Learning to read mammograms: Complex skill acquisition, training and individual differences

Regina Pauli

Thesis submitted in fulfilment of the requirements for the award of Doctor of Philosophy

Department of Psychology
University of Surrey
1993

© Regina Pauli, 1993, all rights reserved
Abstract

The interpretation of mammograms is a complex perceptual cognitive skill involving the visual and interpretative evaluation of breast X-rays. In the context of the National Health Service Breast Cancer Screening Programme, reading mammograms entails distinguishing between large numbers of normal and few abnormal films, and making decisions with regard to further clinical investigation. This thesis is concerned with lay people learning to interpret screening mammograms. Three issues are considered in this context, the nature of mammogram interpretation skill, issues in training and individual differences.

Previous research on radiological skill has involved the study of visual search, on the one hand, and analyses of cognitive expertise in radiology, on the other. In this thesis, it is argued that skill in screening mammogram interpretation involves both, visual analysis and interpretation of detected abnormality. This notion is formalised in a cognitive process model of mammogram interpretation, which incorporates two stages of processing. The model is supported by data from two skill acquisition studies, one longitudinal study of radiographers and one involving students who have received brief training. Performance outcome and error data from both studies confirm the two stage nature of mammogram interpretation skill.

Issues in selection and training considered in the thesis reflect the applied concerns surrounding lay film interpretation in screening mammography. Individual differences in mammogram reading performance were investigated in terms of ability, personality and cognitive skills measures. Training was examined in terms of the effect of different instructional methods on interpretation performance. It was found that a feature oriented method of training lay people for screening mammogram interpretation has some advantages over more traditional training methods. Individual differences in film reading performance after training were best predicted by cognitive style and strategy variables.
For my Mother and Father
Acknowledgements

During the four years of work on this thesis, many people have contributed to its progress in different ways. Some have aided the development of my thinking about theoretical issues, others have contributed to the success of the practical research, and yet others have provided criticism, advice or encouragement.

I would particularly like to thank everyone at the Jarvis Breast Screening Centre in Guildford for supporting this work. Dr Barbara Thomas and Dr Julie Cooke have contributed to the success of this thesis in innumerable ways, not least by making possible the study involving radiographers and by providing access to training materials for the studies involving students, the collation of which they also supported financially. In addition, The Test Agency in High Wycombe provided some of the psychometric tests free of charge, which reduced the financial constraints on this research considerably.

Janet Ansell's continued enthusiasm for lay film reading has ensured that this work is of relevance to breast cancer screening, and I would also like to thank her for her practical support, which has benefitted many aspects of this research. The two radiographers, A.S. and F.A., were a pleasure to work with, always willing to try the impossible and very enthusiastic about reading mammograms. Thanks should also go to all other participants in my research, particularly to those students who willingly gave their time to take part in the training studies.

I would like to thank my supervisors, Sean Hammond and Ian Davies. Ian gave me the opportunity for doing this research in the first place, and was invaluable in his critical reading of this thesis. Sean's supervision was characterised by great skill and sensitivity, in particular in my numerous crises of confidence. In addition, discussions with Peter Simpson and Paul Sowden have significantly contributed to my understanding of theoretical issues concerning mammogram interpretation.
My family and friends have been very supportive throughout. I would particularly like to thank my parents, who have supported me emotionally and financially through difficult periods. My friends have been invaluable sources of advice, friendship and support, and I would like to mention particularly Geoff Hone, Jörg Huber, Rukhsana Kausar, Glynis Laws, Yun Low, Lynne Martin and Clare Twigger. The person, who undoubtedly contributed most significantly to the completion of this thesis, is Stephen Jones, my friend and partner. On a practical level, he wrote the computer programs for the cognitive test battery and supervised the technical aspects of producing this document. On a personal level, he handled me with great patience and consideration, focusing his energy on helping me to complete this thesis.
Table of Contents

1 Introduction: Issues in learning to read mammograms .................. 1
   1.1 Breast cancer screening .................................................. 2
       1.1.1 Problems in breast screening ............................... 2
       1.1.2 Solutions to problems .............................................. 4
   1.2 Perception and interpretation of medical images .................. 5
       1.2.1 Perception and cognition ....................................... 6
       1.2.2 Image Processing ................................................. 7
       1.2.3 Interpretation of X-rays in the medical context ......... 8
       1.2.4 Perception and interpretation in lay film reading ...... 9
   1.3 Selection and training of lay film readers .......................... 10
       1.3.1 Training ............................................................... 11
       1.3.2 Selection ............................................................... 12
   1.4 Layout of the thesis ...................................................... 13
       1.4.1 Part 1: Acquisition of mammogram interpretation skill .... 14
       1.4.2 Part 2: Developing a model of mammogram interpretation skill ........ 15
       1.4.3 Part 3: Prediction of film reading performance ........... 16

Part 1 Learning to interpret X-rays: The case of screening mammography .... 17

2 Radiological Skill and its Acquisition ........................................ 18
   2.1 Radiological skill and visual search ................................ 19
       2.1.1 Experimental studies of radiological skill ............... 19
       2.1.2 Models of visual search in radiology ................. 25
   2.2 Radiological Expertise ............................................... 28
       2.2.1 Empirical investigations of radiological expertise ...... 29
       2.2.2 Models of radiological expertise .......................... 34
   2.3 Acquisition of radiological skill ................................. 36
       2.3.1 Visual search and skill acquisition ....................... 36
       2.3.2 Acquiring radiological expertise.......................... 37
   2.4 Summary ................................................................. 39

3 Radiological skill in mammography ......................................... 41
   3.1 Mammography ......................................................... 41
       3.1.1 Reading screening mammograms .................................. 45
       3.1.2 Measuring observer performance in screening mammography .... 48
       3.1.3 Visual Search in Screening Mammography .................. 51
       3.1.4 Cognitive Expertise and Screening Mammography .......... 54
   3.2 Learning to read mammograms ....................................... 56
       3.2.1 Current training practice for mammography screeners .... 56
       3.2.2 Radiologists or radiographers? ............................... 58
   3.3 Summary ................................................................. 61
4 Acquiring radiological skill: A case study of two radiographers ...
   4.1 Introduction ................................................................. 63
      4.1.1 Training .................................................................. 64
      4.1.2 Mammogram interpretation skills .............................. 66
   4.2 Method ........................................................................... 68
      4.2.1 Mammography Training ........................................... 68
      4.2.2 Performance assessment .......................................... 69
      4.2.3 Triple-blind screening film reading protocol ............... 72
   4.3 Results .......................................................................... 73
      4.3.1 Training and mammogram interpretation test performance .. 74
      4.3.2 Error analysis .......................................................... 76
      4.3.3 Performance in mammography screening ...................... 79
      4.3.4 The effect of dual reading on screening performance ........ 81
      4.3.5 Verbal protocols ....................................................... 83
   4.4 Discussion ....................................................................... 90
      4.4.1 Training radiographers for mammogram interpretation .... 90
      4.4.2 Pre-screening or 'second pair of eyes'? ......................... 92
      4.4.3 Skills implied in screening mammography .................... 94
      4.4.4 Changes in mammogram interpretation after training ...... 95
      4.4.5 Individual differences after training ............................ 96
      4.4.6 Effects of dual reading .............................................. 98
   4.5 Conclusion ...................................................................... 98

Part 2 Modelling mammogram interpretation skill: The influence of training . 101

5 Further issues: Mammography skills, skill acquisition and training .... 102
   5.1 What is learned during skill acquisition in mammogram interpretation? 103
      5.1.1 Visual discrimination and pattern/object recognition ........ 104
      5.1.2 Categorisation of visual and disease feature .................. 108
      5.1.3 Medical decision making with probabilistic features .......... 114
      5.1.4 Summary ............................................................... 116
   5.2 Radiology and Mammography Education ............................... 117
      5.2.1 Initial Radiology and Mammography Training ................. 118
      5.2.2 Maintaining Radiology Skills ...................................... 119
   5.3 What mechanisms are involved in learning mammography? .............. 120
      5.3.1 Perceptual Learning .................................................. 121
      5.3.2 Conceptual learning .................................................. 126
      5.3.3 The role of training and instruction ............................... 129

6 Perceptual Learning or Clinical Reasoning: An Evaluation of two Training Methods for Lay Film Readers ....................... 132
   6.1 Comparing feature training and medical reasoning training .......... 133
      6.1.1 Training lay film readers .......................................... 133
      6.1.2 Comparing two training methods ............................... 134
      6.1.3 The training study ................................................... 136
      6.1.4 Results ................................................................. 143
6.1.5 Discussion: Is feature training more effective than decision training? .......................... 149
6.2 Feature training and individual differences in film reading performance .......................................................... 151
6.2.1 A further training study ........................................................................................................ 152
6.2.2 Results .................................................................................................................................. 153
6.2.3 Discussion: Homogenising effects of feature training ......................................................... 155
6.3 General discussion: Learning to read mammograms with brief instruction .................................................. 156
6.3.1 Wider issues in brief training .............................................................................................. 157
6.3.2 Learning from nothing? ....................................................................................................... 159
6.3.3 Summary .............................................................................................................................. 160

7 A model of screening mammogram interpretation ............................................................... 161
7.1 Modelling cognitive processes in mammogram interpretation .............................................. 162
7.1.1 A task model of mammography screening film interpretation ........................................ 163
7.1.2 Cognitive processes in mammogram interpretation .......................................................... 166
7.2 Validating the process model of mammogram interpretation ........................................... 168
7.2.1 Error analysis .................................................................................................................... 169
7.2.2 Validating the task model in terms of error and training data .......................................... 173
7.2.3 Comparison to previous models of radiological skill .......................................................... 177
7.2.4 Conclusion .......................................................................................................................... 180

Part 3 Individual differences in mammogram interpretation: Issues in selection and performance prediction ............................................................. 181

8 Identifying Sources of Individual Differences in Mammogram Interpretation Skill ................................................................. 182
8.1 Applications of individual differences in screening performance .............................................. 183
8.1.1 Training Design and Individual Differences ....................................................................... 184
8.1.2 Selection and Aptitude for Mammogram Interpretation ...................................................... 185
8.2 Individual differences in radiological interpretation .............................................................. 187
8.2.1 Aptitude for radiological interpretation ........................................................................... 188
8.2.2 Aptitude for mammogram interpretation ............................................................................ 189
8.3 Methodological issues in investigating individual differences in complex skills ...................... 190
8.3.1 Individual differences in skilled performance and skill acquisition ..................................... 191
8.3.2 Psychometric versus information processing approaches ................................................. 196
8.4 Psychometric and cognitive task analysis approaches to individual differences in mammography ........................................................................................................ 198
8.4.1 Aptitude for screening mammography ............................................................................. 198
8.4.2 Cognitive task analysis ....................................................................................................... 201
8.4.3 Prediction of performance and individual differences ......................................................... 202
8.4.4 Summary .............................................................................................................................. 203
9 Searching for a psychometric perceptual personality domain ........................................ 205
  9.1 Perceptual Ability and Mammography ............................................................. 206
    9.1.1 Perception, Personality and Mammogram Interpretation ....................... 207
  9.2 Method ........................................................................................................... 212
    9.2.1 Subjects .................................................................................................. 212
    9.2.2 The Test Battery .................................................................................... 213
    9.2.3 Procedure ............................................................................................. 215
  9.3 Results .......................................................................................................... 216
    9.3.1 Test Reliability ...................................................................................... 218
    9.3.2 Correlational Analysis ......................................................................... 218
    9.3.3 Principal Components Analysis ......................................................... 220
  9.4 Discussion .................................................................................................... 222
    9.4.1 Correlational and factor structures of the test battery ......................... 222
    9.4.2 Psychometric prediction of mammography performance ................. 225

10 Development of a cognitive test battery ......................................................... 226
  10.1 Derivation of cognitive predictor tests ....................................................... 227
    10.1.1 Target detection .................................................................................. 229
    10.1.2 Target analysis ................................................................................... 230
    10.1.3 Decision making and classification ................................................... 232
    10.1.4 The nature of the cognitive tasks and their relationships with
          psychometric predictor tests ..................................................................... 234
  10.2 Method ........................................................................................................ 235
    10.2.1 Subjects ............................................................................................... 235
    10.2.2 Procedure ............................................................................................ 235
    10.2.3 The test battery: Task designs and administration
          procedures ......................................................................................... 236
  10.3 Results ....................................................................................................... 240
    10.3.1 Test development experiments and score derivation of the
          'grid' task ................................................................................................. 241
    10.3.2 Task validation and score derivation for the 'target' task .................. 244
    10.3.3 Validity of the contrast sensitivity and concept learning
          strategy scores ......................................................................................... 250
    10.3.4 Construct validity of the cognitive task battery ..................................... 254
    10.3.5 Concurrent validity of the batteries: Relationships between
          cognitive tasks and psychometric tests .................................................. 258
  10.4 Discussion .................................................................................................. 261
    10.4.1 Validity of cognitive task measures ..................................................... 261
    10.4.2 Overlap between cognitive tasks ......................................................... 263
    10.4.3 Overlap between cognitive and psychometric measures ................. 265
    10.4.4 Representativeness of derived measures of cognitive
          processes in mammogram interpretation .............................................. 267
    10.4.5 Summary ............................................................................................... 269
11 Predicting mammogram interpretation performance ................................... 270
  11.1 Introduction .............................................................................................. 270
    11.1.1 Selection of film readers for breast cancer screening  271
    11.1.2 Assessing readiness to learn in lay film readers 272
    11.1.3 Diagnosis of weaknesses in performance 273
  11.2 Method ........................................................................................................ 274
    11.2.1 Subjects ......................................................................................... 274
  11.3 Results ........................................................................................................ 274
    11.3.1 Predictive validity of psychometric and cognitive variables 275
    11.3.2 Prediction of performance after training ................................. 277
    11.3.3 Prediction of improvement as a function of training 279
    11.3.4 Prediction of good and poor performance ............................... 281
  11.4 Discussion ................................................................................................. 283
    11.4.1 Personality and post-training performance ................................. 285
    11.4.2 Readiness to learn from feature instruction as a function of pre­
    training performance ..................................................................... 287
    11.4.3 Strengths and weaknesses in post-training performance ... 288
  11.4.4 Conclusion .................................................................................... 289

12 Discussion: Complex visual skill acquisition, training and individual
differences ........................................................................................................ 291
  12.1 Theoretical issues in mammogram interpretation ................................. 291
    12.1.1 The nature of mammogram interpretation ............................... 292
    12.1.2 Individual differences and cognitive processes in mammogram
    interpretation ..................................................................................... 294
  12.2 Empirical findings and implications for breast screening ................. 296
    12.2.1 Lay film reading .......................................................................... 296
    12.2.2 Issues in training lay film readers for screening
    mammography ............................................................................... 297
    12.2.3 Selection of lay readers for mammogram interpretation . 301
    12.2.4 Implications for lay film reading: Roles of lay readers 302
  12.3 Limitations of the present research ...................................................... 305
    12.3.1 Practical limitations ..................................................................... 306
    12.3.2 Limitations of theoretical considerations ............................... 307
  12.4 Suggestions for further work ................................................................... 309
    12.4.1 Future work concerning lay film reader selection and training 309
    12.4.2 Future work on mammogram interpretation skill and its
    acquisition .............................................................. 310
    12.4.3 Conclusion .................................................................................... 312

References ....................................................................................................................... 313
List of Figures

Figure 4.1: Recording form for mammogram interpretation .............................. 71
Figure 4.2: Triple-blind film reading protocol for radiographer screening film interpretation. ................................................................. 72
Figure 4.3: Pre- and post training and follow-up performance of the radiographers, compared with pre- and post training performance of a group of radiologist measured with the G-index. ............... 74
Figure 4.4: Probabilities of false and true positives (radiographers) .............. 75
Figure 4.5: Total number of false negatives and recalls for the wrong breast (radiographers) ................................................................. 77
Figure 4.6: Error analysis: Total number of perception and interpretation errors (radiographers) ................................................................. 78
Figure 4.7: Radiographer screening performance over 16 chronological intervals. ................................................................. 79
Figure 4.8: Radiographer true and false positives over 16 chronological intervals with radiologist performance as the criterion. ............... 80
Figure 4.9: Assessment outcome of 326 recalled cases ....................................... 82
Figure 4.10: Assessment outcome of 29 cases not initially recalled by the radiologist ............................................................................................ 82
Figure 4.11: Correspondence analysis plot of A.S.'s verbal data ......................... 87
Figure 4.12: Correspondence analysis plot for F.A.'s verbal data ....................... 87
Figure 6.1: Recording form for mammogram interpretation of features and decision outcomes (students). ................................................................. 143
Figure 6.2: Performance of two student mammogram training groups compared with radiographer and radiologist performance . . . . . . 144
Figure 6.3: Probabilities of true and false positives for feature training and medical reasoning ................................................................. 146
Figure 6.4: Standard deviations for feature training and medical reasoning groups. ................................................................. 148
Figure 6.5: Mean performance and performance range measured by the G-index as a function of feature training (students). ......................... 155
Figure 7.1: Process model of mammogram interpretation ............................... 165
Figure 7.2: Mean number of perception and interpretation errors (students). ................................................................. 170
Figure 7.3: Simplified process model of mammogram interpretation showing the locations of false negative and false positive errors 174
Figure 9.1: Percentage breakdown of student subjects by faculty (psychometric study) ................................................................. 212
Figure 11.1: Correspondence analysis plot between three levels of post-training performance and variables predictive of mammogram interpretation performance after training. ......................... 282
List of Tables

Table 3.1: Decision outcomes in radiological assessment as a function of subjective observer decisions and objective presence of abnormality ................................................................. 45
Table 4.1: Agreement between radiographers and radiologist in screening film recall ................................................................. 80
Table 4.2: Radiographer agreement with final decision, screening and referral outcome (Goodman-Kruskal Gamma) ............................................. 81
Table 4.3: Percentage of decision outcomes in four decision categories (radiographer screening data) .................................................. 81
Table 4.4: Examples of verbal protocols (radiographers) ................................................................. 84
Table 4.5: Verbal features most closely associated with particular decision outcomes before and after training ........................................ 88
Table 6.1: Classification of training mammograms by perceptual difficulty and abnormality type .............................................................. 139
Table 6.2: Analysis of Variance table for overall film reading performance (students) ................................................................. 145
Table 6.3: Analysis of Variance table for true positive (hit) classification performance (students) ................................................ 147
Table 6.4: Analysis of Variance table for false positive (false recall) classification performance (students) ........................................ 147
Table 6.5: Descriptive statistics of film reading performance for all feature method trainees (n = 23) ................................................ 153
Table 6.6: Three one-way analyses of variance for overall performance, false positives and true negatives as a function of feature training (n = 23) ................................................ 154
Table 7.1: Two one-way analyses of variance for perception and interpretation errors (students) ................................................................. 171
Table 8.1: Specific task aspects of screening mammography and possible predictors of performance after training ........................................ 200
Table 9.1: Means, standard deviations and norms of five personality and ability tests ................................................................. 217
Table 9.2: Evaluation of psychometric test reliability with Cronbach's alpha coefficient ................................................................. 218
Table 9.3: Correlations between psychometric tests and subscales ................................................................. 219
Table 9.4: Three factor solution of principal components analysis with oblique rotation (psychometric study) ........................................ 221
Table 9.5: Factor correlation matrix for the principal components analysis (psychometric study) ................................................................. 222
Table 10.1: Summary of analysis of variance investigating the role of depth and diameter of holes in phantom objects in the perceptibility of dots in X-ray grids ........................................ 243
Table 10.2: Means, standard deviations and ranges of scores derived from the grid task ................................................................. 244
Table 10.3: Mean accuracy scores and response times with standard deviations and ranges for the 'target' task .......................................... 245
Table 10.4: Summary of analysis of variance for the number of correct trials on the 'target' task .................................................. 246
Table 10.5: Summary of analysis of variance of response times (correct trials) on the 'target' task .............................................. 246
Table 10.6: Mean accuracy scores and response times for task transfer, background and target effects ........................................ 247
Table 10.7: Two factor solution for principal components analysis of scores derived from the 'target' task ........................................... 249
Table 10.8: Factor correlation matrix for two components solution ................................................................. 250
Table 10.9: Age norms and mean contrast sensitivity scores for both eyes and the poorer eye .......................................................... 251
Table 10.10: Trials to solution for different strategy groups in Johnson's Zaps and Duds task compared to those obtained in the abbreviated 'concept' task .................................................. 253
Table 10.11: Prevalence of blundering after the determination of the solution in the 'concept' task ...................................................... 253
Table 10.12: Correlations between measures derived from the cognitive test battery .............................................................. 255
Table 10.13: Five factor solution of principal components analysis of cognitive tasks ................................................................. 257
Table 10.14: Correlations between cognitive tasks and psychometric tests .................................................. 258
Table 10.15: Item-function correlations between cognitive task and psychometric variables for four canonical variates .................. 260
Table 11.1: Correlations between predictor variables and mammogram interpretation performance before and after training and improvement and retention as a function of training .................................................. 276
Table 11.2: Standard multiple regression of predictor variables associated with performance after training ........................................ 279
Table 11.3: Standard multiple regressions of predictor variables on improvement and retention of mammogram performance as a function of training .................................................. 280
Table 11.4: Mean predictor scores for good, neutral and poor post-training mammogram interpretation ........................................ 282
Table 12.1: Potential roles and characteristics of performance of lay film readers in terms of task level and role definition .................................................. 303
List of Pictures

Picture 1  Normal, oblique view mammogram showing images of right and left breasts, respectively. ............................................. 42

Picture 2  Mammogram with abnormality in the upper right breast. This is represented by a spiky white density indicating a mass. .............. 44

Picture 3  Mammogram showing asymmetry, a more subtle abnormality indicated by the proliferation of white 'fluff' in the lower right breast. ................................................................. 44
Introduction

Chapter 1
Issues in learning to read mammograms

This thesis is concerned with issues relating to the acquisition of X-ray interpretation skill. The research specifically addressed the interpretation of mammograms, which are X-rays of the female breast. Mammography is the medical examination conducted under the auspice of the national Breast Cancer Screening Programme in the UK. The interpretation of screening mammograms is a complex perceptual and interpretative task, which at present is typically conducted by radiologists, specialists in visual medical diagnosis. Screening mammography thus is a consultancy based service in Britain. The present research was motivated by an increasing need to explore alternative models to this as the service is nearing full implementation.

This chapter presents a brief outline of the applied and theoretical issues addressed in the current research. It should become clear that there is a close correlation between practical problems in breast screening and psychological theory. The intention is to deal with the former in close reference to the latter. This chapter begins with a discussion of some issues in breast cancer screening which have led to the re-evaluation of some aspects of the original focus of the Screening Programme. Although this thesis was motivated by applied problems in breast cancer screening, there are a number of important theoretical issues which are pertinent to these. They particularly concern the nature of perception and interpretation in visual medical diagnosis. The introductory discussion of theoretical matters as they pertain to mammogram interpretation is followed by a brief introduction to matters of training and selection. These applied issues partly arise from the current concerns in breast screening and are at the centre of the present research. Finally, a structural outline of the research covered in this thesis is presented.
1.1 Breast cancer screening

In 1986, the Forrest Report on "Breast Cancer Screening" was published. The Forrest Report resulted from the recommendations of a Working Group chaired by Sir Patrick Forrest which had been convened to examine the viability of breast cancer screening in Britain. Their evaluation included a review of breast screening programmes in other countries. They came to the conclusion that

"The information that is already available from the principal overseas studies demonstrates that screening by mammography can lead to prolongation of life for women aged 50 and over with breast cancer. There is a convincing case on clinical grounds for a change in UK policy on the provision of mammographic facilities and the screening of symptomless women." (Forrest Report, 1986, p. 7).

The Forrest Report recommended that all women between the ages of fifty and sixty-four should receive breast cancer screening as part of the National Health Service provisions. This involves an examination by mammography which is be available to every woman in the screening population at three yearly intervals, a screening interval which was considered appropriate by the Forrest Committee. Further recommendations were that the mammographic examination should consist of a single high quality oblique view X-ray. All mammograms were to be interpreted by a single highly qualified radiologist, rendering breast screening in Britain a consultancy based service. The original recommendations of the Forrest Committee were largely endorsed by the NHS Breast Screening Programme's Advisory Committee in 1991. This body emphasized the need for the maintenance of a high quality programme not least achievable through the provision of suitably skilled staff.

1.1.1 Problems in breast screening

Although initial evaluation of the breast screening programme in Britain (NHSBSP, 1991) indicates an overall success in terms of reduction of mortality, a number of
questions have been raised regarding some of the practical recommendations made by Forrest (1986). Like in other imaging based screening programmes such as tuberculosis screening by chest X-ray (Yerushalmy, 1969), problems of observer variability have been encountered in breast screening. There are inter- and intra-individual differences in film readers' interpretation of mammograms which cannot be attributed to the level of training or experience alone (Gale, Walker and Roebuck, 1989). Although the provision of suitably qualified staff is a significant factor in the success of the breast screening programme, procedural factors may also contribute to the reduction of observer error in screening mammography. Two procedural issues which have been widely discussed are two view mammography replacing the current practice of screening with one oblique view mammogram and dual reading of mammograms. In the present context, the latter is of greater interest, because of its potential implication for training of film readers for mammogram interpretation. This issue will be discussed further below.

The provision of suitably qualified staff is largely dependent on the existence of sufficient radiologists who are interested in working in breast cancer screening. There is some evidence that some screening centres have had difficulties in recruiting enough radiologists. At present, some centres employ general practitioners who have received training in mammogram interpretation. There is no evidence that the standard of film reading performance of these individuals is lower than that of trained radiologists, and in practice, these general practitioners are equivalent to radiologists. The problem of staff provision would be exacerbated if procedural changes, such as dual reading, were introduced. This would, in effect, result in a doubling of film reading work load as dual reading implies that each mammogram is read twice, either independently by two different observers or by one observer on two separate occasions.
1.1.2 Solutions to problems

1.1.2.1 Dual film reading

Dual film reading has been suggested as one viable solution to reducing the effects of observer variability in radiological diagnosis. It has been found a particularly useful approach in population screening programmes such as tuberculosis screening (Yerushalmy, Harkness, Cope and Kennedy, 1950). The general effect of dual reading is one of reducing errors in which abnormality is missed, but in its most elementary form such a procedure may lead to an increase in overall true and false identification of abnormality. The latter may be reduced, however, by the employment of an arbitration process in which all cases identified as abnormal by only one reader are subject to further review. Benefits of dual film interpretation have also been noted in mammography screening programmes (Haiart and Henderson, 1991, Kirkpatrick, 1993).

Although, at present, there is no provision for dual film interpretation in the NHS Breast Screening Programme (Forrest, 1986), a number of screening centres have adopted this procedure on an experimental basis with beneficial effects in terms of cancer detection rates. The expense of dual consultant film interpretation exceeds the provisions made for fully implemented national screening, though. One possible solution to the cost problem might be to employ lay film readers alongside consultant radiologists. Dual screening with one lay reader and one consultant would be achievable in terms of costs for most screening centres. This is worth considering, if it results in a significant reduction of observer error, especially with respect to missed abnormality.

1.1.2.2 Lay film reading

Consideration of lay screening in the context of dual film interpretation gives rise to questions regarding the role a lay observer should play. These questions are intrinsically linked to matters such as training and selection which, in turn, depend on the definition of mammogram interpretation skill in the screening context. There
are essentially two alternative models of dual reading involving lay film readers. The first one is a straightforward dual reading scheme in which a suitably trained lay reader interprets mammograms alongside a radiologist. The success of this scheme is based on the assumption that "two pairs of eyes are better than one", i.e. two independent observers are less likely to miss abnormalities than one. The second possible model is a pre-screening one. Some population based screening programmes, for example, cervical cancer screening, employ a pre-screening scheme, whereby fully trained cytological technicians evaluate all cases in the first instance. Only those judged abnormal plus a certain percentage of clear tests are referred to a consultant pathologist for evaluation. A similar scheme might be appropriate for breast cancer screening, if radiographic assistants can be trained to read mammograms.

The issue of lay film interpretation in breast screening is contentious, however. The question is whether individuals without medical background are capable of learning to interpret mammograms adequately. The benefits of dual reading may be compromised, if one of the observers performs below standard. Although there is already some evidence for the viability of lay film reading in mammography (see chapter 3), there is a requirement for an evaluation of the potential of lay screeners. The scope of the present research did not permit an assessment of the impact of dual screening beyond the most exploratory level (see chapter 4). However, it presents an analysis of the viability of lay film interpretation in screening mammography and related issues of training and selection of lay readers. The issue of lay screening is discussed in more detail in chapter 3.

1.2 Perception and interpretation of medical images

The main theoretical concern in the present thesis was to investigate the nature of mammogram interpretation skill and its acquisition. It is, therefore, important to discern the nature of perception and cognition in the understanding of complex visual stimuli such as medical images. Two questions in particular arise in this
context. Firstly, it is a matter of some interest how visual information contained in medical images is extracted and assigned meaning. The second question concerns how information extracted from medical images is applied in making clinical decisions. In this section, the problem of the link between perception and cognition is first addressed on a general level. This is followed by more detailed enquiries into the nature of the skill involved in interpreting mammograms. In particular, mammogram interpretation is discussed briefly in terms of image processing and medical diagnosis.

1.2.1 Perception and cognition

The nature of perception and the cognitive interpretation of sensory information has been an important issue in philosophical and psychological enquiry for centuries. The main question concerns the link between the physical world and the way in which it is experienced by human observers. The idea that the imagination of the observer somehow constitutes the link between perception and thought can be traced back to Aristotle (384-322 BC). But it was Kant's work in the eighteenth century, which reaffirmed the role of imagination as the mediating influence between sensory experience and conceptual recognition (Hendee, 1993). Kant (1724-1804) postulated that causal relationships between perception and recognition arise from previous experience. Imagination serves to conceptualise sensory experience in time and space and is the origin of 'categories', which organise such experience explicitly or implicitly in cognition. Reason coupled with experience thus leads to complete understanding of the world. However, the involvement of observers in achieving such absolute understanding leads to a degree of scepticism regarding their ability to achieve complete rational knowledge.

Although there are alternative philosophical accounts of the link between perception and cognition (see Hendee (1993) for a more complete review), Kant's exposition is perhaps the one that provides the most interesting ideas in the present context, because it gives some indications as to how skills involving the perception
and interpretation of visual stimuli may be acquired. According to Kant perception is based on matching sensory experience with 'categories' of previously stored knowledge, an idea that is equivalent to present day schema theory. Like schemata, categories are the result of internal organisation processes conducted by the observer on the basis of sensory experience. Acquisition of knowledge thus results from the active concern of the observer to make sense of his or her environment.

Similar accounts have been put forward in the more specific context of the perception of visual images. Helmholtz (1821-1894) argued that the perception of visual images results from the imposition of internally organised knowledge upon visual perceptual data. The observer can be misled by visual data, however, and this can result in erroneous inductive inferences and the misinterpretation of visual data. This view is representative of modern theories of perceptual hypotheses, which regard perception as a largely 'top-down' process, whereby the interpretation of sensory information is critically dependent on the presence of appropriate stored knowledge. One alternative to this approach has been presented by Gibson (1950, 1979), who argued that perception of images is 'direct'. The observer simply picks up information from an array of light without specific processing or involvement of cognition. Gibson's work is representative of a 'bottom-up' approach to visual perception. Although the Gibsonian approach may be useful in explaining some of the image processing aspect of interpreting mammograms, it fails to explain how errors of perception arise. Interpretation of X-rays involves more complex processes than the mere perception of visual information. A full account of X-ray interpretation must take into account both, image processing mechanisms and cognitive mechanisms relating to the interpretation of images in the medical context. These are discussed in more detail below.

1.2.2 Image Processing

Image processing especially in computer vision has been influenced by the Gibsonian view of perception as a bottom-up process. On the simplest level, the
perceptual task in X-ray interpretation may be regarded as a problem of visual texture discrimination, given the planar nature of traditional X-ray films. Radiologists are typically able to examine X-ray images for internal inconsistencies or inappropriate image elements very rapidly and accurately. This detection process may be pre-attentive and results in the realisation that an anomaly is present in the image (Hendee, 1993). However, such pattern perception processes do not constitute an interpretation of the medical image. Interpretation can be more appropriately construed as a problem akin to image processing in picture perception, involving the recognition and categorisation of visual information representative of the three-dimensional world.

One problem inherent in radiological images, however, is that the images bear little relationship to real scenes. Whilst the comprehension of natural scenes can be explained in terms of image intensities, surface illumination and surface reflectance (Raff, 1993), X-rays are the result of different physical parameters such as the absorption characteristics of different tissue densities and mapping of three-dimensional structures into a plane. Recognition of abnormality in X-rays is thus akin to but different from natural scene perception and requires specific knowledge of the visual properties of imaged structures. The framing of X-ray interpretation as an image processing problem strongly emphasizes the perceptual detection and recognition components of the task. The role of cognition is of relatively little importance in this context as compared to low level perceptual processes. Cognition only gains importance in the later stages of recognition, when visual information needs to be matched with prior knowledge in the Kantian sense in order to arrive at the conscious realisation that abnormality is present.

1.2.3 Interpretation of X-rays in the medical context

With the possible exception of the research laboratory, medical images are rarely interpreted outside the medical context. Usually, the aim of interpreting an X-ray is to make a medical diagnosis, in which the X-ray may only represent one piece of
Medical diagnosis on the basis of radiological evidence can be construed as a complex skill involving knowledge of radiological anatomy, relationships between variations in anatomical structure and patterns seen in X-rays (Lesgold, 1984). The diagnostic task involves evaluation of the visual evidence and interpretation of their significance in the context of alternative disease patterns. In this framework, perception is relatively less important, because it results from the execution of complex cognitive processes such as medical reasoning. The link between cognition and perception in this context is much more akin to Kant's conceptualisations and the theory of perceptual hypotheses whereby cognition guides perception.

Some radiological examinations, such as screening mammography, are not concerned with medical diagnosis but with the detection of early signs of abnormality. In this context image processing approaches may seem more appropriate. However, it should be noted that even in the breast screening context, X-ray interpretation leads to some kind of decision regarding further client management. The observer makes decisions not only with regard to whether an abnormality is present or not, but also with respect to the likely cause of abnormality in terms of benign or malignant disease processes. In this respect, mammogram interpretation in the breast screening context represents a diagnostic reasoning task.

