	-0.05	0.03	0.07	0.02	0.11	-0.02	0.09	-0.03	0.11
37	0.72	-0.02	0.07	-0.04	-0.03	0.01	0.19	0.17	0.05	-0.05	0.15	0.04	0.09
Mean	0.41	0.05	-0.09	0.04	0.05	0.11	-0.10	0.15	0.02	-0.07	0.17	0.05	-0.02
St.dev	0.18	0.11	0.15	0.12	0.07	0.19	0.19	0.20	0.09	0.16	0.22	0.10	0.08
T test p	0.00	0.01	0.00	0.08	0.00	0.00	0.01	0.00	0.16	0.02	0.00	0.01	0.27

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality

Appendix 32

Tacit policies: Standardised Regression Coefficients on the IS task (Study 2)

GP	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	COM	WEI	ATT	FHD	PER
1	**0.33**	**0.16**	-0.02	0.07	0.07	0.07	0.10	**0.21**	0.07	0.03	**0.67**	-0.13	0.10
2	**0.41**	**-0.20**	**-0.29**	**0.32**	-0.01	**0.27**	-0.09	**0.30**	-0.05	0.04	**0.43**	-0.06	-0.06
3	**0.42**	-0.01	**-0.16**	0.10	-0.08	**0.32**	-0.05	0.13	**0.22**	-0.09	**0.40**	0.09	-0.05
4	**0.62**	0.13	0.02	-0.02	0.08	-0.04	-0.01	0.07	-0.07	0.01	0.13	**0.32**	-0.10
5	**0.41**	0.00	**-0.31**	0.04	0.02	-0.07	**-0.21**	**0.60**	**0.11**	0.03	**0.18**	0.05	-0.05
6	**0.70**	-0.01	**-0.26**	0.03	0.12	0.02	-0.03	**0.15**	0.03	-0.03	0.05	-0.07	**-0.15**
7	**0.52**	0.04	-0.12	0.10	0.08	0.05	**-0.37**	**0.27**	0.04	**-0.22**	**0.22**	0.01	0.04
9	**0.65**	**0.18**	0.05	0.02	0.06	**0.18**	-0.07	**0.23**	-0.07	-0.01	**0.26**	0.01	-0.09
11	**0.33**	**-0.20**	**-0.28**	**0.24**	-0.08	0.17	-0.09	-0.17	0.06	-0.03	0.04	-0.11	-0.06
12	**0.33**	**0.20**	-0.09	-0.03	0.05	-0.04	0.08	**0.36**	0.04	0.08	0.15	**0.26**	-0.14
13	**0.25**	0.06	**-0.34**	0.13	0.06	-0.03	**0.30**	**0.24**	0.01	-0.03	**0.17**	**0.20**	0.06
14	0.07	-0.05	-0.09	0.06	-0.01	0.08	-0.19	0.01	0.15	0.01	**0.29**	**0.20**	-0.01
15	0.14	**-0.20**	**-0.29**	**0.28**	0.03	-0.12	**-0.24**	0.01	0.11	**-0.22**	0.06	0.02	-0.09
17	**0.40**	**0.14**	-0.06	**0.21**	0.00	**0.29**	-0.11	**0.35**	0.04	**-0.20**	**0.44**	-0.04	0.12
19	**0.69**	0.01	**-0.13**	-0.01	0.03	0.00	0.07	**0.37**	0.02	0.08	-0.01	0.03	0.05
20	**0.40**	-0.02	0.09	0.01	-0.02	**0.53**	-0.08	0.05	0.07	-0.04	**0.25**	**0.23**	0.10
21	**0.67**	-0.01	-0.07	0.02	0.06	0.09	-0.11	**0.20**	-0.02	-0.05	**0.41**	**-0.14**	-0.11
22	**0.24**	-0.12	-0.13	**0.14**	**0.17**	-0.08	**-0.70**	0.00	-0.08	-0.07	-0.10	**0.30**	-0.12
24	**0.45**	-0.15	-0.04	0.00	0.08	**0.19**	**-0.29**	-0.03	0.04	**-0.22**	**0.33**	0.10	-0.08
26	-0.02	0.11	-0.06	-0.05	0.05	**0.17**	0.02	0.04	0.04	0.04	**0.82**	-0.10	0.11
28	**0.33**	0.09	**-0.15**	-0.05	0.02	**0.12**	0.04	**0.50**	0.08	-0.01	**0.59**	-0.06	0.01
29	**0.36**	0.10	-0.12	0.06	0.06	-0.08	**-0.21**	**0.49**	0.04	-0.09	**0.18**	0.08	-0.04
30	**0.70**	-0.01	-0.03	0.03	0.07	0.01	**-0.38**	0.07	**0.20**	**-0.18**	0.06	0.04	-0.10
31	**0.85**	0.02	-0.07	0.03	0.01	0.07	**0.12**	0.00	-0.03	0.02	0.07	0.07	**0.16**
32	**0.56**	-0.03	**-0.41**	0.09	0.12	0.10	**-0.28**	**0.14**	-0.01	-0.14	-0.01	0.05	-0.11
33	**0.33**	-0.11	-0.02	0.14	0.02	0.03	**-0.41**	0.16	-0.06	-0.16	0.17	-0.02	0.13
34	**0.28**	-0.01	-0.16	**0.44**	0.01	-0.13	-0.12	0.12	-0.01	0.02	**0.41**	0.01	0.06
35	**0.40**	**-0.28**	**-0.19**	0.06	**-0.16**	**0.29**	0.02	0.00	**0.33**	**-0.20**	**0.25**	0.03	0.05
36	0.06	0.04	-0.08	-0.01	0.05	0.11	0.03	-0.03	0.02	-0.05	**0.77**	**-0.17**	0.09
37	**0.50**	-0.04	0.09	0.09	0.08	0.02	**-0.53**	**0.16**	0.02	-0.03	0.03	**0.17**	-0.05
Mean	**0.41**	-0.01	**-0.12**	**0.08**	**0.03**	**0.09**	**-0.13**	**0.17**	**0.04**	**-0.06**	**0.26**	0.05	-0.01
St.dev	0.21	0.12	0.12	0.11	0.06	0.15	0.21	0.18	0.09	0.09	0.23	0.13	0.09
T test p	0.00	0.79	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.51

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality

Appendix 33
Percentage cue selection on IS task (Study 2)

GP	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	COM	WEI	ATT	FHD	PER
1	100	100	100	96	0	0	100	100	37	10	99	0	0
2	100	1	74	70	0	73	2	67	3	0	99	35	0
3	100	22	100	0	1	100	98	98	98	61	100	52	32
4	100	46	54	52	1	46	46	46	42	5	53	75	13
5	100	100	100	100	0	0	100	100	86	2	100	70	38
6	100	49	100	2	1	42	48	48	4	7	1	43	0
7	100	80	100	100	13	20	90	76	33	63	23	25	2
9	100	42	10	29	0	71	30	44	0	0	38	42	0
11	95	57	100	100	18	51	64	66	29	44	32	41	20
12	100	100	96	0	0	11	67	90	10	0	52	100	0
13	100	6	78	0	1	6	69	57	3	1	9	74	5
14	100	99	100	99	1	1	100	100	93	0	34	98	0
15	100	27	100	100	7	100	100	98	92	92	19	42	5
17	100	100	98	96	30	98	82	100	16	44	93	82	24
19	100	0	100	0	0	0	0	100	100	0	0	100	0
20	62	34	31	0	1	100	34	29	35	26	54	100	0
21	100	50	100	0	0	52	53	59	42	50	99	54	0
22	100	100	100	100	3	94	100	100	16	18	5	100	2
24	100	77	100	100	5	41	100	73	32	34	92	83	2
26	45	61	43	0	0	1	1	34	1	0	100	75	0
28	100	58	99	0	0	40	86	99	0	2	100	16	0
29	100	100	100	98	98	2	100	100	8	9	99	99	4
30	100	73	11	0	0	58	81	44	74	61	39	59	0
31	100	89	85	60	2	64	65	66	5	57	7	59	4
32	100	58	100	0	2	52	72	60	18	45	5	55	3
33	100	1	84	0	0	11	81	1	0	2	25	13	0
34	100	0	85	67	1	1	1	3	40	4	50	28	11
35	100	42	100	2	0	87	12	46	90	74	95	58	0
36	86	14	69	6	3	1	4	2	3	2	99	1	0
37	100	100	100	100	100	100	100	100	100	100	99	100	100
Mean	96.3	56.2	83.9	45.9	9.6	44.1	62.9	66.9	37	27.1	57.3	59.3	8.8
St. dev	12.1	36.0	27.1	45.9	25.2	37.8	36.5	32.2	36.4	30.7	39	31.2	19.9

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality

Appendix 34

Average cue selection positions on IS task (Study 2)

GP	CHO	HYP	AGE	GEN	OCC	ART	SMO	DIA	COM	WEI	ATT	FHD	PER
1	12.99	12.99	10.96	9.63	0.00	0.00	9.05	8.05	2.23	0.70	6.92	0.00	0.00
2	12.63	0.10	7.50	7.43	0.00	6.83	0.18	5.54	0.36	0.00	12.21	2.59	0.00
3	13.00	0.50	9.27	0.00	0.12	0.58	7.58	5.70	0.25	0.06	0.72	7.46	0.39
4	13.00	4.07	5.94	5.20	0.02	2.74	3.64	3.17	2.07	0.18	2.79	8.95	0.36
5	13.00	8.99	11.82	11.12	0.00	0.00	8.01	9.99	4.62	0.07	7.01	4.10	1.67
6	13.00	4.10	11.99	0.12	0.09	3.73	4.79	5.16	0.29	0.50	0.03	3.03	0.00
7	12.99	7.66	11.93	10.93	1.39	1.14	7.93	6.18	1.98	4.60	1.01	1.12	0.05
9	12.99	3.97	0.72	2.15	0.00	8.50	2.35	4.30	0.00	0.00	3.16	4.04	0.00
11	10.34	4.52	11.98	13.00	0.28	3.86	5.81	6.43	1.26	2.55	1.07	2.01	0.43
12	13.00	10.97	9.51	0.00	0.00	0.93	5.38	8.16	0.67	0.00	3.65	11.98	0.00
13	13.00	0.50	9.27	0.00	0.12	0.58	7.58	5.70	0.25	0.06	0.72	7.46	0.39
14	13.00	11.85	11.03	9.92	0.09	0.09	9.00	8.00	6.46	0.00	1.71	6.00	0.00
15	12.97	1.25	11.96	10.99	0.32	9.96	8.95	7.81	5.49	6.44	0.63	2.10	0.05
17	8.82	11.20	8.63	9.06	0.93	10.98	5.94	12.97	0.68	2.14	5.05	4.69	0.69
19	13.00	0.00	11.98	0.00	0.00	0.00	0.00	11.01	10.00	0.00	0.00	9.01	0.00
20	6.85	2.88	2.41	0.00	0.03	12.94	2.72	1.83	2.26	1.27	5.01	11.94	0.00
21	13.00	3.88	12.00	0.00	0.00	2.73	5.56	5.31	2.43	3.36	8.45	5.20	0.00
22	9.99	10.90	12.92	11.86	0.36	8.34	8.13	7.13	0.85	0.92	0.23	6.01	0.08
24	12.97	5.78	11.97	10.98	0.41	2.43	9.94	4.45	1.61	1.57	7.78	6.24	0.10
26	5.13	6.57	4.41	0.00	0.00	0.08	0.06	3.14	0.07	0.00	13.00	8.69	0.00
28	13.00	4.76	11.80	0.00	0.00	2.81	7.84	9.92	0.00	0.12	10.81	1.14	0.00
29	13.00	11.99	11.01	9.80	8.84	0.18	7.97	7.02	0.43	0.38	5.87	4.89	0.14
30	13.00	8.62	0.51	0.00	0.00	6.16	7.81	3.53	5.72	5.06	2.28	5.72	0.00
31	12.77	10.43	9.14	5.73	0.10	5.64	5.44	5.16	0.30	3.66	0.35	3.37	0.10
32	11.75	5.58	12.99	0.00	0.06	4.11	7.82	5.88	0.99	2.95	0.24	3.96	0.11
33	13.00	0.12	10.07	0.00	0.00	0.96	8.92	0.09	0.00	0.17	2.47	1.20	0.00
34	13.00	0.00	10.20	7.32	0.04	0.02	0.05	0.21	3.68	0.27	4.97	2.52	0.82
35	13.00	2.22	12.00	0.19	0.00	8.89	0.49	2.78	8.07	5.89	9.98	4.35	0.00
36	10.68	1.53	7.55	0.61	0.30	0.11	0.39	0.21	0.37	0.21	12.36	0.10	0.00
37	13.00	12.00	11.00	9.82	8.46	7.49	7.18	6.13	5.15	4.16	3.18	2.21	1.21
Mean	12.06	5.66	9.48	4.86	0.73	3.76	5.55	5.7	2.29	1.58	4.46	4.74	0.22
St. dev	1.98	4.37	3.47	5.0	2.17	3.87	3.28	3.11	2.67	2.00	4.11	3.2	0.4

Key:- CHO = cholesterol level, HYP = hypertension, AGE = age, GEN = gender, OCC = occupation, EVA = evidence of arteriosclerosis, SMO = smoking behaviour, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, PER = personality

Appendix 35
Instructions for the Policy Recognition task (Study 3, Chapter 7)

Instructions for feedback task

Using judgements you made previously on the likelihood of prescription for a set of cases we have calculated the relative importance of the cues available on each task. This gives, for example, the bearing 'Hypertension' had, in comparison to 'Gender' or any other cue, on the decision to prescribe lipid lowering drugs. These relative importances are shown in terms of statistical weights known as 'regression coefficients' on a series of graphs that each represent one doctor's policy on this decision making.

At the time of the original task you also gave a rating to each cue to indicate the relative influence you felt each cue had on your decisions. These did not correspond very well with the statistical weights. However, it may be the case that although there are problems in explicitly stating a policy, it can be recognised when displayed amongst a group.