1.2.4 Perception and interpretation in lay film reading

In the discussions above, mammogram interpretation skill in the breast cancer screening context has been characterised in terms of two possible theoretical conceptualisations. It is clear that a full examination of film reading in screening mammography must attempt to address both, image processing and medical reasoning. These issues are discussed further in chapters 2 and 5. The matter of adopting an appropriate frame of reference when considering mammogram interpretation skill, is of particular importance in the context of lay film reading.
The problem of what role lay film readers can play in the breast screening programme has already been mentioned. In view of the theoretical discussions above, it is clear that mammogram interpretation skill in lay screeners could take a number of shapes. In the simplest implementation, lay screeners may learn to recognise patterns of abnormality (Swinburne, 1971), reducing the interpretation task to a purely perceptual level. Alternatively, they may be encouraged to learn to differentiate between normal and abnormal variation at a level of interpretation akin to radiologists. The question arises, though, to what extent prior medical knowledge is a prerequisite of the full interpretative task. In essence, the problem is one of training which will be discussed further below.

A further issue concerning the employment of lay film readers is that of individual differences. In recent years, it has become increasingly clear that cognitive processes are subject to individual differences. Although individual differences have been identified across virtually all domains of complex cognitive functioning, they have been particularly pertinent in perceptual cognitive domains such as spatial reasoning (Lohman and Kylönen, 1983) or map reading (Thorndyke and Stasz, 1980). Individual differences have also been identified in more specifically perceptual domains. Gale (1993) has argued that individuals' responses to visual stimuli may be influenced by a number of factors including aptitude for the visual task. In the present context, individual differences in mammogram interpretation were examined in order to investigate the potential for selecting lay film readers with specific aptitude for the task.

1.3 Selection and training of lay film readers

Both issues, training and selection of lay film readers, are inextricably linked with deeper theoretical questions concerning the nature of the skills involved in mammogram interpretation. Thus, if the aim is to train individuals to recognise abnormal patterns, training needs to enable lay film readers to recognise the visual features indicative of abnormality in mammograms. Similarly, selection procedures...
have to focus on specific visual ability characteristics involved in pattern recognition. If lay film readers are to perform more complex interpretative duties, however, training also has to address decision making processes on the basis of visual medical information, whilst individual differences of a much broader spectrum need to be considered for selection purposes.

1.3.1 Training

One of the main questions to be addressed in the context of the present thesis, apart from those regarding the viability of lay film reading in screening mammography, is whether different methods of training lead to differential performance. The contents of training may have an important role to play in interaction with the role definition of the trainee. Although image interpretation in terms of disease schemata is an appropriate activity for radiologists and other doctors, who as part of their medical training have acquired extensive knowledge of anatomy and pathology, it may not be desirable for paramedical film readers to make decisions regarding the status of an observed abnormality unless their training specifically included medical diagnosis of breast disease. At present, film reading training is designed specifically for fully trained radiologists entering breast screening. The appropriateness of such training for lay film readers is addressed in the present research. In addition, a more visually based training method is evaluated. The argument is that training more specifically aimed at visual detection might be more appropriate for training lay readers as pre-screeners. In essence, the question is whether it is possible to train lay readers as pattern perceivers or as film interpreters, respectively. The nature and level of performance after such differential training is of considerable interest in the context of establishing a suitable role for lay readers in breast cancer screening. This issue is addressed with reference to a detailed process model of mammogram interpretation performance in chapter 7.
1.3.2 Selection

Selection of individuals for mammography interpretation training becomes pertinent in the context of training a relatively large number of lay film readers. In this situation, it would be highly desirable to identify those individuals who have an aptitude for mammogram interpretation. In the present research, a systematic attempt was made to relate psychometric variables to performance in mammogram interpretation. At its most complex, mammogram interpretation performance in breast screening may be influenced by personality and ability variables on a number of levels. Perceptual and cognitive ability may be prerequisites of good task performance in recognising and interpreting abnormality correctly. Such performance may also be influenced by differential use of cognitive strategies. Film interpretation in breast screening is subject to a number of special conditions. For example, films are usually viewed in large numbers and the implications of making errors can be severe for others' well-being. These vigilance conditions combined with the social implications of making mistakes give rise to the notion that individual differences in performance may also result from differences in personality. There is some indication that personality traits such as impulsivity and anxiety may lower performance, because observers characterised by these traits produce a higher than average error rate in the recognition of normal structures (Ansell, Davies and Hammond, 1990a).

Even if there is no possibility for positive selection of lay film readers on the basis of aptitude, knowledge of ability, cognitive skill and personality variables predictive of performance after training are useful. It may, for example, be possible to identify specific areas of strengths or weaknesses which affect skilled performance. If the cause of weaknesses can be identified in terms of information processing parameters, it may be possible to selectively support such weaknesses with specific additional training. For this purpose, it is necessary to specify the film reading task in terms of information processing steps, measure performance on these and identify those which are responsible for low performance. In the context
of the present research, opportunity for such an analysis arose with the development of the model of mammogram interpretation presented in chapter 7. The model was employed to develop specific cognitive tasks designed to measure aspects of information processing in mammogram interpretation (chapter 10). In this manner it was possible to investigate the predictive power for mammogram interpretation performance of both, psychometric ability and personality variables and cognitive skill variables.

1.4 Layout of the thesis

The research presented in this thesis approached the issue of lay film reading in screening mammography from three perspectives. In the first instance, a longitudinal study of film reading skill acquisition was conducted, which assessed the progress of two radiographers from initial training to performance during ten months of reading screening mammograms in an every day breast screening environment. This represented a case study investigation of complex skill acquisition in mammogram interpretation. A second line of enquiry involved the collection of psychometric test data from first year university undergraduates. The psychometric test battery initially consisted of paper-and-pencil tests which assessed ability and personality traits that were hypothesized to be associated with aptitude for mammography. A second stage of psychometric assessment involved the employment of tests measuring cognitive components of mammogram interpretation. These were developed on the basis of a detailed cognitive task model. Lastly, a third approach consisted of experiments entailing brief training for novices in mammogram interpretation. The purpose of these were threefold and involved the assessment of alternative training methods, a broader investigation of the skills acquired during mammogram interpretation training and an assessment of the predictability of training outcome on the basis of personality, ability and cognitive task variables.

The three empirical approaches, which broadly relate to skill acquisition, individual
differences and training, are reflected in the three part structure of the present thesis. The first part deals with an analysis of skills involved in screening mammography and their acquisition. The second part presents an expansion of this topic with the aim of developing a detailed process model of film reading skill. In this section, the focus is broadened to include more specific issues of training. The third part addresses the problem of individual differences in terms of aptitude for mammogram interpretation and prediction of performance. From an applied point of view, the second part thus addresses issues of training, and the third part is concerned with matters of personnel selection. Apart from the longitudinal study of lay film reading acquisition reported in the first part, participants in all studies were undergraduate students. All students participating in the training studies had also partaken in the psychometric assessment studies. The presentation of research in this thesis is not chronological in terms of data collection, because the training studies logically follow the case studies on skill acquisition. They are therefore reported in the second part, although they were chronologically preceded by the psychometric data collection. Each of the three parts and the eleven chapters comprising the main body of this thesis are introduced in more detail below. The final chapter, which follows the three main parts, presents a general discussion of the data presented and suggestions for further psychological work in film reading skill acquisition and complex visual tasks in general.

1.4.1 Part 1: Acquisition of mammogram interpretation skill

The first part of this thesis is concerned with an initial analysis of skills involved in mammogram interpretation. This section consists of three chapters, the first of which presents a review of research on radiological skill and its acquisition. The purpose of chapter 2 is to review two psychological approaches to the investigation of radiological skill and its acquisition. One derives from empirical research into visual search phenomena in X-ray interpretation, whilst the other originates from more recent approaches to radiological interpretation as a complex cognitive skill. The second chapter in this section presents an introduction to mammography and
mammogram interpretation in the breast screening context. The aim of this chapter is to assess the applicability of previous research on radiological skill in both, visual perceptual and cognitive frameworks to mammogram interpretation in breast screening. In addition, this chapter serves as an introduction to pertinent issues of lay film reading and training of film readers in screening mammography. The final chapter in this section presents a longitudinal case study of two radiological assistants acquiring mammogram interpretation skills over a ten months period. This addresses the issue of lay film reading and skill acquisition on an empirical level.

1.4.2 Part 2: Developing a model of mammogram interpretation skill

The second part is concerned with issues of skill acquisition and training at a deeper level. The first chapter in this section presents a detailed analysis of mammogram interpretation skill in terms of the pertinent psychological literature. This includes discussions of pattern recognition, categorisation and decision making in the medical context. In addition, issues of training and skill acquisition in radiology and mammography are reviewed, including a psychological analysis of the types of learning involved in acquiring mammogram interpretation skills. The next chapter presents data comparing novice skill acquisition as a function of two training methods, one orientated towards pattern perception and the other concerned with the acquisition of interpretative skills. A further training experiment concerned with investigating the influence of the former training method on individual differences in performance is also reported in this chapter. Participants in these training studies were first year undergraduates who had previously partaken in the psychometric assessment studies reported in the third part. The last chapter in this part is primarily concerned with the presentation of a double two-stage task model of mammogram interpretation, consisting of recognition and decision processes on perceptual and interpretative levels. Data from the training studies and an additional error analysis are used to validate and provide support for the two processes involved in mammogram interpretation which are identified formally in the model.
This chapter concludes with an overview of the implications of the task model for lay film reading in mammography.

1.4.3 Part 3: Prediction of film reading performance

The third part consists of four chapters and is concerned with individual differences in mammogram interpretation performance and their prediction. The first chapter presents a discussion of individual differences in film reading performance, as well as an analysis of methodological approaches to individual differences in complex cognitive tasks. The relative merits of psychometric assessment based on personality and ability measurement and those based on task-orientated assessments are reviewed. The second and third chapters in this section present psychometric data of both kinds. In chapter 9, an attempt to identify a psychometric aptitude domain for mammogram interpretation on the basis of personality and ability variables is reported. This is followed in chapter 10 by the development of a test battery based on cognitive elements of the task model presented in the previous section. The last chapter is concerned with prediction of performance. For this purpose performance data from the training studies were employed as criteria, and personality/ability and cognitive performance data from the psychometric studies served as predictor variables. The aim of this chapter was to assess the potential for predicting performance after training as well as improvement as a function of training. These are major issues addressing the problem of individual differences in the context of lay film reader selection and training. This chapter also concludes the main part of this thesis.
Part 1

Learning to interpret X-rays:
The case of screening mammography

In the first part of this thesis, X-ray interpretation skill in general and screening mammogram interpretation in particular are discussed. The aim is to present a discussion of previous research on visual search and expertise in radiology and to provide a preliminary analysis of the skills involved in interpreting screening mammograms. The empirical work presented in this section addresses the issue of lay film reading. The present part consists of three chapters. The first of these presents a literature review covering two broad approaches to radiological skill and its acquisition. Much of the previous work on radiological has been exclusively concerned with the visual perceptual aspects of X-ray interpretation, in particular with visual search. More recently attempts have been made to examine radiological skill from a cognitive expertise perspective. This approach has been characterised by an emphasis of the diagnostic reasoning and decision making skills of the task.

In chapter 3, screening mammography is introduced as a specific discipline within radiology. The applicability of previous research to this radiological specialisation and issues of observer performance measurement in breast cancer screening are examined. The third chapter in this part presents a longitudinal study of two radiological assistants acquiring mammogram interpretation skill over a two-week training course and ten months of subsequent experience in reading screening mammograms. Both, objective performance measured with a test set of mammograms and screening performance measured against radiologist performance and screening outcome were assessed. In addition, qualitative verbal data were presented in an attempt to understand the changes in cognitive processing which occur as a result of skill acquisition. The data presented in chapter 4 thus represent a comprehensive attempt to study lay reader skill acquisition in the applied context.
Chapter 2
Radiological Skill and its Acquisition

The task facing the radiologist when interpreting X-rays is distinctive in its combination of elements of visual search, image perception, object recognition and decision making. The observer must scan the display for abnormal features, resolving ambiguity with regard to the genuinely abnormal nature of detected features, while attempting to reconstruct a three-dimensional image from a two-dimensional display and deciding on further action on the basis of the visual evidence. Examples of normal and abnormal mammograms, illustrating the ambiguity involved in their interpretation, are shown in chapter 3. The task of interpreting mammograms involves the disembedding of ambiguous features from an ambiguous background and applying probabilistic mapping rules to both recognition of abnormal features and diagnostic criteria in order to make a diagnostic decision. The expert radiologist is highly skilled in performing this complex visual task, rapidly detecting and processing features that distinguish abnormal from normal (Myles-Worsley, Johnston and Simons, 1988), albeit not always without error (Garland, 1959, Lusted, 1960).

In this chapter, empirical investigations of radiological performance and resulting theoretical models of radiological skill are discussed. Most of the previous studies on radiological skill have been concerned with detection and interpretation of abnormality in chest X-rays. Although the present research is concerned with skill in interpreting a different kind of radiograph, the mammogram, there are a number of similarities between chest X-rays and mammograms that warrant the consideration of research on interpretation of chest radiographs.

Although much of the early research on radiological skill was concerned with visual search and target detection (Swensson, Hessel and Herman, 1977, Kundel, Nodine and Carmody, 1978), there has been a shift towards an interest in the cognitive evaluation elements of the radiological diagnostic task (Lesgold, 1984,
Lesgold, Glaser, Rubinson, Klopfer, Feltovich and Wang, 1988) more recently. This reflects a general shift of focus towards studying the cognitive aspects of expertise (for a review, see Ericsson and Smith, 1991). In this chapter, the literature on detection and interpretation of abnormality in chest X-rays and other radiographs is reviewed both, with respect to empirical investigations and with regard to the theories of radiological skill arising from them. The relevance of previous research to breast cancer screening is discussed in more detail in the next chapter. A further focus of interest in the present chapter is the acquisition of the complex skills involved in interpreting radiographs. Research investigating the kind of knowledge which is prerequisite for expert performance is discussed from both, visual search and cognitive expertise perspectives.

2.1 Radiological skill and visual search

Much of the early work on radiological skill was motivated by a desire to understand observer error in expert radiologists. Garland (1959) noted that expert observers missed 25 to 30% of pulmonary abnormalities on chest X-rays and, if presented with the same films after a period of time, were liable to change their minds in approximately 20% of positive cases. There was also over-reading of negative films by 1 to 2%. Similar error rates were reported by Tuddenham (1962), Herman and Hessel (1975) and Kundel et al (1978). It is, therefore, not surprising that many early empirical investigations of observer error have focused on the visual detection aspects of the task, given the unique bearing that visual perception appears to have on radiological diagnosis in comparison with other medical disciplines.

2.1.1 Experimental studies of radiological skill

Tuddenham (1962) has argued that failure of perception must account for a substantial part of diagnostic errors because "one cannot interpret a shadow he has not perceived ..." (p. 694). The implication is that visual search is a critical element
in radiological diagnosis and a major source of error. The focus of visual search research in radiology has been the detection of low contrast 'targets' in radiographs. These are usually artificially imposed low contrast round shadows which are designed to simulate lung nodules in chest radiographs. Detection performance is measured in terms of the accuracy in detecting these artificial lung nodules.

Visual search performance in radiology has been investigated with a variety of aims and employing a number of different methodologies. Some studies have utilised signal detection theory and Receiver Operating Characteristic (ROC) analysis in order to investigate detection performance under various experimental conditions (Revesz, Kundel and Graber, 1974, Swensson, Hessel and Herman, 1977, Kundel and Revesz, 1980, Swensson, Hessel and Herman, 1980, Swensson, Hessel and Herman, 1985, Kundel, Nodine, Thickman, Carmody and Toto, 1985). Application of the technique to the detection of radiological abnormality is based on the assumption that the problem can be framed as one of detecting a signal on a 'noisy' background. ROC analysis permits the simultaneous measurement of two aspects of performance, sensitivity or the probability of detecting an abnormality (p(true positive)) and specificity or the inverse probability of distinguishing normal from abnormal (1 - p(true negative)).

Another common method for studying observer error in radiology has been the study of selective visual attention by quantification of eye movements recorded during visual search. Eye movement studies can be broadly classified into two types, those concerned with identifying sources of error in terms of eye movement parameters (Kundel, Nodine and Carmody, 1978, Kundel and Nodine, 1978, Carmody, Nodine and Kundel, 1980) and those investigating scanning strategies as a means for reducing observer error (Tuddenham and Calvert, 1961, Kundel and Wright, 1969, Kundel and LaFollette, 1972, Gale and Worthington, 1983, De Valk and Eijkman, 1984).
2.1.1.1 Measuring performance in the detection of radiological abnormality

The use of ROC analysis and associated accuracy indices in the assessment of diagnostic judgements has been described in detail by Swets and his co-workers (Swets, 1986, Swets and Pickett, 1982). Metz (1980) has discussed the application of ROC analysis to observer performance in radiology. It is in the nature of two-alternative decision tasks, such as deciding whether an abnormality is present in an X-ray or not, that there is a trade-off between sensitivity and specificity. The Receiver Operating Characteristic (ROC) curve is a description of the functional relationship between various combinations of sensitivity and specificity when different decision criteria are applied. ROC analysis can be used to determine decision criteria which characterise that point on the curve that combines maximum sensitivity with the smallest possible reduction in specificity. Another application of the technique is the measurement of observer performance, the accuracy of which can be quantified in terms of a number of indices derived from ROC analysis (Swets, 1986). It has also been employed to investigate experimentally induced differences in detection performance.

Although the ROC technique has been widely applied in the measurement of radiological detection performance, there are some problems with the approach. Radiological diagnosis does not only involve visual detection of abnormality but also involves making clinical decisions. The skills involved are, therefore, not pure detection skills but involve classification. Since ROC analysis is primarily a detection methodology, certain anomalies can arise when it is applied to a classification problem (Berbaum, Dorfman and Franken, 1989, Kundel et al., 1985). It is, for example, possible to detect an abnormality in an X-ray correctly but to evaluate it wrongly as insignificant and thus classify it wrongly as normal. This is not a great problem when the target is relatively unidimensional, i.e. detection and classification represent the same process, as in the case of artificial lung nodule targets. However, with increased complexity of the target abnormality the radiological task becomes more of a classification task than a detection task. In screening radiology, there are a number of additional problems since ROC analysis
is based on parametric assumptions. These are invariably violated when the ROC technique is applied to performance measurement in the screening context, because the great majority of screening radiographs are normal, resulting in a skewed distribution of normal and abnormal X-rays.

2.1.1.2 Factors influencing target detection in radiology

Despite the limitations of the ROC technique, a number of studies have employed the method successfully in experimental contexts. In particular, conditions influencing simple target detection in X-rays have been investigated. Kundel et al (1985) employed ROC analysis to compare the detection of simulated lung nodules in chest images with those on a uniform background at three different 'noise' levels. They found that detection performance was impoverished in images with detailed anatomical background. The conclusion was that anatomical structure in X-rays is not just background noise, but actually interferes with the structural integrity of the target image.

Swensson, Hessel and Herman (1977,1980,1985) examined the effect of different modes of search on detection performance. In their first study (Swensson et al, 1977), radiologists read difficult chest X-rays first under normal free search conditions and then under focused search conditions, during which subjects' attention was directed at frequently omitted findings. It was found that focused search increased true positive reports but simultaneously resulted in a higher incidence of false positives. They concluded that focused search caused the radiologists to adopt less stringent decision criteria rather than to improved detection performance. That visual search actually seems to produce positive effects was shown in further experiments comparing free and focused search with a no search condition (Swensson et al, 1980,1985). Radiologists performed much worse when they were asked to evaluate specified locations in the lung fields of experimental films. This was interpreted as indicating that the search process involves perceptual mechanisms that contribute directly to the detection of abnormality. However, an alternative interpretation of these results may be given in
terms of task demands. It might be possible that the interpretation task without search is simply more difficult and thus yields poorer results (see discussion of this point in chapter 7).

The influence of target properties on detection performance has been investigated by Revesz, Kundel and Graber (1974) and Kundel and Revesz (1980). They introduced the concept of target conspicuity, defined as the ratio of feature contrast to surround complexity. It was found that the thus quantified measure of target conspicuity correlated well with detection performance. Revesz and Kundel (1977) showed that conspicuity could be described in terms of a number of surround and lesion parameters in such a way that the detectability of a lesion by a skilled reader could be predicted with a probability of more than 75%.

2.1.1.3 Eye movement studies identifying sources of error

It has been argued that a better understanding of the sources of observer error in radiology might be useful in designing training with the aim of substantially reducing perceptual and judgement errors (Tuddenham, 1962). A number of studies have attempted to relate error sources to measurable eye movement parameters. Kundel et al (1978) have distinguished between four types of errors in small target detection (lung nodules). Orientation errors arise only in inexperienced observers and are due to lack of familiarity with the target. Observers fail to segregate stimulus and ground features, i.e. they do not know what to look for and where to look. Search errors result from a failure to scan the target with the useful visual field (estimated at 5 - 6° for nodule detection by Kundel and Nodine, 1978). Recognition errors occur when the target area has been scanned but the potentially abnormal features have not been recognised. Decision errors are due to wrong evaluation of ambiguous features, false negatives (misses) being false rejections of abnormal features and false positives resulting from false acceptance of normal features as abnormal.

The relative frequency of the latter three error types in experienced search has been
estimated using single eye fixation and fixation cluster parameters (Kundel et al., 1978). It was found that approximately 10 to 30% of misses (depending on the definition of the useful field of vision) were due to scanning errors, i.e. the target area was not fixated. A further 25% to 45% false negative errors occurred when the target area had been fixated but there was no clustering of fixations. These errors were thought to represent pattern recognition errors and were defined in terms of below average dwell times (less than .48 sec). The majority of errors were decision making errors (45%). These were errors which arose despite the fact that the target area had been scanned and dwell time had exceeded .48 seconds, but were due to the wrong classification of the abnormality as insignificant. The error analysis method was found to be too crude for analysing false positive errors.

Further evidence for the differentiation of recognition and decision errors has been presented by Carmody et al. (1980). In two experiments they investigated firstly, the effect of dwell time on detection performance and secondly, the role of peripheral vision in detecting nodules. They found that longer dwell times improved decision accuracy but not recognition accuracy. Peripheral visual information also appeared to have an effect. Although detection performance was much poorer when the target was 5° off axis, it was shown that search is to some degree informed by peripheral inputs.

2.1.1.4 Visual Scanning Strategies
One of the more pervasive notions with regard to reducing reader error in radiology has been the suggestion that the adoption of systematic viewing procedures might reduce perceptually based misses (Tuddenham, 1957, Garland, 1959, Lusted, 1960, Tuddenham, 1962). The idea was that differences in scanning strategies might be at the heart of large individual differences in detecting abnormality, and it was thought that experts might have developed an 'ideal' scan path on the basis of experience. Tuddenham and Calvert (1961) studied the scan paths of trained radiologists using a spot light illumination method. They found that there was wide individual variation in search patterns which did not seem to be greatly influenced
by training. As far as accuracy was concerned, there was no evidence that the adoption of orderly search patterns was in any way prerequisite for good performance. There was even some indication that excessive concern with systematic search interfered with detection performance. Thomas and Lansdown (1963) also found widely varying search patterns amongst novice radiologists entering training. Their observers were particularly attracted to edges but tended to fail to fixate less structured parts of the radiograph.

Kundel and Wright (1969), however, were able to isolate some common search strategies in clearly specified, limited radiological search tasks using eye movement recordings. They distinguished between two search modes, an initial global one and a more detailed local one. There was some evidence that early search proceeded in a circumferential manner in a similar way as in viewing works of art (Buswell, 1935), but the nature of the observed eye movements was to some degree task specific. The purpose of this global search was thought to be the selection of local areas for detailed analysis. Local search, on the other hand, served the purpose of resolving ambiguity. They tentatively concluded that observer error may result from the selection of an inappropriate search strategy.

2.1.2 Models of visual search in radiology

Visual search research has led to the development of a number of models attempting to describe the processes involved in radiological diagnosis. The basic structure of the majority of models of experienced radiological search is similar in that radiological diagnosis seems to consist essentially of three aspects. These have been described by Tuddenham (1962) as the "... three basic steps: the recording, the perception, and the interpretation of critical roentgen shadows." (p. 694). There are, however, some models which have drawn on psychological picture perception theory. These models, on the whole, present a more 'top down' view of radiological search.
2.1.2.1 Process models of visual search

Tuddenham's model is a process model of film reading which is influenced by his basic contention that image detection in radiology is largely explicable in terms of 'the predictable activity of the cones of the central retina' (Tuddenham, 1957), so that the observed visual phenomena can be entirely explained in terms of the underlying physiology. Emphasis is thus on the visual scanning of the display. Decisions are made by comparing critical shadows identified during search and fixation with memories of prior experience. If abnormality is detected an involuntary search for meaning is initiated which leads rightly or wrongly to the 'perception' (recognition) of a shadow as abnormal. Correct perception of abnormality is thought to be critically dependent on previous visual experience of the same kind.

Kundel and Nodine (1978) proposed a model based on their studies of error analysis of eye movement data. It differs from Tuddenham's model only in the addition of an orienting stage, which drives the sampling distribution utilised during search, and in an interpretation stage, which follows decision making. This serves to evaluate abnormalities in the context of the general disease pattern and culminates in a diagnosis. Blesser and Ozonoff (1972) put forward a similar but more comprehensive model of the radiological process, taking into account psychophysical, psychological and nosological factors of the diagnostic radiologic situation. The psychological stage is characterised by three aspects, search, recognition and decision making.

Later models have opted to regard the radiological process as less linear. Swensson (1980) acknowledged the interdependence of the three steps and formulated his model in terms of two stages, a perceptual-recognition stage and a decision making stage. He suggested that the perceptual-recognition stage acted as a pre-attentive filter selecting features for the observer's attention, which were then cognitively evaluated during the decision-making stage (see also Swensson, Hessel and Herman, 1980). Swensson points out that the two stages are not independent of
each other as the physical characteristics underlying the selection of pattern features for visual attention in a particular situation probably overlap those used to evaluate features as potential targets. The model formalizes the idea that skilled observers use visual strategies which render exhaustive evaluation of every pattern feature superfluous.

Nodine and Kundel (1987) reformulated their earlier orientation and scanning model in a similar vein. Although their interpretation of eye-fixation data still results in a model comprising four stages of search, these are seen as basic perceptual-cognitive units feeding into an interpretative decision making process which is not specified in detail. The search stages are global search, discovery search, foveal verification and reflective search respectively. The first stage performs the same function as Swensson's (1980) pre-attentive filter, selecting image features and potential target sites for attention and evaluation during the second stage. The remaining two stages relate target acquisition to interpretation and decision making processes. The model is, thus, essentially an elaboration of Swensson's (1980) two-stage model which also takes into consideration prior knowledge, training and experience as factors in the radiological process.

2.1.2.2 'Top-down' models of visual search
A number of researchers who have examined radiological skill from a visual perspective have postulated visual cognitive structures as the basis for interpretation of X-rays. Kundel and Nodine (1983), for example, concluded that experimental data on detection during very brief exposure suggests that radiologists perceive abnormality in radiographs in a top-down fashion. They have used theories of picture perception and in particular Arnheim's (1969) 'visual concept' to explain their findings. Arnheim argued that picture perception is "... an active concern of the mind" (p.37) and that categories, which he calls visual concepts, guide perception. Kundel and Nodine have postulated that the visual concept determines what is perceived in the radiograph and acts as a hypothesis about what can be expected on the basis of previous visual input. Visual data from the retina serve to
either support the hypothesis or to disconfirm it. In the latter case the visual concept is modified. X-ray interpretation is, thus, dependent on the meaning the observer can derive from the image on the basis of expectations and current visual input. Errors in image interpretation arise when the radiologist operates on either inappropriate visual concepts or does not have a suitable concept at all.

Gale, Johnson and Worthington (1979) have argued that a more dynamic approach is required to explain the processes involved in radiological search. They draw on Neisser's (1976) 'perceptual cycle' theory in order to characterize the dynamic interaction between expectations and visual input (see also Gale, 1993). Neisser views perception as a continuous cycle between the selection of aspects of the visual stimulus for exploration, which is guided by an anticipatory schema and stimulus information which modifies the schema. Gale et al conclude that adequate detection of abnormality and diagnosis depend on this dynamic process. Like Kundel and Nodine, they argue that the appropriate selection of an anticipatory schema is dependent on the radiologist's prior experience.

2.2 Radiological Expertise

Much of the work on radiological skill discussed so far has been primarily concerned with the visual aspects of radiological skill. Findings such as Kundel et al's (1978) that possibly only as few as 10% of errors in radiological diagnosis are due to failures in visual search have precipitated a shift of interest towards the more cognitive, knowledge-based aspects of radiological diagnosis as, for example, evidenced in the work of Kundel and Nodine (1983). They introduced the notion of a visual concept that shapes perception of the radiograph in a top-down fashion. The implication is that past experience is critically important in the perception of abnormality. In recent years, however, a different approach to studying complex skills in a variety of domains has been developed which draws on the theories and methods of cognitive psychology. A number of studies on radiological expertise have been conducted within this framework.
2.2.1 Empirical investigations of radiological expertise

Expertise research is typically concerned with the learning of complex skills as evidenced in the cross-sectional analysis of skill displayed at different levels of expertise. A number of studies concerned with the analysis of the cognitive elements of radiological expertise have been conducted within this framework (Lesgold, 1984, Lesgold et al, 1988, Myles-Worsley et al, 1988). Another source of interest in expert performance has come from Artificial Intelligence research, which aims to build machine models of expertise in order to understand the types of knowledge and processes involved in performing complex skills. Practical outcomes are, for example, the development of decision support systems for radiologists (Gale, Roebuck, Riley and Worthington, 1987, Getty, Pickett, D'Orsi and Swets, 1988). In this section, empirical investigations of expert-novice differences in radiological diagnosis and AI approaches to radiological skill are discussed. In addition, theoretical orientations arising from this work are introduced.

2.2.1.1 Expert-novice differences in radiological skill

The most comprehensive cognitive skill analysis of radiological diagnosis has been conducted by Lesgold and his co-workers (1984, 1988). Like much of the visual search research, they employed chest radiographs as stimuli, but the methods and aims of this research are quite different from those of visual detection research. The primary concern is with the interpretation of abnormality in X-rays rather than its detection. The idea is that the assignment of meaning to individual perceptual features has to be seen as resulting from a complex decision making process. The expert radiologist is thought to build a mental representation of the patient during diagnostic reasoning. As in other medical reasoning, knowledge of disease processes is thought to play a central role in interpreting visual medical information. The nature of such medical knowledge has been conceptualised with reference to schema theory (Patel and Groen, 1991).

Lesgold et al used a naturalistic diagnostic task for which subjects were given X-
rays to examine and then asked to make an oral diagnosis. In a second experiment, radiologists were asked to trace critical features of the X-ray on an acetate overlay in addition to reporting verbally what they were seeing. Verbal protocols were content-analysed with respect to the number of findings, causes and effects reported. Chains and clusters of these were also quantified. It was found that experts, who had typically more than 10 years of experience, outperformed novices (1st and 2nd year radiology residents) and intermediaries (3rd and 4th year radiology residents) on all measures. Experts were also found to engage in more inferential thinking, which caused them to have a more coherent model or 'schema' of the patient.

According to Lesgold et al, the central role of schemata in radiological diagnostic reasoning also becomes apparent in other subtle differences between experts and novices. Firstly, schemata are important for the storage of anatomical knowledge and thus serve to identify abnormal features. They found that experts were more able to adjust their expectations of what constitutes abnormality to a particular case, and were therefore better able to distinguish abnormality from normality. A related finding was that experts evoked schemata more quickly. Lesgold et al have presented protocol data regarding the consistency of reported findings with the relevant disease schema which show not only that preliminary schemata play a role in shaping experts' diagnostic behaviour, but also that these are triggered very early. Such schemata, which are triggered within the first few seconds of search, guide the further evaluation of the film. A further difference between experts and novices is that schemata of novices are more rigid and less well tuned than those of experts especially when presented with ambiguous or confusing cases. Experts are also more opportunistic with regard to new information. It was shown that they were able to deal more efficiently with new clinical information about a case by adjusting their diagnoses accordingly. Novices, on the other hand, often stuck with wrong diagnoses even in view of contrary additional information. Lesgold et al have suggested that rigid schema use may either be due to inefficiency of diagnostic subprocesses, such as the simultaneous consideration of alternative
hypotheses, or to a failure to generate sufficient hypotheses. A third explanation might be that novice lack the fine-tuned feature discrimination ability of experts. That schemata do indeed drive perception is evidenced in the fact that experts actually see things differently. The data from the tracing task showed that novices were liable to mis-perceive abnormalities and assign them to normal categories. The drawing data demonstrated that features were seen differently by those subjects who did not produce an accurate diagnosis.

Myles-Worsley et al (1988) set out to investigate experts' apparent ability to see more detailed feature information in ambiguous X-rays than novices. They argued that expertise in radiology is characterised by the ability to attend selectively to distinguishing features in X-ray films rather than variations of normal features. The hypothesized perceptual advantages of expert radiologists were tested in a memory task using both normal and abnormal films. A face recognition task was employed as a control condition. Comparison with less experienced radiologists and novices revealed that recognition of abnormal films increased with expertise, whilst recognition of normal films decreased to chance level for experts. Face recognition was high for all levels of experience, and expert radiologists recognised abnormal films with approximately the same accuracy as they recognised faces. Myles-Worsley et al concluded that there are two types of knowledge engendered in radiological expertise. Knowledge of characteristic features of normal exemplars permits rapid and automatic processing of normal radiograph features, while knowledge of abnormal features allows the fast detection of pathological conditions. Expertise renders the observer selectively sensitive to abnormal features.

2.2.1.2 Artificial Intelligence approaches to radiological expertise

Nodine (1992) has argued that the way to reduce error in radiological diagnosis is to make radiologists better decision makers. That decisions in radiological diagnosis are subject not only to inter-individual differences but also to intra-individual differences has been well documented (Garland, 1959, Gale et al, 1989). While the former can often been explained in terms of variability in the decision
criterion, the case of intra-individual differences is not as clear. The problem is amplified in diagnostic radiograph interpretation without the search component of the task (Swensson et al, 1980, 1985).

Although a number of computer-based techniques for extracting 'hidden' information from images have been developed, none of these have been primarily concerned with modelling radiological expertise. This is because, the image processing skills implied in such expertise are difficult to implement given currently available, essentially serial hardware (Garnham, 1988, Raff, 1993). Nevertheless, artificial intelligence techniques have been applied to the problem of radiological diagnosis, in particular with regard to making decisions in image interpretation. There have been a number of attempts to develop decision support systems for radiologists. This involves extracting knowledge about features and their relationships with radiological abnormality from experts which is then implemented in a system designed to aid radiological decision making. Two approaches to modelling expert knowledge for decision support in mammography are those by Gale et al (1987) and Getty et al (1988), which are discussed further below.

2.2.1.3 Diagnostic Decision Aids

Complex decision making is particularly pertinent in radiological subfield which involve the interpretation of relatively complex abnormalities. A major problem in mammography, for example, is the distinction between benign and malignant disease. Gale et al's (1987) research was concerned with increasing specificity in mammographic diagnosis, i.e. the correct classification of benign lesions as not malignant. Mammograms of patients who had been biopsied on clinical or mammographic grounds were re-examined for the presence of a set of pre-determined features. Discriminant analysis was used to reveal which features discriminated between benign and malignant cases. Resulting feature patterns were classified as representing benign or malignant disease on the basis of biopsy results. They were then used as a basis for predicting biopsy outcome in new patients.
Comparison with radiologist performance showed that the feature approach accounted for much higher specificity but at the cost of some sensitivity. A computer decision aid, the Mammographic Computer Assisted Diagnosis program (MAMCAD) was developed, which provides the radiologist with instant decision feedback based on previous cases.

 Getty et al (1988) essentially pursued the same aim in their approach to error reduction in mammographic image interpretation, the outcome being the development of a computer decision support device, the computer classifier. The main difference to Gale's study is Getty's much more elaborate procedure in extracting experts' knowledge. An initial feature list was generated from interviews with experts which was subsequently subjected to scaling analysis of perceptual-similarity ratings using multidimensional scaling and hierarchical clustering. The outcome of these analyses were discussed by a group of experts, culminating in a master list of relatively few features. These were discriminant analysed to determine a small set of necessary and sufficient features and their relative weights, when merged into decisions. These features and corresponding weights were implemented in a computer classifier. Performance comparisons revealed that the computer classifier performed better than general radiologists under standard interpretation conditions. Enhanced reading conditions, which involved using the computer classifier as an aid, improved the average performance of mammographically inexperienced radiologists to an extent that it was comparable to that of mammography specialists. ROC analysis revealed that moving from standard to enhanced reading increased both specificity and sensitivity by .07 at the negative diagonal, where sensitivity and specificity are equal. This stands in contrast to Gale et al's decision aid, which improved only specificity. Getty et al concluded that the computer classifier provides an important tool in the human performance of diagnostic interpretation, which is, at present, difficult to automate fully.
2.2.2 Models of radiological expertise

As with visual search models, two alternative types of models are apparent in the work on radiological expertise. One approach emphasizes reasoning and decision making processes while the other places more importance on feature evaluation and recognition. As each model accepts that both, perceptual and higher cognitive processes, have a role to play in radiological expertise, the differences between them are not simply the result of differential emphasis on the perceptual or decision making aspects of the task. The different accounts arise from the assumption of different cognitive processes underlying radiological diagnosis. The 'top-down' or experience-driven models assume that radiological interpretation proceeds on the basis of expectations hard-wired into cognitive structures, while the 'bottom-up' or data-driven models presume that an analytic feature-based approach informs radiological decision making.

2.2.2.1 The medical expertise model
Lesgold et al (1984, 1988) have made extensive use of schema theory in their analysis of radiological expertise. Not unlike the visual top-down models of radiological skill, they postulate that schemata guide the diagnostic process. Lesgold's schemata, however, are not visual schemata in Arnheim's (1969) or Neisser's (1976) sense but rather knowledge structures incorporating both visual and medical knowledge. Perceptual knowledge in radiology, according to Lesgold, can be best described in terms of the pandemonium model of perception (Selfridge, 1959). Pandemonium can describe the simple perceptual properties of a radiograph which form the basis for diagnosis. If there is ambiguity with regard to alternative diseases, which may be represented by similar sets of features, further information can be accessed by 'looking again'. The cognitive aspects of radiological expertise are dependent on perceptual capability, but the more important aspect of radiological expertise in Lesgold's model is the diagnostic decision making process. This involves distinguishing between alternative disease patterns. The knowledge about different disease patterns and associated visual
features is stored in the form of schemata or sets of rules, which guide diagnostic
decision making.

Lesgold's work, thus, draws to a large extent on the literature on expertise in
complex cognitive skills. Research on medical diagnosis, one of the domains which
has been studied extensively within this framework, has recently identified a
number of special features of medical knowledge and the diagnostic task that
distinguish medical expertise from expertise in other domains (Gilhooly, 1990). For
example, novices and experts do not remember new information in the medical
domain as well as intermediaries, while in other expertise domains, there is a linear
relationship between skill level and memory. One of the reasons for this may be
the ill-defined nature of diagnostic decision making. Schmidt, Norman and
Boshuizen (1990) have argued that medical knowledge is stored in terms of ‘illness
scripts', which are prototypical representations of particular diseases. Diagnostic
decisions are made on the basis of similarity to previously encountered cases. The
prototype theory of medical expertise can account for otherwise counter-intuitive
findings in the visual medical domains, such as errors being associated with longer
viewing times (Kundel et al, 1978). Thus, errors do not arise because the
radiologist is unable to apply appropriate rules, but because abnormality is not
recognised due to a lack of prior experience with similar cases.

2.2.2.2 Bottom-up models of radiological expertise

Some studies of radiological expertise have emphasized the importance of low level
perceptual processes in radiological diagnosis. The model of expertise underlying
such work is one which assumes that the identification of significant features is
sufficient to build a representation of a case for diagnosis. Getty et al's (1988) work
is an example of this approach. They have argued that facilitating the visual
assessments of relevant image features and the merging of these assessments into a
diagnostic decision is sufficient for error-free diagnosis. Consequently, much of
their work is concerned with the identification of such features.
Myles-Worsley et al's (1988) have also argued that feature identification is at the heart of the diagnostic process. Although they believe that the perception of normality is guided by expectations based on knowledge of normal chest anatomy, abnormal features serve to build a representation of disease. In effect, Myles-Worsley et al subscribe to a hybrid model which accounts for the rapid and automatic processing of normal structures in terms of top-down processes, while postulating bottom-up processes for the diagnosis of abnormality.

2.3 Acquisition of radiological skill

Many of the ideas regarding radiological skill acquisition are implied in the discussion of work on visual search in radiology and radiological expertise. This section serves to make explicit the salient points regarding the acquisition of radiological skill. Radiologists' education has been a concern for a number of years (see Calhoun, Vydareny, Ten Haken and Blane, 1988 for a review). Empirical research on radiological skill has served to inform educationalists' design of training programs for radiologists. In this section, implications of empirical evidence and models of radiological skill for the acquisition of the skill are examined. The impact of these theoretical orientations on training design and practical radiology education are discussed in chapter 8.

2.3.1 Visual search and skill acquisition

One of the profound problems of radiological education is the extent of individual differences in performance even after training. Garland (1959) has suggested that correct viewing of films especially with regard to systematic viewing procedures may serve to reduce the problem. Thomas (1976) has developed a list of advice to the observer from the evidence provided by eye-movement studies which is primarily concerned with improving search strategy. The outcome of visual search research with regard to improving radiology performance, thus, seems to centre on an attempt to elaborate visual scanning strategies in order to persuade the observer
to visually 'cover' more of the primary visual data.

The potential of such strategy training was investigated by Kundel and LaFollette (1972). They found that there was a definite evolution of search strategy between the untrained observer to that of the expert radiologist. This did not, however, depend on formal training, but rather on knowledge of radiological anatomy, pathology and clinical medicine. In addition, familiarity with the features constituting abnormality and the nature of the background were found to be critical for the efficient utilisation of search mechanisms. They concluded that there is more utility in the development of clear and unambiguous definitions of normal and abnormal than in search strategy training, since such knowledge is more effective in improving search patterns. Similar conclusions were reached by Gale and Worthington (1983). Although they found that observers could be trained to follow simplified optimal search strategies, the benefit from such training was thought to be doubtful for two reasons. Firstly, using a prescribed strategy seemed to cause readers to adopt a more liberal criterion as to what constitutes abnormality. Secondly, prescribed search patterns may disrupt adequate search (see also Swensson et al, 1985), because search is driven to some degree by visual information contained in the display.

It thus seems that the acquisition of visual search skill depends on learning about features, their locations and their relationship to normal and abnormal anatomy. Such knowledge results automatically in more informed and therefore more efficient visual search of the display. Nodine (1992) has argued that in order to improve observers' performance in radiology, training needs to be focused on making them better decision makers rather than better searchers.

### 2.3.2 Acquiring radiological expertise

Top-down and bottom-up models of radiological expertise result in different predictions as to how such expertise is acquired. If radiological skill is essentially
based on the acquisition of complex cognitive knowledge structures, these are likely to be developed through extensive experience in radiology. According to this model, little can be done to aid the development of task-specific expertise, which is typically acquired over a number of years. On the other hand, if appropriate feature analysis is responsible for accurate radiological diagnosis it should be possible to teach radiologists an understanding of the nature, structure and location of significant features which can be used to build up a representation of disease processes observed in radiographs.

An example of a top-down model of radiological expertise is Lesgold et al's (1988) description of the acquisition of radiological skill as the fine-tuning of schemata brought about by "... a cognitively deep form of generalization and discrimination." (p.340), a process which has been likened to general cognitive development. This process brings about the development of expertise through interaction with the object of expertise, the radiograph. Radiologists learn to diagnose radiographs through practice. An expert radiologist has inspected approximately 200,000 X-rays in his or her professional life. Experience with and exposure to radiographs is, thus, at the heart of acquiring radiological expertise, according to the medical model. Such experiential learning is even more important if the knowledge structures implicated in radiological expertise are not rule-based but analogous to 'illness scripts' (Schmidt et al, 1990). Knowledge representation of this kind is more flexible with regard to facilitating rapid diagnosis, but it is also 'fuzzier' in that it is prototypical rather than well defined. Gilhooly (1990) has argued that clinical experience is probably the best route to the acquisition of the prototypical medical knowledge required for medical diagnostic skill. The question remains, though, whether this is true for radiological diagnostic skill which depends not only on medical knowledge, but also on the interpretation of tangible visual data.

Getty et al (1988) have proposed that the acquisition of radiological skill involves two elements, the appropriate visual assessment of relevant features of the radiograph and the combination of these assessments into a diagnostic decision.
Prior experience does not play such a great role in this approach, and knowledge structures are about radiographic features rather than disease. Radiological skill acquisition is, thus, about acquiring feature knowledge, and expertise in radiology is evidenced in extensive knowledge about features and their weights in indicating normality and abnormality. Getty et al's work is based on the belief that extracting such knowledge from experts and making it available to less experienced observers can, in fact, improve interpretation performance of the latter.

The question of whether high level cognitive knowledge or low-level perceptual knowledge is of greater importance in radiology may to some extent depend on the radiological speciality studied. Rubin (1989) has argued that in most radiological disciplines both types of expertise are likely to operate side by side. One way in which this might be conceptualized is, for example, Myles-Worsley et al's (1988) hybrid approach to radiological skill acquisition, whereby the processes involved in acquiring knowledge are different for normality and abnormality. It is more likely, however, that accurate radiograph interpretation is dependent on both bottom-up and top-down processes, each proceeding in a recursive and interactive cycle until a decision is reached.

2.4 Summary

In this chapter, major theoretical and empirical approaches to radiological skill and its acquisition were discussed. In essence, radiological skill appears to comprise two major aspects, perceptual skill and medical expertise. While early research into radiological skill was characterised by a pre-occupation with visual perceptual processes in X-ray interpretation, such as simple target detection, more recent research has focused on the cognitive expertise implied in complex diagnostic decision making in the visual medical domains. Theoretical explanations of radiological skill have largely mirrored these two major empirical approaches. Thus, both basic visual processes and complex visual medical knowledge have been
emphasised as important prerequisites for radiological skill. Similarly, different predictions have been made with regard to acquisition of radiological expertise. While some have argued that diagnostic performance is based on prototypical knowledge which is acquired through prior experience with similar cases, the alternative explanation is that radiologists learn about visual features and their relationships with various disease patterns.

It is clearly the case that the different theoretical and empirical approaches to radiological skill do not so much reflect a difference in substance between the different radiological subfields studied, but a difference in emphasis. Most radiological interpretation involves the evaluation of visual data and their interpretation as normal or abnormal structures in the context of knowledge about the appearance of disease patterns. In the following chapter, mammography is introduced as the radiological subfield of particular interest in the current research. One of the concerns of the next chapter is to establish the extent to which research on visual search on the one hand, and on cognitive expertise on the other, is of relevance to the interpretation of mammograms, in particular in the breast cancer screening context.
Chapter 3

Radiological skill in mammography

The majority of research on radiological skill, with the exception of the development of decision aids for mammography, has been concerned with performance in reading chest X-rays. This is probably because chest X-rays are a kind of 'prototypical' radiograph which most radiologists routinely encounter throughout the course of their professional lives. A chest radiograph images an area of the human body which contains bone structures and a number of vital organs. It is, therefore, possible to diagnose a range of diseases from abnormalities detected in chest radiographs.

The present research is, however, concerned with a different kind of radiograph, the mammogram. Mammograms are X-rays of the, usually, female breast and as such represent an image of one organ, the mammary gland. As less anatomy is imaged, disease patterns found in mammograms are less varied than those in chest X-rays. Nevertheless, there are certain similarities between the detection and interpretation of abnormality found in chest radiographs and mammograms, so that many of the studies discussed in the previous chapter are of direct relevance to mammography. This chapter will consider in detail the task of reading mammograms, in particular screening mammograms, and the applicability of results from visual search and cognitive skill studies to mammography. In addition, questions regarding the acquisition of mammogram interpretation skills are discussed. Current training practice is reviewed, and the possibility of training lay film readers for the screening programme is investigated.

3.1 Mammography

Mammograms are taken by compressing the breast vertically (cranio-caudal view), horizontally (lateral view) or diagonally (oblique view), resulting in a two-dimensional image of a three-dimensional anatomical structure. A set of
mammograms consists of two complementary images of the right and left breasts respectively. Like other X-rays, mammograms are very complex images which represent internal structures as shadows of varying brightness (approximately 400 shades of grey) depending on their radiation absorbing qualities. In contrast to chest X-rays, the breast image is of soft glandular, fibrous and fatty tissue only, comprising anatomical features such as lobules, ducts, connective tissue, blood vessels and portions of the chest muscle, but no bone structures or organs. Picture 1 shows a photograph of a normal, oblique view mammogram.

Picture 1 Normal, oblique view mammogram showing images of right and left breasts respectively.

The accuracy of mammography is dependent on a number of factors. In addition to observer factors, Andersson (1986) has distinguished between technical, patient and examination technique factors as independent contributors to mammographic accuracy. Ambiguities in the image may, for example, arise from blurry exposure due to faulty positioning, lack of compression or movement during radiation exposure or from patient factors, such as a high ratio of fibroglandular tissue with
resulting overlapping or merging of structures. Mammography is particularly sensitive to technical factors, as the structures imaged are relatively similar in their absorption characteristics. Thus the use of dedicated equipment is essential in order to achieve maximum contrast.

Breast disease is essentially of two kinds, malignant disease and benign changes. Diagnostic mammography is concerned with differentiating between benignity and malignancy of radiographic features. This can be problematic for two reasons, firstly because features may be ambiguous in their appearance and secondly, because features are only probabilistically related to one or the other decision outcome. The effect of feature indeterminacy on radiologist performance has been documented by Gale, Walker and Roebruck (1989), who found that expert radiologists agreed on the presence of a particular feature in a particular mammogram in only 3.3% of cases. In addition, radiologists' performance may be subject to decision biases. Gale et al (1987), for example, found that radiologists presented with abnormal mammograms were biased towards recommending biopsy.

Radiographic signs for the main types of abnormalities can be indicated by locally increased brightness, distortion in the parenchymal pattern, radial linear structures, asymmetry between the two breast images or a combination of these. Calcifications, which are very common abnormalities in mammograms, can be indicative of benign changes or malignant disease. These appear as small speckles of white and are usually found in a localised area. Masses are represented as local areas of decreased radiolucency which arise from increased tissue density. Some abnormalities are much vaguer in their early stages and may appear as pattern distortions or asymmetry when comparing the two breasts. Two of the main types of abnormality are shown in Pictures 2 and 3.
Picture 2  Mammogram showing abnormality in the upper right breast. This is represented by the spiky white density indicating a mass.

Picture 3  Mammogram showing asymmetry, a more subtle abnormality indicated by the proliferation of white 'fluff' in the lower right breast.
3.1.1 Reading screening mammograms

The decision processes and possible decision outcomes involved in the radiological assessment of X-rays can be conceptualised in terms of a fourfold table which is shown below (Table 3.1). It can be seen that subjective decisions of the observer (recall or no recall) and the objective presence or absence of radiological abnormality can be cross-classified to arrive at the four possible decision outcomes. True classifications occur when the subjective observer assessment of the radiograph matches the actual presence or absence of abnormalities. Errors are due to observer misclassification and arise either when an abnormality is missed (false negative) or when normal structures are misinterpreted as abnormal (false positive).

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>no Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiological</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>abnormality present</td>
<td>(TP)</td>
<td>(FN)</td>
</tr>
<tr>
<td>Radiological</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
<tr>
<td>abnormality absent</td>
<td>(FP)</td>
<td>(TN)</td>
</tr>
</tbody>
</table>

When interpreting screening mammograms two kinds of decision have to be made. Firstly, the observer must decide whether a lesion is present or not and secondly, whether a lesion, which is present, requires further investigation. Reading screening mammograms, thus, involves feature detection on the one hand and two-category classification of the mammogram on the other. Although the problems of
feature indeterminacy and decision bias are as relevant to screening mammography as to diagnostic mammography, there are other additional pressures in screening mammography. It is different from diagnostic mammography because there are a number of additional factors bearing on decision making in the screening situation, which will be discussed further below.

The incidence of breast cancer in the screening population is quite low at between 0.4% and 0.7%. Although many more women undergo further investigation than turn out to have cancer (approximately 5% to 10% women are recalled), the expectation of finding an abnormality that warrants recall in a particular set of mammograms is quite low. Radiologists in breast screening may interpret up to 500 screening films per week of which approximately twenty-five may be recalled. Films are usually examined in sets of between 80 and 100 sets of mammograms, one "roller" per session. They are typically displayed on a 'roller viewer', a large image display device incorporating a 'conveyer belt' display screen. Technicians are responsible for placing the X-ray films on the viewer, enabling the radiologist to view the films without having to handle them. Turn over is, thus, quite high and, on average, a skilled observer will spend less than one minute on each set of mammograms.

One of the advantages of mammographic screening over, for example, breast self examination or clinical examination is its ability to pick up breast disease at a much earlier stage. The implication is that the visual features in mammograms, which are indicative of malignant changes, can be quite minimal and may be easily missed. The combination of high speed mass processing, the low incidence of abnormal mammograms and the potentially missable signs of abnormality bear out the particular difficulties engendered in the screening task. The observer needs to be able to maintain performance under vigilance conditions for a considerable period of time.

A further implication of the mass reading method is that decisions may be
influenced by expectations about base rates. It is well established that base rate information influences perception and decision making under uncertainty in the medical domain, a topic which is discussed in more detail in chapter 5. In screening mammography, the problem arises because significant abnormality is relatively rare. There is also pressure to keep false recalls to a minimum because they decrease the cost-effectiveness of the screening programme, whilst maintenance of a high true positive rate is of paramount importance, if the screening programme is to have any impact on long term public health. Both, decisions to recall and decisions not to recall a particular case may thus be influenced by expectations based on a knowledge of the respective base rates.

Another difference between diagnostic mammography and screening mammography lies in the accessibility of further information. Decisions in screening mammography are normally only informed by one set of oblique projection mammograms and minimal clinical information collected by the radiographic technician at the time of X-raying. Some screening centres in Britain routinely produce two mammograms in two different projections which increases the opportunity for resolving ambiguity at the perceptual level. However, there is no opportunity to resolve observed ambiguities in the X-rays in consultation with the patient or by accessing additional information from medical files.

In summary, it can be said that the interpretation of the visual data is of relatively greater importance than diagnostic decision making in screening mammography. Decisions in screening mammography are largely about the presence or absence of significant abnormal features and about whether or not these features warrant further investigation. There is no need to make a final decision as to what is the most likely disease pattern underlying a particular mammographic abnormality, since this is established during a clinical session with additional mammograms following recall. On the other hand, film reading in screening mammography is not only a visual perceptual task. At the very least, it also involves performance under vigilance conditions and making potentially life-threatening or life-saving decisions.
in a cost-benefit context. Mammogram interpretation can thus be seen to be not simply a matter of visual perceptual ability, but is also influenced by individual personality characteristics. The need to consider personality in selecting observers for screening mammography is discussed in more detail in chapters 8 and 9.

3.1.2 Measuring observer performance in screening mammography

It was argued in the previous section that screening film interpretation is a task with a strong perceptual element. This justifies the use of signal detection methodology for measuring observer performance, although mammographic diagnosis also involves classification, which can render the employment of ROC methodology problematic (see discussion in the previous chapter). Assessment of observer performance in screening mammography poses a number of specific problems which warrant careful selection of an appropriate performance measure. A particular problem with measuring performance in radiological screening programmes is the imbalance of positive and negative cases. In screening mammography clear films outweigh those with an abnormality by a margin of one to ten. In addition, the establishment of the normality or abnormality of a mammogram by the process of radiologic interpretation is not objective. The mere fact that an observer is involved means that radiological diagnosis is based on a subjective evaluation. A radiological abnormality has to be verified by histopathological diagnosis, before there is certainty that the appearance is not just an unusual presentation of normal structures. As this involves surgical biopsy and pathological evaluation, there is frequently a long delay between radiological assessment of a patient and the disclosure of objective outcome of the case. Normal mammograms pose an even more serious problem regarding objective verification. This is because of the problem of 'interval cancers', patients whose mammograms were 'normal' at the last screening, but who presented with cancer in the screening interval. With hindsight, such mammograms sometimes show vague signs of abnormality despite having been classified as normal at the time of screening.
One way of by-passing some of the problems of assessment with ROC analysis posed by mammography screening is to forego measurement of performance in the actual screening situation and to use a well established test set of histopathologically validated mammograms. This reduces the problem of objective verification of radiological abnormality because mammograms with proven histopathological abnormality can be selected for inclusion. 'Normal' mammograms can be chosen after a delay of two to three years in order to exclude the possibility of interval cancers being included. Construction of such a test set also permits the inclusion of a greater number of abnormal films than would normally be found in a screening roller, thus reducing the imbalance between normal and abnormal.

Previous research has frequently employed ROC analysis to analyse performance in mammography (Getty et al, 1988, Gale, Towle, Wilson and Roebuck, 1992). In screening mammography there have been two specific applications of the methodology to performance assessment, quality control and the evaluation of training. Gale et al (1992) have established a "gold standard" set of mammograms, the interpretation of which has been agreed by a panel of mammography experts. This set has been used to monitor the performance of radiologists in the NHS Breast Screening Programme. A similar set was used by Ansell, Davies and Hammond (1990b) to evaluate the screening radiology training programme at one of the NHS training centres. In this study, radiologists read the test set prior to the commencement of training and again after the completion of training. A number of factors have to be taken into consideration, however, if such artificial performance assessment is to have any validity regarding performance in the real screening situation.

In order to compare individual radiologists' performance over time or with each other the locus on the ROC can be measured to give a one-value index of accuracy. There are a number of alternative indices, which measure performance independently of any particular decision criterion. The most common indices in radiological accuracy measurement are variations of the area under the ROC (e.g. 49
 Getty et al, 1988). The fact that there are special problems of performance measurement with the ROC methodology in screening mammography has already been mentioned. In particular, ROCs of screening performance are based on only two points on the curve. While it is possible to express alternative decisions in terms of confidence judgements, which permits the establishment of several points on the ROC, decisions in screening mammography are essentially dichotomous, recall or no recall. In practice, screeners may be asked to provide confidence ratings for assessment purposes, but it has to be borne in mind that this situation does not accurately reflect screening routine. The second problem is the imbalance in positive and negative cases. A number of performance indices are sensitive to such imbalances. This problem is rectified to some extent by the use of a test set with an inflated proportion of abnormalities to normal mammograms. However, the imbalance is still likely to be considerable if the test is to simulate the screening task in a realistic manner. Accurate performance measurement in screening mammography thus depends on the selection of a performance index which is not based on parametric assumptions.

There are some sixteen different accuracy indices which have been identified as suitable for use in measuring accuracy in screening mammography (Ansell, Davies and Hammond, 1990a). Some of these have been applied in other radiological domains, whilst others were originally devised for other purposes. Ansell et al (1990a) report a series of Monte-Carlo studies which aimed to assess the value of these indices in terms of discriminatory power under conditions of varying positive-negative ratios and different sample sizes. These studies identified two indices as suitable in the screening situation, Grier's index of bias (Grier's B) and, in particular, the G index or point symmetry adjusted phi coefficient (Holley and Guilford, 1964, Lienert, von Eye and Rovine, 1987, Hammond and Lienert, 1992).

The G index was selected as the most suitable performance index for the present research. As a phi coefficient it is a correlational measure designed for measuring relationships between dichotomous variables. The G index is a simple measure of
the relative incidence of true and false classifications (see Table 1) and is expressed as follows:

\[ G = \frac{(TP + TN) - (FP + FN)}{N} \]

The G-index is particularly suitable for performance measurement in screening mammography because it always has a common range between -1 and +1 irrespective of the skew in the incidence of abnormality. This means that performance measured on different occasions or for different individuals can be compared even if the incidence varies. Although this is not the case in established test sets, performance on these can be compared directly with performance in a real screening situation. It is thus possible to generalise from individual performance measured as G on a mammogram test to performance in the screening environment.

### 3.1.3 Visual Search in Screening Mammography

The importance of the visual detection aspect of the mammography screening task can be examined in terms of the concepts discussed in connection with previous research on visual search in radiology. Four issues in visual search are of particular interest in connection with screening mammography. These are firstly, the effect of target and background patterns on detection performance, secondly, the effect of different search modes on accuracy, thirdly, the identification of different error types and, fourthly, visual scanning strategies.

Some abnormalities in screening mammography are similar to the lung nodule targets used in many experimental studies employing chest X-rays. In both cases the targets are relatively small and embedded in a complex background. The main difference is that there is a wider variety of appearance in mammographic abnormality compared with lung nodules, which, by all accounts, are fairly uniform densities. Thus, although masses in mammograms can appear as densities similar to lung nodules, there are other abnormalities in mammograms which are entirely
Unlike lung nodules. Calcifications, for example, are very small white flecks, the visibility of which may be more dependent on visual acuity. Similarly, the background properties of chest X-rays are quite different from those in mammograms. Chest X-rays contain various anatomical structures including bone structure, whilst mammograms image one organ and are therefore less complex in terms of content.

Revesz et al (1974) introduced the concept of target conspicuity in order to give a measure of the discriminability of the lung nodule from the lung field. Although the problem of detecting abnormality in mammograms can also be construed as one of figure-ground discrimination, it should be noted that target abnormalities and background patterns are more variable in screening mammography. The observer is required to consider a number of features ranging from local increases in brightness to pattern distortions when attempting to extract abnormalities from mammograms (Moskowitz, 1983). Similarly the background varies in terms of complexity depending on the ratio of fatty to fibroglandular tissue (Wolfe, 1976). In addition, both, targets and backgrounds can be very variable from mammogram to mammogram. The usefulness of the concept of target conspicuity in screening mammography is thus somewhat limited by the complexity its definition involves. This may be one of the reasons why mammogram interpretation is one domain which has largely failed to benefit from advances in image processing (Garnham, 1989, Astley, Taylor, Boggis, Asbury and Wilson, 1993, Raff, 1993).

Target detection has been found to be not only influenced by target-background properties but also by the search mode. Swensson et al's (1980, 1985) counter-intuitive finding that radiologists performed better under search conditions than under no search conditions are of interest in the mammography context. It would be predicted from these results, for example, that diagnostic mammography should be more difficult and less accurate than screening mammography, as it usually involves making decisions about already detected lesions. Swensson et al explained their findings in terms of perceptual processes positively contributing to
detection performance in radiology. However, there is possibly an alternative explanation for their findings in terms of task demands. It has been argued that the decision making part of radiological interpretation is the most difficult aspect of the task (Getty et al, 1988, Nodine, 1992). Thus, poor performance in the no search condition of Swensson et al's experiments may simply be due to the fact that the interpretation task had been cut down to the most difficult element. If, however, Swensson et al's interpretation of their findings were true, perceptual ability would be of great importance for performance in screening mammography, since interpretation of X-rays simply depends on the accurate execution of perceptual processes. An alternative explanation might be that decision making is the weakest element in radiological interpretation skill and, therefore, needs the most support either in terms of training or by means of decision aids. This is the argument made by Getty et al (1988) and Gale et al (1987) for providing decision support in mammography.

Error analysis (Kundel et al, 1978) has also identified decision making as the weak link in radiological interpretation. Although research has been limited to perceptual decision making regarding the presence or absence of abnormality, a number of different error sources could be identified in this context. Kundel et al (1978) found that the majority of perceptual errors in lung nodule detection were decision errors (45%), while search and recognition errors jointly accounted for the remaining 55% errors. These findings may be directly applicable to screening mammography as the task, small target search is similar in some respects, although targets tend to be more complex (see discussion above). Gale and Walker (1993) examined errors made by expert radiologists in mammography. They found that false negative errors (misses) tended to be interpretation rather than detection errors. The extent to which accuracy in reading mammograms is an interpretation and decision making problem rather than a perceptual pattern recognition one is a central issue, which will be discussed in detail in the context of training mammography (chapters 4 to 7).
A number of studies have investigated the extent to which systematic visual scanning can reduce errors in radiology. Despite the fact that experimental investigation has thrown doubt on the benefit of training radiologists to use such strategies (Kundel and LaFollette, 1972, Gale and Worthington, 1983), there is evidence that in mammography advice on viewing strategy is prevalent. Tabar and Dean (1983), for example, devote a considerable section of their mammography teaching atlas to systematic viewing techniques. It is possible that systematic viewing is of greater importance in mammography than it appears to be in other radiological domains because adequate evaluation of mammograms involves a comparison of the two breasts. The observer has to view the two radiographs simultaneously, and abnormality may only be discernible as differences between the two X-rays. For this reason, Tabar et al advocated the use of a sequential viewing strategy and have discussed a number of different masking techniques for detailed comparison of the left and right breasts.

3.1.4 Cognitive Expertise and Screening Mammography

Screening mammography involves making two types of decision, whether an abnormality is present or not and whether an abnormality that is present is sufficiently suspicious to warrant recall. The former decision is the result of largely perceptual processes such as visual search and recognition. This can be conceptualised in terms of image processing mechanisms and may be analogous to scene recognition (Raff, 1993). The second decision is more akin to a diagnostic judgement, although it is an essentially dichotomous decision between benignity and malignancy rather than a decision between multiple diseases. It is this 'diagnostic' classification decision that warrants closer examination in terms of the concepts of cognitive expertise.

It is justified to pose the question to what extent concepts such as schemata or prototypes are useful in mammography. It is clear from Lesgold et al's (1988) work that schema theory has some utility in investigating radiological diagnosis of
chest X-rays, but much of the work on mammography (Getty et al, 1988, Gale et al, 1987) has opted for a more perceptually based approach to expertise. Nevertheless, there is some evidence for the operation of schemata in mammography. Gale et al (1989) concluded that individual differences in the mammographic features reported by experts were due to differences in schemata brought to bear during interpretation of mammograms. Although both Gale et al and Lesgold et al use the same language, they refer to qualitatively different cognitive structures. As both, recent theories of perception and of cognitive expertise, have their roots in general information processing theory, their terminology overlaps to some extent. It is likely, however, that Gale et al's schemata are based on perceptual features and are therefore perceptual schemata, whereas Lesgold et al refer to knowledge structures involved in making medical diagnoses. Lesgold's schemata are thus based on disease features.

Rubin (1989) has argued that both, medical and perceptual expertise are likely to operate in most radiological disciplines. In screening mammography radiologists are likely to bring to bear their knowledge of breast anatomy when making decisions regarding the benignity or malignancy of certain perceptual features. Since breast anatomy is not as complex as anatomy contained in, for example, chest X-rays it might be argued that radiologists' cognitive expertise regarding the relationships between perceptual features, anatomy and malignancy is relatively less important in making a decision than in other radiological disciplines. This is the case especially in view of the fact that positive diagnosis is very rarely made from the mammogram alone, so that the decision to recall for further investigation is often made on the basis of indeterminacy of perceptual features. Nevertheless, in distinguishing routinely between clearly benign cases and cases that are indeterminate on the basis of radiological features, mammographic film readers require a considerable degree of cognitive expertise in radiology based on medical, anatomical and radiological knowledge.
3.2 Learning to read mammograms

The question of what skills are implicated in reading mammograms is of great importance in the context of training observers for the screening programme. Forrest (1989) reports that the manpower requirements for administering breast screening in Britain amount to some 48 radiologists and over 200 radiographers, a considerable pressure on resources. It is clear that the NHS Breast Screening Programme can only succeed if training is appropriate and of the highest standard.

In this section, the current approach to training breast screeners is discussed. The extensive literature on the training of radiological skill will be reviewed in detail in chapter 5. Here, it suffices to describe current training procedures in screening mammography. In addition, the question of cost-effectiveness of radiologists' reporting on all screening mammograms is discussed. The issue is whether mammography screening as a consultancy based programme is more efficient than a programme which involves dual reading of mammograms but employs non-radiologist film readers. Research has indicated that dual reading, i.e. the viewing of all screening mammograms by two independent observers, may serve to reduce observer error (Yerushalmy, Harkness, Cope and Kennedy, 1950, Haiart and Henderson, 1991). There is some evidence that medical technicians may be able to learn to read mammograms in a dual film reading context.

3.2.1 Current training practice for mammography screeners

There are currently four training centres in Britain whose brief it is to train breast screening staff to "... appropriate levels of competence." (NHSBSP, 1991, p.15). Training is conducted by secondment to one of the centres for a two-week period. The main premise of the current training protocol, which claims to be based on the principles of adult education (NHSBSP, 1991), is that screeners only require a small amount of initial training followed by support during their initial time in screening. This stands in contrast to the guidelines on training provided by the
European Group for Breast Cancer Screening (1987) who have advocated an initial training period of two to three months based on experience with the Swedish national breast screening programme.