You will be shown a selection of graphs including one representing your own policy on this task. We would like you to select three of these graphs that you feel may represent your policy and put these in order of likelihood. You may arrange the graphs how you wish for this purpose.

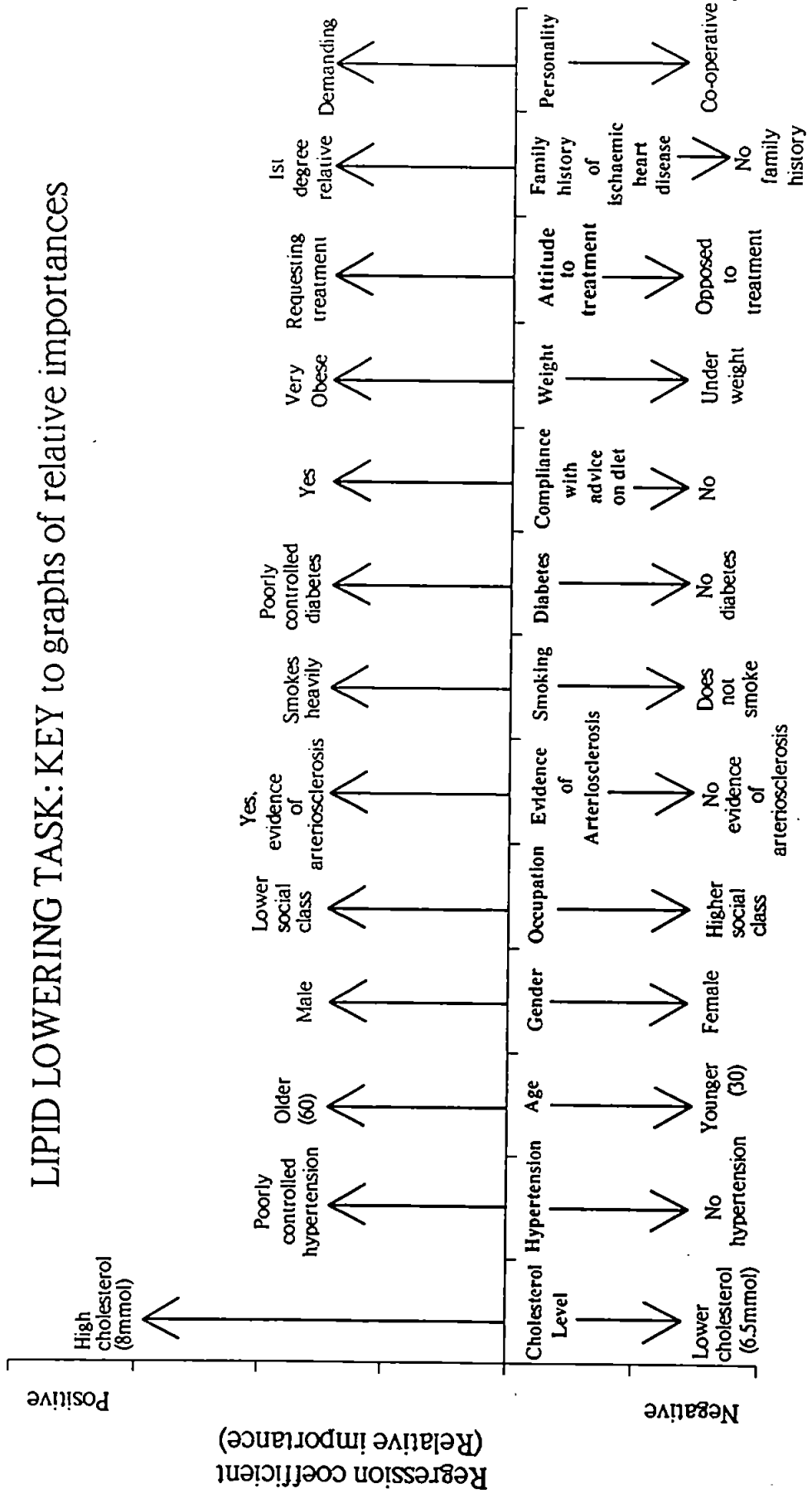
The accompanying key indicates what positive or negative weights for each cue actually represent. For example when the rating is negative for 'Weight' that doctor was more likely to prescribe for underweight rather than overweight people. When the rating for 'Gender' is positive that doctor was more likely to prescribe for males than females. Please pay more attention to darker bars as these show cues that were statistically significant.

There is also a second set of graphs for each computer task for you to look at. These are based on the verbal description of your policy which you gave after you had worked through the cases. Again please pick the graph which you think is yours and choose two reserves.

When you have completed these two tasks for both the Lipid lowering decisions and the Migraine prophylaxis decisions we will give you a copy of both your statistically calculated policy and your stated policy at the time. You will be free to look at the range of policies used by the whole sample of doctors.

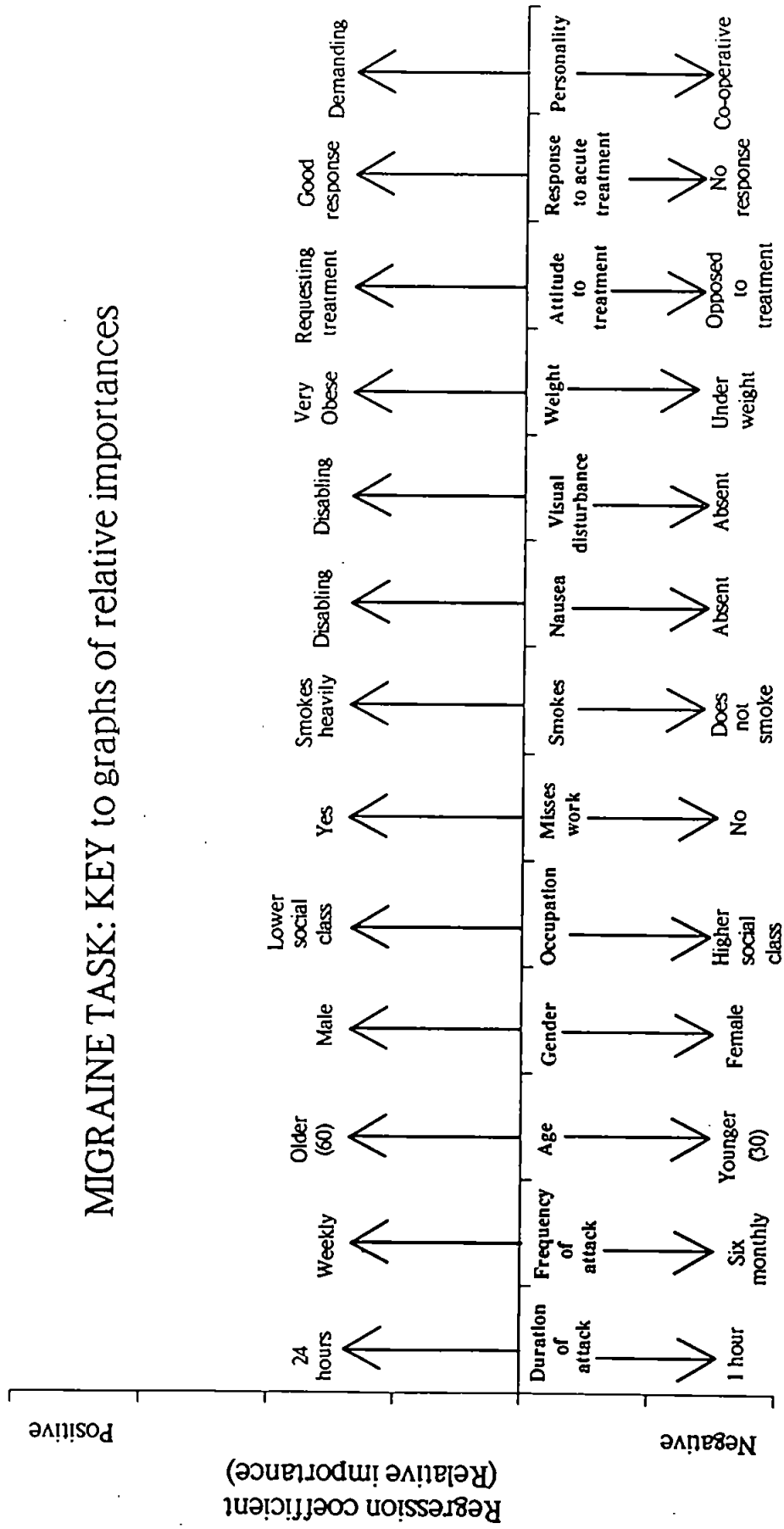
Appendix 36

Key to aid interpretation of LIPID policy bar charts (Study 3)



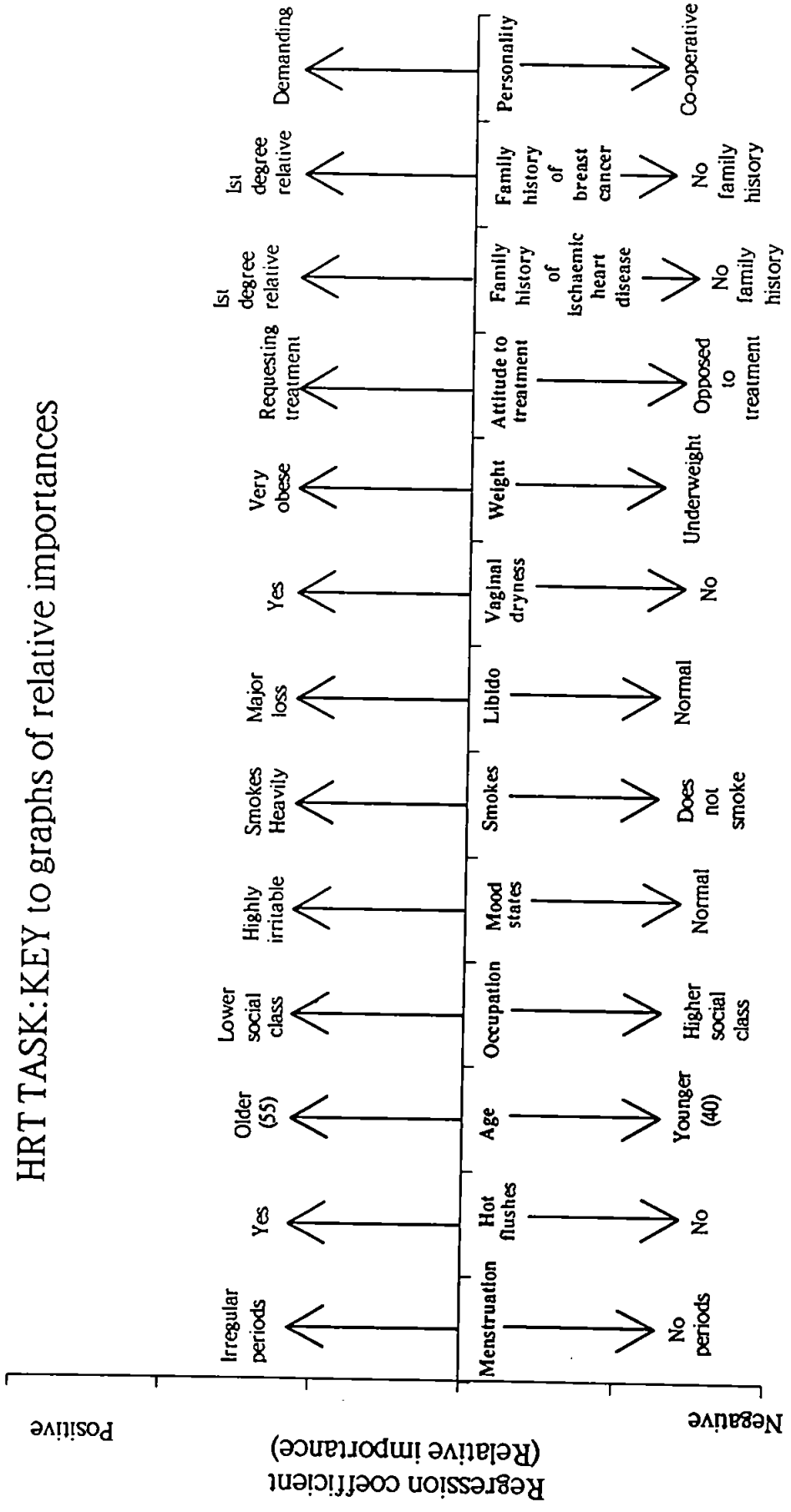
Appendix 37

Key to aid interpretation of MIGRAINE policy bar charts (Study 3)