3.2.1.1 The training schedule

At the present time, screeners trained for the British breast screening programme are almost exclusively fully trained radiologists, since the breast screening service aspires to be a consultancy based service. Radiologists are likely to have some degree of prior experience with mammography, albeit diagnostic mammography. For this reason training does not directly focus on the perceptual features of abnormality but rather on distinguishing between normal and abnormal mammograms and in particular on detecting minimal abnormality. Learning is essentially experiential in nature.

During their attendance at the training centre trainee screeners partake in all aspects of the daily screening routine. Thus, they read rollers of screening mammograms, discuss cases with resident expert screening radiologists and attend assessment clinics and pathology meetings. In addition, they have the opportunity to study files of abnormal teaching mammograms and standard mammography teaching material such as Tabar's mammography teaching atlas (Tabar and Dean, 1983). In all, the training is conducted in a 'tutorial' fashion whereby the trainee acts initially as an observer and later as an apprentice. Performance is thought to continue improving after training through exposure to large amounts of 'real' screening mammograms.

3.2.1.2 Problems with current training

Although there is some evidence that current training improves performance (Ansell, Davies and Hammond, 1990b), the overall credentials of the training protocol have not been established in a systematic manner. Recently, there have been some moves towards quality control and radiologist performance monitoring in the NHS breast screening programme (Gale et al, 1992). However, the long term effect of training and the benefit of daily screening experience are not proven. This
presents a problem, especially because NHS training requirements are substantially less than those in other national screening programmes. In addition, the skills acquired during training are not well defined. One characteristic of the tutorial approach is that trainees learn to emulate the expert. The success of such a method is critically dependent on the material to be learned. It is a feature of expert performance that much of the skilled behaviour is automatic and possibly consciously inaccessible. It may, therefore, be very difficult if not impossible for them to communicate their skills to a trainee. The trainee, on the other hand, is not likely to benefit from observing an expert if the skills in question are largely perceptual or cognitive. This is acceptable in radiologist training who might reasonably be expected to possess the perceptual skills prerequisite for mammography. However, there is some indication that shortages in radiological expertise in screening mammography have led to the recruitment and training of general medical practitioners who may require more specific instruction on perceptual aspects of film reading.

In practice the tutorial method of training in screening mammography probably results in improving clinical reasoning about screening mammograms. Although current training is closely modelled on the manner in which medical professionals best acquire their expertise (Gilhooly, 1990) it is not clear that this is the most effective procedure in learning to distinguish between normal and abnormal mammograms, especially when non-radiologist trainees are involved. On the other hand, screening radiologists spend only a small part of their time reading screening rollers. The skills acquired during training may well be very beneficial for the clinical management of women with abnormalities.

3.2.2 Radiologists or radiographers?

Reading screening rollers is one of the screening radiologist's routine tasks. The question is whether radiologists are uniquely equipped to perform this task or whether it could be performed by other suitably selected and trained individuals. As
indicated above, medical practitioners have already taken on mammogram interpretation duties in some screening centres. The argument for lay film reading becomes particularly pertinent in the context of dual film reading. It has been argued above that the skills involved in this mammogram interpretation are probably dependent to some extent on medical expertise. Nevertheless, it might be possible to train lay film readers in pattern perception skills required for identification of abnormality. One of the main problems in radiological mass screening programmes are errors due to missing abnormality. Yerushalmy et al (1950) and Garland (1959) have discussed this problem with reference to the tuberculosis screening programme. They have suggested that accuracy can be increased by dual reading of radiographs either by two independent observers or by the same observer on two occasions.

As discussed in chapter 1, dual reading of mammograms is not amongst the provisions of the original breast screening protocol (Forrest, 1986). However, some screening centres have adopted dual reading procedures, resources permitting. As the screening programme is nearing full implementation, this status is difficult to maintain in view of the shortages of radiological expertise in screening mammography. One solution to the problem is the selection and training of technical staff, e.g. radiographers, to take on some of the routine evaluation of screening mammograms. Saxton (1992), for example, has argued that screening mammography is one of the areas that could benefit from training suitable radiographers for film reading. This would both, enhance radiographers’ jobs (Swinburne, 1971) and free radiologists from some of the routine work for the highly skilled clinical management for which they were trained.

3.2.2.1 Lay film readers in breast screening

There is evidence from breast screening in other countries that training non-radiologists to interpret screening mammograms is feasible. Breast screening in America, for example, encountered early problems with expert radiologist resources. Dowdy, Lagasse, Roach and Wilson (1970) argued that, despite the
theoretical benefits breast screening would have in terms of mortality from breast cancer, there were simply not enough radiologists to guarantee these benefits in practice. However, they were able to demonstrate that it was possible to train a secretary and a radiographer in mammography film reading by the tutorial method. Both lay trainees showed correlations of 90% with radiologists' film interpretations.

A more comprehensive investigation of training lay film readers for mammography was conducted by Alcorn and his co-workers (Alcorn and O'Donnell, 1968, 1969, Alcorn, O'Donnell and Ackerman, 1971). They used a flow chart method (Tuddenham, 1968, Tuddenham, Houser, Tuddenham, Booth and Matthews, 1969) to train radiographers to recognise abnormal patterns in mammograms. Comparison with an expert radiologist's decision revealed that the non-radiologist observers detected as many abnormalities as the expert but at a cost of a high false positive rate. The outcome of comparing radiographer performance to pathological results yielded similar results. Alcorn et al concluded that the training of radiographers as mammography screeners was feasible, although the high false positive rate needed to be addressed. In the dual screening context, high false positive rates by lay film readers would not necessarily present a problem, but would indeed be expected. Radiographers trained in detecting patterns of abnormality might be able to function as filters whereby all such patterns are referred to a radiologist for interpretation and rejection of all benign abnormalities. This would be expected to result in a higher true positive rate, but not necessarily in a higher false positive rate.

The training of radiographers as mammogram screeners has also been advocated in the Dutch breast screening programme. Rombach (1980) suggested the use of radiographers as pre-screeners in analogy to cervical cancer screening. Cytological analyses of Pap tests are routinely conducted by cyto-technicians. Only those classified as abnormal plus a percentage of the overall screening sample are examined by an expert pathologist. Rombach argued that suitably trained radiographers could do as well as experienced radiologists in recording
mammographic signs but that actual decisions regarding abnormality have to be made by experts with knowledge of "...all kinds of mammary cancer images on the mammogram." (p.97).

In practice, training of lay film readers for a mammogram interpretation function has been successfully completed in some countries. In Holland, for example, radiographers have been routinely trained in mammogram interpretation, so that they can make independent decisions on the presence or absence of significant abnormality. This enables them to implement the Dutch policy of taking additional mammograms for clients with an abnormality without having to refer to a radiologist first. In Sweden, radiographers have also been successfully trained to read mammograms (Bjurstam, 1993). The successful employment of radiographers as mammogram film readers in other national breast screening programmes demonstrates that there is considerable potential for training of lay screeners in mammogram interpretation in the NHS Breast Screening Programme.

3.3 Summary

This chapter was concerned with introducing specific issues of relevance to X-ray interpretation in the context of breast cancer screening. In particular, special characteristics of mammograms and mammographic abnormality indicative of breast cancer were discussed with reference to previous research on perceptual skill and medical expertise in radiograph interpretation. The indication was that film interpretation in screening mammography was based on both, visual search skills and interpretation skills mediated by medical knowledge. The issue of the skills implied in mammogram interpretation performance is investigated on a case study basis in the following chapter, but will be dealt with in more detail in chapters 5 to 7.

Observer performance measurement in screening mammography was another issue to be examined in some detail in the present chapter. Although X-ray interpretation
has often been construed as a visual detection task which can be analysed by methodologies derived from signal detection theory, the task also has clear classification elements. Screening mammography yields particular problems for the employment of ROC methodology in assessing observer performance. The G-index or point-symmetry adjusted phi-coefficient was identified as the most appropriate performance measure in screening mammography. For this reason, the G-index was used as a performance measure in all empirical studies involving mammogram interpretation which are reported in this thesis.

Another issue of considerable importance to the success of breast screening in Britain relates to training of skilled observers to interpret mammograms. One way of improving observer performance is to read all mammograms twice, which might be achievable financially, if lay film readers can be trained for mammography film reading. Previous research and practice in other national screening programmes appear to indicate that lay film reading is a viable option. The applicability of this in Britain is examined in the next chapter, in which an in depth longitudinal study of two radiographers learning to interpret mammograms in Britain is presented. A major issue in this context is the extent to which current British breast screening training practice is suitable for teaching lay people mammogram interpretation skills.
Chapter 4

Acquiring radiological skill: A case study of two radiographers

In this chapter, a case study of two radiographers, who were trained to read mammograms at one of the NHS breast screening training centres, is reported. The research was conducted, firstly, to establish the utility of current training procedures for the instruction of technicians with no medical background and, secondly, to present an assessment of the long term benefit of training. In addition, the study sought to investigate the nature of the skills involved in reading mammograms and how they develop during training.

4.1 Introduction

The role of radiographers in breast screening is currently a highly skilled but purely technical one. It is their job to carry out the X-ray examination, including all aspects of dealing with the women under examination and the processing of the film material. In the majority of cases the radiographer is the only person to see the woman. Thus, they represent the human interface of the breast screening service. Nevertheless, there is a distinct lack of possibilities with regard to career development in radiography, particularly in breast screening radiography. The problem is that, on the one hand, highly skilled, highly motivated and committed radiography staff are required to run an efficient breast screening service, whilst, on the other hand, there is no reward system that promotes staff satisfaction (Swinburne, 1971).

Interpretation of screening films represents one possible enhancement to the radiographer's job. As discussed in the previous chapter, the feasibility of training technicians to read routine X-ray films has been established in other national breast screening programmes. Apart from enhancing the radiographer's career path, there are other benefits in terms of cost-effectiveness of the screening programme. Dual reading, with all its associated advantages of higher cancer detection rates, could be
considered as a routine option at virtually no extra cost to the overall breast screening programme. In addition, radiologists can reduce their considerable routine workload with the possibility of redeploying their efforts in the highly skilled diagnostic work for which they were trained.

4.1.1 Training

Current training procedures have been described in some detail in the previous chapter. The main feature of NHS training compared to that of other countries' programmes is the brevity of initial instruction. The two-week secondment to a training centre represents only a quarter or less of the time spent on training by breast screening radiologists in other European countries, e.g. Sweden. There have been no long term follow-up studies to demonstrate that cancer detection performance of radiologists actually improves further after initial training. Although the acquisition of expertise is thought to be a function of continued exposure to normal and abnormal mammograms during the daily screening routine, there is no direct evidence that this is the case.

One question arising in the context of training technicians for mammogram interpretation regards the suitability of the present tutorial approach for their training needs. As discussed previously, the tutorial approach to training in breast screening may primarily result in the development of clinical management skills rather than good discrimination performance in film reading. Although Dowdy et al (1970) have shown that the tutorial method can produce good detection performance in non-radiologists, there may be more efficient training practices for lay film readers. As radiographers are not expected to partake in the clinical management of women with breast cancer, the training time relating directly to such activities may be better spent on a more structured approach with regard to detecting abnormalities in mammograms (Alcorn et al, 1971).
This issue is related to a second question regarding the desired training outcome. As indicated in the previous chapter, there are at least two alternative models of radiographer involvement in mammogram interpretation, the pre-screening model (Rombach, 1980, Alcorn et al, 1971) and the dual reading model (Dowdy et al, 1970, Haiart and Henderson, 1991). The former places great emphasis on feature detection, while discrimination is of less importance. In other words, it does not matter if the radiographer over-reads as long as she produces a very high detection rate. The idea is that all clearly normal cases are discarded before the radiologist makes decisions about abnormal, difficult or ambiguous cases. In this model the radiographer acts as a pre-screener. The dual reading model, on the other hand, regards the radiographer simply as 'a second pair of eyes'. The radiographer is expected to achieve a similar ratio of true to false classifications as the radiologist, requiring the adoption of an appropriate diagnostic decision criterion. Discrimination is as important as detection for both radiographers and radiologists in this scenario. The reasoning behind this approach is that dual reading yields a higher pick-up rate than single reading.

The present research was designed to evaluate the long term benefit of NHS breast screening training and screening experience respectively. The two radiographers were followed up over a ten months period during which performance was measured objectively with a test set of mammograms at regular intervals. In addition, actual screening performance was monitored over the same period. This longitudinal design permitted the measurement of improvement in mammogram interpretation performance as a function of training and during subsequent screening experience. It also facilitated the evaluation of the tutorial approach with respect to training outcome. If radiographers were to fulfil a pre-screening function they would have to learn to pick up even very subtle lesions. Performance would be characterised by a much higher recall rate, and subsequently by a higher false positive rate, than that normally obtained by radiologists, i.e. radiographers would have to adopt a lax decision criterion. If, on the other hand, training was to produce 'second' readers, radiographers' performance would have to approximate that of
radiologists, whereby a balance must be struck between good pick-up rates (true positives) and low recall rates.

4.1.2 Mammogram interpretation skills

A further focus of interest in this study was an examination of the skills which develop during mammography film reading training. Although radiographers working in breast screening are typically very experienced in examining the technical adequacy of mammograms, they have little or no experience in detecting subtle abnormality. This distinguishes them from radiologists entering training in film reading for screening mammography. Radiologists typically have some experience in interpreting mammograms, although diagnostic mammography is different from screening mammography in terms of the number and the extent of abnormality encountered.

It has been argued before, that mammogram interpretation involves both, detection and classification skills. The nature of the processes entailed in acquiring and performing these skills is somewhat contentious. Although most previous research acknowledges the co-existence of bottom-up and top-down processes in radiological skill, one tends to be emphasized to the exclusion of the other. This may simply be because one kind of process is of greater importance in the radiological subfield studied (Rubin, 1989). In mammography, bottom-up processes have been emphasized (Getty et al, 1988), because mammograms contain a number of easily accessible visual features which lend themselves to bottom-up analysis. This does not, however, imply that learning mammogram interpretation is not essentially the acquisition of normal and abnormal prototypes and that classifications are not made on the basis of similarity to such prototypical representations. Myles-Worsley et al (1988), for example, showed that the acquisition of radiological expertise is evidenced in the ability to recognize normality instantly and that only abnormal features are available for bottom-up analysis.
The present case study of two radiographers acquiring mammogram interpretation skills over a ten months period has provided the opportunity to examine some of the questions regarding radiological skill in mammography and its acquisition. Previous research on interpreting radiographs has employed a number of different methodologies for eliciting the skills and learning processes involved. Kundel et al (1978) recorded eye movements in order to identify sources of detection error. Lesgold et al (1988) elicited verbal protocols which they used to study cross-sectionally the process of acquiring X-ray interpretation skills. The present study is concerned with identifying sources of error in mammogram interpretation in terms of the responses given on performance tests and with elucidating the relationship between feature recognition and classification performance as evidenced in verbal data. Performance was measured objectively in terms of a test set of mammograms, and in comparison with radiologist performance. The latter involved the evaluation of screening performance measured against both, a radiologist's decisions and final screening outcome.

As the current study is longitudinal, it is also possible to measure qualitative change in the interpretation of mammograms as a function of training and screening experience. The processes involved in learning radiological skill can thus be examined in a direct fashion. Lesgold et al (1988) argued that perceptual skills precede cognitive skills in X-ray interpretation. In terms of the present study, this contention would lead to the prediction that perceptual errors should become less frequent as a function of training and expertise. A further issue, which can be directly investigated, is Myles-Worsley et al's (1988) suggestion that conscious access to normal features 'fossilizes' as expertise develops. This leads to the prediction that it should be possible to observe qualitative change in the verbal protocols as a function of exposure to mammograms.
The study was designed to examine in depth the acquisition of mammographic screening skills in two radiographers. The research followed a longitudinal test-training-test design with further performance tests at approximately two- to three-monthly intervals after every 1000 mammograms read during screening routine. For the assessment of actual screening performance a triple-blind film reading protocol was implemented. Training, performance tests and the screening film protocol are described in detail below. The opportunity for this study arose when one of the NHSBSP training centres, the Jarvis Breast Screening Clinic in Guildford, embarked on the training of a small sample of radiographers as a feasibility study for a larger trial. The intention was to train mammography pre-screeners in analogy to the cyto-technician's role in cervical cancer screening.

The two radiographers selected to take part had previously expressed interest in learning to read mammograms and were highly motivated. Radiographer A.S. was forty-two years old and had 21 years radiographic experience prior to her employment in screening mammography. She had been working in breast screening for three years and three months at the commencement of film reading training. Radiographer F. A. was twenty-six and had worked in breast screening for six months. Although both radiographers were experienced in the technical evaluation of screening mammograms, neither had previously interpreted mammograms in terms of clinical abnormality.

### 4.2.1 Mammography Training

The training for the two radiographers was the same as that radiologists would normally receive at an NHS Breast screening training centre. A.S. and F.A. took part in a two-day introductory course which covered all aspects of breast screening and clinical management of women with abnormalities. This was followed by a two-week secondment period which consisted of intensive on-the-job training.
including periods of self study. During this time the radiographers participated in all aspects of the radiologists' daily routine, including reading screening rollers and attending assessment clinics and pathology meetings. They also studied standard mammography training material such as Tabar's teaching atlas of mammography (Tabar and Dean, 1983) and teaching material compiled by the Jarvis Breast Screening Centre, which consisted of abnormal X-rays with descriptions of mammographic findings and pathology reports. The radiographers were encouraged to discuss cases which they found difficult with one of the mammography experts, who would explain why decisions to recall or not to recall had been made in a particular case.

After the two-week secondment the radiographers spent two sessions a week reading screening mammograms (approximately 150 mammograms per week). During this post-training period they received support from more experienced screening staff in accordance with NHS training guidelines. Both radiographers independently read the same films which were then interpreted again by an expert radiologist without knowledge of the radiographers' decisions. The three independent outcomes were compared and any mammograms which had caused disagreements were discussed. In cases of continued disagreement between the radiographers and the radiologist, a second expert opinion was sought for arbitration. The radiographers, thus, continued to receive performance feedback even after their initial training had been completed.

4.2.2 Performance assessment

Performance was assessed with a test set of seventy-nine pairs of mammograms of known histopathological outcome on six occasions, prior to training, after training and on four follow-up sessions after every 1000 mammograms read during normal weekly screening. No feedback on test performance was given at any point. The test set has been employed for the measurement of radiologists' improvement of mammogram interpretation performance as a function of training since 1989.
(Ansell et al, 1990b). A verbal description of the seventy-nine mammograms in the test set by an expert radiologist can be found in Appendix IV. Assessment of radiographer performance with this particular test set was convenient because it was well established and, more importantly, because performance data for nine radiologists were available for comparison. The radiologists had completed the test set on two occasions during their two week training period, prior to the commencement of training and after the completion of the two week secondment. Pre-and post-training performance of the two radiographers could thus be directly compared to a group of radiologists who had received the same training. The test set consisted of sixty-three normal mammograms and sixteen radiologically abnormal ones, including eleven cancers, three benign biopsies and two recalls which further mammograms showed to be clear.

The radiographers recorded their responses on an answer sheet modelled on the recording sheets used during normal screening. This required them to report whether any of the four main classes of abnormalities in mammograms, asymmetry, calcification, rounded and irregular masses, were present and to record a graded decision outcome. In addition, a small schematic diagram of the two breasts as seen in an oblique projection mammogram was provided, in which the location of any reported features had to be indicated. An example of a record for one mammogram is shown in Figure 4.1. Although each record provided space for two radiographer reports and a radiologist report, in practice each of the three observers reported mammograms independently on separate sheets during normal screening. The additional space facilitated transcriptions for computer entry of the data and is irrelevant to reporting on mammograms in the test set.
In addition to the information recorded for each mammogram, verbal protocols of the performance test assessment sessions were obtained. The radiographers were instructed to 'think aloud' when interpreting the mammograms according to the instructions provided by Ericsson and Simon (1984). In their review of protocol analysis, Ericsson and Simon (1984) have presented extensive evidence that verbalisation does not change task performance, if the instructions do not induce such a change. There might be a problem with verbalisations about mammograms, because of the explicitly perceptual nature of the information they contain.

However, the recording of features requires the radiographers to translated this visual information into verbal codes, so that verbalisation of thoughts should not pose an additional burden on processing during mammogram interpretation. Verbal protocols were obtained by tape-recording radiographers whilst 'thinking aloud', when reading the test roller on all six occasions.
4.2.3 Triple-blind screening film reading protocol

In order to assess performance in the real screening situation, data from the first ten months of screening were collected according to a triple-blind film reading protocol. This is outlined in Figure 4.2. The two radiographers were encouraged to read the same screening mammograms each week, which were also interpreted by a radiologist. Both, the radiographers and the radiologist made decisions on each mammogram independently of the other two and without knowledge of the others' decisions. This procedure effectively implements a double reading scheme in which the two trainee readers simultaneously represent a second reader. This permits a comparison of the radiographers' performance with each other as well as with the radiologist. In addition, the effectiveness of dual film reading with radiographers can be assessed in terms of screening outcome. Final decisions were made on the basis of consensus or arbitration by a second radiologist in case of disagreements or queries.

Figure 4.2: Triple-blind film reading protocol for radiographer screening film interpretation.
4.3 Results

Two kinds of data were obtained in the objective performance assessment part of this study. Firstly, performance was measured quantitatively using the established test set of mammograms. The radiographers' decision responses were compared to the verified outcome of the mammograms and classified into one of four categories, true or false positives and true or false negatives (see Table 3.1). In addition, all responses were analysed in terms of features identified either in the abnormal category boxes or as drawn into the schematic diagram (see Figure 4.1 above). By considering the feature information contained on the record sheet, it was possible to examine false negative errors, including cases which had been recalled for the wrong breast, in terms of detailed feature reports. This analysis permitted a closer examination of the nature of the more serious kind of error in mammography screening, the missing of a potentially malignant abnormality. The outcome of objective performance measurement during and after training is considered in section 4.3.1, while the analysis of false positive errors is presented in section 4.3.2.

Screening performance assessment was more subjective than the performance measurement with a well-established test set, because of the indeterminate nature of the criterion against which such performance can be measured. In the present study, radiographer performance was assessed in terms of agreement with the radiologist on the one hand, and in terms of final decision outcome on the other. In addition, the benefits of dual reading were considered in comparison with the decisions made by the radiologist alone. The analysis on the agreement between radiographers and the radiologist is presented in section 4.3.3, whilst the effect of dual reading is examined in section 4.3.4.

The second type of data obtained was of a qualitative nature. 'Thinking aloud' protocols were collected in order to investigate the kinds of processes and strategies involved in making decisions about mammograms. The protocols were content-
analysed to examine the relationship between features and decision outcomes as a function of training. This was achieved by means of correspondence analysis (Hammond, 1988). This qualitative analysis of the verbal data is reported in section 4.3.5.

4.3.1 Training and mammogram interpretation test performance

Performance on the mammogram test set was measured with the G-Index (see chapter 3). Figure 4.3 (see also Table Ia.1 in Appendix Ia) shows the G-index for the two radiographers graphed over the six test sessions in comparison with the pre- and post-training performance of the nine radiologists for whom pre- and post-training performance data were available (Ansell et al, 1990b). The first test represents pre-training assessment, the second test was taken after the completion of the two week training period. Tests 3, 4, 5 and 6 were completed after each one thousand mammograms read during normal screening routine.

![Figure 4.3: Pre- and post training and follow-up performance of the radiographers, compared with pre- and post training performance of a group of radiologist, measured with the G-index.](image)

It can be seen that both radiographers reached a similar level of performance after training (test 2) and remained similar throughout subsequent tests. The greatest
change in performance takes place between the first and second test session during the training period. Performance on the four subsequent readings remain more or less at the same level for both radiographers. Comparison between radiologists and radiographers clearly demonstrates that although radiologists start out with better performance before training, both, radiologists and radiographers achieve the same performance level after training.

A more detailed idea of the changes occurring during training and subsequent screening experience can be gained from an examination of the respective true positive and false positive data. These are plotted separately for each radiographer in Figure 4.4 (see also Tables 1a.2 and 1a.3 in Appendix 1a). The main improvement after training is manifest in a much lower false positive rate in both cases. The proportion of false positives is relatively stable over the remainder of the follow-up period. True positive rates are not so directly influenced by training and remain much more variable during screening experience. Although A.S. initially shows steady improvement in picking up abnormalities as a function of training and experience, F.A.'s true positive rate goes down immediately after training but increases with further experience. In summary, the true positive rates of both radiographers are less stable than the post-training false positive rates.

Figure 4.4: Probabilities of false and true positives

Figure 4.4a: A.S.  
Figure 4.4b: F.S.
Comparison of the true and false positive rates achieved after training (test 2) by the radiographers with those obtained on average by radiologists (Ansell et al., 1990b) shows that the two technicians perform at a level similar to that of the doctors. The radiographers on average picked up 72% of abnormalities compared with 83% identified by the radiologists. False positives were lower for the radiographers with 16% of classifications compared to radiologists' 18%. The slightly lower pick-up rate of radiographers can be explained in terms of a lower recall rate at 27% compared with radiologists, who recalled on average 31% of cases. In comparison with the radiologists, the radiographers appeared to employ a slightly more stringent decision criterion after training.

4.3.2 Error analysis

None of the radiographers or radiologists achieved a 100% pick-up rate of abnormalities. This may be a function of the difficulty of the mammograms used in the test. However, an analysis of the mammograms which had produced either false negative or false positive errors for the two radiographers after training revealed both intra- and inter-individual differences for the two radiographers. Overall 76.5% of abnormalities were missed on at least one occasion by at least one radiographer, but the overlap between those abnormal mammograms that were missed on three or more occasions by both radiographers amounted to only 11.7%. False positives yielded a similar pattern with 52% of clear mammograms recalled on at least one occasion by at least one of the radiographers but only 9% commonality between normal mammograms recalled by both radiographers twice or more. It, thus, seems that variability in performance can not be attributed to variability in the degree of difficulty of individual mammograms alone.

In training radiographers as pre-screeners, it is highly desirable that they should not miss any abnormalities. It is clear from the performance data above that a high true positive rate can be achieved to some extend by trade-off with a higher false positive rate as is, for example, evident in A.S.'s sixth test performance. Film
readers have to achieve a balance between missing as few abnormalities as possible whilst keeping recalls low. However, in training pre-screeners the false positive rate is of much less concern as all cases recalled by the radiographer are viewed again by a radiologist who might be more competent at deciding which abnormalities need to be recalled and which do not. The false negative rate, on the other hand, must be kept to an absolute minimum as misses at the pre-screening level are unacceptable.

Figure 4.5 graphically represent the number of false negatives accumulated by the two radiographers in each test session (see also Tables Ia.4 and Ia.5 in Appendix Ia). For the purposes of this analysis some true positives were included if the woman had been recalled for the wrong breast. The graph clearly shows how little effect training has on reducing false negatives. In the post-training test (test 2) A.S.'s miss rate is reduced negligibly, but substantially increased in F.A.'s case. In addition, both radiographers produce a considerable number of recalls for the wrong breast over the six test sessions. These are correctly identified abnormal mammograms (true positives), but analysis of the feature report revealed that the radiographer has not identified the critical abnormality but has instead based her decision to recall on minor, insignificant lesions in the wrong breast.

Figure 4.5: Total number of false negatives and recalls for the wrong breast

Figure 4.5a: A.S.  Figure 4.5b: F.A.
In Figure 4.6 (Table Ia.4 and Ia.5, Appendix Ia), the sources of the false negative errors including recalls for the wrong breast have been identified. An error due to detection failure was recorded if there was no mention of the relevant abnormal features on either the record sheets or in the verbal protocols. An interpretation error had occurred when the relevant features had been noted but were then dismissed as insignificant. For example, if an abnormal mammogram contained a mass in the lower part of the left breast, the observer might or might not have indicated this in the report box relating to masses in the left breast (see Figure 4.1). Similarly she might or might not have indicated the abnormality in the schematic drawing as located in the lower portion of the left breast. If neither, the schematic drawing nor the feature recording box contained any indication of the presence of the abnormality, it was assumed that the radiographer had failed to notice it at a perceptual recognition level. This was classified as a perception error. If, on the other hand, the abnormality had been indicated, but the decision outcome was not to recall, the radiographer had evidently seen the abnormality and the error had occurred at the interpretation level. Figure 4.6 shows that perception errors thus classified are less frequent than interpretation errors. There is little change over the six tests with regard to the overall frequency of detection and interpretation errors.

**Figure 4.6:** Error analysis: Total number of perception and interpretation errors
4.3.3 Performance in mammography screening

During their ten months of film reading the radiographers interpreted in excess of 4000 screening mammograms each. The triple screening protocol was only applicable in 3167 cases, since it was not always possible for the radiographers to read the same films. The present analysis is based on the 3167 screening mammograms which were interpreted by both radiographers and a radiologist according to the protocol in Figure 4.2. In order to examine the radiographers' progress over time, these cases were split into 16 chronological groups of 200. The criterion against which radiographers' performance was measured is the radiologist's decision. This is a subjective criterion, because the radiologist's decision may be wrong in individual cases, but it presents a rough measure of actual screening performance. Figure 4.7 shows radiographer performance measured by the G-index, and in Figure 4.8 probabilities of true and false positives are graphically presented (see also Table Ib.1 in Appendix Ib).

Figure 4.7: Radiographer screening performance over 16 chronological intervals, measured by the G-Index with radiologist performance as the criterion.
Figure 4.8: Radiographer true and false positives over 16 chronological intervals with radiologist performance as the criterion.

Figure 4.8a: A.S. 

Figure 4.8b: F.A.

Agreement between the two radiographers and the radiologist can be assessed in terms of the correlation between their respective performance indices. This is shown in Table 4.1. In the present case, a non-parametric monotonicity coefficient, the Goodman-Kruskal Gamma index, was employed. This index is less sensitive to the skew in the screening data than other more commonly used correlation indices.

Table 4.1: Agreement between radiographers and radiologist in screening film recall measured with the Goodman-Kruskal Gamma index.

<table>
<thead>
<tr>
<th></th>
<th>A.S.</th>
<th>F.A.</th>
<th>Radiologist</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.S.</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F.A.</td>
<td>.75</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Radiologist</td>
<td>.74</td>
<td>.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Since radiologist performance is a subjective performance criterion, it may be more appropriate to measure radiographer performance in terms of actual screening outcome. In Table 4.2, radiographers' agreement with three more objective performance criteria is shown. The final decision criterion represents the actual decision to recall or not to recall after arbitration by a second radiologist. Screening
outcome is an index of the outcome after recall assessment. Most of the women attending assessment clinics are returned to routine screening after further mammograms. Very few cases are referred for surgical assessment. The referral criterion thus represents the distinction between benign and malignant abnormality.

Table 4.2: Radiographer agreement with final decision, screening and referral outcome (Goodman-Kruskal Gamma).

<table>
<thead>
<tr>
<th></th>
<th>A.S.</th>
<th>F.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Decision</td>
<td>.63</td>
<td>.75</td>
</tr>
<tr>
<td>Screening Outcome</td>
<td>.79</td>
<td>.92</td>
</tr>
<tr>
<td>Referral Outcome</td>
<td>.77</td>
<td>.86</td>
</tr>
</tbody>
</table>

4.3.4 The effect of dual reading on screening performance

The decisions made by radiologists in breast screening relate to whether to recall a case or not. The radiographers were permitted to grade their recall decisions into clear recall or no recall or query versions of these two categories (see Figure 4.1 for recording form). The screening decisions made by the radiographers and the radiologist are shown in terms of the four decision outcomes in Table 4.3.

Table 4.3: Percentage of decision outcomes in four decision categories.

<table>
<thead>
<tr>
<th>Decision</th>
<th>Radiographer</th>
<th>Radiologist</th>
<th>Final Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>No recall</td>
<td>75%</td>
<td>90.5%</td>
<td>89.6%</td>
</tr>
<tr>
<td>No recall?</td>
<td>12.9%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Recall?</td>
<td>5.9%</td>
<td>&lt; 0.1%</td>
<td>0%</td>
</tr>
<tr>
<td>Recall</td>
<td>6.2%</td>
<td>9.4%</td>
<td>10.3%</td>
</tr>
</tbody>
</table>
Table 4.3 shows that the radiographers used the query categories, whereas the radiologist did not. This is entirely expected given the trainee status of the radiographers. Radiologists are used to making clear binary decisions in the screening context. On the whole, radiographers have a slight tendency to recall more cases than the radiologist. This is reflected in a final recall rate which is 0.9% higher than that of the radiologist. Altogether 326 women were recalled for further assessment (10.3%). Figure 4.9 presents a breakdown of the outcome of these recalls. The contribution of the radiographers to the overall recall rate was an additional 29 cases over and above the radiologist’s recall (0.9%).

**Figure 4.9:** Assessment outcome of 326 recalled cases

![Pie chart](image)

Figure 4.10 shows a breakdown of the outcome of these cases. It can be seen that two women were offered early repeat screening, and one was referred for surgical biopsy. Since this biopsy confirmed malignancy, it can be said that the radiographers increased the cancer detection rate by 6.25% at the cost of an increased recall rate of 0.9%.

**Figure 4.10:** Assessment outcome of 29 cases not initially recalled by the radiologist

![Pie chart](image)
4.3.5 Verbal protocols

Verbal data from both radiographers were collected during the six test sessions by tape-recording their 'thinking aloud' protocols. The protocols were classified according to the four decision outcomes into true or false positives and true or false negatives (see Table 3.1). Examples of typical protocols from each of the radiographers are shown in Table 4.4. The example protocols are taken from the fifth test (after 3000 screening mammograms), and with the exception of the false negatives they are based on the same mammograms for the two radiographers. It can be seen that the verbalisations are closely based on features identified during visual search of the mammogram.

The verbal protocols also give some insight in the strategies used to examine mammograms by the two radiographers. Ericsson and Simon (1984) have advocated a retrospective verbal report method with perceptual tasks. Both radiographers were asked to describe "what they were doing when interpreting a mammogram" after the third test session. A.S., who did not seem to be following any particular strategy other than looking for evidence of abnormality, was unable to describe her strategy. F.A, however, described a fairly structured and complex search strategy:

"First of all I look at them [mammograms] from a distance and just compare the two. If there is anything obvious that jumps out I make a note of it. Then I look at each of them from a distance of about two or three feet to see if they are symmetrical or to see how they differ. And then I look closely with the magnifying glass starting at the top of the right one and work my way in zigzag lines all the way to the bottom, mentally noting anything that might be worth noting, whether it is distorted or calcium or whatever. Then I do the same on the left, start at the top and work my way down to the bottom. And then I sit back again and look at them both together, and then I make a decision based on what I have seen. If one side has, say, got a density that's not on the left then I'll decide whether that is just a normal variance or whether it should be called back. It's all experience really, the more you do, the more you know what is normal and what's abnormal...". (F.A. after the third test).
Table 4.4: Examples of verbal protocols for each of the four types of decision outcomes for the two radiographers.

<table>
<thead>
<tr>
<th>A.S.</th>
<th>F.A.</th>
</tr>
</thead>
</table>
| **true negative** (no. 3) | "Nothing really very remarkable in either breast, some benign calcification on the left ... so I wouldn't recall that."
| **true positive** (no. 6) | "Benign calcium on both sides, but there is - looks like an irregular mass on the left in the upper part which you can't really undress by imagining vessels going through it, 'cause they seem to radiate from it. So I'm definitely going to recall that lady."
| **false negative** (no. 63) | "Some benign looking lesions on the right. Very messy films, lots of artefacts and things. I think the breasts are ok though."
| **false positive** (no. 2) | "Some benign calcification on the left, quite dense breasts, there is a little something down there, a little asymmetry on the left on the lower aspect, which looks like it could be worrying, although there is something vaguely matched on the other side. I would query and recall that for an irregular mass on the left." |
| **true negative** (no. 3) | "Very fatty breasts, perfectly ok from a distance, right breast - some contorted vessels, other than that I can't see anything of interest at all. On the left, I can see absolutely nothing."
| **true positive** (no. 6) | "Fairly fatty breasts, looking at it to start with, I can see a spiky area on the upper part of the left breast plus two or three flecks of benign calcium. Under the magnifying glass ... right side looks fine apart from two benign flecks of calcium, very rounded. Left side ... there is the distorted spiky area on the upper part of the left breast and some benign calcium as well. I'll definitely bring that back for the distorted area." |
| **false negative** (no. 70) | "Again very symmetrical looking breasts, couple of bits of benign looking calcium on the left side. Right side ... central part of the breast is fairly dense and fatty around the edge. All I can see really is the calcium within it. The left breast... very similar looking appearances, can't see anything worrying."
| **false positive** (no. 2) | "Fairly dense breasts with nice positioning, they look very symmetrical. Under the magnifying glass, right hand breast... I can't see anything abnormal at all. On the left one small dot of benign calcium. On the bottom of the left breast there is a small spiky looking area which is different to the right, very low density. I think that should be recalled just because it is different to the other side." |
Content analysis of the verbal protocols revealed that the great majority of the propositions contained in the verbalisations could be cross-classified into seven categories, specific abnormalities (Spec), unspecified abnormalities (Unspec), features relating to previous history (Hist), normality (Normal), background (Backgnd), technical features (Technical) and causal associations (Causal). Comments classified in the two abnormality categories (Spec and Unspec) included both benign and malignant features. Specific abnormalities were those which fitted one of the usual descriptions of abnormalities, spiky or rounded masses, calcifications, distortions or asymmetry. Other abnormalities e.g. 'density', 'a smudgy area', 'a little something' etc. were classified as unspecified abnormalities. References to previous surgery, scars or congenital abnormalities were classified as features relating to previous history. Normality and remarks referring to the absence of specific abnormal features were coded as 'normal'. Descriptions of breast background features, including the parenchymal pattern, vessels and other examples of normal breast anatomy were recorded in the background category, whilst verbalisations referring to specifically technical features and processes, e.g. the positioning of the breasts, were categorised as 'technical'. The causal associations category included all instances of giving reasons either by providing secondary features or by providing explanations or diagnoses for primary features. The content analysis of the protocols in terms of the seven feature categories can be justified because they cover the entire observed spectrum of verbalisations including a crude measure of causal reasoning. Apart from remarks relating to the decision outcome (i.e. recall or no recall), which were discarded for the purpose of the correspondence analysis, only very few verbalizations proved to be unclassifiable within the present scheme.

The cross-classified features and decision outcomes were analysed by means of correspondence analysis, an exploratory metric scaling technique (Hammond, 1987, Weller and Romney, 1990). Correspondence analysis (CA) permits the simultaneous representation of decision outcomes and features in a two-dimensional space, thus providing a means to directly describe the relationship between and
amongst them. The plots arising from CA are straight visual representations of the associations between the variables. The closer two points are graphically, the closer the respective variables are associated with each other.

For the purpose of the present analysis, feature classifications before training (test 1) and after training (means of tests 2 to 6) were treated as separate variables. This provides a way to assess the impact of training in terms of the relationship between features recorded and decisions made. Separate analyses were performed for the verbal data produced by A.S. and F.A. Both CA analyses yielded two principal components, accounting for 54.38% and 36.3% of inertia in A.S.'s case and for 71.5% and 20.5% of inertia for F.A. The CA plots for both radiographers can thus be interpreted across two dimensions. The graphical representations of the four outcomes and seven feature categories before (1) and after training (2) are reproduced in Figures 4.11 and 4.12. The locations of decision outcomes are represented by star markers whereas the feature categories are indicated by triangles. The corresponding coordinates of the CA plots are shown in Tables Ic.1 and Ic.2 in Appendix Ic.

It can be seen that the correspondence analyses produce quite different plots for the two radiographers. Whilst the points representing decision outcomes and pre- and post-training features in F.A.'s plot are distributed throughout the plot, they are quite closely clustered in A.S.'s plot. In both plots, the decision outcomes are reasonably well differentiated, although true and false positives appear fairly similar in A.S.'s case. Out of the four decision outcomes, true positives and true negatives are the most dissimilar categories, which is represented by the distance between them across the vertical dimension in A.S.'s case and horizontally for F.A.

The impact of training on the features categories can be examined with respect to the distance between pre- (1) and post-training (2) features of the same type. For A.S. the greatest change is in the use of normal features which becomes associated with false negatives, followed by a marked shift in technical features. There is
Figure 4.11: Correspondence analysis plot of A.S.' verbal data cross-classified by decision outcomes.

Figure 4.12: Correspondence analysis plot for F.A.'s verbal feature categories (triangles) cross-classified by decision outcomes (stars).
relatively little change with respect to the location of specific and unspecified abnormal features, background and causal features in the plot before and after training. In F.A.'s case, the main changes concern prior medical history, which is associated with true positives prior to training but looses importance in terms of all decision outcomes after training. Shifts in specified and unspecified features, normal and causal features, on the other hand, are only slight.

The strength of association of particular features with particular decision outcomes can be determined by partitioning the plot into regions representative of specific decision outcomes. Verbal categories falling into these regions can be seen to be associated most closely with the decision outcome representative of the region in question. Weller and Romney (1990) have suggested that correspondence analysis plots may be more satisfactorily interpreted by drawing circles of increasing size around the points representing decision outcomes. The verbal categories falling inside the smallest circles are most closely associated with the decision outcome in the centre. Table 4.5 summarises the closest relationships between decision outcomes and verbal categories before and after training.

Table 4.5: Verbal features most closely associated with particular decision outcomes before and after training.

<table>
<thead>
<tr>
<th></th>
<th>A.S.</th>
<th>F.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test 1</td>
<td>test 2 to 6</td>
</tr>
<tr>
<td>false negatives</td>
<td>prior medical history</td>
<td>normal, specific abnormal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>false positives</td>
<td>unspecified abnormality</td>
<td>unspecified abnormality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>true negatives</td>
<td>background, technical</td>
<td>background</td>
</tr>
<tr>
<td>true positives</td>
<td>specific abnormal</td>
<td>unspecified abnormality,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>causal</td>
</tr>
</tbody>
</table>

88
Table 4.5 and Figure 4.11 show that for A.S. false negatives are most closely associated with prior medical history before training and with normal features and to a lesser degree specific abnormalities after training. Unspecified abnormal features are closest to false positives before and after training. Background and technical features associate with true negatives prior to training but only background features retain this position after training. Initially, specific abnormalities relate to true positives, but after training there is a stronger association between true positives and unspecified abnormalities and causal features. Before training normal features do not associate with any particular decision outcome, and causal features are equi-distant from true positives, false positives and false negatives, while after training previous history and technical features are equally associated with them.

Associations between decision outcomes and verbal feature categories in F.A.'s case are graphically represented in Figure 4.12 and summarised in Table 4.5. In this case, unspecified abnormal features are most closely related to false negatives before and after training. Technical features are of some importance before training but are more associated with false positives after training. None of the features are particularly associated with false positives before training, the closest being unspecified abnormality which is closer to false negatives, causal features which is equally associated with true positives, and background features. After training unspecified abnormality, causal and technical features are more directly associated with false positives. True negatives are characterised by the close proximity of normal features before training and normal and background features after training, while true positives are closely associated with specific abnormalities both before and after training. Prior to training, true positives are also characterised by the proximity of medical history and causal features. After training normal features are equi-distant from true positives, and false negatives and medical history features are no longer associated with any particular decision outcome.
4.4 Discussion

The data presented in this chapter were case study data based on two individual subjects. Such case studies yield rich, in depth data, but there is little opportunity for statistical inference because of the small sample size. Although the results from this study are descriptively informative, generalisation to all radiographers or lay film readers would be premature. Nevertheless, a number of conclusions can be drawn from the descriptive data of the radiographer film reading study with regard to training outcome and skills and processes involved in learning to read mammograms. In this section, the implications of the radiographer data for training design, modelling mammogram interpretation skill and the identification of individual differences in skill implementation are discussed.

4.4.1 Training radiographers for mammogram interpretation

The performance data in Figure 4.3 clearly demonstrate that the radiographers are trainable as mammography screeners. This notion is supported by uniformly high performance indices in actual screening (Figure 4.7). The unequivocal nature of the criterion (radiologist performance), however, means that these subjective performance indices have to be interpreted with caution. In the objective test assessment (Figure 4.3), it can be seen that radiologists perform better prior to training, but the radiographers' post-training performance on the mammogram test roller is virtually identically to the radiologists' mean performance. These findings support Dowdy et al's (1970) and Alcorn et al's (1968, 1969 and 1971) claims that it is feasible to train lay film readers for screening mammography. Radiologists' better performance before training is probably due to the fact that radiologists usually have prior experience in clinical mammography and are, therefore, familiar with what constitutes abnormality in mammograms. In other words, even before training to read screening mammograms, radiologists, on average, have well-developed schemata of mammographic abnormality.
For the radiographers, there does not seem to be much improvement after initial training. The performance curve for tests 3 to 6 in Figure 4.3 is essentially flat. Although there is very slight improvement over the first two follow-ups (test 3 and 4), this is matched by an equally slight drop of performance over the third and fourth follow-up (test 5 and 6). Figure 4.7 shows that there is no improvement in real screening either. The fluctuation in performance observed over time are a clear indication that performance remains variable, although some of this variation may be attributable to the subjective criterion. This lack of significant improvement in the post-training period is of particular interest because the NHSBSP (1991) assumes that support during this period brings about further refinement of mammogram interpretation skill. The longitudinal radiographer data do not support this notion, and there is no reason to assume that this would be different for radiologists.

It is possible that the radiographers do not improve their mammogram interpretation performance during normal screening because their involvement in mammogram reading is only part time. Alternatively, ten months of screening experience may be too short an interval to show significant shifts in performance. After all, expertise in a task such as radiological interpretation probably takes 10 years or more to acquire, while the radiographers were studied for less than 1/10 of that time. However, if mammogram interpretation did improve as a function of experience, there should be some indication of such a trend even over the first few months. The assumptions made by the NHSBSP (1991) about continued learning after initial training merely on the strength of exposure to day-to-day screening mammograms, thus, seem optimistic in the light of the present data. This issue requires further examination in the shape of research into the longitudinal effects of training and daily screening routine on film reading performance in radiologists and lay film readers.
4.4.2 Pre-screening or second pair of eyes?

The true positive and false positive data shown in Figures 4.3a and 4.3b show that the main reason for improvement after training is a large drop in false positive rates for both radiographers. The low false positive rate, with the exception of one blip in Figure 4.4a, is maintained throughout the post-training period. True positives rates, on the other hand, are much more variable over the six test occasion, a finding which is supported by the actual screening data. Figure 4.8 shows that the false positive are uniformly very low indicating an emphasis on discrimination in radiographer screening performance. True positives on the other hand are highly variable in the real screening context (note reservations about criterion as above, though). The radiographers' failure to reach a 100% pick-up rate or even a stable high true positive rate leads to the conclusion that the training they received is not adequate to equip them for a pre-screening function, since the production of any false negatives is unacceptable in this context.

It is possible that the tutorial approach is unsuitable for training pre-screeners. The task in pre-screening is essentially one of detecting abnormal patterns. The tutorial approach, however, encourages trainees to make judgements on the status of detected abnormalities as benign or malignant. Skills acquired with this training method are more analogous to medical reasoning than to pattern detection as trainees learn to emulate making diagnoses of detected abnormality. If trainees were to fulfil a pre-screening function after training, it might be better to emphasize the detection and recognition of abnormalities. Some studies on the acquisition of perceptual expertise, e.g. Biederman & Shiffar's (1987) work on chick sexing, have demonstrated that relatively brief instruction on critical features and their interrelations can improve novice performance on the perceptual aspects of the task to approximate that of experts. Similarly, there have been promising results from a study on the effect of focused instruction on mammography performance (Farjado et. al., 1987). These issues are discussed in more detail in the chapter 5.
It can be said, however, that the present training is clearly succeeding in equipping radiographers to be second readers in dual reading of screening mammograms. Both radiographers achieve the same post-training performance as the mean performance of a group of nine radiologists. In addition, their agreement with the radiologist (Table 4.1) and with more objective screening outcomes (Table 4.2) in the actual screening context indicates that they probably are as competent in interpreting screening mammograms as most radiologists. The highest agreement coefficients for both radiographers were obtained with screening outcome (.92 and .79, respectively), which may represent the most appropriate criterion against which to measure performance.

Unlike Alcorn et al's (1971) lay screeners, the radiographers did not produce high false positive rates in return for relatively high true positive rates but found a balance similar to that of the radiologists. The similarities between average true and false positive rates achieved by radiologists and radiographers after training provide support for the notion that the radiographers are emulating radiologists in their film reading performance. This appears to be a direct result of giving the radiographers the same training as radiologists. The question remains, though, whether it is appropriate to train radiographers, who have no medical training, to make decisions emulating those of doctors. It might be preferable to adopt a pre-screening model and train radiographers to adopt a lax decision criterion producing more false positives but picking up the very larger majority of abnormalities. Alcorn et al's (1971) programmed learning approach which focused on abnormal features appeared to achieve this. The high false positive rate is not a problem in the pre-screening context, since radiologists trained in diagnostic interpretation of mammograms review all cases classified as abnormal by the lay reader. This should normally result in an overall increase in the true positive rate, but not at the cost of a high false positive rate.
4.4.3 Skills implied in screening mammography

One source of information about the processes involved in learning to interpret mammograms is the analysis of errors made in mammogram interpretation tests. Of particular interest in the context of training pre-screeners are the false negative errors. Figure 4.5 showed that both radiographers consistently produce misses and recalls for the wrong breast over the six test rollers. False positive errors were not analysed in this particular study because their analysis presents interpretation problems, and they are only of limited interest in the context of pre-screening. In 'second reading' the false positive rate is of importance only if the radiographers' probability of false recall was substantially higher than that of the radiologists. As this is not the case, error analysis only considered false negative classification and detection.

Previous studies of radiological performance have conducted error analysis in terms of eye movement parameters in order to identify some of the perceptual processes underlying detection and classification of abnormality (Kundel et al, 1978). Recording of eye movements was not within the scope of the present study. Nevertheless, it was possible to distinguish clearly between two different types of errors, detection and interpretation errors, in terms of the features identified in the recorded responses. In the present study, it was not, however, possible to distinguish between specific perceptual error types such as search and recognition errors (Kundel et al., 1978, Carmody et al, 1980). This is not a major problem, though, as even studies employing eye movement recording have not been able to make a clear distinction between search and recognition errors because their definition is operationalised in terms of arbitrary eye movement parameters. Such work has shown that the most frequent errors are interpretation errors, which occur when an abnormality has been fixated and dwelt upon for an adequate time span yet the observer has made a wrong decision.
Figure 4.6 shows that both detection and interpretation errors occur in mammography screening. This is an indication that the distinction between perceptual and cognitive expertise in mammography is justified. Figures 4.6a and 4.6b show that both radiographers produce interpretation errors more frequently than perceptual errors. This is in keeping with Getty et al.'s (1988) assertion that mammography experts are most error-prone when combining features to make decisions as well as Kundel et al's (1978) identification of errors in decision making as the most frequent ones.

It has been argued before, that perceptual and cognitive expertise are likely to operate side by side in most types of radiology (Rubin, 1989), but the inter-relationship of the two types of processes has remained somewhat contentious. Although interpretation errors are more frequent overall for both radiographers, there is no indication that there is significant change in either perception or decision errors as a function of training or experience. Thus, there is no support in the present error data for the notion that there is a developmental sequence progressing from perceptual to cognitive skills as has been suggested by Lesgold et al (1988). The fact that perceptual and decision errors continue to co-occur even after several months of experience seems to indicate that the perceptual and interpretative processes underlying mammogram interpretation are more inter-dependent than Lesgold et al acknowledged in their work.

4.4.4 Changes in mammogram interpretation after training

Verbal protocols of the test sessions were collected in order to monitor qualitative changes in mammogram interpretation as a function of training. The verbal protocols were difficult to obtain, because the radiographers initially found it difficult to verbalise their thoughts. This is not altogether surprising, since the type of skill under examination is essentially perceptually based. Ericsson & Simon (1984) have discussed the problems inherent in obtaining verbal data with tasks that are not verbally coded. Some of the information processed in such tasks may
be inaccessible to conscious thought. However, there are aspects of mammogram interpretation that require the information to be transposed into a verbal format, as abnormal findings have to be communicated in terms of diagnostic reports. It would, therefore, be expected that the 'thinking aloud' protocols would reflect the verbal transpositions of visual features noted in the mammogram, which was indeed the case.

Comparison of pre- and post training feature categories in Figures 4.11 and 4.12 reveals that there was some change in the way these categories related to decision outcomes before and after training. The main changes seemed to occur in the use of those features which were unrelated to abnormality in mammograms. In neither case, there was much change in either specified or unspecified abnormal features, but there were considerable changes in how categories of normality, previous history, technical aspects and background were used. A.S. (Figure 4.11), for instance, did not think in terms of normal features prior to training. After training normal features became closely associated with false negatives, indicating that she employs normal features to explain away specific abnormality, which is also closely associated with false negatives, thus producing misses. The most significant change in F.A.'s verbal reports is the complete drop out of previous history features after training which are closely associated with true positives prior to training. In addition, background features, which are initially more related to false positives become more closely associated with true negatives. The features most closely related with both false positives and false negatives after training are unspecified abnormalities. F.A., thus, makes errors because she misinterprets the significance of ambiguous perceptual features. The direction of change varies for individual feature categories and is different for the two radiographers, indicating that there are differences in strategies employed by the radiographers.

4.4.5 Individual differences after training

Differences in strategies between the two radiographers can be identified in the
examples of verbalisations in Table 4.4. F.A.'s verbal protocols are longer and more detailed. She follows a structured viewing procedure as is apparent from her retrospective description of how she interprets mammograms. A.S., on the other hand, is unable to describe any structured viewing strategy. Her verbalizations are shorter and essentially orientated towards abnormal feature detection.

Individual differences in viewing strategy are also evident in the different plots of decision outcomes and feature categories produced by correspondence analysis. F.A. takes a very structured approach which might be described as the 'ticking of a mental checklist'. She actively searches for the different categories of abnormality, which also accounts for the close association between specific abnormal features and true positives and normal features and true negatives. Errors seem to arise in the face of uncertainty or ambiguity (unspecified abnormal features). A.S., on the other hand, takes a more opportunistic approach, essentially attempting to identify abnormality. True and false positives are closely associated in her case and characterised in terms of similar verbal categories. The identification of abnormality is, thus, not clearly defined in terms of verbal features. True and false negatives are more effectively separated in terms of feature categories. True negatives are essentially characterised in terms of background features. False negatives, on the other hand, arise when specified abnormal features are in conflict with normal features.

Further evidence for individual differences can be identified in terms of screening performance. The agreement between F.A. and the radiologist is much better than that between A.S. and the radiologist (Table 4.1). The indication is that this may be due to A.S.'s higher recall rate as indicated in Table 4.3. If this indeed the case, A.S. could be seen to perform more like a pre-screener than F.A. In terms of the test set, however, individual differences do not seem to affect the efficiency of overall performance. Figures 4.3 and 4.7 show that there is no real differentiation between the two radiographers with regard to mammogram interpretation performance. Although F.A. approaches the mammogram interpretation task in a
more structured manner, this is not reflected in better detection of abnormality in the objective test. This is in keeping with previous findings that structuring search strategy does not necessarily result in better detection performance (Gale and Worthington, 1983).

4.4.6 Effects of dual reading

From the actual screening data presented in this study, it would appear that real benefits may be gained from radiographer/radiologist dual reading schemes in terms of identification of true positives. Figure 4.10 shows that one additional cancer was detected on account of radiographer film reading. This represents an increase in cancer detection of 6.25%, a sizable increase by any standard. Kirkpatrick (1993) has estimated that the benefit of dual reading is an increase of between 2.5% and 15%. The increase of 6.25% observed in the present study was gained at a cost of an extra 0.9% women being recalled. It must be noted, however, that the numbers involved in this study are very small indeed. In order to assess the value of dual screening, a much more powerful study involving a very large sample is required, mainly because the incidence of cancer is so low. A further issue in the present study is that the screening protocol, in fact, involved triple reading of mammograms. The question is whether one radiographer and a radiologist can be as effective as two radiographers and a radiologist. In summary, there is some evidence for benefits in cancer detection of dual reading involving radiographers, but this has to be assessed in terms of overall costs and benefits and requires verification in a larger study.

4.5 Conclusion

In summary, the case studies presented in this chapter have demonstrated the feasibility of training radiographers as lay film readers in mammography supporting the results of a number of previous studies (Dowdy et al, 1970, Alcorn et al, 1968, 1969, 1971, Haiart and Henderson, 1991, Bjurstam, 1993). This was shown in terms
of both, objective test data and real screening data. There are two questions which remain to be answered though. The first refers to the role trained lay film readers should occupy, and the second concerns individual differences in trained film reading performance.

The question regarding the role lay film readers should play in breast screening needs to be addressed with reference to the nature of the perceptual and cognitive skills involved in the interpretation of mammograms. In addition, an exploration of how training can support the development of such skills is required. The present case studies showed that lay reader performance approximated that of fully trained radiologists after training. The data also indicated that the radiographers true positive rate immediately after training was marginally lower than that of radiologists. Although this is offset, to some extent, by a lower false positive rate, the indication is that radiographers may be less skilled in identifying true abnormality. This may be a direct result of their lack of medical knowledge. In the following three chapters, the skills involved in mammogram interpretation and how they are influenced by training are examined. At the simplest level, mammogram interpretation can be construed as a pure pattern perception task which can be supported by training directed at the perceptual recognition of deviant appearances. In screening mammography, recognition of abnormality is not sufficient, though, as high level decisions regarding the benignity or malignancy of abnormal patterns need to be made at some point in time. The question is whether lay film readers should be encouraged to make such high level decisions. The influence of training methods on lay reader performance is examined in chapter 6. In chapter 7, the question of skills involved in mammogram interpretation and training will be discussed in greater depth with reference to a model of mammogram interpretation performance.

The second question arising from the present case study regards individual differences. There was evidence that the two radiographers differed in the way they approached the film reading task. Differences were particularly evident in the
search strategies employed and in the extent to which mammograms were evaluated in a methodological manner.

Although the two radiographers achieved similar performance standards after training, it likely that individual differences in performance would arise in larger groups of trainees. Saxton (1991), for example, has argued for a careful selection of lay observers in radiology. These issues are addressed in detail in the third part of this thesis. The question whether an aptitude domain for mammography can be identified, and the extent to which trained film reading performance is predictable in terms of personality, ability and cognitive skills variables will be examined in chapters 8 to 11.
Part 2

Modelling mammogram interpretation skill:

The influence of training

The three chapters in this section combine to investigate at a deeper level the questions regarding skill acquisition and the influence of training first raised the previous part. This is achieved in terms of a review of the pertinent psychological literature and, on an empirical level, in terms of longitudinal training studies with complete novices. Training decisions, regardless of the type of skill to be learned, should be based on careful task analysis in order to determine what the training content should be and what training methods best facilitate skill acquisition. In the first chapter of this part, mammography skill is first discussed in terms of the various task aspects involved in radiograph interpretation. The aim is to develop an understanding of the skills required for lay film readers in mammography. This is followed by a brief review of the literature on training and education in radiology and mammography. The final section of the first chapter is concerned with the learning mechanisms involved in complex visual skill acquisition. The resulting analysis of mammography screening skill and its acquisition may be used to make informed decisions about training design. Alternative training schemes can then be examined in terms of their efficiency, an issue which will be addressed empirically in a training study reported in the subsequent chapter. A further concern in connection with training is the prevalence of observer variability. The influence of training on individual differences in mammogram interpretation is examined in more detail in a second training study. In chapter 7, the skills analysis and the empirical training data from chapter 6 are combined to build a model of mammogram interpretation. This double two-stage process model is examined in terms of its ability to account for the data in the training studies, including error data. Other pertinent findings in radiological performance are also discussed in terms of the model.
Chapter 5
Further issues: Mammography skills, skill acquisition and training

Training screeners is an important issue for the NHS breast screening programme, since the quality of the programme depends critically on film reader quality. Some of the salient issues in this context have already been addressed in the previous three chapters. These related particularly to the kind of training most beneficial for training radiographers as either pre-screeners or second readers (’second pair of eyes’). In chapter 4, the conclusion was that radiographers could be trained as ’second pairs of eyes’ to support consultants in a dual reading situation. However, current tutorial based training methods were unsuitable for training pre-screeners, the role initially envisaged for radiographers in mammogram interpretation (see method in the previous chapter).

In this chapter, the skills involved in reading mammograms and the acquisition of these skills are explored in more detail with reference to training procedures. Decisions on how to train screeners should be based on an understanding of the skills involved in the task to be learned. For this purpose, the major skill elements of radiographic diagnosis, visual discrimination, acquisition of perceptual and diagnostic categories and probabilistic decision making are examined in terms of their relevance to learning to read mammograms. The second part of this chapter is concerned with radiology and mammography education. Although this is an issue of considerable importance, a review of the radiology education literature indicates that there are few studies directly evaluating particular training schemes for radiology. In the last section, potential learning mechanisms involved in learning to read mammograms are explored. In particular, the role of practice and experience, induction, and focused instruction in acquiring mammography skills are discussed.
5.1 What is learned during skill acquisition in mammogram interpretation?

At present, training in mammography is not based on well-established research. It is assumed that skills develop as part of on-the-job experience after initial brief training (NHSBSP, 1991). The longitudinal radiographer case study reported in the previous chapter did not, however, lend much support to this notion. Experience in daily screening did not seem to be enough to improve either sensitivity or specificity any further after initial training. One possible explanation for the radiographers' failure to profit from continued exposure to mammograms is that the experience gained is irrelevant to learning mammography. The skills required for the accurate reading of mammograms may not be sufficiently practised in simple exposure to screening mammograms. One reason for this may be that only a very small number of positive cases are seen in daily screening.

As discussed previously, the skills lay film readers in mammography must acquire are dependent on the eventual function they are to fulfil in the screening service. In chapter 3, two alternative models of lay film reader functioning were outlined, the 'second pair of eyes' or dual screening model and the pre-screening model. As 'second pairs of eyes' lay readers must be able to perform at a level analogous to radiologists, including visual analysis of the radiograph and diagnostic decision making. In the alternative pre-screening model the emphasis is on visual detection of abnormality. In this model, the lay reader is required to adopt a more lax criterion for recall than the radiologist. For both, second screening and pre-screening skill, the primary learning task involves the acquisition of categorisation skills on the basis of visual data. In this context, lay readers have to acquire skills in probabilistic decision making, as they learn to evaluate the relative importance of observed visual features and to make decisions to recall or not to recall particular cases. In second screening, the lay observer needs to learn about disease categories in addition to visual categories, as she is required to make decisions about probable malignancy or benignity of lesions. Pattern/object recognition,
categorisation on the basis of visual and conceptual features and probabilistic decision making, thus, all contribute to the skill displayed by the fully trained mammogram interpreter and are discussed in turn. The influence of training on the acquisition of these skills is discussed in the final section of this chapter.

5.1.1 Visual discrimination and pattern/object recognition

One important aspect of radiological skill is the ability to identify perceptually patterns or objects which are important for making diagnostic decisions about radiographs (Getty et al, 1988). The question arises whether reading X-rays involves simple two-dimensional pattern recognition or more complicated object recognition. Radiologists may build up three-dimensional mental models of the structures imaged in the two-dimensional radiograph (Smoker, Berbaum, Luebke and Jacoby, 1984). The importance of three-dimensional perception in radiology for a true interpretation of spatial relations has been discussed by Sewerin (1983), for example. Diagnosis proceeds in terms of knowledge of healthy anatomy and observed deviations in the 3D model built from the image. An alternative view of radiological diagnosis is that it proceeds in terms of two-dimensional pattern perception. Radiological knowledge may be represented in terms of two-dimensional images of anatomy, diagnosis being based on the interpretation of abnormal image features which have been abstracted from comparisons with prototypical normal images.

It has been argued before that the nature of the processes involved in radiological interpretation may be to some extent dependent on the type of radiograph under investigation (Rubin, 1989). Mammograms which are images of one organ comprised of soft tissue may be more appropriately viewed in terms of patterns than radiographs imaging a number of different anatomical structures. In the next chapter, it will be argued that it might be possible to influence with training whether a radiograph is perceived as a two-dimensional pattern or as a two-dimensional representation from which a three-dimensional model is built. For the
present discussion, it suffices say that the visual processes in either case are fairly similar, as both pattern perception and object recognition are based on similar principles, object recognition representing the more complex case. Pattern and object recognition and their role in mammogram interpretation are discussed further below.

5.1.1.1 Pattern and object recognition
The question of how radiologists learn to recognise visual features critical to making a diagnosis can be addressed with reference to psychological pattern perception theories. These theories provide clues regarding the processes involved in matching visual input with knowledge representations stored in memory. They can be categorised as belonging to one of two classes, template based models or feature based models. Template theories of pattern recognition are based on the assumption that previous knowledge is stored in terms of either templates or more abstract prototypes, and that novel patterns are recognised by comparison to stored representations and identification of a 'best fit'. Such theories have frequently run into problems on account of the potentially very large number of representations that need to be stored in memory. The alternative class of models has been based on the notion that patterns are best described in terms of critical local features. Feature models of pattern recognition may be more viable in terms of stored knowledge. However, they frequently fail to account for the spatial relations between critical features which may be of vital importance for the classification of patterns.

In the radiological expertise literature, feature models have frequently been employed in order to account for perceptual refinement during skill acquisition (Getty et al, 1988, Lesgold et al, 1988). Lesgold et al (1988) have specifically evoked the Pandemonium model of pattern perception (Selfridge, 1959) in order to explain how radiologists learn to discriminate perceptual features. However, the Pandemonium model does not account for the relationship between features or visual context. There is some evidence that pattern recognition does not proceed in
the exclusively bottom-up fashion that is implicit in the Pandemonium model.

Earlier recognition may influence the interpretation of structures currently being recognised. Palmer (1975), for example, described the parsing paradox whereby it is possible to recognize a whole figure even if its parts are not recognisable on their own. In object recognition, Biederman, Mezzanotte and Rabinowitz (1982) found that the recognition of individual objects within the scene are facilitated by context. In radiology, this has been acknowledged in most recent models of visual search (Swensson, 1980, 1985, Nodine and Kundel, 1987), which postulate a recursive action between current pre-attentive search and structures already recognised.

Lesgold et al's (1988) work, thus, fails to examine the processes which lead to the acquisition of visual categories which can be used in diagnostic decision making in a satisfactory manner. Theories of visual search in radiology, which are primarily concerned with visual aspects of radiological interpretation, have fared better in terms of accounting for pattern perception phenomena. As far as screening mammography is concerned, visual aspects of the task are very important in view of the fact that the task involves detection of minimal abnormality in symptomless women. The acquisition of visual feature knowledge in mammography is, thus, likely to play an important role and may, for example, involve making a distinction between visual and semantic knowledge of visual features (Humphreys and Bruce, 1989). Knowledge of critical visual features in radiographs is about knowing the physical properties of such features, i.e. their shape and appearance, whereas semantic knowledge refers to the meaning of features and their function in determining normality or abnormality. The role of training in the acquisition of such visual feature knowledge is discussed in the third part of this chapter.

5.1.1.2 Categorical visual perception

Categorisation involves the treatment of a set of discriminable or indiscriminable stimuli as equivalent (Bornstein, 1987). Categorical perception entails an enhancement in discrimination between categories as compared to within-category
discrimination of physical differences of a comparable magnitude. Visual
categorical perception has often been investigated with respect to colour perception
because the colour spectrum provides a means of accurate measurement of physical
between and within category differences. The mechanisms involved in categorical
perception may provide a conceptualisation of the processes involved in
discriminating normal and abnormal categories in mammography rapidly.

Physical properties of features in radiographs are not as physically unambiguous as
those observed in colours as their manifestation and intensity depends on the
overall quality of the radiograph and the general breast pattern. Nevertheless, visual
evaluation of features in a radiograph may proceed in terms of rapid assignment to
visual categories. There is some anecdotal evidence that experienced radiologists
encounter instant recognition of abnormality, i.e. they claim that abnormal features
'jump out' of the image. Hendee (1993) has described the ability of radiologists to
identify abnormality rapidly by means of pre-attentive visual processing. Similarly,
Myles-Worsley et al (1988) found that expert radiologists are able to attend
selectively to abnormal features, thus enabling rapid recognition and categorisation
of radiographs as normal or abnormal.