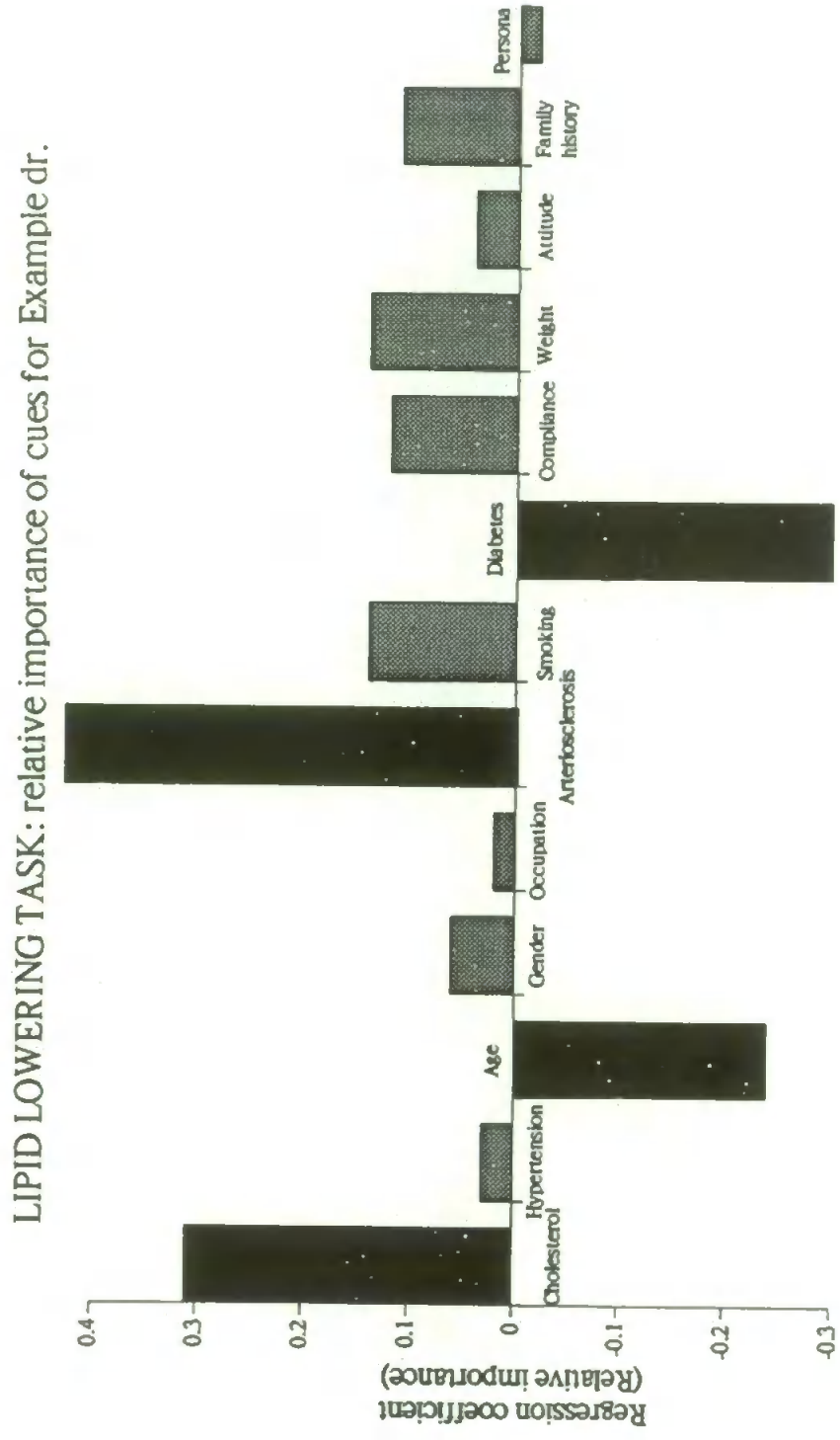


Appendix 38

Key to aid interpretation of HRT policy bar charts (Study 3)



Appendix 39
Example LIPID policy bar chart (Study 3)



Appendix 40
Policy Recognition scores (Study 3)

GP	LIP. T	LIP. E	MIG. T	MIG. E	HRT T	HRT E	GROUP A/B
1	1	3	3	0	*	*	B
2	0	0	0	0	*	*	B
3	2	3	2	3	*	*	B
4	0	0	3	3	*	*	B
5	3	2	0	3	*	*	B
6	0	0	1	0	*	*	B
7	3	2	0	0	*	*	B
9	0	3	1	2	0	1	B
11	1	*	0	2	*	*	A
12	1	2	0	3	*	*	B
13	1	0	0	0	0	0	A
14	0	0	0	3	*	*	A
15	3	2	3	2	*	*	B
17	1	3	0	0	*	*	B
19	3	3	0	0	*	*	A
20	2	3	1	0	*	*	A
21	0	3	0	0	*	*	B
22	0	3	3	0	*	*	B
23	*	*	0	0	*	*	B
24	0	0	*	*	*	*	B
26	0	2	1	0	*	3	B
27	*	0	0	0	0	1	B
28	0	0	0	1	1	2	A
29	0	3	0	0	*	*	B
30	0	0	0	0	0	0	A
31	2	0	1	0	0	0	B
32	2	0	0	3	*	*	B
33	2	2	0	1	0	3	B
34	3	3	*	*	0	0	B
35	3	3	0	0	*	*	B
36	3	0	3	2	*	*	B
37	0	3	2	0	*	*	A

Key:- LIP. T = LIPID TACIT policy
LIP. E = LIPID EXPLICIT policy
MIG. T = MIGRAINE TACIT policy
MIG. E = MIGRAINE EXPLICIT policy
HRT T = HRT TACIT policy
HRTE = HRT EXPLICIT policy

3 = chosen first choice, 2 = 2nd choice, 1 = 3rd choice, 0 = failed to select own policy, * = no data.

Group refers to whether the experimenter may have known the subject's policy (A) or not (B).

Appendix 41
Inter-cue correlation in Study 4, Chapter 8

	Cho	Hyp	Age	Gen	EvA	Dri	Smo	Dia	Com	Wei	Att	FH
Cho	1	-0.05	-0.07	-0.02	0.00	-0.01	-0.08	-0.02	-0.09	0.14	-0.10	-0.05
Hyp	-0.05	1	-0.10	0.10	0.05	0.04	0.03	-0.12	0.10	-0.05	-0.07	-0.08
Age	-0.07	-0.10	1	-0.14	-0.04	-0.03	0.11	-0.17	0.08	-0.07	0.09	-0.10
Gen	-0.02	0.10	-0.14	1	0.02	0.10	-0.11	-0.15	-0.05	0.05	0.12	-0.01
EvA	0.00	0.05	-0.04	0.02	1	-0.08	0.07	0.00	0.07	-0.04	-0.06	-0.04
Dri	-0.01	0.04	-0.03	0.10	-0.08	1	-0.11	-0.07	-0.11	-0.03	-0.10	-0.08
Smo	-0.08	0.03	0.11	-0.11	0.07	-0.11	1	-0.03	-0.05	0.08	-0.10	0.06
Dia	-0.02	-0.12	-0.17	-0.15	0.00	-0.07	-0.03	1	0.03	-0.14	-0.03	0.01
Com	-0.09	0.10	0.08	-0.05	0.07	-0.11	-0.05	0.03	1	0	-0.06	-0.05
Wei	0.14	-0.05	-0.06	0.05	-0.04	-0.02	0.07	-0.14	0	1	0.04	0.00
Att	-0.10	-0.07	0.09	0.12	-0.06	-0.10	-0.10	-0.03	-0.06	0.04	1	0.05
FH	-0.05	-0.08	-0.10	-0.01	-0.04	-0.08	0.06	0.01	-0.05	0.00	0.05	1

Key:- Cho = cholesterol level, Hyp = hypertension, Age = age, Gen = gender, EvA = evidence of arteriosclerosis, Dri = drinking behaviour, Smo = smoking behaviour, Dia = diabetes, Com = compliance with advice on diet, Wei = weight, Att = attitude to treatment, FH = family history of ischaemic heart disease.

Appendix 42
Instructions for the PRESCRIBE task (Study 4, Chapter 8)

You will be presented with a series of 130 hypothetical cases in which you have to decide whether or not to prescribe a lipid lowering agent. Assume that the option to refer is not available. A number of details about the patient and their medical history will be shown. Given all the information, you have to decide whether or not you would now prescribe a lipid lowering drug.

In each case you originally tested the patient's blood cholesterol level at least six months previously. The triglyceride levels on the returned lipid profile were normal. No patient is on a medication which might otherwise affect their lipid levels. You have offered the usual advice on alterations to diet and have recommended, where appropriate, that the patient should give up smoking cigarettes. They already exercise to a degree that is appropriate for their age and general state of health. The blood cholesterol level given in the problem is the current one and reflects any changes in the patient's life-style that he or she has made or is likely to make.

We realize that the cases may not be clear cut and that in practice you might seek further information. However, you can express your decision in terms of the chance of you prescribing the drug. In each case you can indicate this by moving the mouse to the left (less likelihood) or right (greater likelihood of prescribing). When the bar is in the correct position please press any button on the mouse to proceed to the next problem. The first case shown will be an example.

PRESS ANY MOUSE BUTTON TO START

Appendix 43
Instructions for the RISK task (Study 4, Chapter 8)

You will be presented with 130 hypothetical cases. We would like you to make a judgement about the patient's risk of coronary heart disease. A number of details about the patient and their medical history will be shown and certain assumptions can be made about all the patients.

In each case you originally tested the patient's blood cholesterol level at least six months previously. The triglyceride levels on the returned lipid profile were normal. No patient is on a medication which might otherwise affect their lipid levels. You have offered the usual advice on alterations to diet and have recommended, where appropriate, that the patient should give up smoking cigarettes. They already exercise to a degree that is appropriate for their age and general state of health. The blood cholesterol level given in the problem is the current one and reflects any changes in the patients life-style that he or she has made or is likely to make.

To make a numerical estimate of the patient's level of the risk you might wish to have more information. However, here only an approximate estimate is required. Please indicate the risk you feel this patient is at by moving the mouse to the left (less risk) or right (greater risk). Please use points all along the bar to indicate the different risks of different patients. When the bar shows your feeling of the patient's level of risk then please press any button on the mouse to proceed to the next problem.

PRESS ANY MOUSE BUTTON TO START

Appendix 44
Verbal instructions to ascertain self-insight on Study 4

I would like you to be as honest as possible in trying to estimate what affected your judgements on the task you have just completed. All the following questions are trying to ascertain what you feel *did* affect, not what you feel *should have* affected your behaviour, or what might affect it in other circumstances.

Firstly, we will go through each piece of information that would have been available to you on a case. I will give you two descriptions that may have come up for each information. Please indicate which of these would be more likely to make you prescribe by circling it or, if there is no difference between the two descriptions, put a line through both.

After that I would like you to rate the importance of each piece of information on a 0-10 scale. If a piece of information was of maximum importance: its value had a large effect on your judgement please rate it highly. If a piece of information did not have a bearing on your judgements please rate it at 0. Intermediate values should be assigned as appropriate to attributes which you feel have some affect on your decision. You can give the same rating to different items.

Now I'd like you to describe in any other terms the process you were going through in making judgements. Was the importance of any cue dependent on the value of any other cue?

How does this relate to your decision making in real life?

Do you prescribe lipid lowering drugs in real life?

Do you refer patients?

Would you require any other information?

Is there any information that was presented that you wouldn't know?

Appendix 45

Sample sheet seen by GP when rating relative importance of cues in Study 4

Doctor code:- GP_____ **PRESCRIPTION or RISK ASSESSMENT**

Circle description more likely to make you prescribe or put a line through both if there was no difference. Go through all items indicating this. Then go through all items again assigning an importance rating of any number between 0 (no difference) and 10 (very important).

Item	Value 1	Value 2	Rating
Cholesterol level	8	6.5	
Hypertension	Poorly controlled	No	
Age	60	30	
Gender	Female	Male	
Evidence of Arteriosclerosis	Yes	No	
Drinks	Heavily	No	
Smokes	No	Heavily	
Diabetes	No	Poorly controlled	
Compliance with advice on diet	No	Yes	
Weight	Very obese	Under	
Attitude to treatment	Requesting	Opposed	
Family History of IHD	No	1st degree relative	

Appendix 46

Indices on the PRESCRIBE task: linear fit, consistency, mean judgement and mean latency

GP	R^2	r	Mean Judgement	St. dev.	Mean Latency	St. dev.	Correlation Judgement & latency
40	0.64	0.82	44.3	20.93	2024.7	646.2	0.1
41	0.71	0.57	47.67	25.22	1511.3	513	0.3
42	0.40	0.61	62.04	20.21	2372.2	1285.5	0.02
43	0.47	0.49	57.56	23.89	1773.9	873.8	0.18
44	0.46	0.35	28.2	20.73	1946.5	1003.1	0.25
45	0.28	0.05	53.91	14.64	1875.6	616.1	-0.06
46	0.60	0.73	45.75	19.15	1865.9	1163.1	0.26
48	0.48	-0.06	26.9	18.96	1540.9	637.7	0.06
49	0.42	0.69	78.87	13.37	1183.2	417.8	-0.37
50	0.72	0.70	41.32	30.27	652.2	377.9	0.5
51	0.44	0.49	49.34	13.79	1124.1	407.2	-0.12
52	0.34	0.19	7.09	14.63	644.2	335.3	0.43
53	0.57	0.63	44.16	34.37	1440.6	323.7	-0.06
54	0.35	0.73	71.36	12.39	1449.9	944	0.06
55	0.58	0.72	34.12	14.82	2301.4	921.3	0.27
56	0.55	0.42	44.48	22.79	1616.4	778.4	0
57	0.56	0.36	38.72	15.4	1812.8	663.6	-0.21
58	0.43	0.60	36.55	25.68	1462.9	391.1	0.02
59	0.46	0.52	44.55	20.78	1930.1	1059.5	-0.09
60	0.62	0.77	70.87	19.16	2649.3	766.1	0.23
61	0.58	0.68	44.8	27.23	1034	538	0.02
62	0.49	0.44	49.57	22.14	2181.1	1237.3	0.22
63	0.27	0.21	27.39	27.33	1362.3	649.1	0.3
64	0.41	0.50	46.15	19.68	1152	467.3	0.14
65	0.75	-0.10	58.52	41.64	1191.9	593.6	0.16
68	0.69	0.58	53.15	17.31	1815	666.3	0.05
69	0.54	0.54	16.53	23.88	1087	512	0.15
71	0.40	0.68	62.02	16.48	2347.1	1438.5	-0.07
72	0.47	0.38	22.56	19.69	855.9	353	0.05
73	0.50	0.77	46.89	12.79	1463	446.4	0.03
74	0.48	0.67	42.5	18.89	1404.3	525.7	0.6
76	0.63	0.67	53.48	24.27	2736.1	1354.2	0.05
77	0.73	0.77	42.79	21.73	1740.4	746.6	0.11
78	0.77	0.79	61.67	20.65	1704.1	791.7	0.02
79	0.51	0.50	51.63	37.84	1115.1	643	-0.28
80	0.31	0.31	19.98	26.23	2707.9	1595.8	0.24
Mean	0.52	0.52	45.21	21.64	1641.0	741.2	0.10
St Dev	0.13	0.24	15.65	6.75	545.0	337.2	0.20