Although it is difficult to investigate visual categorical perception in radiographs
directly because of the ambiguity in physical descriptions of the images, it can be
said that the processes underlying such perception are likely to be similar to other
categorisation phenomena. Medin and Barsalou (1987) have examined the
differences between sensory perception and generic knowledge categories and have
come to the conclusion that there are many similarities in terms of how
categorisation proceeds in perceptual and knowledge domains. In both domains, for
example, most categories are fuzzy ones, classification can be based on rules,
prototypes, exemplars or boundaries, and the general purpose of category formation
is classification. Category acquisition is thought to be based on both, innate
learning mechanisms and experience. These mechanisms are discussed in more
detail later on.
5.1.2 Categorisation of visual and disease feature

Making a decision on whether a particular pattern of imaged structures is normal or abnormal depends on the observer's ability to classify image features with respect to their meaning in the context of the image. Although expert radiologists may perform this part of the task with a high degree of automaticity at a perceptual level (Myles-Worsley et al, 1988, Hendee, 1993), it is likely that during the early stages of skill acquisition visual features are processed involving conscious effort. Categorisation of mammograms can be seen to require even greater sophistication, if additional decisions on the probable benignity or malignancy of abnormal features have to be made. Although the information required for such decisions is largely visual perceptual, generic knowledge has to be developed regarding the semantic properties of such perceptual features in the context of the radiograph (see discussion above). The question of whether knowledge in mammogram interpretation is perceptual or generic in nature can thus be seen to be largely irrelevant, since it is undoubtedly true that both kinds of knowledge are required. Their relative importance for accurate mammogram interpretation may vary depending on the skill level of the observer, on the one hand, and the screening task she is required to perform, on the other.

5.1.2.1 Categorising visual features as normal or abnormal

Getty et al (1988) have claimed that learning mammography involves 'the visual assessment of relevant features of the image' (p.240). This implies that features are extracted perceptually and then evaluated with respect to their significance regarding normality or abnormality of the image. Film readers must, therefore, know what relevant features are (perceptual knowledge) and how to interpret these features in the context of the whole image (semantic knowledge). Classification of mammograms as normal or abnormal thus requires recognition as well as feature categorisation. In this section, the kinds of processes involved in categorising features semantically are discussed.
As in categorical perception, mechanisms in knowledge category acquisition have been explained in terms of feature acquisition or prototype abstraction. The feature model of knowledge acquisition is strongly associated with Bruner's work on category learning. Bruner, Goodnow and Austin (1956) conducted a series of experiments into the learning of categories. They concluded that subjects acquired categories by making hypotheses about the rules defining a particular category, seeking confirmatory evidence for the currently held hypothesis and only changing it if it is disconfirmed by feedback. One criticism frequently levelled at Bruner et al's approach is that they did not take account of the probabilistic nature of many natural categories. However, the hypothesis testing approach is easily adapted to such natural categories by replacing defining features with characteristic features. Bruner et al (1956), in fact, discuss the case of probabilistic features in concept attainment in some detail (p.182-230). There is another problem with the feature model of concept attainment as related to probabilistic features, though. On the whole, human beings are not very good at entertaining probabilistic hypotheses (Brehmer, 1980). Laboratory experiments have shown that subjects have great difficulty in learning probabilistic rules by testing hypotheses leading to the conclusion that there may be other ways of learning categories, such as prototype or exemplar abstraction (Lakoff, 1987, Baron, 1988).

Prototype abstraction involves remembering the salient features of each example of a particular category without remembering specific features in specific examples. Thus a strong memory of the most common features relevant to a particular category evolves. Laboratory experiments have shown that such rule learning about categories may occur implicitly (Baron, 1988). An alternative explanation for this learning phenomenon is that subjects do not learn rules but simply make similarity judgements of new examples with previous ones which have been remembered as exemplars. Given a relatively complex, probabilistic categorisation as in the case of classifying features in mammograms, prototype learning or the use of exemplars may be a more efficient process for acquiring perceptual knowledge than rule abstraction by hypothesis testing.
The structure and acquisition of visual schemata has been investigated extensively in the laboratory context. Homa, Rhoads and Chambliss (1979), for instance, found that prior experience determines the degree of category identity with dot patterns. In addition people with more expertise in categorising such patterns had developed tighter category clusters, i.e. category members were rated as more similar to one another and more different to other categories. Vandierendonck (1984) also used a similarity judgement methodology in order to investigate subjects' classification of dot matrix patterns. He found that schema formation was based on those features that are most common to most exemplars of a category. This kind of research leads to the conclusion that expertise in categorising features in mammograms may be dependent on the ability to discriminate features based on detailed feature knowledge, and that some of this knowledge is well-defined and highly organised in terms of visual schemata (Neisser, 1976) which, in turn, influence search processes in mammogram interpretation (Gale, 1993).

5.1.2.2 Categorising disease features

Visual feature extraction and classification with respect to normality and abnormality may be sufficient in a pre-screening context, where all abnormal appearances are referred to a consultant radiologist. In second screening, however, there is a degree of diagnostic decision making. Abnormality can be further analysed in terms of benign and malignant features, and clearly benign cases may not be recalled for further investigation. In Getty et al's (1988) terminology this involves 'the merging of those assessments [the visual assessment(s) of relevant features] to arrive at a diagnostic decision' (p. 240). Clearly, diagnostic reasoning as well as the categorisation of abnormal mammograms involves semantic knowledge of disease features.

The nature of category structures in medical and radiological decision making has often been conceptualised in terms of schemata (Abercrombie, 1960, Lesgold, 1982, Lesgold et al, 1988). Bartlett (1932), one of the earliest psychologists to utilise schema theory for explaining memory phenomena, has defined a schema as '... an
active organization of past reactions, or of past experiences, which must always be supposed to be operating in any well-adapted organic response.' (p. 201). There are no universally accepted definitions of schemata, but the knowledge structures underlying the classification of abnormal mammograms as benign or malignant can be considered analogous to disease or illness scripts which have been employed to characterise knowledge in general medical diagnosis (Gilhooly, 1990, Schmidt et al, 1990). The knowledge involved in medical schema acquisition thus refers to illness and disease features, and, in the context of radiology, to the semantic properties of certain visual features which relate to specific abnormalities.

Research on the formation of diagnostic categories in real life medical diagnosis has put into question some of the assumptions regarding linear relationships between experience and detailed knowledge in medical expertise. Murphy and Wright (1984) found that expert clinical psychologists sometimes used fuzzy prototypes for rapid identification of a diagnostic category rather than highly refined schemata. Novices, on the other hand, seemed to rely heavily on conscious reasoning. This is consistent with findings by Myles-Worsley et al (1988) that expert radiologists made very rapid decisions with respect to the presence of abnormal features. This discrepancy between other expertise domains and medical expertise may be due to training in the medical domain being essentially prototype based, which leads to an expectation of a high degree of category distinctiveness in newly trained diagnosticians. Diagnostic experience, on the other hand, serves to decrease category distinctiveness and leads to the development of heuristics, rules of plausible reasoning and theories constraining thinking, that enable considered decision making in the medical context (Murphy and Wright, 1984).

The purpose of medical schemata is to guide thinking and aid interpretation when attempting to make sense of medical information. Lesgold et al (1988) argued that disease schemata aid the interpretation of visual information in radiographs on the basis of expectations from previous experience with similar cases. As in Murphy and Wright's study on clinical psychologists, radiological knowledge organised as
schemata can thus be seen as powerful way of making sense of visual medical information instantly in terms of well-established heuristics, which indicate how medical symptoms, or, in the case of radiology, visual features, co-occur and influence each other. However, Abercrombie (1960) has pointed out that schematically organised knowledge may be a source of error in making medical diagnosis, because the use of schemata can lead to rigidness in decision making. Given the link between visual perception of features in radiographs and their interpretation, the employment of rigid schemata may create false expectations leading to either misperception of critical visual features or to the misinterpretation of the meaning of observed structures (Lesgold et al, 1988).

5.1.2.3 Strategies and rules in perceptual and conceptual categorisation

Both, visual perceptual and conceptual categorisation, are subject to the application of rules and strategies by the observer when merging observations into decisions. Getty et al (1988) have characterised this decision making process as the weak link in human mammogram interpretation, as this is where most errors occur (see also Kundel et al, 1978). The decision rules bearing on the perception and categorisation of complex stimuli such as medical symptoms have been studied in a number of laboratory based experiments. Medin, Altom, Edelson and Freko (1982), for example, used a simulated medical classification task in order to investigate what kinds of rules are used in complex decision making. In particular, they focused on the question of whether medical classification was based on a simple additive summation of the available information or whether configural information was also utilised. They found that subjects were able to take account of the fact that some symptoms were correlated in their decision making. Ashby and Gott (1988) using artificial, but highly complex multidimensional stimuli also found that subjects were able to integrate information from separate stimulus dimensions, thus arriving at near optimal decision rules. Ashby and Maddox (1992) investigated the acquisition of such complex decision rules and found that most subjects employ highly non-linear, but deterministic rules in decision making. The difference between experienced subjects and novices is not so much the type of rule applied
but rather the consistency with which it is applied.

A further question concerns whether different decision rules come to bear in visual and conceptual categorisation. This is important in the context of mammogram interpretation, because accurate performance depends on both types of categorisation except in the very simplest of task definitions, in which the appropriate categorisation of visual features suffices. Bruner (1957) has argued that there is a continuity in the rules of inference used at both perceptual and conceptual levels. According to Bruner, both involve an act of identification by placing stimuli into a certain class on the basis of defining or characteristic attributes. The rules for classifying objects have to specify the critical attributes of the category, the manner in which attributes are combined, the weight assigned to each attribute and the acceptance limits of each category. Reed and Friedman (1973) investigated experimentally whether Bruner's contention that people use the same decision strategies in perceptual and conceptual tasks was tenable. They found that if the categories in both tasks were immediately given and specified in terms of well-defined attributes, both types of categorisation behaviour could be explained in terms of simple linear decision rules which were predictable by employing a prototype model of categorisation.

Bruner's approach to categorisation has been criticised for not taking into account the properties of natural categories which tend to be not well defined. In mammography, in particular, feature categories tend to be fuzzy and are only probabilistically related to medical outcome. However, the notion that there is a degree of continuity between perceptual and conceptual categorisation persists even in the absence of Bruner's own theories of perception and categorisation. Neisser (1987) argued on the basis of Gibson's theory of direct perception (Gibson, 1966, 1979) and Rosch's work on categorisation (Rosch, 1978) that although perception is more direct, categorisation at the basic level is so closely associated with perceptual properties and affordances that categories may appear to be perceptually given, thus implying a degree of continuity between perceptual and conceptual
categorisation processes.

Making a decision regarding the benignity or malignancy of an abnormality involves the visual assessment of multi-feature images and the diagnostic evaluation of complex combinations of features. The research discussed above seems to indicate that people learning to make such difficult decisions do not have problems merging even complex visual and conceptual information in an appropriate way. There is evidence, however, that information tends to be combined in a deterministic rather than a probabilistic way. As most medical information, including radiological features, is probabilistic the problem of diagnostic reasoning with probabilistic features will be discussed in more detail in the following section.

5.1.3 Medical decision making with probabilistic features

Making a medical diagnosis involves reasoning on the basis of probabilistic information. There is extensive evidence that people are not very good at making optimal decisions when probabilistic information has to be taken into account (Tversky and Kahneman, 1974). In experiments involving probabilistic reasoning, subjects tend to treat probabilistic problems as if they were deterministic, and they generally follow heuristics rather than logical reasoning patterns. This has also been found to be the case with medical practitioners making medical diagnoses (Baron, 1988, Brehmer, 1980). Reasoning on the basis of heuristics can be an appropriate activity in medical diagnosis. Boreham (1989), for example, has modelled the heuristic knowledge expert physicians employ when deciding on the dosage of drugs required by individual patients. He found that production systems modelling such knowledge perform at an effective level and that they can be usefully employed in medical training by elucidating the thought processes behind difficult probabilistic decisions.

The extensive use of heuristics in medical diagnosis in combination with the
treatment of probabilistic information as if it were deterministic leads to a number of decision biases, however, which may influence the accuracy of diagnosis (Baron, 1988, Brehmer, 1980). Baron, Beattie and Hershey (1988), for example, found that doctors are subject to a confirmation bias in medical diagnosis. Even if an additional test will not add to deciding on alternative treatments, doctors will tend to request the test to be performed for the purpose of confirming their current diagnosis. In breast screening, this can mean that film readers sometimes opt to recall women for confirmatory evaluation, even if the radiological evidence indicates clearly benign abnormalities. Brehmer (1980) has discussed two other biases, assumptions of causality (determinism) and disregard of negative information, in the context of learning from clinical experience. He has argued that such biases make it impossible for physicians to learn from clinical experience. This contention and its implication for learning screening mammography will be considered in more detail later.

A further finding is that doctors lack statistical sophistication when combining contingent probabilities. Eddy (1982) investigated how physicians use the clinical information from mammography examinations in combination with clinical examination results in deciding whether a patient should or should not undergo a breast biopsy. He found that typically there was a neglect of base rate information. In addition, about 95% of physicians had difficulties distinguishing between the probability of cancer given a positive X-ray (relatively low) and the probability of a positive X-ray in a patient who has cancer (relatively high). The latter is commonly reported in medical test evaluations and most doctors are, therefore, more familiar with it. Ayton (1991) found that base rate effects operate on doctors' interpretations of neonatal heart beat traces during birth. Symptoms of distress in a number of traces were discounted on account of a low incidence of serious complications during birth. Weber (1992) also found that both base rate information and availability of the diagnosis (having diagnosed a similar case before) influence diagnostic decisions in probabilistic situations.
In screening mammography, both decisions on whether an abnormality is present and diagnostic decisions about the benignity or malignancy of an abnormality are based on both, the appropriate evaluation of probabilistic visual evidence and probabilistic abnormal features. The prevalence of decision support systems for mammography (Getty et al, 1988, Gale et al, 1987) is an indication of the importance for the accuracy of mammography of combining perceptual and conceptual features in an appropriate manner. Making decisions at this level has also been shown to be the major cause of radiographer error in screening mammography (see chapter 4). Learning to make decisions on the basis of the available evidence can thus be seen to be an important aspect of learning to read screening mammograms.

5.1.4 Summary

In this section, visual discrimination, conceptual categorisation and decision making have been identified and characterised as the three main elements of mammogram interpretation performance. It was argued that the recognition of significant visual features is not well accounted for by simple feature models of pattern perception. A better conceptualisation of perceptual processes in radiological interpretation is presented by more complex visual search mechanisms, which account for the influence of visual context and prior experience. Some aspects of visual feature recognition, in particular in highly practised performance, has been characterised as highly automatic and pre-attentive. These processes may be the result of visual categorical perception. It was further argued that conceptual categorisation involves two separate processes in screening mammography. Firstly, visual features have to be classified semantically as belonging to normal or abnormal categories, and, secondly, any abnormal findings have to be further classified as being indicative of benign changes or malignant processes. Both types of conceptual classification entail dealing with features which are only probabilistically associated with one or the other category. The final decision involves merging the information from visual features classified as normal or abnormal and abnormal features classified as benign.
or malignant. It was pointed out that such probabilistic decision making may be subject to decision biases, which may also affect the development of knowledge categories on the basis of experience. The analysis of mammogram interpretation presented in this section will be formalised in terms of a process model of the task in chapter 7. The following section presents a review of the literature on radiology education and training, which is followed in the final section by an analysis of learning mechanisms involved in acquiring mammogram interpretation skill.

5.2 Radiology and Mammography Education

Given the complexity of the skills involved in reading X-ray films, it is clear that observers should be carefully trained. However, Calhoun et al (1988) have noted 'a substantial lack in empirical research' (p. 62) in their review of the radiology education literature from 1966 to 1986. Of the 194 journal articles reviewed only 8.7% were experimental or quasi-experimental in nature. They concluded that the paucity of research regarding the effectiveness and efficiency of training programmes and pedagogical methods was particularly noteworthy in view of the unique aspects of radiology education. There are, however, some empirical studies which have addressed issues in radiology and mammography training. These are reviewed later on in this section.

Another question arising in the context of radiology education is whether initial training and continued exposure to radiographs is sufficient for the maintenance of high skill levels. There is general agreement in the literature that continuing medical education is necessary for radiologists in order to maintain and develop their skills. In addition, there is some evidence that radiologists' competence assessment is beneficial in order to establish the efficacy of programmes for skill maintenance or for specific changes in practice (Stevenson and Cockshott, 1988, Fajardo, Hillman, Hunter and Bjelland, 1987).
5.2.1 Initial Radiology and Mammography Training

The question of how radiology and, more specifically, mammography should be taught to novices has been addressed by a number of studies. As early research into radiology performance was essentially directed at the visual aspects of the diagnostic task, it is not surprising that educational programmes focused on the application of programmed learning principles to teaching the visual evaluation of radiographs as is evidenced in work by Tuddenham and his colleagues (1968, 1969) in general radiology and Alcorn and his colleagues (1969, 1971) in mammography. They suggested the use of flow charts or decision trees for teaching roentgen diagnosis, a hierarchical system of progressively more specific questions leading to the most specific diagnostic output, which was justified by the evidence in any given radiograph. These decision trees could be used in self-instruction and served to guide search and to facilitate the analysis of observed abnormalities.

Alcorn and O'Donnell (1969) argued that as the information on mammograms is based more on pattern recognition than on knowledge of anatomy, physiology and pathology, trainees were required to develop a memory store of patterns, which could be facilitated by the use of decision trees incorporating the characteristics of normal and abnormal breast patterns. Students were, thus, taught to search for specific findings and to develop a concept of normal breast background.

More recently, concordant with a shift in attention towards the more cognitive aspects of radiological diagnosis (see chapter 2), there has been increased interest in cognitive performance assessment (Curtis, Amis, Cruess and Riordan, 1985). Similarly, the focus of radiology training has shifted towards cognitive problem-solving skills in radiological diagnosis. Blane, Vydareny, Ten Haken and Calhoun (1989), for example, have attempted to teach radiologists to think independently. Their training scheme involved medical students solving diagnostic problems in small groups. Training was based on a task in which radiologists had to develop a logical imaging model for a particular problem in the absence of any actual imaging data.
5.2.1.1 The role of instruction

The question of whether instruction can facilitate radiology performance has been addressed with respect to computer-assisted instruction (Aronberg, Rodewald and Jost, 1985, Jacoby, Smith and Albanese, 1984). There has also been some research on the relative benefits of different instructional methods in radiology. Smith, Berbaum, Franken and Ell (1986), for example, compared lecture versus case presentation formats on residents' radiology performance. In mammography, there is also some evidence for the benefit of instruction on performance. Fajardo et al (1987) found that a seven hour course of focused instruction in mammography improved radiologists' factual knowledge, interpretative skills with respect to benignity, specificity and decision-making ability in respect of patient disposition. Overall there was a significant improvement in performance as measured by the area under the ROC. Amongst radiologists there was no improvement in sensitivity (see physician assistants below though).

5.2.2 Maintaining Radiology Skills

There is little research on how skills in radiology are maintained after initial training. There is a general notion that expertise in radiology as a cognitive skill is a function of experience (Lesgold et al, 1988), which results in a linear relationship between experience and performance. Comparisons between expert and novice performance in radiology seem to support this contention (Myles-Worsley et al, 1988). However, research by Herman and Hessel (1975) raises questions about the tenability of such claims. They found that after initial training, individual reader characteristics were more important in determining chest radiograph interpretation accuracy than either experience or length of training.

Nevertheless, training in screening mammography as advocated by the European Group for Breast Cancer Screening (1987) and by the NHSBSP (1991) is essentially based on providing the trainee with relevant experience by secondment to a training centre. An indication that experience might be important with respect
to at least some aspects of learning mammography was provided by Fajardo et al (1987). They found that, although focused instruction improved overall mammography performance in radiologists, there was no improvement in cancer sensitivity. A group of physician assistants, however, who received continued practical exposure in interpreting mammograms after focused instruction were found to increase their cancer pick-up rate significantly over a twelve week period. Data in the present research did not support this finding, though. Neither of the radiographers (chapter 4) achieved a stable true positive rate during ten months of breast screening experience. There is thus no question of experience having improved sensitivity in these two individuals. The question thus remains by what mechanisms mammogram interpretation skills are acquired and whether training can support the acquisition and maintenance of such skills. These issues are addressed in the following section.

5.3 What mechanisms are involved in learning mammography?

In order to investigate whether specifically designed training can be of use in learning mammography it is important to specify the learning mechanisms involved in mammogram skill acquisition. It can be concluded from the previous discussion of radiology skill that the interpretation of screening mammograms involves two quite different tasks, one being primarily visual and the other primarily conceptual. The two tasks are similar in that their performance is influenced by expectations and prior knowledge, but the cognitive systems involved in task performance are different. Learning to perform visual search and conceptual analysis of the radiograph may, therefore, involve different learning mechanisms. The separate discussion of these task aspects for the purpose of training analysis does not, however, imply that they are necessarily separable in skilled performance (see Swensson, 1980).

In this section, potential learning mechanisms for acquiring the perceptual and conceptual skills involved in mammogram interpretation are reviewed. Perceptual
learning is discussed with respect to the role played by experience and practice in
visual category acquisition, whereas conceptual learning is considered in terms of
induction, schema acquisition and learning by analogy. In addition, the role of
training and instruction in supporting the acquisition of mammogram interpretation
skills is addressed.

5.3.1 Perceptual Learning

Walk (1978) has defined perceptual learning as any 'change in perception over
time, which is based on experience'. In the visual domain perceptual learning is
based on the assumption that the visual system possesses a degree of plasticity, i.e.
that it can be modulated through experience or practice so that visual stimuli are
discriminated on the basis of certain salient attributes which were not discriminable
before exposure to these stimuli. In terms of radiological skill, perceptual learning
occurs when the radiograph reader discriminates between shadows on the X-ray
which he or she would previously have considered indistinguishable. In other
words, visual perceptual categories are developed which serve to distinguish, for
example, between normal and abnormal appearances in mammograms on the basis
of a number of attributes which have not previously been recognised as
distinguishing feature attributes by the observer. The mechanisms involved in such
perceptual learning are less than well understood, but learning in any perceptual
modality is by definition a function of continued exposure, professional interest or
practice (Gibson, 1953, 1969, Epstein, Hughes, Schneider and Bach-y-Rita, 1989).

Theories of visual perception have difficulties explaining the mechanisms of
perceptual learning not least because the concept of perceptual learning is only
loosely defined, and the term has been used in different meanings by different
researchers. Even with respect to radiology, perceptual learning as an underlying
mechanism for visual skill acquisition can have at least two different meanings.
Firstly, it might refer to stimulus differentiation and generalisation in a manner
suggested by Lesgold et al (1988) and, secondly, perceptual learning might result in
selective visual attention to critical features as is implicit in Getty et al's (1988) and Myles-Worsley et al's (1988) work.

5.3.1.1 Differentiation and generalisation

Two varieties of perceptual learning of particular interest to mammogram interpretation are those of differentiation and generalisation. Lesgold et al (1988) argued that radiological expertise is acquired 'through a cognitively deep form of generalisation and discrimination' (p. 340). Perceptual learning results in 'the formation and tuning of lower-order demons [in the pandemonium model of perception]' (p. 337), implying that in addition to schema development, which is about conceptual knowledge acquisition, there is a degree of learning which occurs purely at the level of visual analysis. This recourse to schema theory and the pandemonium model identifies Lesgold et al's position as somewhat mentalistic. The problem is that it is difficult to define precisely the perceptual learning mechanisms within such an approach. As indicated before, Lesgold et al's work falls short of providing an explanation of what exactly generalised and discriminated and how.

A different kind of approach to discriminative perceptual learning has its roots in the ecological theory of perception (Gibson, 1979) which postulates that all the information required for accurate perception is contained within the stimulus itself. There is no need to make a distinction between 'sensations which are bare and meaningless, and perceptions, which arise when sensations are integrated and supplemented with information derived from some source other than the stimulus' (E. Gibson, 1969, p.75). The mechanism of perceptual learning within this approach is one of stimulus differentiation and generalisation which comes about by 'creative enrichment' of the stimulus and its qualities and relationships with other stimuli (Gibson and Gibson, 1955, Gibson, 1969). The stimulus is conceived as an influx of subtle variations of energy which contains all the necessary information without a need for recourse to mentally enriched 'percepts'. Perceptual learning encompasses both, the discrimination of previously undetected differences
and the discovery of invariant relations and structures within the stimulus. In relation to perceptual learning in radiology, there is no need to specify any specific learning mechanisms in this framework as all the information is contained in the radiograph. Perceptual learning in the Gibsonian sense comes about purely through exposure or practice (Gibson, 1953, Gibson and Gibson, 1955), when the observer begins to notice subtle but meaningful variations in the appearance of the radiograph and uses them to generalise across or to discriminate between visual categories.

5.3.1.2 Visual attention

Another way in which perceptual learning may affect radiological performance is in terms of directing selective visual attention. Myles-Worsley et al (1988), for example, found that radiological experience has the effect of making the observer selectively more sensitive to clinically significant deviations in the X-ray at the same time as reducing sensitivity to clinically non-significant deviations. This implies that practice with radiograph interpretation serves to set up visual categories which can be recognised very rapidly by selective visual attention being directed at the relevant visual features (Getty et al, 1988).

The mechanisms involved in this kind of perceptual learning must be explained in terms other than the differences in the salience of stimulus properties. The differentiation theory of perceptual learning put forward by Gibson and Gibson (1955) does not account for how an individual comes to attend selectively to some stimulus attributes but not to others. Postman (1955), for instance, has argued that the theory fails to describe the processes that lead to increased differentiation. He suggested that association learning is a necessary mechanism in increased perceptual differentiation, an argument recently supported by experiments conducted by Tetewsky and Garner (1986). They found that the differentiation theory may not provide a complete explanation of perceptual learning in the classification of Hebrew and Roman characters. This leads to the conclusion that perceptual learning causes the perception of characters to become enhanced by
memory associations acquired during learning, which, firstly, permit greater stimulus differentiation and, secondly, facilitate selective attention to certain learned stimulus attributes. Whether association theory can account for very complex perceptual learning such as visual category acquisition in radiology is not at all clear, however.

A more plausible view of the mechanisms involved in complex perceptual learning is that put forward by Neisser (1976). He has also argued that the Gibsonian position is too limited to explain perceptual learning in an adequate manner. According to Neisser, anticipatory schemata direct visual attention to objects of interest, perception of which in turn modify the guiding schema. Such a 'perceptual cycle' is better able to account for the complex interactions of visual knowledge and what is perceived than Postman's (1955) simple memory associations. In radiology, Gale et al (1979) and Kundel and Nodine (1983) have argued that experience serves to build up a set of anticipatory visual schemata in Neisser's sense which guide the exploration of the X-ray. This explains how significant deviations can be focused on rapidly and that there is no need for an exhaustive search of the image, since critical features are attended to first. The perceptual cycle framework can also account for the misperception of critical features since perception is only as accurate as the anticipating schema that guides it, i.e. if the visual concept directing search is inappropriate or none is available, errors in both, perception and interpretation of the X-ray, may occur. Gale (1993) has argued that gross misinterpretations of radiographic features can be explained in terms of inappropriate visual schemata.

**5.3.1.3 The role of practice, training and feedback in mammography**

Perceptual learning is, by definition, brought about by practice in the sense of repeated opportunity to detect informative stimulus structure. Any training programme aiming to improve the visual search aspects of mammogram interpretation should, therefore, involve exposing the learner to large amounts of representative examples from the visual categories to be acquired. This is, of
course, what current screening mammography training procedures for radiologists in the NHS focus on. During their two-week secondment to a training centre, trainee film readers are provided with 'exposure to a wide variety of normal and abnormal mammograms' (European Group for Breast Cancer Screening, 1987). Current training in film reading is thus essentially concerned with the perceptual differentiation of mammograms. The influence of practice and training on such perceptual learning has been reviewed in detail by Gibson (1953).

Whilst the radiographer study reported in the previous chapter shows that training has a measurable effect on performance, there is no further measurable improvement during everyday screening, despite the fact that mammography is considered to be a skill with a learning curve improving over several years (NHSBSP, 1991). One of the reasons for the radiographers' failure to improve their performance during screening experience might be the lack of immediate feedback. Gibson (1953) has discussed the beneficial effects of reinforcement in perceptual learning. Although limited learning may occur even under no feedback conditions, reinforcement which is rich in information is much more effective in bringing about perceptual differentiation and generalisation. In addition, the more conceptual aspects of mammogram interpretation may not be learnable by experience alone (Brehmer, 1980). This point is discussed further below.

Whilst conditions for perceptual learning are impoverished in daily screening, the circumstances for knowledge acquisition in terms of semantic properties of visual features and their combinations may be even less favourable. One of the problems with learning from experience in screening mammography is that the outcome of recall decisions (i.e. whether the recalled women were returned to routine screening or whether they were referred for biopsy) is, at best, obtainable a few days after the initial recall decision has been made. In the case of lay film readers, this feedback is not usually available at all, since they are not involved in the assessment of recalled women. Thus, although lay readers receive immediate feedback during training (either by answers provided on teaching mammograms or from the
consultant radiologist in tutorial sessions), such favourable conditions for perceptual learning were not maintained in the post-training period.

5.3.2 Conceptual learning

In order to interpret mammograms correctly, the observer must possess semantic knowledge about features, their meaning and how they relate to each other. Such generic knowledge is, firstly, about visual features that are indicative of normality and abnormality and, secondly, about the likelihood with which an observed feature combination is indicative of benign or malignant disease. A clear distinction can be drawn between this semantic knowledge about features and their properties and knowledge about the physical nature of the visual features discussed in the previous section. Whilst the latter are likely to be highly individual and unique for different observers, semantic feature knowledge refers to those aspects of visual features and their attributes which can be abstracted and shared by different observers. Thus, experienced mammography screeners are able to discuss an observed mammographic abnormality in terms of broad linguistic categories such as a 'mass' or an 'asymmetry', yet complete consensus on which visual features are of primary importance in defining the particular abnormality is quite rare even amongst experts (Gale et al, 1989).

As indicated above, there are two kinds of generic knowledge that can be distinguished in mammography, semantic feature knowledge and knowledge of disease categories. The former is essential for all film readers, whilst the latter is not important in pre-screening. As a 'second pair of eyes', however, even the lay reader must acquire some idea of which disease features indicate benign processes and which are indicative of malignancy. This may be difficult since they lack the detailed anatomical and pathological knowledge fully trained medical practitioners would be expected to possess. Although the application of feature knowledge is essentially data-driven, whilst disease knowledge tends to be applied in a hypothesis-driven manner, the representations in memory of both are likely to be
quite similar in that they are probably rule-based (Medin and Barsalou, 1987). Mechanisms involved in the acquisition of both kinds of knowledge are addressed in this section.

5.3.2.1 Acquiring knowledge by induction
The acquisition of feature knowledge is largely a matter of abstracting significant perceptual feature attributes in terms of broad conceptual categories as has been discussed previously. Acquiring disease schemata is more complex and involves the development of integrated conceptual patterns which combine knowledge of perceptual features, anatomy and pathology in meaningful ways. It has been argued that there is only one coherent account of how such concept learning is achieved. Fodor (1976) postulated that all conceptual learning is a process of inductive extrapolation involving the formation, testing and confirmation of hypotheses. Although this is an extreme position, there is some truth in that the acquisition of new concepts is critically dependent on the state of the learner's individual representational system. Thus external agents for learning such as training, practice or instruction are of use only if the learner is able to assimilate the material presented. Such assimilation of training material can be construed to be an inductive process even under conditions of directed instruction.

The mechanisms involved in the acquisition of feature and disease knowledge for mammogram interpretation can thus be conceptualised as an inductive process of specialisation and generalisation of knowledge structures (Holyoak and Nisbett, 1988). This description is consistent with Lesgold et al's (1988) account of the acquisition of radiological expertise. Specialisation is about developing knowledge structures which take account of the variability observed in radiographs, whereas generalisation serves to exploit this variability in a manner that reduces variability. An alternative inductive mechanism is learning by analogy (Winston, 1980). This is implicit in tutorial based radiology training which provides the trainee with case histories of previous cases. New cases, which may contain some new information, can be solved by analogy with previous cases. Such analogical reasoning processes
lead to the adjustment of feature and disease schemata in the light of novel information from new cases.

Both kinds of inductive processes, analogical reasoning and generalisation/specialisation, thus lead to the formation of rules about the conceptual properties of visual and disease features. This does not imply, however, that rules acquired by induction are necessarily explicit. Increased knowledge may, in fact, serve to obscure such explicit knowledge as it is automatically applied in appropriate situations. Film readers in mammography find it difficult, for example, to express the rules they apply in interpreting mammograms verbally, as was evident in the radiographer study in chapter 4. Hunt, Marin and Stone (1966) conducted a series of experiments on induction using naturalistic, complex concepts. They showed that subjects have difficulties in abstracting certain rules explicitly, even if they were given feedback on why a particular instance belonged to a particular category. The indication is that rules may be tacitly abstracted, since subjects were able to apply them given a specific example. This approach by Hunt et al illustrates that induction is about the discovery of both, which properties to look for and how to best look for them.

5.3.2.2 The role of experience

Winston (1978) has characterised induction as relatively autonomous learning. The indication is that experience with many examples is the most efficient learning mechanism in order to achieve adequate generalisation and specification of knowledge structures required for medical diagnosis (Gilhooly, 1990). However, Brehmer (1980) has argued that certain complex skills, in particular skills in medical diagnosis, cannot arise from pure exposure to cases. There is a considerable body of literature which indicates that experience does not increase the practitioner's competence in making diagnoses. Apparently, experience only serves to increase doctors' confidence in the correctness of their decision (Einhorn and Hogarth, 1978). Brehmer (1980) argues that this is the case because the information contained in clinical experience is too minimal to be beneficial in
further learning. The reason for this is that people are not cognitively equipped to
deal with probabilistic information for the purpose of reasoning and, therefore,
cannot benefit from such information in terms of increasing their knowledge.

Brehmer (1980) also discusses the case of learning, when decision outcomes are
dependent on judgement as is the case in most medical decision tasks. When the
outcome of a particular clinical decision is dependent on the treatment
recommended, a number of factors may lead practitioners to conclude falsely that
their judgement was valid. In screening mammography, for example, the breast
 screener only sees the outcome of true positives and false positives. As these
represent only a small proportion of all screening cases most of the information on
which learning could be based is not available. In the absence of knowledge about
their false negatives, film readers may come to believe that the accuracy of their
referral decisions is increasing as they are becoming more confident in making
these decisions.

5.3.3 The role of training and instruction

Brehmer's work indicates that experience is insufficient as a learning mechanism in
most medical diagnosis. In mammography screening, visual information is of much
greater importance than in other medical diagnosis. It was argued above, that
practice, especially in conjunction with appropriate feedback, plays an important
role in learning to deal with the perceptual variation observed in mammograms.
With respect to knowledge about the meaning of features in screening
mammography, however, be it visual knowledge or disease knowledge, the
indication is that specific feature training and instruction may be more beneficial
for improving skilled performance than experience alone.