Key:- R^2 = linear fit; r = consistency

Appendix 47

Indices on the RISK task: linear fit, consistency, mean judgement and mean latency

GP	R^2	r	Mean Judgement	St. dev.	Mean Latency	St. dev.	Correlation Judgement & latency
40	0.61	0.64	79.52	13.3	1224.6	417.7	-0.27
41	0.51	0.76	69.66	19.43	1438.9	575.9	-0.14
42	0.67	0.77	76.35	8.7	1760.9	885.3	-0.03
43	0.62	0.78	74.57	14.12	1164.6	512.3	-0.14
44	0.25	0.22	56.18	11.69	1246.4	716	0.08
45	0.56	0.75	76.23	18.03	1212	528.7	0.19
46	0.51	0.77	62.78	9.36	1079.2	437.3	-0.03
48	0.78	0.77	76.42	20.87	1173.8	379.3	-0.13
49	0.60	0.78	79.61	18.54	1015.4	438.8	-0.19
50	0.52	0.73	57.25	21.01	863.3	307.1	0.03
51	0.56	0.49	56.68	12.66	1283.1	523.4	-0.14
52	0.59	0.57	73.34	22.92	1110.1	390.2	-0.23
53	0.58	0.74	53.83	29.23	1251.5	324.2	-0.09
54	0.46	0.46	65.75	11.17	808	430.1	-0.05
55	0.54	0.48	82.34	11.75	2361.3	1007.3	0.04
56	0.33	0.14	80.47	9.23	2091.8	876.3	0.00
57	0.52	0.50	68.18	14.99	1552.5	712.9	-0.17
58	0.60	0.66	71.65	15.1	1168	433.2	-0.24
59	0.47	0.62	61.93	11.48	1243.3	576.8	-0.25
60	0.51	0.73	86.28	12.96	2101.4	872.1	-0.28
61	0.48	0.56	70.19	20.3	1269.4	780.3	-0.05
62	0.68	0.80	88.62	12.56	908.1	1146	-0.44
63	0.33	0.56	70.45	16.47	1166.8	491.2	-0.02
64	0.54	0.57	67.66	15.85	968.7	419	-0.04
65	0.59	0.70	75.36	28.64	1086.8	653.2	-0.39
68	0.61	0.72	69.63	6.77	1497.2	508.4	-0.04
69	0.54	0.54	45.88	30.05	1060.9	436.5	-0.06
71	0.55	0.42	71.12	9.46	2414.4	2161.7	-0.08
72	0.41	0.55	64.7	12.7	916.5	550.9	-0.12
73	0.55	0.61	59.81	12.51	975.4	340.6	-0.01
74	0.68	0.70	83.72	11.27	888.1	346.6	-0.34
76	0.53	0.72	72.79	11.48	1618.5	660.3	0.12
77	0.53	0.60	73.65	11.26	1422.6	653.6	0.08
78	0.51	0.77	66.59	15.1	1246.7	836.1	0.11
79	0.43	0.63	60.4	23.21	1125	581	-0.45
80	0.45	0.51	43.08	14.83	2037.9	841.6	0.06
Mean	0.53	0.62	69.24	15.5	1326.5	632.0	-0.10
St Dev	0.10	0.15	10.57	5.9	418.2	334.3	0.16

Key:- R^2 = linear fit; r = consistency

Appendix 48
Tacit policies: Standardised Regression Coefficients on the PRESCRIBE task (Study 4)

GP	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH	No. sig
40	**0.55**	0.08	**-0.16**	0.05	**0.28**	**-0.15**	**-0.26**	**0.20**	**0.14**	**-0.26**	**0.28**	0.06	9
41	**0.16**	**0.13**	0.06	**0.63**	0.06	-0.02	0.03	-0.01	0.05	-0.08	**0.41**	**0.28**	5
42	**0.33**	**0.18**	**-0.29**	-0.03	**0.42**	-0.10	-0.04	0.14	-0.02	-0.12	**0.17**	0.15	5
43	**0.22**	0.00	**-0.22**	0.03	**0.36**	-0.02	**-0.21**	**0.41**	0.09	**-0.21**	**0.19**	0.08	7
44	**0.41**	0.14	-0.03	0.03	0.06	0.12	-0.07	0.00	0.14	**-0.48**	**0.38**	0.10	3
45	**0.27**	0.02	-0.10	0.06	0.15	-0.01	-0.10	0.10	**0.37**	**-0.31**	0.13	**0.19**	4
46	**0.46**	**0.20**	**-0.49**	-0.01	0.06	-0.05	-0.06	**0.28**	-0.08	0.01	-0.02	-0.04	4
48	**0.60**	0.05	-0.11	-0.08	0.03	-0.02	**-0.21**	0.08	**0.19**	**-0.21**	0.03	**-0.17**	5
49	0.14	**0.36**	-0.09	0.07	0.09	**0.18**	**0.28**	**0.43**	0.01	-0.09	**0.20**	**0.26**	6
50	**0.85**	-0.02	0.00	-0.01	0.09	**0.11**	0.01	0.05	0.01	0.00	0.08	0.05	2
51	**0.61**	0.05	**-0.23**	-0.01	-0.03	-0.06	-0.12	0.06	0.09	-0.08	**0.23**	0.11	3
52	**0.39**	0.07	0.01	0.07	-0.06	0.15	-0.16	0.04	0.12	-0.11	**0.39**	**0.26**	3
53	**0.24**	0.06	**-0.15**	0.13	**0.21**	0.05	-0.08	0.01	-0.01	0.06	0.02	**0.69**	4
54	**0.29**	0.05	-0.05	0.01	0.02	-0.02	**-0.41**	**0.28**	0.04	-0.04	0.09	**0.28**	4
55	**0.28**	-0.03	-0.07	0.03	0.01	**-0.52**	**-0.23**	-0.05	-0.03	**-0.21**	**0.41**	0.08	5
56	**0.44**	**0.15**	**-0.16**	-0.01	0.07	-0.03	**-0.38**	0.06	-0.06	0.00	**0.42**	0.12	5
57	**0.61**	-0.01	**-0.19**	0.00	0.04	0.02	0.04	0.07	-0.01	-0.08	**0.21**	**0.43**	4
58	**0.39**	0.10	-0.13	-0.04	-0.04	-0.13	-0.12	**0.20**	0.02	0.04	**0.51**	0.05	3
59	**0.63**	**0.21**	0.00	0.08	0.01	-0.06	-0.05	**-0.19**	**0.20**	0.01	0.02	0.14	4
60	**0.49**	0.02	**-0.48**	0.07	**0.26**	-0.04	**-0.19**	-0.03	0.11	**-0.16**	**0.16**	0.09	6
61	**0.69**	-0.03	-0.02	-0.02	-0.09	0.02	-0.07	-0.07	0.01	-0.09	**0.35**	**0.16**	3
62	**0.52**	**0.18**	-0.13	0.03	**0.32**	0.02	-0.09	0.08	0.08	**-0.17**	**0.28**	**0.27**	6
63	**0.19**	0.11	0.08	-0.03	0.12	-0.05	-0.29	0.15	0.07	**-0.20**	**0.35**	0.13	4
64	**0.38**	**0.25**	**-0.31**	0.00	**0.22**	-0.04	-0.02	0.10	-0.01	0.07	**0.33**	0.13	5
65	**0.14**	0.06	**0.17**	**0.22**	-0.04	-0.05	**0.20**	**0.18**	0.04	**-0.14**	**0.75**	**0.20**	8
68	**0.60**	0.10	**-0.12**	-0.04	**0.16**	**-0.14**	0.02	**0.16**	0.05	-0.02	**0.56**	0.09	6
69	**0.71**	0.03	-0.11	-0.02	0.08	0.04	0.02	0.06	-0.01	-0.07	0.00	**0.30**	2
71	**0.40**	-0.01	-0.13	**0.16**	**0.24**	0.06	0.13	**-0.24**	0.15	-0.12	**0.38**	0.00	5
72	**0.18**	-0.02	-0.03	-0.03	0.13	**-0.45**	**-0.19**	0.04	0.10	**-0.29**	**0.37**	-0.12	5
73	**0.31**	0.03	-0.06	0.03	0.04	-0.03	**-0.56**	-0.01	0.11	**-0.26**	0.14	-0.01	3
74	0.07	0.09	-0.11	-0.04	0.05	**-0.21**	**-0.49**	-0.01	**0.29**	**-0.29**	**0.23**	-0.01	5
76	**0.46**	0.10	-0.03	**0.15**	0.11	0.00	-0.04	**0.18**	**0.16**	-0.05	**0.52**	**0.38**	6
77	**0.69**	-0.02	-0.10	0.10	**0.19**	**0.19**	**0.36**	**0.12**	-0.06	-0.08	-0.01	**0.34**	6
78	**0.87**	0.07	-0.09	0.01	0.04	-0.07	0.01	0.09	0.04	-0.01	**0.15**	**0.17**	3
79	**0.62**	**0.24**	0.03	0.00	0.03	0.06	-0.09	**0.35**	-0.12	-0.04	0.04	**0.23**	4
80	**0.40**	0.07	**-0.19**	**0.26**	**0.31**	0.00	0.09	0.12	0.09	-0.03	0.09	0.00	4
Mean	0.43	0.09	-0.11	0.05	0.11	-0.04	-0.09	010	0.07	-0.11	0.25	0.15	4.6
St Dev	0.20	0.09	0.14	0.12	0.12	0.14	0.19	0.14	0.10	0.12	0.18	0.16	1.6

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease. Bold indicates significant ($p < 0.05$) value; No. sig. = the number of significant cues.

Appendix 49

Tacit policies: Standardised Regression Coefficients on the RISK task (Study 4)

GP	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH	No. sig
40	**0.17**	**0.28**	-0.10	0.13	**0.50**	-0.03	**0.18**	**0.48**	0.05	-0.13	-0.01	**0.16**	6
41	**0.19**	0.09	0.14	**0.59**	**0.17**	0.13	**0.25**	0.07	0.04	-0.08	0.01	**0.36**	5
42	**0.14**	**0.33**	**-0.43**	**0.15**	**0.43**	0.06	**0.18**	**0.21**	-0.01	-0.05	0.00	**0.28**	8
43	**0.59**	-0.03	0.04	0.08	0.04	0.03	**0.28**	**0.43**	0.02	-0.07	-0.02	**0.35**	4
44	**0.35**	**0.19**	**0.22**	0.08	-0.04	-0.03	**0.27**	**0.18**	0.08	**-0.27**	-0.03	0.08	6
45	0.09	0.07	**-0.18**	0.03	**0.63**	-0.03	0.06	**0.36**	-0.08	-0.01	-0.07	**0.17**	4
46	**0.22**	**0.24**	**-0.25**	-0.01	-0.04	0.13	**0.43**	**0.34**	0.00	**0.31**	-0.02	-0.01	6
48	**0.22**	**0.26**	-0.04	-0.06	**0.57**	-0.01	**0.46**	**0.30**	-0.02	**0.23**	0.08	**0.10**	7
49	-0.02	**0.15**	-0.13	-0.04	-0.08	**0.23**	**0.64**	**0.40**	0.09	-0.05	-0.02	0.09	4
50	**0.31**	**0.32**	0.01	0.08	0.08	0.09	**0.52**	**0.27**	-0.06	-0.02	0.02	**0.29**	5
51	**0.32**	**0.18**	**-0.31**	0.13	**0.23**	0.15	**0.39**	**0.40**	-0.02	-0.01	0.03	0.12	6
52	0.02	**0.33**	-0.10	**0.41**	0.04	0.10	**0.51**	**0.32**	0.04	**0.16**	-0.08	0.00	5
53	**0.15**	0.10	-0.07	0.14	**0.17**	0.06	**0.41**	-0.02	0.03	0.04	0.13	**0.61**	4
54	**0.40**	0.14	-0.06	0.08	-0.07	0.04	**0.39**	0.15	0.07	-0.07	0.08	**0.46**	3
55	**0.51**	-0.05	-0.04	**0.22**	0.05	**0.28**	**0.34**	**0.25**	0.03	**0.22**	0.00	0.12	6
56	0.10	**0.34**	0.02	0.06	**0.25**	-0.01	0.15	**0.45**	0.07	-0.09	0.03	**0.18**	4
57	**0.21**	**0.16**	0.05	-0.09	0.02	**0.18**	**0.47**	0.10	0.08	**0.33**	0.04	**0.37**	6
58	**0.23**	**0.37**	-0.11	-0.08	0.01	-0.13	**0.46**	**0.46**	-0.06	0.00	-0.04	**0.13**	5
59	**0.26**	**0.30**	**-0.33**	**0.20**	0.05	0.06	**0.42**	0.15	-0.06	-0.02	-0.06	0.14	5
60	**0.20**	0.14	**-0.23**	0.08	**0.48**	-0.11	**0.34**	**0.24**	0.06	0.08	0.04	0.14	5
61	**0.16**	0.13	-0.08	-0.03	**0.27**	0.07	**0.53**	**0.19**	0.14	0.11	0.11	**0.28**	5
62	**0.12**	**0.15**	0.01	0.04	**0.80**	0.05	0.04	**0.18**	-0.03	0.02	0.06	0.05	4
63	**0.27**	0.10	0.03	**0.21**	0.07	0.14	**0.51**	0.07	0.08	0.13	-0.02	0.13	3
64	**0.24**	**0.26**	-0.14	-0.13	-0.01	**0.14**	**0.53**	**0.36**	0.04	-0.05	0.07	**0.27**	6
65	0.07	**0.14**	-0.03	-0.05	0.04	0.09	**0.67**	**0.25**	-0.03	-0.04	0.00	**0.30**	4
68	**0.34**	**0.44**	-0.07	**0.19**	**0.32**	0.10	0.03	**0.38**	-0.06	-0.04	0.06	**0.36**	6
69	**0.26**	**0.19**	0.00	0.05	0.00	0.04	**0.43**	**0.34**	-0.02	0.05	0.04	**0.48**	5
71	**0.38**	0.11	0.00	0.09	**0.34**	0.05	**0.46**	**0.34**	0.02	0.13	0.10	0.09	4
72	0.02	-0.03	0.07	0.02	**0.25**	0.15	**0.59**	0.11	0.07	0.12	-0.06	0.13	2
73	0.02	**0.16**	-0.13	**0.25**	-0.06	0.08	**0.63**	**0.34**	-0.02	**0.18**	0.03	-0.01	5
74	-0.02	**0.15**	-0.09	-0.02	**0.25**	0.03	**0.66**	**0.13**	-0.02	**0.35**	0.03	**-0.13**	6
76	**0.16**	**0.17**	**-0.18**	0.08	**0.39**	0.03	**0.41**	0.13	-0.02	0.03	0.06	**0.42**	6
77	0.11	**0.17**	-0.13	0.11	**0.24**	**0.22**	**0.34**	**0.31**	0.14	0.14	0.10	**0.46**	6
78	**0.49**	**0.17**	0.01	0.08	**0.16**	0.13	**0.48**	**0.31**	0.06	-0.02	0.01	0.04	5
79	0.06	0.08	0.01	**0.20**	-0.12	0.11	**0.28**	**0.60**	0.00	-0.04	-0.04	**0.19**	4
80	**0.24**	**0.29**	-0.04	**0.25**	**0.21**	0.06	**0.26**	**0.42**	-0.06	**0.27**	0.07	0.06	7
Mean	0.21	0.18	-0.07	0.10	0.18	0.07	0.39	0.28	0.02	0.05	0.02	0.20	5.1
St Dev	0.15	0.11	0.13	0.14	0.22	0.09	0.17	0.14	0.06	0.14	0.05	0.17	1.2

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease. Bold indicates significant ($p < 0.05$) value; No. sig. = the number of significant cues.

Appendix 50

Correlations between cue values and decisions about prescription for cases with risk judgement greater than the median risk judgement (for that doctor). If N = 50, significant correlations ($p < 0.05$) are > 0.231 (one-sided).

GP	Cho	Att	Dri	Smo	Wei
40	**0.6**	**0.32**	0.08	**-0.58**	**-0.26**
41	0.16	**0.55**	-0.07	-0.23	-0.09
42	**0.32**	-0.19	0.00	-0.12	-0.06
43	0.18	**0.44**	0.08	-0.23	-0.09
44	**0.39**	**0.25**	0.07	-0.23	**-0.41**
45	0.20	-0.10	0.02	-0.16	**-0.39**
46	**0.40**	0.00	0.05	**-0.37**	-0.03
48	**0.55**	0.01	0.09	**-0.41**	**-0.31**
49	**0.26**	0.23	0.00	0.05	-0.06
50	**0.88**	0.19	0.01	**-0.24**	0.05
51	**0.68**	0.23	-0.02	**-0.49**	0.01
52	**0.27**	**0.36**	0.11	**-0.29**	-0.03
53	**0.24**	-0.05	**0.34**	**-0.41**	-0.13
54	0.20	0.22	-0.09	**-0.43**	0.07
55	**0.30**	**0.35**	**-0.54**	-0.17	-0.1
56	**0.39**	**0.39**	0.03	**-0.54**	**0.25**
57	**0.57**	0.12	0.12	**-0.24**	-0.19
58	**0.43**	**0.50**	-0.12	**-0.32**	0.20
59	**0.68**	-0.05	-0.04	-0.12	0.04
60	**0.55**	0.13	0.10	**-0.38**	-0.11
61	**0.70**	0.10	0.04	-0.15	0.08
62	**0.43**	0.22	0.09	-0.18	-0.16
63	0.10	**0.29**	-0.05	**-0.36**	-0.30
64	**0.46**	0.21	0.08	-0.19	0.18
65	-0.10	**0.80**	-0.22	0.05	-0.06
68	**0.45**	**0.72**	-0.22	-0.07	-0.02
69	**0.73**	-0.07	0.20	-0.20	-0.03
71	**0.35**	**0.27**	0.18	-0.10	0.06
72	0.22	**0.51**	**-0.49**	**-0.25**	-0.11
73	**0.28**	0.04	0.11	**-0.64**	**-0.32**
74	**-0.24**	0.13	-0.04	**-0.51**	-0.05
76	**0.41**	**0.65**	-0.12	-0.18	0.16
77	**0.73**	-0.06	0.19	0.05	-0.19
78	**0.83**	0.05	**0.29**	**-0.29**	0.10
79	**0.73**	0.21	-0.20	-0.14	0.12
80	**0.45**	-0.01	0.13	-0.03	-0.04

Key:- Cho = cholesterol level, Att = attitude to treatment, Dri = drinking behaviour, Smo = smoking behaviour, Wei = weight.

Appendix 51

Explicit policies: Subjective ratings on the PRESCRIBE task (Study 4)

GP	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH
40	8	3.0	-5.0	5.0	8.0	0	*	7.0	6	-5	*	6.0
41	7	8.0	-7.0	7.0	7.0	5	5.0	-7.0	-7	*	10.0	9.0
42	8	8.0	-7.0	5.0	6.0	5	7.0	7.0	6	-6	4.0	7.0
43	10	6.0	4.0	0.0	10.0	-4	-8.0	10.0	8	-6	3.0	10.0
44	7	*	-9.0	7.0	6.0	-9	-7.0	-8.0	9	-9	9.0	8.0
45	10	8.0	-8.0	2.0	8.0	-3	4.0	7.0	7	-7	0.0	10.0
46	8	6.0	-8.0	0.0	8.0	6	6.0	7.0	0	3	3.0	6.0
48	8	6.0	0.0	0.0	5.0	-4	-4.0	6.0	8	-5	2.0	7.0
49	7	7.0	0.0	0.0	7.0	0	7.0	7.0	2	0	2.0	7.0
50	10	8.0	-8.0	4.0	3.0	0	8.0	9.0	0	-1	5.0	7.0
51	9	7.0	-8.0	8.0	6.0	0	-8.0	8.0	6	-6	7.0	8.0
52	10	0.0	-8.0	8.0	5.0	-4	-8.0	4.0	7	-3	7.0	5.0
53	6	0.0	-6.0	7.5	7.0	-4	6.5	5.0	6	-7	1.0	9.0
54	10	5.0	-10.0	0.0	6.0	-6	-10.0	10.0	10	-6	8.0	8.0
55	9	6.0	-5.0	2.0	2.0	-8	-8.0	7.0	0	6	10.0	7.0
56	9	6.0	-8.0	0.0	0.0	0	-7.0	5.0	5	0	7.0	7.0
57	8	5.0	-7.0	0.0	7.0	0	-8.0	6.0	7	-6	7.0	8.0
58	8	6.0	-5.0	6.0	8.0	-4	-6.0	8.0	5	-5	10.0	5.0
59	7	6.0	-7.0	4.0	3.0	0	3.0	*	7	-4	1.0	4.0
60	10	0.0	-8.0	0.0	9.0	0	-3.0	-2.0	5	-6	5.0	7.0
61	8	5.0	-9.0	2.0	0.0	2	3.0	3.0	8	-6	6.0	7.0
62	7	7.0	-7.0	0.0	7.0	0	-3.0	5.0	4	0	6.0	6.0
63	9	7.0	-4.0	5.0	8.0	-7	-9.0	8.0	8	-7	8.0	9.0
64	10	6.0	-3.0	2.0	10.0	0	4.0	7.0	3	-2	2.0	5.0
65	7	7.0	0.0	6.0	3.0	0	10.0	7.0	3	0	10.0	7.0
68	7	7.0	-7.0	6.0	8.0	7	7.0	8.0	-6	7	8.0	8.0
69	10	7.0	-5.0	5.0	8.0	0	8.0	6.0	-8	-5	0.0	8.0
71	10	6.0	-8.0	7.0	9.0	0	0.0	*	0	0	7.0	6.0
72	10	0.0	-7.0	2.0	5.0	-10	-5.0	4.0	7	-7	7.0	4.0
73	6	7.0	-7.0	2.0	0.0	-6	-9.0	6.0	0	-6	1.0	2.0
74	9	7.0	-5.0	0.0	8.0	-8	-10.0	*	8	-8	4.0	7.0
76	8	4.0	-6.0	2.0	7.0	0	0.0	8.0	5	0	9.0	8.0
77	8	7.0	-7.0	6.0	7.0	0	7.0	7.0	0	0	0.0	7.0
78	10	6.5	-7.5	0.0	4.5	0	-1.0	7.5	0	0	8.5	8.5
79	10	7.0	0.0	0.0	3.0	0	0.0	9.5	0	0	3.0	6.5
80	9	7.0	-4.0	2.0	10.0	1	8.0	8.0	2	2	2.0	7.0
Mean	8.53	5.67	-5.74	3.13	6.07	-1.42	-0.59	5.76	3.64	-3.00	5.21	6.97
St. dev.	1.30	2.33	3.12	2.87	2.78	4.09	6.65	4.09	4.49	3.99	3.29	1.68

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease.

Appendix 52

Explicit policies: Subjective ratings on the RISK task (Study 4)

GP	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH
40	8	8	-8	5.0	9.0	4.0	7	9	-0.5	2.0	0	6.0
41	7	6	-5	7.0	7.0	5.0	8	8	0.0	5.0	0	9.0
42	7	8	-7	6.5	9.0	5.0	7	7	-5.0	4.0	0	7.5
43	10	4	2	2.0	2.0	2.0	10	10	0.0	2.0	0	10.0
44	7	8	-9	7.0	8.0	0.0	7	8	-7.0	7.0	0	8.0
45	6	8	-7	-5.0	10.0	6.0	7	9	0.0	5.0	0	8.0
46	6	9	7	8.0	8.0	6.0	9	8	-5.0	6.0	-7	7.0
48	8	6	0	-1.0	10.0	1.0	8	6	-2.0	4.0	0	4.0
49	2	4	0	0.0	0.0	0.0	6	6	0.0	0.0	0	2.0
50	7	7	4	5.0	1.0	0.0	9	7	0.0	0.0	0	3.0
51	5	7	-6	5.0	7.0	3.0	8	8	0.0	2.0	0	4.0
52	3	8	-8	8.0	0.0	6.0	10	9	0.0	6.0	0	3.0
53	6	5	-6	7.0	8.0	-4.0	8	7	0.0	5.0	0	9.0
54	7	6	-8	7.0	6.0	8.0	10	7	0.0	5.0	0	9.0
55	8	8	-4	3.0	2.0	7.0	9	8	0.0	5.0	0	7.0
56	6	7	7	6.0	10.0	6.0	8	8	-5.0	5.0	0	7.0
57	7	7	-7	6.0	7.0	6.0	9	8	0.0	6.0	0	7.0
58	6	8	0	0.0	6.0	3.0	10	8	-3.0	4.0	0	6.0
59	6	7	-3	5.0	7.0	0.0	7	4	-1.0	2.0	0	5.0
60	4	4	-6	3.0	10.0	0.0	7	8	-2.0	4.0	2	8.0
61	3	6	1	3.0	7.0	4.0	10	6	-1.0	4.0	0	7.0
62	4	6	0	0.0	10.0	6.0	8	7	-4.0	4.0	0	5.0
63	6	8	7	6.0	9.0	7.0	9	9	-7.0	6.0	0	8.0
64	2	7	-2	0.0	7.0	4.0	10	10	0.0	3.0	0	5.0
65	7	7	0	3.0	0.0	0.0	10	8	0.0	4.0	0	7.0
68	7	8	-7	6.0	8.0	6.0	7	8	-6.0	6.0	-7	8.0
69	7	8	2	5.0	6.0	2.0	10	10	0.0	0.0	0	9.0
71	6	7	7	7.0	9.0	2.0	9	7	0.0	4.0	0	6.0
72	3	5	3	3.0	10.0	2.0	8	7	0.0	0.0	0	10.0
73	3	7	-3	6.0	1.0	1.0	8	7	1.0	5.0	0	1.0
74	8	8	0	2.0	8.0	2.0	10	8	0.0	8.0	0	7.0
76	9	7	0	5.0	6.0	3.0	9	6	0.0	0.0	0	9.0
77	4	7	-6	0.0	6.0	6.0	8	8	0.0	7.0	0	8.0
78	6	7	4	0.0	9.0	9.0	10	8	0.0	6.0	0	8.0
79	5	0	0	1.5	1.5	1.5	10	10	-1.5	2.5	0	10.0
80	4	6	1	2.0	8.0	2.0	10	9	0.0	3.0	0	8.0
Mean	5.83	6.64	-1.58	3.72	6.46	3.38	8.61	7.81	-1.36	3.93	-0.33	6.82
St Dev	1.96	1.69	4.88	3.08	3.27	2.88	1.22	1.31	2.27	2.17	1.67	2.27

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease.