Evidence for the notion that visual feature knowledge can be trained by instruction
comes from the study of perceptual learning in a task that shares many features of
screening mammography, although the it is set in a completely different domain.
Biederman and Shiffar (1987) investigated the impact of focused instruction on chick sexing performance, a task which involves the discrimination of ambiguous visual features of day-old chicks' genitals in order to classify them as male or female. Like mammography screening this task is normally performed by highly skilled individuals who have been in exposure training for a number of years. Biederman and Shiffar found that brief instruction, consisting of an instruction sheet which could be read in approximately one minute, improved previously naïve subjects' classification performance to a level comparable to that of professional chick sexers. A control group which had not received any instruction did not improve their chick sexing skill. These dramatic results have been explained in terms of enabling the trainee where to look and what to look for (see also Hunt et al, 1966), which is typically difficult to work out in complex visual displays. Alcorn et al's (1969, 1971) approach to mammography instruction was also aimed at enabling the trainee to search for features in a goal-directed manner. Programmed instruction of this kind guides the learner to look for specific features in specified locations.

Although high accuracy in visual discrimination tasks can be achieved with relatively brief instruction, it has been noted that a much longer learning period may be required to learn to deal with rare or difficult cases. This is also implicit in the notion that mammography skill is learned best with relatively brief initial training with the assumption that there is a prolonged learning curve which typically improves over some years (NHSBSP, 1991). Biederman and Shiffar have argued, though, that difficult and rare case classification may also be aided in terms of 'an album of such instances' (p. 644). In mammography training centres, collections of past recall cases are available for training, and their use as instructional media is strongly encouraged. In this way continual instruction or self-instruction could be implemented in order to maintain high levels of mammogram interpretation skill in all screeners.

In summary, different learning mechanisms may be involved in mammogram
interpretation skill acquisition. While visual discrimination of features is possibly best achieved by experience with relevant visual displays, semantic feature knowledge and knowledge about how features combine to indicate malignant abnormality is more difficult to acquire by exposure alone. Specific feature training may be particularly helpful in aiding novices to find out rapidly what to look for and where. In the next chapter, the question of how useful such feature training is in learning mammography interpretation compared with current NHS training methods is addressed on an empirical level. Feature training might be particularly appropriate for lay film readers, because they lack knowledge of anatomy and pathology which might represent the basis on which medical practitioners build their radiological knowledge during exposure training.
Chapter 6

Perceptual Learning or Clinical Reasoning: An Evaluation of two Training Methods for Lay Film Readers

The previous chapter was concerned with an exploration of the theories of relevance to the skills involved in mammogram interpretation and their training. It was argued that there are two main aspects to X-ray film interpretation, a visual perceptual one, which involves searching the film for relevant visual features and a conceptual, knowledge based one, which entails the evaluation of visual features in terms of their semantic properties in the context of the whole image. In screening mammography, the latter pertains to the analysis of features with respect to normality or abnormality, but it also includes further evaluation of abnormal features with respect to whether they are likely to represent benign or malignant processes in the full interpretation task.

It was demonstrated in chapter 4 that two radiographers could be trained to perform mammogram interpretation to a standard comparable to radiologists. The data reported in that chapter were case study data, however, which may not be representative of lay film readers in general. In this chapter, the question of whether it is possible to train lay people to read screening mammograms is addressed on a more general level. Radiographers are the obvious candidates for lay film reading because of their proximity to the screening programme. However, it is a legitimate question whether radiographers' previous training makes them ideal lay film interpreters or whether other lay people unrelated to the breast screening programme can be trained to perform the same skills. This question is similar to the current debate on whether consultant breast screeners should be radiologists or whether any medical practitioner can perform breast screening tasks.
6.1 Comparing feature training and medical reasoning training

The kind of training that lay film readers should receive is the primary issue addressed in this chapter. In the previous chapter, a range of possible learning mechanisms for acquiring screening skills were discussed. It was argued that even visual skill acquisition need not be entirely dependent on experiential learning which may not, in fact, be the most efficient way of learning about the properties of visual features. Biederman and Shiffrar (1987) showed that very brief instruction on visual features can dramatically improve performance in complex visual skills. In this chapter, a tutorial method of instruction is compared with one which emphasizes feature recognition. It must be stated again, though, that the training that lay observers should receive depends on a clear definition of their role in the screening context. It is, therefore, clear that there may be more than one appropriate approach to training.

6.1.1 Training lay film readers

In this chapter, a training study is reported in which undergraduate students were given brief training in mammogram interpretation. The participants in this study represented a group of people of similar educational level to radiographers who have typically been educated to 'A' level standard. Recent moves to professionalise nursing and other paramedical occupations have also led to a large number of radiographers attending part-time degree courses. Unlike radiographers, the student participants did not have any medical background and little or no knowledge of breast anatomy. One of the issues addressed in this study was whether students without any prior knowledge can learn to read mammograms with relatively brief training. Training for radiographers entering breast screening includes some aspects of breast anatomy as it relates to the technical aspects of taking the breast X-ray. It is possible that such related knowledge is a prerequisite for acquiring film reading skills, thus excluding lay people without prior medical knowledge from being able to acquire mammogram interpretation skills. However, since mammogram
interpretation in the breast screening context can be performed purely as a visual
discrimination task, it was predicted that even students without any medical
knowledge could be trained to make simple recall decisions on the basis of whether
an abnormality was perceived or not.

6.1.2 Comparing two training methods

In the last chapter, it was argued that training schedules for lay film readers must
take account of the fact that they lack the prior knowledge that present trainees
from medical backgrounds possess. Effective training needs to be specifically
designed for the target group taking into account their characteristics and prior
knowledge (Stammers, 1987). The tutorial method, currently employed in breast
screening, was designed for training radiologists and other medical practitioners and
may not yield the best results when training lay film readers. The question of what
represents 'best results' depends, to a large extent, on how the role of lay observer
is defined in the context of the whole breast screening programme. As discussed
before, there are essentially two kinds of role the lay reader could take, a pre­
screening role or a dual screening role ('second pair of eyes'). Whilst the former
demands the detection of abnormality only, the latter involves emulating radiologist
performance entailing some diagnostic decision making.

Although previous research has shown that lay-screeners can generally attain
performance levels similar to those of medical practitioners, the indication is that
the nature of film reading performance may differ with different training methods.
Alcorn et al (1971) used a training method orientated towards feature detection and
found that lay readers had a higher recall rate than medical practitioners, and they
produced more false positives. Their true positive rate matched that of the
radiologist, though. Dowdy et al (1970) and, more recently, Haiart and Henderson
(1991) found that lay readers produced true and false positive rates, which were
similar in proportion to those of radiologists after receiving tutorial based training,
but their overall performance was slightly lower than that of the radiologist. It
would, thus, seem that a feature based instructional method would be more suited to training pre-screeners, whilst the tutorial method encourages lay film readers to emulate radiologist performance.

As the two training methods under discussion appear to produce different performance outcomes, it is likely that the skills acquired are different. The tutorial method is aimed at generating increased differentiation between normality and abnormality by exposure to a wide variety of normal and abnormal mammograms. This exposure approach is consistent with the mechanisms of perceptual learning (see, for example, Gibson, 1953). In addition, this method seeks to develop the trainee's ability to make clinical decisions about perceived abnormality in terms of benignity or malignancy. Radiologists and medical practitioners with the prerequisite anatomical and pathological knowledge probably learn to build mental 'patient representations', which are three-dimensional models of the two-dimensional visual data as a result of such instruction, as this provides them with richer information on which to base their decisions. Lay readers, however, are less likely to possess the medical knowledge necessary to form such representations, unless they have received specific instruction in the anatomy and pathology of the breast. Without such instruction, lay film reading performance may be impoverished because it is based on limited information contained directly in the two-dimensional visual data.

The idea that feature instruction might be a more efficient agent for training lay film readers than pure exposure training, derives from Biederman and Shiffar's (1987) work on chick sexing, which was discussed in the previous chapter. Feature training aims to provide the trainee with a clear idea of what to look for and where in their search of the X-ray. There is no need for building three-dimensional representations because trainees learn to maximise the information available from the visual data. There is some evidence that exposure training in the visual medical domains leads to automatic recognition of normality rather than abnormality (Schmidt et al, 1990, Myles-Worsley et al, 1988). Since the detection of
abnormality (true positives) is of such vital importance in breast cancer screening, it makes sense to direct lay reader training more specifically at the recognition of abnormal features. The comparison of the feature method with the tutorial or medical reasoning method was expected to demonstrate that participants trained to recognise abnormal features would yield a higher true positive rate albeit at the cost of a higher false positive rate. This would render the method superior for training pre-screeners.

6.1.3 The training study

The training study was of a longitudinal test-training-test design. Performance was measured prior to and after training and one week after the completion of training. The experiment on training methods was conceived as a 2x3 factorial study with a mixed design. The between subjects factor represented the two training methods to be compared, feature training and medical reasoning training. The second factor was mammogram interpretation performance measured at specified intervals on three occasions before and after training (see below). This was a within subjects factor.

6.1.3.1 Subjects
Seventeen undergraduate students, eleven female and six male, from the University of Surrey took part in the mammography learning study. Their ages ranged from eighteen to thirty-seven with a median age of nineteen. Subjects were assigned to one of the two training methods under investigation (see below for details). The feature training group consisted of eight undergraduates, five female and three male with a median age of twenty. There were nine people in the tutorial-based group, six of which were female. Their median age was nineteen. All participants took part in the experiment voluntarily. They were required to attend for approximately ten hours of training and performance assessment over a two week period and were paid £15 each. Such a payment can only represent a token of appreciation rather than remuneration. The commitment of all participants was thus based on intrinsic
motivation to contribute to the research. This notion is supported by the fact that none of the subjects withdrew from participation in the experiment during the two-week period.

### 6.1.3.2 Training and assessment regime

The study was scheduled to run over two weeks. Subjects were required to attend sessions every day during the first week and a follow-up test session at the end of the second week. On the first day, all subjects attended an introductory lecture (Appendix IIa), which is described in more detail below in the section on instructional methods. In addition, instructions were given with regard to the pre-training performance assessment (see Appendix IIa). Subjects were then required to complete the mammogram test for the first time.

After the introductory day, participants were assigned to one of the two experimental groups. The first group received three sessions of structured, hierarchical feature instruction on three consecutive days. A schedule and detailed contents of the feature training sessions can be found in Appendix IIb. The second group received the same amount of tutorial-style instruction based on specific cases (see detailed descriptions in the instructional methods section). Both training groups were asked to complete a self-assessed mammogram test at the end of each of the three instructional sessions. This test was the same for both groups and was followed by immediate feedback.

On the last day of the first week, all participants again completed the mammogram test in the same way as on the first day. This was the post-training assessment. Subjects returned one week later to evaluate the seventy-nine mammograms once more. This represented the follow-up assessment in order to ascertain retention of mammogram interpretation skills in the longer term. There had been no further training or exposure to mammograms in the time between the post-training assessment and the follow-up assessment.
6.1.3.3 Training Materials

Mammograms for training were drawn from a pool of 182 breast X-rays depicting sets of left and right oblique view (45° angle) mammograms. Eighty-seven of these were abnormal mammograms, and for twenty-nine of these a second slide showing the abnormality in a different view or in magnification was available. The other sixty-six mammograms showed either no abnormality or clearly benign features. These normal mammograms had been used in a mammogram test set (similar to the present test set of mammograms described in chapter 4) at the Jarvis Screening Centre until 1989. Forty-five of these were randomly selected for use during training. The abnormal mammograms were taken from the teaching files of the Jarvis Breast Screening Centre in Guildford. They were the mammograms of most of those women who had been recalled for a mammographic abnormality and had been recommended for biopsy between December 1990 and July 1991.

As the detection of some mammographic abnormalities is critically dependent on image resolution, the slides depicting abnormal mammograms were assessed by two fully trained film readers (one radiologist and one radiographer). Ratings of slide quality with respect to the presence and visibility of an abnormality and of the difficulty of both, perceiving the abnormality and making a decision about its significance were obtained. Only slides which were correctly assessed (the abnormality had been correctly identified by both raters) and considered suitable for training by both observers were used in the training sessions. This selection procedure resulted in forty-six abnormal mammograms, ten of which had a second slide showing the abnormality in magnification, being employed for training. Magnification slides were considered necessary to demonstrate abnormalities such as microcalcification in a satisfactory manner. Such abnormalities are often only visible on original mammograms when viewed under a magnifying glass, which is a piece of standard equipment for mammographic film readers. All magnification views employed in the present training study showed microcalcifications.

The difficulty ratings obtained from the two raters were used to categorise the
abnormal training mammograms as easy, medium or difficult depending on the perceptual difficulty ratings. They were also classified in terms of the type of abnormality they exemplified. Each abnormal mammogram was categorised as predominately representative of a mass, calcification or minimal feature abnormalities. The minimal feature category comprised all vague mammographic abnormalities, specifically asymmetry and distortion of parenchymal pattern. This cross-classification of abnormal mammograms resulted in nine categories of mammograms as shown in Table 6.1. It can be seen that calcifications were mostly categorised as 'perceptually easy' while minimal abnormalities represent the most difficult mammograms.

Table 6.1: Number of abnormal training mammograms in each category after cross-classification in terms of perceptual difficulty and abnormality type.

<table>
<thead>
<tr>
<th>perceptual difficulty</th>
<th>easy</th>
<th>medium</th>
<th>hard</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>mass</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>calcification</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>minimal feature</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>total</td>
<td>24</td>
<td>16</td>
<td>16</td>
<td>46</td>
</tr>
</tbody>
</table>

In addition to mammograms displayed on a slide projector, thirty real mammograms were utilised in training for self-administered testing (see instructional principles below). Half of these exhibited an abnormality whilst the other half were normal. The abnormal mammograms were all derived from the Jarvis Breast Screening Centre teaching files. These were cases that had been referred for a biopsy between September 1990 and November 1990. The normal mammograms in the test were taken from the mammogram test set which was used for performance assessment.
6.1.3.4 Instructional methods

As the aim of this study was to compare two training methods, subjects were split into two groups for the majority of the training as described below. Some aspects of training were the same for both groups, however, namely the introductory lecture and the self-assessed mammogram test which followed each training session. The introductory lecture served the purpose of introducing the purpose of the study and the basic principles of breast screening. This included brief introductions to breast screening in Britain, mammograms as soft tissue x-rays, technical aspects of mammography, breast anatomy and how it appears in mammograms, structured viewing methods and a description of basic categories of abnormality (mass, calcification and minimal abnormality). A transcript of the introductory lecture can be found in Appendix IIa. This introduction was deemed a necessary prerequisite for students the majority of whom had had no previous contact with breast screening.

For the self-assessed mammogram test, participants examined ten sets of real mammograms immediately following each instructional session. These were displayed on light boxes under the same conditions as in the performance assessment test (see performance assessment below). Half of these mammograms had an abnormality, and half were normal. On completion of the test, subjects were given a sheet with the clinical descriptions of each mammogram and were provided with an opportunity to review each case. The inclusion of the self-assessed mammogram test in the training regime served two purposes. Firstly, it provided participants with an opportunity to view mammographic abnormalities and normal mammograms as they appear in the screening context, i.e. as X-rays rather than on slides. This was important because the images may be degraded when displayed as slides, thus preventing optimal opportunities for perceptual learning. The second reason for including the mammogram test was that it provided an element of active learning. Participants experienced the interpretation process in a non-threatening, non-assessment situation and were able to apply newly gained information. The principle of active involvement in learning has been stressed in developmental,
educational and instructional contexts by, for example, Piaget, Bruner and Gagne.

6.1.3.5 Hierarchical feature training
Training for the feature instruction group was based on Biederman and Shiffar's (1987) principles of instruction in complex visual tasks on the one hand, and on information processing approaches to instruction (Gagne, 1985) on the other. Thus, the aim was to present material which enabled participants to know what they were looking for and where to look for it in the first instance, and secondly to structure the material hierarchically, so that it could be assimilated more easily. The three training sessions conducted under this method each involved a half hour lecture about the visual characteristics and their meaning of one of the three types of abnormalities (see Appendix IIb). This was followed by the showing of a set of slides of that kind of abnormality. The example mammograms were ordered in terms of perceptual difficulty, the ones rated as easy being shown first. Each mammogram was analysed in terms of the features described in the lecture. The feature training group, thus, saw all the mammograms described in Table 6.1, masses being the subject of the first training session, calcifications that of the second and minimal abnormalities that of the third.

6.1.3.6 Tutorial-based clinical reasoning training
Tutorial training was based on currently prevailing training methods in breast screening, which consist largely of exposure to normal and abnormal mammograms and discussions between expert film readers and trainees. The aim of this method was to facilitate perceptual learning and to enable participants to reason about visual medical evidence, thus emphasising the decision making aspects of the film reading task. Subjects undergoing training according to this method saw all the abnormal cases presented to the feature training group but, in addition, were shown a matched number of mammographically normal X-rays. There was no introduction to abnormal features and their characteristics other than the very brief remarks in the introductory lecture on the first training day. As in the feature training group, all examples were presented on slides. Subjects in the tutorial group were invited
to voice an opinion with regard to whether they thought an abnormality was present. Active involvement in making decisions about example cases were encouraged. After a group consensus had been reached, they were read a clinical description of the case which had been written by an expert screener. All medical jargon was removed from descriptions prior to reading them out, but the contents of the clinical descriptions was preserved. For this group, the slides were presented in order of difficulty over the three training days. They saw all mammograms classified as easy in Table 6.1 on the first training day, the medium ones on the second and the difficult ones on the third.

6.1.3.7 Performance assessment
Performance was assessed before and after training with the same test set of seventy-nine mammograms that was employed in the radiographer study. This has already been described in some detail in chapter 4 (see also Appendix IV). In addition, participants were required to complete the test set for a third time one week after the completion of training. This served to assess the long-term retention of film reading skill without further training or exposure to mammograms. As in the radiographer study, participants did not receive any feedback on their performance on the mammogram test at any time during the study. The use of the same assessment set in the present training study renders performance measures directly comparable to those of radiographers and radiologists reported in chapter 4.

The test mammograms were viewed on ten light boxes, and magnifying glasses were available for detailed examination. Half of the seventy-nine sets of mammograms were displayed on light boxes, the other half were lying in folders in front of the light boxes. Subjects were required to change over the viewed x-rays for those that had not yet been assessed. The order in which mammograms were viewed was random for each subject. They were required to record whether a mass, calcifications or other abnormalities were present and whether they would recall the woman on the basis of mammographic evidence. There were four possible decision outcomes, recall, probably recall, probably no action and no action. Subjects were
also requested to record the location of all observed abnormalities in a schematic diagram of two obliquely projected breasts. Responses were recorded on simplified mammogram reporting sheets similar to those employed in the radiographer study. An example is shown in Figure 6.1.

**Figure 6.1:** Recording form for mammogram interpretation of features and decision outcomes. The form is a simplified version of the one used in the radiographer study.

![Recording form for mammogram interpretation](image)

6.1.4 Results

Film reading performance for both groups was measured in a manner identical to that in the radiographer study, using the G-index (Lienert et al, 1987, Hammond and Lienert, 1992) as the main index of overall classification performance irrespective of a particular decision criterion, and true and false positive data as indicators of underlying decision performance. Since performance was assessed using the same test set of mammograms as in the radiographer study, students' performance could be directly compared to radiographers on the one hand, and radiologists (Ansell et al, 1990b) on the other. Students' performance data were further analysed with Analysis of Variance with training group as the between
subjects variable and assessment time (pre-, post- and follow-up assessment) as the within subjects variable. Both, overall performance (G-index) and underlying decision performance (true and false positives) were examined with respect to differences in the two training methods. The G-index is primarily a correlation but, alternatively, it can justifiably be interpreted as the proportion of true to false classifications. Moreover, the false and true positive quantities are expressed as proportions of total normal and abnormal mammograms. Since proportional data may violate the assumption of a normal distribution, which is prerequisite of the validity of ANOVA, the analyses were also performed with arcsine transformations (Howell, 1987). The analyses of variance reported in this chapter are all based on original data, but summary tables for equivalent analyses on the transformed data can be found in Appendix Id. Comparison of the analyses on transformed and original data revealed that the outcomes are similar and statistical significance remains unaffected.

6.1.4.1 Students' film reading performance as a function of training
Mammogram interpretation performance of students from both instructional groups is shown in Figure 6.2 in comparison with radiographers' and radiologists' performance on the same test set as previously reported in chapter 4 (see Table Ia.1, Appendix Ia).

Figure 6.2: Performance of two student mammogram reading groups compared with radiographer and radiologist performance as measured by the G-index.
Test 1 and test 2, respectively, represent pre- and post-training performance for students, radiographers and radiologists. The third test represents the one week follow-up test without further intervention for the students. The third and subsequent tests for the radiographers were two months follow-ups with intervening screening experience as described in chapter 4.

It can be seen that overall performance increases over the course of the three assessments, but the differences between the two training methods in terms of overall performance are minimal. Analysis of variance confirmed that performance increases significantly over the course of the three assessments ($F = 8.4$, $df = 2, 30$, $p < .002$). There were no effects of either training method or the interaction between training method and assessment. The full analysis of variance table is shown in Table 6.2 (see Table Id.1, Appendix Id for Anova on arcsine transformed data).

Table 6.2: Analysis of Variance table for overall film reading performance measured with the G-index.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>&lt;.01</td>
<td>1</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>1.47</td>
<td>15</td>
<td>.098</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.42</td>
<td>2</td>
<td>.21</td>
<td>8.4</td>
<td>&lt;.002</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>.02</td>
<td>2</td>
<td>.01</td>
<td>.4</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.75</td>
<td>30</td>
<td>.025</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.1.4.2 Decision making performance as a function of training

As in the radiographer study, the underlying nature of overall improvement can be examined in terms of the true positive and false positive quantities. Figure 6.3
shows these for both, the feature training (n=8) and the medical reasoning (n=9) groups (see Tables a.2 and Ia.3 in Appendix Ia).

**Figure 6.3:** Probabilities of true and false positives for the feature training group (n = 8) and the medical reasoning group (n = 9).

It can be seen that in both groups, improvement is predominantly due to a decrease in false positives. The true positive rate only increases marginally in the post-training assessment, but while this increase is maintained in the follow-up assessment by the feature training group, there is a drop-off of the hit rate in the medical reasoning group. Analysis of variance revealed that there were no effects of either assessment or training group with respect to true positive performance (Table 6.3, see also Table Id.2, Appendix Id). The decrease of false positives over the three assessments was significant, however (F = 6.66, df = 2, 30, p < .005). Neither training group nor the interaction between training and assessment yielded significant effects. The full analysis of variance table is shown in Table 6.4 (see also Table Id.3, Appendix Id). An unrelated t-test was performed to examine the difference in true positive rate in the follow-up assessment between the two groups. There was no significant difference between the groups (t = .76, df = 15, p > .05).
Table 6.3: Analysis of Variance table for true positive (hit) classification performance.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>.01</td>
<td>1</td>
<td>.01</td>
<td>.189</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.80</td>
<td>15</td>
<td>.053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.02</td>
<td>2</td>
<td>.01</td>
<td>.4</td>
<td>n.s.</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>.01</td>
<td>2</td>
<td>.005</td>
<td>.2</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.75</td>
<td>30</td>
<td>.025</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: Analysis of Variance table for false positive (false recall) classification performance.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>&gt;.01</td>
<td>1</td>
<td>&gt;.01</td>
<td>.00</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.73</td>
<td>15</td>
<td>.049</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.16</td>
<td>2</td>
<td>.08</td>
<td>6.66</td>
<td>&lt;.005</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>.01</td>
<td>2</td>
<td>.005</td>
<td>.41</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.37</td>
<td>30</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.1.4.3 Variability in performance as a function of training method

Figure 6.4 shows the standard deviations from mean film reading performance as measured by the G-index plotted separately for the two training groups (see Appendix Ia). It can be seen that there is a large difference between the variances on pre-training assessment. This is a significant effect ($F = 6.72$, df = 7, 8, $p < .01$), which is maintained to a lesser extent during post-training performance.
(F = 4.48, df = 7, 8, p < .05), but not in the follow-up assessment. Feature training appears to reduce variability between learners, whilst variability in the medical reasoning group remains the same. The reduction of variability in the feature training group between pre-training and follow-up assessment is a significant one (t = 5.74, df = 6, p < .005), whereas there are no significant changes in variability in the medical reasoning group.

Figure 6.4: Standard deviations from mean performance over the three assessments for feature training and medical reasoning groups.

The chance finding of heterogeneity of variance between the two training method groups puts into question the validity of the analyses of variance reported above on account of the assumption of homogeneity of variance underlying analysis of variance. Box (1954) has suggested a conservative test of significance by comparing the obtained F ratio with F (1, n-1) to compensate for differences in variance between the groups (see Howell (1987) for a discussion of this). The appropriate conservative degrees of freedom in the present case are 1 and 16. It can be seen that with a critical value of F_{0.05} (1, 16) = 4.94, all previously reported significant differences remain unaffected.
6.1.5 Discussion: Is feature training more effective than decision training?

The main issue addressed by this training study was the question of whether different training methods result in different performance in film reading performance. In particular, it presented a direct comparison between a feature centred method based on instructional principles in complex perceptual learning (Biederman et al, 1987) and a decision centred method, which had been developed for training radiologists. It was argued that the appropriateness of a particular training method might depend on whether the lay film reader role is one of prescreening or of second screening as the decision criterion adopted for the former role should be substantially lower than that adopted for the latter. Feature training was thought to lead more readily to the adoption of such a lax criterion (Alcorn et al, 1971).

Figure 6.2 and Table 6.2 clearly show that performance in both training groups was very similar. There was no main effect for training method nor any interaction between training method and assessments in the analysis of variance. Closer examination of performance in the two groups in terms of true and false positives (Figure 6.3 and Tables 6.3 and 6.4) confirm the overall performance results. There are no significant differences between the groups in terms of either hit rate or false recall rate. In both cases, there is a significant reduction in false positives over time (Table 6.4). Figure 6.3 shows that the true positive rate rises steadily, but very slightly over all three assessments for the feature group, whilst for the medical reasoning group an initial increase in hit rate after training is followed by a sharp drop-off of the true positive rate. No significant effects were observed between or within groups with respect to true positive performance.

The results with respect to true and false positive performance lead to the conclusion that the two training methods under investigation do not result in differential performance in mammogram interpretation. Although, a tendency to maintain a higher hit rate and a very slight indication of a higher recall rate (as
evidenced in the higher false positive rate) may be observed for the feature training group (n=8) in Figure 6.3, none of these differences are sufficiently large to permit even tentative conclusions in the predicted direction. Failure of the two training methods to differentiate between the groups is contrary to what would be expected from previous research. Mammography training studies conducted by Alcorn and O'Donnell (1968, 1969) and Alcorn et al (1971) have shown that feature-orientated training leads to the adoption of a relatively laxer decision criterion of lay film readers when compared with radiologists.

One of the reasons for failing to replicate this finding might be a less stringent adherence to feature learning principles in the present study as compared to Alcorn's programmed learning approach. Although features were of prime importance during the instructional sessions, subjects had an opportunity to view normal and abnormal mammograms side by side in the self-assessed test following each instructional period. This may have induced an element of traditional focus on decision making based on visual prototype formation to the feature training. The self-assessed mammogram test had been introduced to the training schedule for didactic reasons. It was considered important that participants should have an opportunity to take an active part in the learning process. In addition, this was a way of ensuring that subjects knew what they were looking for in real mammograms as mammograms presented as slides do not necessarily match the resolution of real X-rays. As the self-assessed mammogram test was part of both groups' training, differences between them effected by instructional intervention may have been obscured to some extent by the addition of the didactic mammogram test.

6.1.5.1 Differences in variability

Although there were no differences in terms of mean performance between the training groups, there was a notable difference in variance prior to training. Since assignment of subjects to the training groups had been at random, this heterogeneity of variance is a result of chance fluctuation. There were no subject
characteristics, which could conceivably account for the larger variance in performance in the feature training group. However, the effect of training on variability in the feature training group is an interesting one. It appears that such training has a homogenising influence on performance variability. Figure 6.4 demonstrates that there is a large drop in variability in the feature training group, while medical reasoning training has no variability reducing effect. On the contrary, variability between subjects had increased slightly at follow-up compared with pre-training variability. The reduction in variance in the feature training group represents a statistically significant effect ($t = 5.74, p < .002$, two-tailed). Feature training can, thus, be seen to have a homogenising effect on trainees, i.e. their performance becomes more similar and individual differences become less important.

Individual variability in film reading performance presents a problem in mammography screening, because it is more difficult to maintain even performance standards, if some individuals perform far below the average, whilst others operate on a much better level. In particular, lay observers in a pre-screening function need to maintain a consistently high performance level. From the point of view of lay film reader training, the feature approach might thus present a distinct advantage in terms of performance homogenisation over the more traditional tutorial approach, which might even increase individual differences. In order to establish the reliability of the homogenising effect of feature training in mammogram interpretation training, a further training experiment was conducted with the aim of replicating the reduction in variability as a function of such training.

6.2 Feature training and individual differences in film reading performance

Although the data presented in the previous section led to the conclusion that feature training results in a significant reduction in post-training inter-observer variability, substantial individual differences are likely to remain even after such
training. In the present section, an attempt is made to assess systematically the outcome of feature training with respect to overall performance and with specific emphasis on individual differences in mammography film reading. For this purpose, a further training study employing the feature instruction method was conducted. The main aim of this study was to increase the number of people trained by this method, so that the conclusions drawn on film reading performance as a function of such training would be more reliable. Performance data from the present training study were added to those obtained in the feature instruction group in the training methods study reported in the previous section. In this manner, feature training could be examined more reliably in terms of outcome, both with respect to classification performance and with regard to individual differences.

6.2.1 A further training study

Fifteen undergraduate students from the University of Surrey took part in the second mammography learning study. The age range in this group was between eighteen and forty-two, with a median age of nineteen. There were five male participants and ten female ones. As in the previous study, people taking part in the study required to attend for approximately ten hours and were paid £15 each as a token of appreciation.

The training and assessment regime in this study was the same as in the training method study reported in the previous section. The same training materials were employed and the same procedure followed. All participants received three sessions of hierarchical feature training. No tutorial training was conducted in this study. Thus, subjects were trained using only abnormal example mammograms during the instructional sessions. As the previous two groups, they were given an opportunity to attempt the self-assessed mammogram test after each of the three instructional sessions. Performance assessment was conducted employing the same test set of seventy-nine mammograms as in the training method study. As before, there were three assessments, one prior to training, one immediately after training and a third
follow-up assessment one week after the completion of training.

6.2.2 Results

The performance data from the present training study were combined with those from the training method study reported above for the purpose of data analysis. Two aspects of performance as a function of feature training were examined. Firstly, performance was analysed in terms of general improvement and with respect to false positive and true positive data as in the previous training study. Secondly, the extent of individual differences after feature training was examined in terms of variance as a function of training intervention. Only performance data of feature method trainees (n=23) were included in this analysis, because the homogenisation effect observed for feature training in the first training experiment might be obscured if subjects from the other training method were included.

6.2.2.1 Film reading performance and feature training

Mean performance and standard deviation for the participants of both feature training groups (n=23) are shown in Table 6.5. It can be seen that whilst mean performance as measured by the G-index increases over the three assessments, the standard deviations decrease as in the previous training study.

Table 6.5: Descriptive statistics of film reading performance for all feature method trainees (n = 23) as measured by the G-index.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>before training</td>
<td>.291</td>
<td>.257</td>
<td>-.316</td>
<td>.636</td>
</tr>
<tr>
<td>after training</td>
<td>.394</td>
<td>.180</td>
<td>-.114</td>
<td>.636</td>
</tr>
<tr>
<td>follow up</td>
<td>.449</td>
<td>.160</td>
<td>-.114</td>
<td>.696</td>
</tr>
</tbody>
</table>

One-way analysis of variance was performed in order to assess the significance of the increase in performance as a function of feature training. Table 6.6 (see also Table Id.4, Appendix Id) shows that the increase in overall performance (G-index)
is a significant effect ($F = 7.5$, $df = 2, 44$, $p < .005$). As in the smaller feature
method subsample ($n=8$), the overall improvement in performance can be explained
in terms of a significant reduction in false positives ($F = 4.38$, $df = 2, 44$, $p < .05$),
but, as in the previous study, there is no significant increase in true positives. As
previously, a conservative test of significance (Box, 1954) can be applied to the
analyses of variance to compensate for the violation of the homogeneity of variance
assumption. Both, the improvement in overall performance and the decrease in
false positive remain statistically significant when evaluated against $F_{0.05} (1, 22) =
4.30$.

Table 6.6: Three oneway analyses of variance for overall performance (G-
index), false positives and true negatives as a function of feature
training ($n = 23$) over the three assessments.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.30</td>
<td>2</td>
<td>.15</td>
<td>7.5</td>
<td>&lt; .005</td>
</tr>
<tr>
<td>Residual</td>
<td>.85</td>
<td>44</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.07</td>
<td>2</td>
<td>.035</td>
<td>4.38</td>
<td>&lt; .05</td>
</tr>
<tr>
<td>Residual</td>
<td>.37</td>
<td>44</td>
<td>.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>.11</td>
<td>2</td>
<td>.06</td>
<td>3</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>.99</td>
<td>44</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.2.2.2 Feature training and individual differences

In Figure 6.5, the mean performance, minimum and maximum (performance of the
worst and best trainee) are plotted to give an indication of the range of variability
between individuals. It can be seen that variability is reduced particularly with
respect to very poor performance, while maximum performance remains relatively
stable over the three assessments. The homogenising effect of feature training
observed in the subsample ($n = 8$) was confirmed with the larger sample ($n = 23$).
A test of difference between the variances of pre- and post-training performance
was significant at the 2.5% level ($t = 2.19$, $df = 21$, $p < .025$), whilst the difference
in variance between pre-training and follow-up performance was even larger
(t = 4.02, df = 21, p < .0001).

**Figure 6.5:** Mean performance and performance range measured by the G-index
as a function of feature training (n = 23).

### 6.2.3 Discussion: Homogenising effects of feature training

The second training study largely confirmed the findings of the training methods
experiment reported in the previous section. Increasing the sample from eight to
twenty-three feature trainees did not result in any differences in the parameters
underlying overall performance. Thus, performance improvement remains a
function of false positive reduction rather than an increase in true positives. The
verification of the homogenisation effect of feature training further confirms the
general advantage of this method in lay reader training. The method appears to fail
to accommodate those people who already perform well before training
commences, however. Figure 6.5 shows that although trainees with very poor
performance (in some cases worse than chance) seem to benefit from such training,
high performers do not improve beyond their initial level. The indication is that
people with specific aptitude for film reading (see chapter 8) do not benefit from
feature training. It was argued earlier that visual feature training helps people to
develop appropriate visual schemata. It might be possible that high pre-training performers either have appropriate visual schemata already in place, or that they develop them very speedily and with great ease. Such able individuals might profit more from a different approach to training. It might be possible to identify these people through selection procedures aiming to distinguish individuals with aptitude for mammogram interpretation from those with no aptitude and assign them to a more suitable training regime. These issues will be discussed in more detail in chapter 8.