Appendix 53
Subjective Weights on the PRESCRIBE task (Study 4)

GP	r	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH
40	#0.86	0.45	0.17	-0.28	0.28	0.45	0.00	*	0.40	0.34	-0.28	*	0.34
41	#0.51	0.27	0.31	-0.27	0.27	0.27	0.20	0.20	-0.27	-0.27	*	0.39	0.35
42	0.68	0.39	0.39	-0.34	0.24	0.29	0.24	0.34	0.34	0.29	-0.29	0.19	0.34
43	0.75	0.42	0.25	0.17	0.00	0.42	-0.17	-0.33	0.42	0.33	-0.25	0.13	0.42
44	#0.64	0.27	*	-0.34	0.27	0.23	-0.34	-0.27	-0.31	0.34	-0.34	0.34	0.31
45	0.75	0.44	0.35	-0.35	0.09	0.35	-0.13	0.17	0.30	0.30	-0.30	0.00	0.44
46	0.80	0.41	0.31	-0.41	0.00	0.41	0.31	0.31	0.36	0.00	0.15	0.15	0.31
48	0.64	0.46	0.34	0.00	0.00	0.28	-0.23	-0.23	0.34	0.46	-0.28	0.11	0.40
49	0.70	0.43	0.43	0.00	0.00	0.43	0.00	0.43	0.43	0.12	0.00	0.12	0.43
50	0.38	0.48	0.38	-0.38	0.19	0.14	0.00	0.38	0.43	0.00	-0.05	0.24	0.34
51	0.67	0.36	0.28	-0.32	0.32	0.24	0.00	-0.32	0.32	0.24	-0.24	0.28	0.32
52	0.64	0.45	0.00	-0.36	0.36	0.23	-0.18	-0.36	0.18	0.32	-0.14	0.32	0.23
53	0.51	0.31	0.00	-0.31	0.39	0.36	-0.21	0.34	0.26	0.31	-0.36	0.05	0.47
54	0.79	0.37	0.18	-0.37	0.00	0.22	-0.22	-0.37	0.37	0.37	-0.22	0.29	0.29
55	0.76	0.40	0.27	-0.22	0.09	0.09	-0.36	-0.36	0.31	0.00	0.27	0.44	0.31
56	0.83	0.47	0.31	-0.42	0.00	0.00	0.00	-0.36	0.26	0.26	0.00	0.36	0.36
57	0.61	0.37	0.23	-0.32	0.00	0.32	0.00	-0.37	0.28	0.32	-0.28	0.32	0.37
58	0.68	0.36	0.27	-0.22	0.27	0.36	-0.18	-0.27	0.36	0.22	-0.22	0.44	0.22
59	#0.59	0.46	0.39	-0.46	0.26	0.20	0.00	0.20	*	0.46	-0.26	0.07	0.26
60	0.92	0.53	0.00	-0.43	0.00	0.48	0.00	-0.16	-0.11	0.27	-0.32	0.27	0.37
61	0.51	0.45	0.28	-0.51	0.11	0.00	0.11	0.17	0.17	0.45	-0.34	0.34	0.40
62	0.83	0.40	0.40	-0.40	0.00	0.40	0.00	-0.17	0.29	0.23	0.00	0.34	0.34
63	0.80	0.35	0.27	-0.16	0.19	0.31	-0.27	-0.35	0.31	0.31	-0.27	0.31	0.35
64	0.70	0.56	0.33	-0.17	0.11	0.56	0.00	0.22	0.39	0.17	-0.11	0.11	0.28
65	0.68	0.36	0.36	0.00	0.30	0.15	0.00	0.51	0.36	0.15	0.00	0.51	0.36
68	0.34	0.28	0.28	-0.28	0.24	0.32	0.28	0.28	0.32	-0.24	0.28	0.32	0.32
69	0.59	0.45	0.31	-0.22	0.22	0.36	0.00	0.36	0.27	-0.36	-0.22	0.00	0.36
71	#0.70	0.47	0.28	-0.38	0.33	0.43	0.00	0.00	*	0.00	0.00	0.33	0.28
72	0.84	0.47	0.00	-0.33	0.09	0.24	-0.47	-0.24	0.19	0.33	-0.33	0.33	0.19
73	0.72	0.33	0.38	-0.38	0.11	0.00	-0.33	-0.49	0.33	0.00	-0.33	0.05	0.11
74	#0.87	0.38	0.30	-0.21	0.00	0.34	-0.34	-0.42	*	0.34	-0.34	0.17	0.30
76	0.81	0.42	0.21	-0.31	0.10	0.36	0.00	0.00	0.42	0.26	0.00	0.47	0.42
77	0.64	0.41	0.36	-0.36	0.31	0.36	0.00	0.36	0.36	0.00	0.00	0.00	0.36
78	0.65	0.51	0.33	-0.38	0.00	0.23	0.00	-0.05	0.38	0.00	0.00	0.43	0.43
79	0.92	0.63	0.44	0.00	0.00	0.19	0.00	0.00	0.60	0.00	0.00	0.19	0.41
80	0.67	0.45	0.35	-0.20	0.10	0.49	0.05	0.40	0.40	0.10	0.10	0.10	0.35
Mean	0.69	0.42	0.28	-0.28	0.15	0.29	-0.06	-0.01	0.29	0.18	-0.14	0.24	0.34
St.dev.	0.14	0.08	0.12	0.15	0.13	0.14	0.18	0.31	0.19	0.20	0.18	0.15	0.07

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, r = correlation between standardised regression coefficients and subjective ratings (# = some subjective ratings missing).

Appendix 54
Subjective Weights on the RISK task (Study 4)

GP	r	CHO	HYP	AGE	GEN	EVA	DRI	SMO	DIA	COM	WEI	ATT	FH
40	0.77	0.37	0.37	-0.37	0.23	0.41	0.18	0.32	0.41	-0.02	0.09	0.00	0.27
41	0.40	0.33	0.29	-0.24	0.33	0.33	0.24	0.38	0.38	0.00	0.24	0.00	0.43
42	0.87	0.31	0.35	-0.31	0.29	0.40	0.22	0.31	0.31	-0.22	0.18	0.00	0.33
43	0.90	0.51	0.20	0.10	0.10	0.10	0.10	0.51	0.51	0.00	0.10	0.00	0.51
44	-0.01	0.29	0.33	-0.37	0.29	0.33	0.00	0.29	0.33	-0.29	0.29	0.00	0.33
45	0.69	0.27	0.36	-0.31	-0.22	0.45	0.27	0.31	0.4	0.00	0.22	0.00	0.36
46	0.32	0.26	0.39	0.31	0.35	0.35	0.26	0.39	0.35	-0.22	0.26	-0.31	0.31
48	0.92	0.45	0.33	0.00	-0.06	0.56	0.06	0.45	0.33	-0.11	0.22	0.00	0.22
49	0.83	0.21	0.42	0.00	0.00	0.00	0.00	0.64	0.64	0.00	0.00	0.00	0.21
50	0.84	0.46	0.46	0.26	0.33	0.07	0.00	0.59	0.46	0.00	0.00	0.00	0.20
51	0.94	0.27	0.38	-0.33	0.27	0.38	0.16	0.44	0.44	0.00	0.11	0.00	0.22
52	0.86	0.14	0.38	-0.38	0.38	0.00	0.28	0.47	0.42	0.00	0.28	0.00	0.14
53	0.58	0.28	0.24	-0.28	0.33	0.38	-0.19	0.38	0.33	0.00	0.24	0.00	0.43
54	0.55	0.30	0.26	-0.34	0.30	0.26	0.34	0.43	0.30	0.00	0.22	0.00	0.39
55	0.64	0.41	0.41	-0.20	0.15	0.10	0.36	0.46	0.41	0.00	0.25	0.00	0.36
56	0.40	0.29	0.33	0.33	0.29	0.48	0.29	0.38	0.38	-0.24	0.24	0.00	0.33
57	0.44	0.32	0.32	-0.32	0.27	0.32	0.27	0.41	0.36	0.00	0.27	0.00	0.32
58	0.85	0.33	0.44	0.00	0.00	0.33	0.16	0.55	0.44	-0.16	0.22	0.00	0.33
59	0.87	0.37	0.44	-0.19	0.31	0.44	0.00	0.44	0.25	-0.06	0.12	0.00	0.31
60	0.88	0.21	0.21	-0.31	0.16	0.52	0.00	0.36	0.42	-0.10	0.21	0.10	0.42
61	0.74	0.18	0.35	0.06	0.18	0.41	0.23	0.59	0.35	-0.06	0.23	0.00	0.41
62	0.61	0.22	0.33	0.00	0.00	0.56	0.33	0.45	0.39	-0.22	0.22	0.00	0.28
63	0.32	0.26	0.35	0.31	0.26	0.40	0.31	0.40	0.40	-0.31	0.26	0.00	0.35
64	0.78	0.11	0.38	-0.11	0.00	0.38	0.22	0.55	0.55	0.00	0.16	0.00	0.27
65	0.75	0.40	0.40	0.00	0.17	0.00	0.00	0.57	0.46	0.00	0.23	0.00	0.40
68	0.68	0.28	0.32	-0.28	0.24	0.32	0.24	0.28	0.32	-0.24	0.24	-0.28	0.32
69	0.84	0.35	0.40	0.10	0.25	0.30	0.10	0.51	0.51	0.00	0.00	0.00	0.45
71	0.58	0.31	0.36	0.36	0.36	0.46	0.10	0.46	0.36	0.00	0.20	0.00	0.31
72	0.52	0.17	0.28	0.17	0.17	0.55	0.11	0.44	0.39	0.00	0.00	0.00	0.55
73	0.86	0.20	0.46	-0.20	0.40	0.07	0.07	0.53	0.46	0.07	0.33	0.00	0.07
74	0.62	0.38	0.38	0.00	0.09	0.38	0.09	0.47	0.38	0.00	0.38	0.00	0.33
76	0.81	0.48	0.38	0.00	0.27	0.32	0.16	0.48	0.32	0.00	0.00	0.00	0.48
77	0.84	0.20	0.35	-0.30	0.00	0.30	0.30	0.40	0.40	0.00	0.35	0.00	0.4
78	0.49	0.28	0.33	0.19	0.00	0.42	0.42	0.47	0.38	0.00	0.28	0.00	0.38
79	0.73	0.28	0.00	0.00	0.08	0.08	0.08	0.55	0.55	-0.08	0.14	0.00	0.55
80	0.63	0.22	0.32	0.05	0.11	0.43	0.11	0.54	0.49	0.00	0.16	0.00	0.43
Mean	0.68	0.30	0.34	-0.07	0.19	0.32	0.16	0.45	0.40	-0.06	0.19	-0.01	0.34
St.dev.	0.21	0.10	0.09	0.23	0.15	0.16	0.13	0.09	0.08	0.10	0.10	0.07	0.11

Key:- CHO = Cholesterol, HYP = hypertension, AGE = age, GEN = gender, EVA = evidence of arteriosclerosis, DRI = drinks, SMO = smokes, DIA = diabetes, COM = compliance with advice on diet, WEI = weight, ATT = attitude to treatment, FH = family history of ischaemic heart disease, r = correlation between standardised regression coefficients and subjective ratings.

Appendix 55
Relationship between RISK and PRESCRIBE

GP	RawJ	RawL	Objective	Subjective	Over
40	0.36	0.01	0.48	#0.71	*
41	0.52	-0.16	0.68	#0.48	*
42	0.48	-0.07	0.78	0.58	-0.02
43	0.32	0.02	0.39	0.19	0.39
44	0.22	0.49	0.48	#0.07	*
45	0.08	-0.12	0.24	0.56	0.86
46	0.39	-0.08	0.62	0.23	0.35
48	0.00	0.08	0.00	0.16	0.17
49	0.48	0.23	0.68	0.71	0.24
50	0.3	0.19	0.24	0.57	0.36
51	0.27	-0.07	0.36	0.37	0.29
52	-0.08	0.05	-0.53	0.00	0.29
53	0.6	0.09	0.73	0.69	-0.18
54	0.10	0.06	0.2	0.14	-0.59
55	-0.14	0.08	-0.16	0.11	0.42
56	-0.01	0.26	0.08	-0.34	0.19
57	0.27	0.09	0.28	0.19	-0.11
58	0.20	0.15	0.19	-0.05	0.25
59	0.32	0.10	0.23	#0.58	*
60	0.36	-0.07	0.50	0.39	0.42
61	0.10	0.12	-0.05	0.02	-0.21
62	0.35	-0.08	0.42	0.21	0.06
63	-0.26	0.10	-0.69	-0.26	0.61
64	0.21	0.00	0.29	0.53	0.59
65	0.17	0.03	0.08	0.59	0.73
68	0.36	0.04	0.34	0.74	-0.03
69	0.34	-0.14	0.46	0.87	0.13
71	0.27	0.04	0.22	#0.12	*
72	-0.17	-0.05	-0.42	0.11	-0.06
73	-0.25	0.00	-0.7	0.23	0.31
74	-0.41	0.20	-0.66	#0.01	*
76	0.33	0.00	0.19	0.36	-0.09
77	0.41	0.12	0.44	0.70	0.16
78	0.44	-0.05	0.55	0.13	0.04
79	0.28	-0.02	0.30	0.44	0.19
80	0.53	0.13	0.45	0.80	0.87
Mean	0.22	0.05	0.21	0.33	0.22
St.dev.	0.24	0.13	0.40	0.31	0.32

Key:- RawJ = correlations between a doctors judgements on the PRESCRIBE task and their judgements on the RISK task (last 100 cases); RawL = correlation between latencies on the tasks (last 100 cases); objective = correlation between regression coefficients on the two tasks, subjective = correlation between subjective ratings on the two tasks (# = some subjective ratings missing), Over = correlation between over-rating on the PRESCRIBE task and subjective ratings on the RISK task.

Appendix 56

Summary of Comments made by GPs on non-linear cue use on the PRESCRIBE task

GP40	Heavy smokers who were requesting = no treatment unless other major risk factors But Heavy smokers with other major risk factors = treat Non-smokers, minimal risks, mildly raised cholesterol = no treatment Opposed less likely to treat Requesting more unless other RF they could deal with themselves Then if requesting no treat until dealt with those. E.g. smoking (major), drinkers, smokers, poorly controlled in every way. Favoured young if medium raised cholesterol. Look at sex and age If EvA &/or diabetes then if cholesterol level 7-8 reasonably inclined to prescribe dep. on other RF, e.g. hypertension. Requesting = more inclined unless heavy smokers. Ist rel + EvA + young age + even just moderately raised = persuade that necessary Poorly compliant + Heavy drinkers + opposed = waste of time = Evidence of not having taken previous advice. Look at 1st relative: If no diabetes or EvA no prescription unless c8 and 1st rel. But none here particularly high.
GP41	Hypertension wc & pc as same Age & sex: YM = v likely; OM less likely; OF even less likely. (O = post menopausal age) Weight v. impt if outside BMI range (over and under) more likely to prescribe. Chances of CV event plus likelihood of compliance: Assessment of former based on e.g. hypertension, EvA, smoking,..diabertes,..weight.?If high risk then prescribe. ?If not high risk then look at attitude as well as compliance
GP42	Attitude to treatment only when decision made but strong effect. (if opposed then down a few %)
GP43	Age: 40-50 year olds more likely than 30 or 60 year olds. If going to treat but opposed = less likely.
GP44	Hypertension: pc less likely to prescribe, wc = more likely to prescribe.
GP45	Age and Gender: YF more than YM, OF and OM same. But Y more pres than O. RF: as increase imptce of each increases.
GP46	The greater the age the less the significance of the cholesterol level. Increased imptce of cholesterol level with diabetes.
GP48	Many RF increases importance. ?control e.g. diabetes, smoking, drinking, overweight, hyp. Bands 5.2, 6.5, 7.8. Below 7.8 reluctant. ? Number of RF. ?avoidable RF. Not comp = no pre.
GP49	
GP50	More RF = more likely to prescribe. Ch < 7 never treat
GP51	If not compliant and smoking much less likely. Ch < 7 unlikely. Allow F to have higher levels before prescribing
GP52	If Ch < 8 and F then no treat. 60 year olds no treat
GP53	
GP54	Smoking and Diabetes: If smokers and Not diabetic = no pres. If smokers but diabetic pres.
GP55	
GP56	Attitude: if don't want treatment don't prescribe
GP57	High Ch and EvA = Ch more important
GP58	Multiple RF increased likelohood. But Opposed = vito.
GP59	Diabetes: wc more likely, no and pc equal.
GP60	Age ad Ch.: Given Ch level more impt if person younger. Given age, higher Ch mre impt.
GP61	Age, High Ch and compliance = positive interaction.
GP62	
GP63	
GP64	Ch and requesting increase. Obese and smoker (and drinker) less likely to prescribe. Smoking postive otherwise. Age O trying to ignore some of the RF. e.g. FH
GP65	
GP68	
GP69	Age and gender: YM more likely than YF but OF more likely than YF, OM less likely than YM, OM = OF.

Appendix 56 continued.

GP71	Diabetes: pc = less likely to prescribe, wc = more likely to prescribe. Requesting = increase unless no other factors. PC led to reevaluation of importance of Ch. (less because deal + d)
GP72	Pc indicated as more but "looking to control their diabetes first". Ch < 7.2 unlikely to pres. Ch > 7.2 &/or Dri &/or smokes not prescribe. Ch > 7.2 &/or EvA, &/or FH, &/or diabetic pre.
GP73	
GP74	? less for pc hyp. Moderate and no drinkers no problem. Diabetes and compliance: If pc and not compl then less likely to pre, but if pc and comp with diet then more likely
GP76	FH and Hyp : if FH hyp increase pre. If no FH no difference. Gender and Age: No diffce if Y. If O less for F
GP77	O: could mean factors unimportant that important if Y
GP78	If requesting more likely to give to the non-smoking, than smokers (difference in handling) Diabetes and Comp: If diabetic and not comp more likely (compl N.D.)
GP79	Any RF should have Ch < 7.5. No RF leave alone even if > 7.5. Ch > 7.5 with diabetes pre. I < 7.5 even if have diabetes, leave.
GP80	M more than F in Y, even more than F in O. (didn't treat post-menopausal F). All linked.

Appendix 57

Summary of Comments made by GPs on non-linear cue use on the RISK task

GP40	Ch mildly raised but no FH and requesting less likely. If FH then more likely. Mildly raised Ch + No FH + req = no treat. High Ch v. Medium Ch. + FH = up
GP41	Age and Gender: male 30 stuck, female not. Female 60 less difference.Combination fo FH, ist degr. rel. stablished path critical.
GP42	Not really .. Ev A come in quite high i all other things equal
GP43	YM higher than YF but n.d. OM and OF. J shaped curve of drinking. Muliplicative effects of RF, not just cumulative. But you ran out of space very quickly on your..
GP44	Downrated imptce of Ch if overweight; heavy drinkers and Ch; if heavy drinkers and hyp. less impt; diabetic and overweight = diabetes less important; weight and hypertension: overweight hyp less important.
GP45	Compounding affect of imptce of several factors. Age and Gender: EvA & F more than EvA and M. But n.d if O.
GP46	Multifactorial decision. all factors interact. pc diabetes and pc hyp magnify. Smoking = most significant multifier. Age and diabetes: Y + diabetes more at risk than O + diabetes. and more than O without. But no diabetes O more at risk than Y.
GP48	More RF more compounding:Single not much notice of it. The more they had, the more I'd worry.
GP49	
GP50	More RF = increased risk. OF (over 50) higher risk than YF (nearing M?)
GP51	Smoking and diabetes = increased risk
GP52	hypertensive, diabetic, smoke, summate
GP53	Combination of smoking, Ch, diet, hypetension (poss) much more impt than any alone. Age, sex and fertility status. Can tolerate a highish Ch in F of childbearing age. Risk for F increases with age
GP54	Whole combination, summation of things: Y, bad FH, smoke, drink, raised Ch, overweight
GP55	
GP56	Age and Gender: below 50 M more at risk, after 50 F catch up with men. RF all add up.
GP57	Age and sex: more risk in F over 50. But generally Y age more risk
GP58	Smoking, hypertension, diabetes, (plus Ch important if everything else low risk).
GP59	All inter-relate at once. Hypertension, smoking, diabetes, Ch. If O then discount hyp.
GP60	All interact. I mean I really think that's a funny question. They all interact with each other. Surely that's what I'm suppposed to be doing: weighing up the different um factor - risk factor that each , that each er piece of information that I was given made into the equation?
GP61	Many combinations: BP and smoking (positively), smoking plus any other positive factor e.g. pc diabetes, hypertension; EvA and any other positive factor- BP, pc hypertension, pc diabetes, smoking. Hyp and diab, all pair with each other. (?exacerbation of importance?) Yes, yes.
GP62	
GP63	Age and Gender: postmenopausal F (55+) risk climbs towards M. EvA and Smoking amplified risk.
GP64	Smoking, diabetes, hypertension doubled influence.
GP65	Multifactorial: many things contributing to increased risk. Smoked who had a high cholesterol
GP68	Positive interactions between hypertension, diabetes, raised Ch, FH. More weight on Ch in absence of other RF ? + EvA
GP69	Age and Gender? Take them all independently
GP71	
GP72	Not really but one very abnormal finding doesn't have as much bearing as 3 or 4 lesser abnormal findings
GP73	Diabetes or Hyp weights value of whether they smoke or not. Obviously all multiple RF
GP74	All part of a jigsaw basically
GP76	Sort of:FH & smoking & pc hyp = higher risk, + EvA
GP77	
GP78	EvA, smoking, & drinking more impt, +BP, +diabetes
GP79	
GP80	

General practitioners' tacit and stated policies in the prescription of lipid lowering agents

JONATHAN ST B T EVANS

CLARE HARRIES

IAN DENNIS

JOHN DEAN

SUMMARY

Background. *Research into general practitioners' prescribing behaviour with regard to lipid lowering agents has relied on survey methods which presume that doctors have insight into their prescribing behaviour and can describe it accurately.*

Aim. *This study set out to measure the tacit policies used by general practitioners in prescribing lipid lowering agents and to compare these with their stated policies.*

Method. *Effects of 13 separate cues on decisions to prescribe were examined. The cues included cholesterol levels and a number of associated risk factors for coronary heart disease. Doctors rated 130 imaginary cases presented by a computer. Thirty five general practitioners in the Plymouth area participated in the study. Their ages ranged from 31 to 55 years and all but four were men. The raw data in each case was a rating of the likelihood that the doctor would prescribe for the patient described. These were converted into statistical weightings by use of multiple linear regression. The pattern of (standardized) weights constituted the tacit policy for each doctor. Stated policies were measured in a subsequent interview by asking doctors to rate the influence of each cue.*

Results. *Both tacit and stated policies diverged widely between different doctors. Most doctors overestimated the number of cues that had actually influenced their decisions, and many believed that they had taken into account associated factors for coronary heart disease when they had not. On lifestyle related risks doctors were generally less likely to treat overweight people and most stated this as their policy. Most were also less likely to treat smokers but some had the opposite policy. Those less likely to treat smokers were also less likely to treat obese patients. There was also considerable variation in the extent to which the doctors took account of the attitude of the patient to receiving treatment.*

Conclusion. *Doctors' policies are highly variable and particularly inconsistent in the treatment of smokers. Relevant risk factors may be ignored — even though they are understood — because the risk assessment involved is too psychologically complex a task to be performed intuitively. Decision aids and clear protocols are needed in this area.*

J St B T Evans, BSc, PhD, ABPsS, professor; C Harries, BA, research student; and I Dennis, MA, PhD, principal lecturer, Department of Psychology, University of Plymouth. J Dean, MRCGP, honorary lecturer in general practice, Plymouth Postgraduate Medical School.

Submitted: 25 November 1993; accepted: 25 July 1994.

Keywords: serum cholesterol level; coronary risk factors; drug therapy; medical decision making.

Introduction

THE decision to prescribe lipid lowering agents for patients with raised blood cholesterol levels involves a complex and difficult judgement requiring the doctor to weigh the risk of coronary heart disease against the continuing costs and possible side effects of long-term treatment.[1-4] There is evidence showing that doctors' assessments of cardiac risk may be quite inaccurate.[5] The assessment of risk is complicated by the need to consider a wide range of associated risk factors and the complexity of the medical evidence on the dangers of hyperlipidaemia itself.[1-4] In addition, there are some associated risk factors — particularly cigarette smoking — where the medical risks are well understood, but where there has been some well publicized disagreement in the general media about the policy of treatment which should be followed.

It is important to discover the policies which general practitioners currently follow in the prescription of lipid lowering agents, and the degree to which these policies are sensitive to the relevant risk factors. Previous research into this problem has relied on survey methods[6] which presume that doctors have insight into their prescribing behaviour and can describe it accurately. In contrast, the research described in this paper is based on the methods of social judgement theory[7] which derive tacit policies, in the form of statistical weights, by analysing decisions made over a large number of hypothetical cases in which cues are allowed to vary. Insight can be assessed by interviewing the subjects and asking them to identify the cues which they believe to be influencing their judgements. The method has been used quite widely in medical contexts with consultant groups and with general practitioners in other countries.[8,9] In general this kind of research shows that experts have widely differing policies and rather low levels of insight. Both findings have been demonstrated in a medical context by Kirwan and colleagues' studies of rheumatologists.[10,11]

It is important to explore the application of the methodology of social judgement theory to decision making in British general practitioners, especially in areas such as the treatment of hyperlipidaemia where complex judgements are required. The present study set out to discover both the tacit and stated policies of each member of a sample of British general practitioners in the prescription of lipid lowering agents. The aims were to determine what these policies were, the extent to which different doctors agreed and disagreed with one another, and the degree of insight that individual doctors had into their own decision making, as measured by the correspondence between their tacit and stated policies.

Method

Participating doctors

Thirty five general practitioners practising in or close to the city of Plymouth were recruited as participants in the study, which was conducted between March and June 1993. Recruitment was

achieved by a mailshot to all local doctors followed by personal visits by members of the research team to those who responded. Approximately a quarter of those mailed eventually participated in the study. The doctors varied in age from 31 to 55 years and all but four were men. They were recruited from a variety of different types of practice in both urban and rural settings.

Judgement task

The judgement task was administered via a portable Acorn A4® computer which was custom programmed by J E in BBC *BASIC 5*. Each case was displayed on a separate screen with a series of cue labels on the left (for example, sex, age) and corresponding cue values on the right (for example, male, 45 years of age). The doctors were given on-screen instructions informing them that they would receive 130 cases with information about blood cholesterol levels and a number of other factors. The instructions stated: 'You can assume that in each case you originally tested the patient's blood cholesterol level at least six months previously and have offered the usual advice on alterations to diet and have recommended, where appropriate, that the patient should give up smoking cigarettes. The blood cholesterol level given in the problem is the current one and reflects any changes in the patient's lifestyle that he or she has made or is likely to make.' Doctors were also told by the investigator (C H) to assume that the option of referral to a consultant was not available and thay they had to decide themselves whether or not to prescribe.

The doctors were told that they should indicate the likelihood that they would prescribe in each case by clicking with a mouse on a line between 0% and 100% shown at the bottom of the screen. After clicking, the screen cleared and the next case was presented. Each of the cases which followed provided information on 13 separate cues. These cues and the range of values for each are shown in Appendix 1. Although in real life some of these cues would be correlated with each other, the set of 130 cases which were presented to each doctor was devised by a process of random generation of cue values in such a way as to minimize their intercorrelation (the maximum r between any two cues was less than 0.20). This was done to simplify the interpretation of the weights derived in the subsequent multiple linear regression analysis. The cue values were generated prior to the study by a separate programme with each doctor receiving the same cases in the same order. The occupation cue was coded as social class 1 to 5 but displayed as an example occupation randomly generated from a set of alternatives.

The need for independence in the cues led to the decision to use total cholesterol levels without separate information on the balance between low density and high density lipoproteins which would normally be inspected prior to prescribing lipid lowering agents. The justifications for this were that levels of low density lipoproteins are highly correlated with total cholesterol levels and that doctors are in any case asked only to judge the probability that they would prescribe. Hence, the extra information which they would achieve from lipid profiles is one of the uncertainty factors which their judgements could reflect. No doctors objected to performing the task with the data presented.

Assessment of tacit policies

An initial check for consistency among the doctors was made by computing the Kendall coefficient of concordance (W) across the 130 raw judgements of each doctor. This statistic is a non-parametric test of multiple correlation and can range from 0 (no agreement) to 1 (perfect agreement).

Each doctor's judgements were then entered individually into a multiple linear regression analysis. This is an extension of the familiar regression analysis in which a model is derived in which one dependent variable (in this case the judgement made) is predicted as a function of a number of independent variables (in this case the 13 cues). The analysis results in a set of weights, each indicating the degree of influence of a cue. Where the cues are intercorrelated stepwise linear regression is often preferred. This was avoided here because such intercorrelations had been minimized and the analysis used facilitates comparison between doctors in their usage of cues. The 13 regression weights derived were standardized so as to remove the influence of differences in the original scale values. The weights could thus vary from −1 (maximum negative use of a cue) through 0 (no use of a cue) to +1 (maximum positive use of a cue, that is to the exclusion of all others).

The regression weights were also intercorrelated between cues and across doctors in order to see whether doctors who took account of a given cue were more or less likely to be influenced by another cue.

Assessment of stated policies

On completion of the judgement task the doctors were shown a list of the cues used. For each cue in turn C H described the two endpoints and asked the doctors which would make them more likely to prescribe or if the cue would make no difference. They were then asked to rate each cue to indicate how much of a bearing it had on their decisions from 0 indicating that it had no weight to 10 indicating that it had maximum weight. These weights were assigned a positive or negative sign by C H according to the previous indication of the direction in which the cue endpoints affected judgements. The sign of the subjective decision weights (as for the objective ones) is as indicated in Appendix 1. For example, a positive weighting for 'smoker' means that doctors were more likely to prescribe to a smoker whereas a negative weight means they were less likely to prescribe to a smoker and so on.

The stated policies were elicited by directly asking the doctors which cues they had used on the task just performed. Hence, the degree of correspondence with the objective regression weights is a direct measure of insight or self-awareness on the task.

Doctors were also invited to add any further comments about the task, their reasons for using or ignoring cues and about their general approach to the treatment of hyperlipidaemia.

Results

Objective decision weights (tacit policy)

The consistency check on raw judgements yielded a Kendall W of 0.25 which though significantly non-zero ($P<0.001$) indicates a substantial degree of variation between the doctors in the priority they assigned to treating the different cases presented.

Of the 35 doctors participating 33 provided usable data on the computer presented judgement task, and the results of the regression analyses are shown in Table 1. For each cue the number of doctors who had statistically significant weights (at $P<0.05$) in a positive or negative direction is shown and also the mean weight for all 33 doctors. The cholesterol level was the biggest single predictor of decisions to prescribe lipid lowering agents with a mean weight of 0.38 and a significantly positive weight for 31 of the 33 respondents. There was a considerable variation across the sample in the extent to which other cues were used (Table 1). For example, nine doctors were significantly less likely to prescribe to older patients but the majority were not influenced by the age cue.

Of the associated risk factors for coronary heart disease, only the presence of diabetes (mean weight 0.15) had a substantial effect on decision to prescribe and even here fewer than two thirds of the sample had a significantly positive weight (Table 1). Although the mean weights were positive for hypertension, evid-

Table 1. Objective (regression) and subjective weights given to cues.

Cue	Objective weights: Number of doctors* Significant negative direction	Not significant	Significant positive direction	Mean	Subjective weights: Number of doctors Negative weight	Zero weight	Positive weight	Median
Cholesterol level	0	2	31	0.38	0	0	34	9.0
Hypertension	1	25	7	0.06	5	2	27	6.0
Age	9	23	1	-0.07	26	8	0	-6.5
Sex	0	28	5	0.05	0	9	25	2.5
Occupation	0	30	3	0.03	7	24	3	0
Evidence of arteriosclerosis	0	27	6	0.09	0	6	28	6.0
Smoker	11	19	3	-0.09	19	4	11	-3.5
Diabetes	2	11	20	0.15	3	3	28	7.0
Compliance with advice on diet	2	25	6	0.03	1	11	22	5.0
Weight	8	25	0	-0.07	21	8	5	-3.5
Attitude to treatment	0	17	16	0.17	0	7	27	6.0
Family history of coronary heart disease	0	26	7	0.06	0	3	31	7.0
Personality	0	31	2	0	7	20	7	0

*Significance levels are computed at 5% two tailed.

ence of arteriosclerosis, family history of coronary heart disease and sex, they were low and each of these cues exerted significant positive influence on decisions for at most seven doctors in the sample.

There were several significant correlations between cues of which the most interesting was a correlation of $r = 0.334$ ($P<0.05$) between smoking and weight. In general, this means that the same minority of doctors who would be less likely to prescribe to smokers would also be less likely to prescribe to overweight patients.

Subjective decision weights (stated policy) and insight

The results of the post-task interview for 34 doctors are also shown in Table 1. These results should be treated as the stated policies of the doctors, that is what they believe to be the cues which determined their decisions. The data are broken down by frequency of negative, zero and positive weights with the median weighting for all 34 doctors also being shown. As with the objective weights, cholesterol level was the most important cue.

Figure 1 presents a scatter chart in which each point represents the subjective and objective weights given by a particular doctor to a particular cue (the scores of all 33 doctors are included). The result is a triangular pattern for both positively and negatively weighted cues. Many cues have both subjective and objective weights near zero, resulting in many points near the origin. As subjective weights rise towards maximum (+10 or −10), however, the distribution of objective weights becomes increasingly spread out. In simple terms, this means that if the doctors say they are not using a cue, they are likely not to be using it. If they say they are using a cue, they may or may not be using it.

Inspection of Table 1 shows that the pattern of insight is different for different cues. For example, while both evidence of arteriosclerosis and family history of coronary heart disease received high subjective weightings, they significantly influenced fewer than a quarter of the doctors' tacit policies. On the other hand, diabetes which received a similar subjective weighting influenced a substantial majority of the tacit policies. Thus, although this cue was much more influential in judgements, doctors appear to lack awareness that this is the case.

Comments of doctors

Table 1 shows that 19 doctors said they were less likely to treat smokers and 11 had significantly negative weightings in the regression analyses. Since smoking is a positive risk factor for coronary heart disease the comments offered in justification of these policies are of interest. The central theme appeared to be that they considered smoking to be a much more serious risk factor than hyperlipidaemia. This led some to say that it was a waste of money and effort to treat the lipid problems because this would be outweighed by the smoking. However, a number of doctors saw the hyperlipidaemia as an opportunity to control the patient's smoking by withholding treatment until their behaviour had been altered. Another commonly expressed view amounted to 'why help someone who is unwilling to help themselves?' On the basis of the reports it also appears that some doctors thought that an acute ethical dilemma was involved in the treatment of smokers whereas others apparently did not.

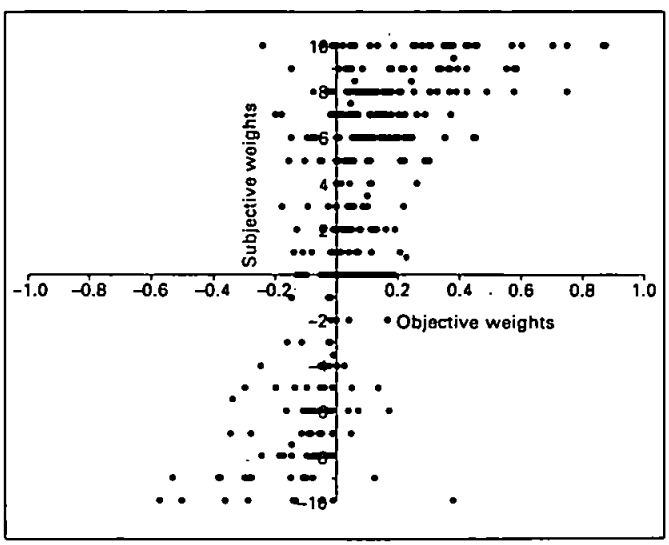


Figure 1. *A plot of subjective against objective weights for each of the 13 cues, for each of the 33 doctors.*

Doctor's comments on the use of protocols or decision aids for treatment of raised blood cholesterol are also of interest. The great majority commented either that they had no such protocol or that they did not follow the ones that were in their possession.

Discussion

This investigation of general practitioners' decisions to prescribe lipid lowering agents using the social judgement theory approach has produced findings broadly compatible with previous research using this methodology.[8] Wide variations between doctors in both their tacit and stated policies for prescribing and a fairly low level of insight have been found. In general doctors believed they used many more cues than they actually did. However, this lack of insight does not account for the policy differences. Even in stated policies, there was considerable variation between doctors in their beliefs about how information should be used with regard to prescribing.

For those associated risk factors for coronary heart disease which are not directly within the control of the patient, namely hypertension, evidence of arteriosclerosis, sex, diabetes and family history of coronary heart disease, while a substantial majority of doctors stated that they took these factors into account, with the exception of diabetes fewer than a quarter of the sample had significant objective weights for any of these cues. This disparity between tacit and stated policies is of considerable importance. A plausible interpretation is that while doctors know that these risk factors are medically relevant, the task of weighing all these factors when making their decisions to prescribe is simply too psychologically complex and defeats them. It is known from psychological research that people, including experts, can only consider a limited number of cues in making judgements[12] and that people have great difficulty in forming accurate assessments of risks and probabilities.[13]

Two associated risk factors which are lifestyle related are smoking and obesity. Both tacit and stated policies were markedly different for these factors compared with those considered above. No doctor was more likely to prescribe to an overweight patient and eight had significantly negative weightings, meaning that they were less likely to prescribe to an overweight patient who was equal on other risk factors (21 doctors also stated that this was their policy). The justification for this might be that the patient could lower his or her cholesterol level by losing weight, without the need for drug treatment. The case of cigarette smoking is more interesting, however, since smoking raises associated risk but does not affect the blood cholesterol level directly. Since smoking is seen as a choice of the patient, the doctor might decide not to prescribe lipid lowering agents simply because the smoking behaviour in itself takes the patient over some threshold level of risk. This would result in smoking being effectively ignored as a risk factor and given a zero weighting. In fact, three doctors had a significantly positive weighting for smoking, treating it as risk factor which increased the argument for drug treatment, 19 doctors ignored smoking as a factor and 11 had a significantly negative weighting. Subjectively, 19 doctors said they would be less likely to treat smokers, although 11 said they would be more likely to do so.

The suggestion of a moral dimension in doctors' attitudes is supported by the finding of a significant correlation between the doctors who were less likely to prescribe to smokers and those who were less likely to prescribe to patients who were overweight. Doctors therefore differed with respect to their treatment of self-inflicted risks. A variety of explanations were offered in the post-task interview by those doctors who penalized smokers. The other indication of difference between doctors on non-medical grounds was the way they responded to the 'attitude to treatment' cue. The sample was split in half here — 16 were more likely to prescribe to patients who were requesting the treatment while 17 ignored the wishes of the patient. The occupation and personality cues were largely ignored.

These findings have serious implications for medical practice in this area and the training of general practitioners. First, it is hard to reconcile both the wide variation in policy and the general neglect of associated risk factors that have been observed with a generally effective treatment of hyperlipidaemia by doctors. Secondly, it has been shown that while education by normal verbal communication (for example, lectures, textbooks) may increase doctors' awareness of relevant factors this may not necessarily impact on their actual prescribing decisions. There is a strong case for implementation of well defined clinical guidelines, preferably backed by risk assessment aids.

Appendix 1. Cues presented to doctors in order shown.

Cue	*Range of values (negative to positive)*
Cholesterol level	6.5–8.0 mmol l^{-1} (step 0.1mmol l^{-1})
Hypertension	No/yes, well controlled/yes, poorly controlled
Age	30–60 years (step 1 year)
Sex	Female/male
Occupation	Social class 1–5
Evidence of arteriosclerosis	No/yes
Smoker	No/occasional/regular/heavy
Diabetes	No/yes, well controlled/yes, poorly controlled
Compliance with advice on diet	No/some/yes
Weight	Under/normal/over/obese/very obese
Attitude to treatment	Opposed/cautious/open to advice/requesting treatment
Family history of coronary heart disease	No/second degree relative/first degree relative
Personality	Cooperative/passive/demanding

References

1. Marmot M. The cholesterol papers [editorial]. *BMJ* 1994; 308: 351-352.
2. Law MR, Wald NJ, Wu T, *et al.* Systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. *BMJ* 1994; 308: 363-366.
3. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ* 1994; 308: 367-373.
4. Law MR, Thompson SG, Wald NJ. Assessing possible hazards of reducing serum cholesterol. *BMJ* 1994; 308: 373-379.
5. Tape TG, Wigton RS. Medical students' and residents' estimates of cardiac risk. *Med Decis Making* 1987; 9: 170-175.
6. GP opinion: cholesterol survey results. *MIMS Magazine* 1993; February: suppl.
7. Brehmer B, Joyce CRB (eds). *Human judgement: the SJT view.* Amsterdam, Netherlands: Elsevier, 1988.
8. Wigton RS. Applications of judgement analysis and cognitive feedback to medicine. In: Brehmer B, Joyce CRB (eds). *Human judgement: the SJT view.* Amsterdam, Netherlands: Elsevier, 1988.
9. Engel JD, Wigton R, LaDuca A, Blacklow RS. A social judgement perspective on clinical psychology. *Evaluation Health Professions* 1990; 13: 63-67.
10. Kirwan JR, de Saintonge DMC, Joyce CRB, Currey HLF. Clinical judgement and rheumatoid arthritis. II. Judging the 'current disease activity' in clinical practice. *Ann Rheum Dis* 1983; 45: 648-651.
11. Kirwan JR, de Saintonge DMC, Joyce CRB, *et al.* Inability of rheumatologists to describe their true policies for assessing rheumatoid arthritis. *Ann Rheum Dis* 1986; 45: 151-161.
12. Shanteau J. How much information does an expert use? Is it relevant? *Acta Psychol (Amst)* 1992; 81: 75-86.
13. Kahneman D, Slovic P, Tversky A (eds). *Judgement under uncertainty: heuristics and biases.* Cambridge University Press, 1982.

Address for correspondence

Professor J St B T Evans, Department of Psychology, University of Plymouth, Drake Circus, Plymouth PL4 8AA.