6.3 General discussion: Learning to read mammograms with brief instruction

The first training experiment reported in this chapter showed that contrary to expectations the two training methods under investigation did not result in any differences in mean performance. However, initial differences in variance between the groups, which were due to chance fluctuations, could be exploited to demonstrate that feature training appeared to have a homogenising effect on performance variability. The second training study confirmed the authenticity of the homogenising properties of feature training in initial mammography training. Although individual differences could still be observed in the interpretation performance of the extended feature training group, the method succeeded in reducing the spread of scores, especially with respect to the lower margin, significantly (Figure 6.5).

The investigation of different training methods, thus, leads to the conclusion that it does not matter how lay film readers are trained in terms of overall performance level. Hierarchical feature based training has some advantages over more traditional training methods, though. The homogenisation of performance observed after feature training is likely to be due to the development of more similar perceptual schemata within this group of trainees. In chapter 5, it was argued that the development of visual schematic knowledge is by nature highly individual.
However, feature training based on Biederman and Shiffar's (1987) principles of instruction (teaching people what to look for and where to look) is designed specifically to aid the development of appropriate visual schemata on which visual search is based. The superiority of feature training in terms of homogenising individuals' performance is thus entirely consistent with the aims of such training. Brief feature instruction can, thus, be seen to represent a viable approach to training larger numbers of lay observers without medical background to read films for the NHS Breast Screening Programme.

6.3.1 Wider issues in brief training

The primary issue addressed in this chapter, apart from the outcome of different training methods, was the question whether even brief training could improve novice performance in mammography film reading in a significant way. It was argued that Biederman and Shiffar's (1987) approach to instruction in complex perceptual learning tasks might be generally applicable to the problem of mammogram skill acquisition. The performance data from both training experiments summarised in Figure 6.2 and Table 6.2 for the first study and Table 6.6 and Figure 6.5 for the second one, indicate that students without medical background can significantly improve their performance in a complex visual medical task as a function of brief training. Whether the improvement in performance can justifiably be attributed to training intervention rather than practice effects is discussed further below. If training was at the root of the observed improvement, however, it must be assumed that participants interpreted mammograms according to pattern perception principles, since the students participating in these training studies had no medical background and no prior knowledge of mammography or X-ray interpretation. If this is the case, it has been demonstrated that the interpretation of mammograms can be usefully approached as a pattern perception problem, and detailed medical, anatomical and pathological knowledge does not seem to be a prerequisite for learning this task.
The present data lend broader support to the notion that it is possible and feasible to train lay film readers for mammography film reading. Previous research has shown that focused instruction can have a marked effect on performance in mammography interpretation (Fajardo et al, 1987). In the present context, the question may be put to what extent brief instruction has any role to play in learning mammography. In particular, the role of instruction in improving the perception of features in complex patterns by enabling the observer to know what to look for and where to look (Biederman et al, 1987) was of interest. It is evident from Figure 6.2 (first training study), that neither of the brief instruction groups achieved a level of performance comparable to that of radiographers and radiologists after two weeks of training. Participants in the second training study (Figure 6.5) yielded similar performance levels. It is, therefore, quite clear that very brief perception based training cannot replace current NHS training methods. Nevertheless, such instruction may be an useful starting point in lay film reader training, since it succeeded in enabling complete novices to make significantly more competent decisions in a very brief space of time.

Mean performance improved steadily over the three assessments in all groups (Figures 6.2 and 6.5). The increase between the first (pre-training) assessment and the second (post-training) assessment was expected as a function of the instruction and practice intervention. The further increases in performance in all groups between the post-training and the follow-up assessments is less easily explained, as there was no further training or exposure to mammograms during this period. One possible explanation might be that there was a practice effect in completing the same mammogram test on three occasions. However, since participants did not receive any feedback on their performance or on the status of specific items in the test, it is not very likely that there should be such a significant practice effect. In addition, no such effect has been observed in the radiographers' test performance, despite the fact that they completed the same test on six separate occasions. Further evidence against this possibility has been presented by Unwin (1988), who showed that brief training on mammographic abnormality does not necessarily result in
performance improvements on a mammogram interpretation task. In that study, subjects did not improve in performance on a short mammogram test in repeated testing during and after training (see also Hinchcliffe, 1988). The indication is that it is very unlikely that simple exposure to a test set of mammograms can account for the significant improvement in performance in the present study. An alternative explanation for the improvement between the second and third test, which cannot be attributed to any further intervention, might be that subjects had taken the opportunity to consolidate their newly gained knowledge during the week following training. This is supported by anecdotal evidence from participants who expressed confusion immediately after training, but felt more confident on returning for the follow-up assessment.

6.3.2 Learning from nothing?

As in the radiographer study, improvement in all groups was evidenced in terms of a reduction in false positives (Tables 6.4 and 6.6). This appears to indicate that subjects learn to recognise normality in mammograms. Better recognition of abnormality would be indicated by improved true positive rates rather than lower false positive rates. Whilst the recognition of normality as a function of experience is a very persistent finding in all medical visual domains (Schmidt et al, 1990, Myles-Worsley et al, 1988), it is a particularly curious learning outcome for the feature based group as they were trained almost exclusively on abnormal examples. This poses the question of what exactly these subjects have actually learned during training. As the opportunities for the direct schema acquisition of prototypical normality would appear to be severely limited, the question arises whether such knowledge could be acquired from abnormal examples alone. Hearst (1991) has discussed the difficulty with which both, people and animals, learn from absence, deletion and non-occurrence. Yet, the knowledge acquired by trainees during feature training seems to be of normal features, which were not explicitly contained in the training material. On the contrary, training specifically aimed to emphasize the features of abnormality. The question of how prototypical normality is learned
in such a persistent manner, even in the absence of relevant training materials, is a challenging one.

One possible solution is that the observed performance after training, i.e. the significant reduction in false positives but no increase of true positives (Figure 6.3, Tables 6.3, 6.4 and 6.6), is not so much due to learned perceptual prototypes or visual schemata as to decision making processes. A decision-to-recall, the production of a true positive, may involve much more complex processes and may, therefore, be more difficult and error-prone than no-recall decisions (true negatives). This possibility will be discussed in more detail in the next chapter in the context of a process model of mammogram interpretation.

6.3.3 Summary

The studies reported in the present chapter have shown that it is possible to train complete novices to read mammograms. Although the training was very brief, participants in the training studies on average improved their interpretation performance significantly. Different training methods do not appear to affect overall performance, but feature training has a homogenising effect on group performance. Individual differences are particularly reduced with respect to poor performance.
Chapter 7
A model of screening mammogram interpretation

Most of this thesis so far has been based on an implicit model of the mammogram interpretation task. In chapter 2, two approaches to radiological expertise were discussed, one primarily concerned with the visual perceptual aspects of mammogram interpretation and the other with cognitive reasoning and decision making aspects of the task. These approaches were examined with regard to screening mammography in chapter 3, and it was argued that both, perceptual and interpretative processes, were important for accurate mammogram interpretation. In chapter 4, two case studies demonstrated that lay readers can learn to interpret mammograms to a level comparable with radiologists. Elements of the skills involved in full mammogram interpretation and mechanisms of acquisition were discussed in some detail in chapter 5. In that chapter, the primary task elements were hypothesized to be the visual discrimination of image features, categorisation of visual and disease features and decision making processes involved in the merging of both, visual and conceptual features. Chapter 6 presented learning experiments designed to examine whether particular training methods affect performance differently implying that they operated on different task elements. No such differences were revealed in overall mean performance, but training aimed at improving visual detection of features served to reduce individual variability in performance.

The main concern of the present chapter is the development of an explicit process model of mammogram interpretation. The purpose of such a model is to identify the psychological mechanisms involved in the performance of the task in a methodical manner. The analysis of perceptual and cognitive components in terms of a task model serves to formalise the discussion of skill aspects in mammography film interpretation conducted in chapter 5. In addition, the cognitive task model is of relevance to the further discussion of individual difference and film reader selection (see chapter 8 for a detailed discussion of this issue). It may also be
possible to identify sources of individual differences in terms of the cognitive processes identified in the model, an issue which is explored further in chapters 10 and 11. A second issue in this chapter is the evaluation of the model. The validity of the model is critically dependent on its ability to account for the observed data in the training studies. Further analysis of false negative errors is presented in support of the model.

7.1 Modelling cognitive processes in mammogram interpretation

Models of radiological skill and expertise have already been reviewed in chapter 2. Whilst more recent formal models such as Nodine and Kundel's (1987) interpretative decision making model present a comprehensive account of radiological search and decision making skill in general, screening mammography skill has a number of special features that are not accounted for well in such general models (see chapter 3). One problem with Nodine and Kundel's model in accounting for screening mammography skill, for example, is the close integration of perceptual recognition units with the cognitive evaluation stage. This is problematic because the model strongly assumes the presence of abnormalities as they are in diagnostic radiographs. The task of the observer according to this model is to locate the abnormality in a process involving perceptual-cognitive evaluation of features pre-selected by perceptual processes. Whilst this kind of model undoubtedly accounts for the processes involved in diagnosing abnormal X-rays, it is questionable whether it is a good general description of screening mammogram interpretation. The large majority of mammograms in screening are normal, and their interpretation may not involve any evaluation of pre-selected features. The intention in this chapter is to develop a model of screening mammography skill which takes into account decision outcome as one of the special features in interpreting screening mammograms.

The validity of the model needs to be evaluated in an appropriate manner. The training studies conducted in the previous chapter provide performance data for this
purpose. However, overall performance provides few clues about the processes involved in film reading and is, therefore, not the most appropriate medium for model testing. In chapter 4 it was shown that error analysis, in particular the analysis of false negative errors, was a viable procedure for distinguishing perceptual and interpretative processes in film reading performance. In this chapter, an analysis of false positive errors data similar to that in the radiographer study (chapter 4) was conducted for all participants in both training studies. The error data can be employed to verify aspects of the mammography screening task model, and they are presented following the outline of the mammogram interpretation model in the first part of this chapter.

### 7.1.1 A task model of mammography screening film interpretation

The major skill elements involved in screening mammogram interpretation have already been identified as pattern recognition, categorisation of visual and conceptual features and probabilistic decision making. The purpose of the task model, which is developed in this chapter, is to put these skill aspects into the dynamic context of a cognitive process model. The formal identification of task elements and their locations in the process of mammogram interpretation is of importance for the further discussion of individual differences and selection in chapters 8 to 11. In chapter 10, in particular, an attempt is made to develop predictive tests of mammogram film reading performance based on cognitive task analysis. These are employed as predictors of mammogram interpretation as a function of training in an exploratory prediction study which is reported in chapter 11.

In order to account for processes in mammogram film interpretation, a number of special features of screening mammography need to be considered. In chapter 3, it was argued that screening mammography is different from diagnostic radiological interpretation in that the expectation of finding an abnormality is quite low. Screening outcome which is heavily biased towards findings of normality is, thus,
an important factor that needs to be taken into account. It was also argued that screening mammography involves making two kinds of decision, whether an abnormality is present, and if it is, whether it requires further action. This notion, that screening mammogram interpretation involves making perceptual and conceptual decisions can be supported empirically in the distinction between two types of error, perception and interpretation errors, which were identified in the radiographer study (chapter 4). It will be shown later that the same distinction between perceptual and interpretation errors can be applied to error data from the student training studies.

Figure 7.1 shows a detailed cognitive information flow diagram of screening film interpretation. It can be seen that the model incorporates three kinds of elements, processes, data and decision outcomes. Data and processes are pictured as square boxes and decision outcomes are elliptical. The direction of information flow is indicated by arrows. The first part of the model represents processes in the visual perceptual analysis of a screening mammogram. The mammogram is searched for specific visual features which are indicative of abnormality. These are, specifically, localised areas of increased brightness and differences between the two images (left and right mammogram). Pattern recognition processes involve a cyclical activity attempting to resolve visual ambiguity in potential areas of abnormality with reference to visual knowledge stored in memory as visual schemata. Perceptual schemata in turn guide the visual search of the mammogram. Visual assessment of the mammogram terminates when a decision is made as to whether an abnormality is present or not.

If an abnormality has been detected during visual search cognitive evaluation of the target abnormality commences. Target analysis involves the examination of visual features identified during visual search in terms of their relevance in malignant or benign disease processes. Knowledge about the symptoms of breast disease is stored in memory as disease schemata. As in the perceptual stage, cognitive evaluation processes are modelled in terms of cyclical activity between attempts to
Figure 7.1: Process model of mammogram interpretation
resolve ambiguity, stored knowledge and analysis of the target abnormality. The interpretation process may have one of two outcomes. If ambiguity is successfully resolved a decision is made with regard to the status of the observed abnormality. It can be classified either as significant, requiring further investigation or as insignificant in the context of disease knowledge. In the former case recall of the woman will be recommended, whilst in the latter case the final outcome will be a return to routine screening. If, however, there is no successful resolution of ambiguity, the film reader may wish to reassess the visual information present. In the present model this possibility is represented in terms of a data manager links between conceptual and visual knowledge in memory. Thus, if disease schemata are insufficient for the resolution of cognitive ambiguity, the observer may gain access to visual memory via the data manager, providing an opportunity to reassess the visual data.

7.1.2 **Cognitive processes in mammogram interpretation**

The data flow model in Figure 7.1 implies a two-stage process model, each stage consisting of an evaluation and decision making component. In chapter 5, visual discrimination, categorisation of visual and disease features and probabilistic decision making were identified as the cognitive processes involved in mammogram interpretation. The present section presents an attempt to explore the cognitive processes implied in the model in terms of the relevant psychological literature which was reviewed in chapter 5.

The first stage in mammogram interpretation is concerned with the visual assessment of the X-ray. This stage involves pattern recognition processes which come about through visual search guided by visual memory. The training studies in chapter 6 demonstrated that complete novices can acquire skills in classifying mammograms as normal or abnormal on the basis of very brief training. As these individuals did not possess any knowledge of breast anatomy or pathology, it can be assumed that they were operating on a pattern perception level. This shows
that the processes in the first stage of the model may usefully be construed as two-dimensional pattern perception processes rather than recognition of three-dimensional objects (see discussion in chapter 5). The outcome of the visual stage is the discrimination between normal and abnormal mammograms. In the skilled reader this stage would be expected to consist of largely automatic processes, which are akin to categorical perception. The decision by an expert of whether or not abnormality is present in the X-ray is made almost instantaneously (Myles-Worsley et al, 1988) and is the result of massively parallel visual processes (Hendee, 1993).

The identification of abnormality leads to the initiation of the conceptual evaluation stage of the model. This is characterised by processes leading to the interpretation of the X-ray. These interpretative processes, which involve the evaluation of 'symptoms' (in the case of mammograms these are visual features indicative of benign or malignant breast disease), require effortful processing even in the most expert of radiologists (Getty et al, 1988). Interpretation of abnormalities is likely to proceed with some recourse to visual data when ambiguity cannot be resolved in terms of current semantic properties of visual features. The outcome of the interpretation stage in screening mammography is a further binary categorisation, but it may be much more complex in other, especially diagnostic, radiographic interpretation.

Classification at both stages involves assigning features to fuzzy categories of normal versus abnormal in the visual assessment stage and significant abnormality versus insignificant abnormality in the interpretation stage. Such decision making is probabilistic and based on a high degree of uncertainty as has been discussed in chapter 5. But whilst the former appears to be a simple discrimination decision based on physical feature properties, the later is more complex. Discussion of strategies and rules in such complex combining of semantic feature properties leads to the conclusion that people may not be optimally equipped to learn to make such decisions (Getty et al, 1988, Brehmer, 1980). They are able to learn quite complex
relationships between probabilistic features (Medin et al., 1982, Ashby et al., 1988) but they tend to apply them in a deterministic fashion (Ashby and Maddox, 1992). Failure to take into account probabilistic properties of features in an adequate fashion may thus result in overdiagnosis (false positives) at this stage (Eddy, 1982). This issue will be discussed further below with reference to error analysis.

Memory plays a critical role in both, visual processing and the interpretation of mammograms. In the model, memory processes can be seen to be involved in cyclical interaction with visual and cognitive processing. Both, visual assessment and interpretation of abnormality thus proceeds with reference to stored knowledge. The knowledge required during the two stages of processing is quite different, however. Although both can be construed as schemata, visual schemata are based on the physical properties of features in X-rays, whereas abnormality schemata (or disease schemata) are based on the semantic properties of these visual features. In chapter 5, it was argued that errors may arise from false or absent schemata at both stages. Visual schemata are critical in the recognition of abnormality (Gale, 1993). Interpretation processes can be construed in terms of making a diagnosis (further action is required or not), and most recent work on medical diagnosis has postulated the operation of disease or illness scripts (Schmidt et al., 1990, Gilhooly, 1990). Errors in the interpretation of medical information may arise as a result of inappropriate, absent or rigid schemata (Abercrombie, 1960). Visual and disease schemata are thus critical for the acquisition of both visual search and interpretation skill as conceptualised in the present model.

7.2 Validating the process model of mammogram interpretation

The remainder of this chapter is concerned with the validation of the process model presented in the preceding section. This can be achieved by examining how well the model accounts for the data presented in the radiographer study (chapter 4) and the student training studies (chapter 6). One aspect of performance which has not
yet been addressed within the student training studies is that of false negative errors. It was argued in chapter 4, that error data, particularly false negative errors, are a source of information regarding the processes involved in mammography skill acquisition. It was shown that on the basis of such an error analysis a distinction can be made between perception and interpretation errors. In this section, an analysis of the false negative errors in the student training studies is presented to support the basic assumptions underlying the process model of mammogram interpretation. In the remaining chapter, performance and error data are discussed in terms of their support of the model. In addition, the present model is compared to previous models of radiological search. One issue of particular interest in the context of evaluating the model against the training data is the question of what is learnt during feature training. It will be shown that the issue of an apparent acquisition of 'normality schemata' (see chapter 6) gains a different perspective in the light of the model.

7.2.1 Error analysis

The present error analysis was performed on the student training data in analogy to the radiographer error analysis (chapter 4). In chapter 4, the outcome of the error analysis was also employed to examine the notion of a developmental sequence in radiological skill acquisition (Lesgold et al, 1988). Although this concept was rejected on the basis of the data presented in that chapter, the current analysis presents an opportunity for a re-examination of the idea by comparing student and radiographer error data for cross-sectional evidence of such a sequence. The main purpose of the error analysis is, however, to provide support for the two stages of the mammogram interpretation model.

7.2.1.1 Method

For the purpose of error analysis the data from all training groups were combined (n = 32). This can be justified, since there were no experimentally induced differences in mean performance between the training groups. Error rates would
thus not be expected to differ between the groups. As in the radiographer study, participants in the training experiments were required to record any noted features separately from their decision to recall or not to recall (see Figure 6.1). It was, therefore, possible to distinguish false positive errors in which the relevant feature(s) had been noted from those in which no features had been recorded. The former is labelled an interpretation error as the abnormality has been registered but wrongly interpreted to be of no significance. The latter represent perception errors, as it can be assumed that the trainee has not consciously recognised the relevant abnormality (see chapter 4 for a detailed discussion). The error analysis was conducted in the same manner as in the radiographer study, except that the two sets of mammograms with abnormalities on both sides (no. 16 and no. 55) were recorded as separate sources of error for the students. Thus, instead of sixteen possible misses, students could have reached a maximum of eighteen possible misses.

7.2.1.2 Results

The mean number of perception and interpretation errors produced during the three performance assessments are shown in Figure 7.2 (see Tables Ia.4 and Ia.5 in Appendix Ia).

**Figure 7.2:** Mean number of perception and interpretation errors over three assessments of all training experiment participants (n = 32).
It can be seen that there is little variation in either perception errors or interpretation errors. There is a very slight decrease in the mean number of perception errors in the second assessment, matched by a similarly marginal increase in interpretation errors in the third assessment. Overall, the false negative error rate remains remarkably stable over the three assessments. Two one-way analyses of variance revealed that there were indeed no significant differences in perception or interpretation errors over the three assessments. Table 7.1 shows the summaries of the analyses of variance.

**Table 7.1:** Two one-way analyses of variance for perception and interpretation errors as a function of brief training over three assessments.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perception Errors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>12.56</td>
<td>2</td>
<td>6.28</td>
<td>1.69</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>230.56</td>
<td>62</td>
<td>3.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretation Errors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>1.02</td>
<td>2</td>
<td>.51</td>
<td>.26</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>122.98</td>
<td>62</td>
<td>1.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**7.2.1.3 Discussion**

The observed stability in the error rates over the three assessments (Figure 7.2) was expected in view of the absence of any significant increase in true positive rates in any of the three training groups. A more intriguing aspect of false negative analysis is the much higher frequency of perception errors over interpretation errors. Comparison with the radiographer error data in chapter 4 (Figures 4.6a and 4.6b) reveals that the ratio of perception errors to interpretation errors appears to be reversed for the student sample. Although the average false positive rate is fairly similar when comparing radiographers (approx. 5/16 averaged over six assessments)
and student trainees (approx. 7/18 averaged over three assessments), only 15% of these represent interpretation errors in the student sample whilst 70% of radiographers' errors are interpretation errors.

In chapter 4, Lesgold et al's (1988) notion of a developmental sequence in radiological skill acquisition was discussed. It was argued that there was no evidence of a developmental progression from perceptual skills to interpretative skills in the longitudinal radiographer data. Although there is a very slight increase in interpretation errors accompanied by a very slight decrease in perception errors in the student data, this cannot be used in evidence for Lesgold et al's developmental sequence as the effect is not a significant one (Table 7.1). However, the comparison between student and radiographer error data indicates that there may be some cross-sectional evidence for such a sequence. Students, who have not previously been exposed to mammograms, produce a much larger number of perception errors but few interpretation errors whilst the opposite is true for radiographers, who have had considerable previous exposure to mammograms. The lack of evidence for sequential acquisition of perceptual and interpretative skills in both sets of longitudinal performance data may, therefore, be due to a lack of sensitivity in the performance measurement.

As far as the model is concerned, the differences between radiographers and students can be explained in terms of differences in knowledge. The radiographers may have already had reasonably well developed visual schemata of what constitutes mammographic abnormalities before the commencement of training because of their previous exposure to mammograms. Students on the other hand, only received very brief training and may not have had sufficient time to fully develop appropriate visual schemata. Since according to Lesgold et al (1988) appropriate visual schemata are a prerequisite for radiograph interpretation, inexperienced students would not be expected to have much opportunity for making interpretation errors after only brief training. More direct, longitudinal evidence for Lesgold et al's developmental sequence might have been obtained if the students
had continued their mammogram interpretation training for some length of time.

7.2.2 Validating the task model in terms of error and training data

The preceding discussion indicates that the two postulated memory stores operating in mammogram interpretation, which represent perceptual and conceptual knowledge respectively, account for observed changes in performance as a result of exposure to mammograms. The differences between perception and interpretation errors produced by radiographers, who had prior experience of mammograms and received two weeks of training, and students, who received only brief training, indicates that students operate on a different level than radiographers, despite little difference in total error rate. The indication is that students read mammograms at a level of discrimination between normal and abnormal, whereas the radiographers are performing the full interpretation task. The distinction between student and radiographer performance levels further supports the notion of separate visual and interpretative processes in mammogram interpretation.

The validity of the above task model can also be assessed more generally in terms of how well it accounts for the performance data obtained in the training studies. In the first instance, it is possible to assess the process model in terms of true outcome. In chapter 3, it was explained that outcome can be expressed in terms of four quantities, true and false positives and true and false negatives (see Figure 3.1 in chapter 3). True negatives and false positives add up to the total number of negative mammograms whilst true positives and false negatives represent all positive mammograms. The performance data in all training studies indicate that training succeeds in improving the true negative rate (the reverse of the false positive rate) while true positive classification remains relatively error-prone. The film reading process model can explain this finding in terms of a longer processing path which is required to produce a true positive. Whilst a substantial number of true negatives result directly from the visual processing stage, all true positives must also pass the cognitive evaluation or interpretation stage error-free.
The relationship between processing path and identification of true positives and negatives is presented graphically in Figure 7.3. This shows a simplified film reading process model in which the locations for potential false negative and false positive errors are indicated. It can be seen that a true positive can only result from the longer processing route with more potential for errors to occur. Similarly, false positive errors are cumulative in that errors have to occur at both, perceptual and interpretative decision points. False negative errors, however, are easier to make in that they result directly from wrong decisions at either perceptual evaluation or at interpretation levels.

Figure 7.3: Simplified process model of mammogram interpretation showing the locations of false negative (FN) and false positive (FP) errors.

In essence, the process model of mammogram interpretation would predict an initial reduction in perception errors as a function of very brief training. This is because such training aims to improve visual schemata (see chapter 6) which in turn results in better discrimination performance in the visual assessment stage. It is
not entirely clear from the model why this should not be the same for false negative errors, though.

7.2.2.1 Learning from nothing: schemata or decision processes?

It is now possible to return to the discussion of what is being learnt in feature training which was begun in the previous chapter. It was argued that the observed reduction in false positives under all training conditions was unexpected, particularly when subjects had been trained almost exclusively on abnormal mammograms. This was because an assumption had been made that improvement in mammogram interpretation was brought about by a better understanding of what represents abnormality in mammograms. In other words, it was argued that training, and in particular feature training, essentially serves to improve memory representations of abnormal appearances, i.e. visual schemata of mammographic abnormality were developed. The training study data, however, seem to indicate that participants learned to recognise normality rather than abnormality.

It was argued in the previous chapter that the apparent learning of normality, which is a persistent finding in all visual medical domains (Schmidt et al, 1991), may have more to do with the complexity of processing involved in producing true positives as compared with true negatives than with memory representations. The production of both true positives and true negatives is dependent on the accuracy of visual schemata which guide search. True positives are also critically dependent on accurate disease schemata which guide the cognitive evaluation of mammograms. Only a small proportion of true negatives will be similarly dependent on accurate cognitive evaluation, namely those with clearly benign abnormalities. In terms of the process model presented above there is, therefore, no contradiction between learning to recognise abnormality in training and a decrease in false positives. False positives appear to arise predominantly as a result of failure to discriminate between normal and abnormal, whereas false negative may be due to either mis-perception or mis-interpretation. Since errors due to mis-interpretation of perceived abnormality become more frequent with experience (as indicated by the relatively
large proportion of such errors in the radiographer study), the overall false negative rate remains stable over longer periods of training, whilst the false positive rate is reduced relatively early on in the learning of the interpretation skill.

The two stage decision model also accounts well for Myles-Worsley et al's (1988) radiological performance data. They found that radiographs without an abnormality can be rapidly assigned to a normal category whilst abnormal radiographs require more effortful processing. The verbal data in the radiographer study (chapter 4) clearly indicated that visual processes are not easy to access consciously. This demonstrates that such processes are clear candidates for early automatization.

Cognitive evaluation of radiographs, on the other hand, remains an effortful process, which is consciously accessible, for a considerable period of time (Lesgold et al, 1988) except possibly in the most expert subjects (Myles-Worsley et al, 1988). This would provide an additional explanation of why false negatives are much more persistent than false positives. Not only does the production of a true positive involve a longer processing route, but the kind of additional processing involved is less automatic and, therefore, more error-prone.

7.2.2.2 Training outcome: pre-screening versus 'second pair of eyes'

A further evaluation of the screening mammogram interpretation model can be conducted in terms of its accuracy in predicting training outcome. The issue of whether lay film readers should be trained as pre-screeners or as 'second pairs of eyes' was first discussed in chapter 4. In terms of the model, the difference is between training film readers to conduct visual search without missing any abnormalities (pre-screening) and training readers to perform the whole interpretation process relatively error-free. The argument was that the substance of training may have an important role to play in interaction with the role definition of the trainee. Although image evaluation in terms of disease schemata is an appropriate activity for radiologists and other doctors who as part of their medical training have acquired extensive knowledge of anatomy and pathology, but it may not be desirable for paramedical lay readers to make decisions regarding the status
of an observed abnormality unless their training specifically included medical diagnosis of breast disease.

In chapter 6, and on a more theoretical level in chapter 4, an attempt was made to argue that differential training might have different performance outcomes which relates closely to the function the trainee screener is to fulfil. With respect to the screening model, feature training would be expected to improve visual knowledge, whilst tutorial training is more broadly aimed at improving knowledge and decision making at both, the perceptual and interpretative level. The results of the training study reported in chapter 6, however, did not support the contention that differences in training would be reflected in differences in post-training performance. It was argued that failure to differentiate between the groups might have been caused by a procedural shortcoming. However, another reason for failure to differentiate between the groups may have been the brevity of training.

Error analysis showed that perceptual false negative errors were prevalent even after training. Student trainees thus had not fully developed appropriate visual schemata. Even the radiographers (chapter 4) after much longer training and more prior experience of mammograms made some perception errors. As the students post-training performance level was much lower than that of the radiographers the relative benefits of different training methods may not have become apparent because the changes effected by very brief training were too minimal. The observed reduction in false positives observed in both training groups may well have been due to improved visual schemata in the feature training group as evidenced by decreased variance in group performance, while the same effect might have been due to better decision making mechanisms in the tutorial group. It is not possible to distinguish between the two on the basis of the present training data.

7.2.3 Comparison to previous models of radiological skill

Previous models of visual search and cognitive expertise were discussed in some
detail in chapter 2. Some of these models appear to have similarities with the model of screening mammogram interpretation proposed in the present chapter. The present model assumes two stages of processing in radiograph interpretation. The idea of a two-stage model does not seem to be a new one, as both Swensson (1980) and Nodine and Kundel (1987) have proposed such models incorporating two stages of processing. However, few previous researchers have attempted to build explicit models of both, perceptual and interpretative processing in radiograph interpretation. The majority of models, including the two mentioned above, are only concerned with one or the other aspect of radiograph interpretation. In this respect the model of screening film interpretation represents, in fact, a double two-stage model as will be argued below.

Swensson (1980) proposed a model of visual search in which the first stage acts as a pre-attentive filter which selects features for cognitive evaluation in the second stage of processing. Nodine and Kundel (1987) elaborated this and their own earlier models (Kundel and Nodine, 1975, Kundel et al, 1978) by connecting the search processes to a memory store of prior knowledge, thus enabling the link to be made between training and experience and decision processes leading to target detection. The model of screening mammogram interpretation is less explicit about the precise nature of the search processes involved than, for example, Nodine and Kundel's model. Visual analysis involves the dynamic interaction of search, visual knowledge and detection which in turn leads to a decision with regard to whether an abnormality is present or not. The model, thus, incorporates a process akin to Neisser's (1976) perceptual cycle as the first phase within the visual processing stage, which has been considered an appropriate model for visual search in radiology by Gale et al (1979). This is followed by a second phase of decision making with respect to the visual data. The perceptual stage of the present model is, thus, very similar to both Swensson's and Nodine and Kundel's models of visual search, although the precise search mechanisms are not presented in as much detail as in the previous models. The kind of data collected in this research do not permit the detailed analysis of visual processes conducted by Swensson (1980) and Kundel
Models of radiological expertise such as Lesgold et al's (1988) schema model have emphasised the interpretation aspects of radiological skill. Interpretation is dealt with by the second stage of the screening mammogram interpretation model. This stage is initiated by the occurrence of a target abnormality identified during visual processing. Lesgold et al (1988) have argued that such interpretation is schema-driven. The present model involves a cyclical interchange between properties of the target abnormality and knowledge of breast disease which is a similar process to the perceptual cycle in visual search. Evaluation of the target is an act of interpretation rather than detection, though, and is, therefore, likely to involve higher cognitive functions. As in the visual evaluation stage, there are two phases of interpretative evaluation, one concerned with making sense of the target data and the second with decision making.

Lesgold et al's model has been criticised previously for not accounting for visual processes in target acquisition in a satisfactory manner (chapter 5). Visual search models of radiological skill can be similarly charged with not accounting for interpretative processes. Although both kinds of processes have a role to play in most areas of radiology (Rubin, 1989), their interaction is probably uniquely important in screening mammography. Visual processes are essential for the identification of potential, usually not well-defined, abnormality and interpretative processes are required to distinguish between clearly benign and possibly malignant appearances. High accuracy in screening mammography would be impossible without either process. The model presented in this chapter, thus, formalises Myles-Worsley et al's (1988) notion of different processes being involved in processing normal and abnormal radiographs. They have argued that rapid visual processing identifies abnormality, which is evaluated in more effortful top-down processing. This is the case in the present model where true negative (normal) mammograms are assigned to their final outcome category on the basis of purely visual assessment. True positive (abnormal) mammograms, on the other hand, require
further interpretative evaluation.

7.2.4 Conclusion

The double-two stage model of screening mammogram interpretation presented in this chapter aimed to account for both, visual and interpretative processes in mammography. The model presents an attempt at formalising the psychological processes underlying mammogram interpretation which were outlined in chapter 5. It was shown that the model accounts well for several empirical findings in both longitudinal and experimental training and skill acquisition studies such as why true positives are more difficult to produce than true negatives. The model also inspired some thoughts about the skills involved in mammogram interpretation and how they may be acquired. This led to some speculation on the effect of different training methods and a re-analysis of why differences had not arisen in the training methods study. The indication is that brief training, whatever the content, must by definition address the development of appropriate visual schemata, as these are prerequisite to the performance of the more complex interpretative processes involved in evaluating mammograms, with regard to the significance of observed abnormality. The double-two stage model was also evaluated in terms of previous models of radiological skill. It was shown that previous models focused either on the visual aspects of radiological diagnosis or on the interpretative ones. The present model thus presents a comprehensive attempt to account for the entirety of the X-ray interpretation process. Further evaluation of the model can be conducted in terms of in how far the postulated processes are actually representative of screening film performance. It can be hypothesized that specific processes within the task model, such as perceptual evaluation, target interpretation and decision making may be predictive of overall film reading performance. This is an issue which will be explored in more detail in chapters 10 and 11.
Part 3

Individual differences in mammogram interpretation: Issues in selection and performance prediction

The final part of this thesis is concerned with the identification and prediction of individual differences in mammogram interpretation. Observer variability is a problem encountered in all visual medical diagnosis, but is of particular relevance in the context of radiological screening programmes, because their success is critically dependent on unchanging application of diagnostic criteria. With respect to lay film reading, selection procedures may be implemented to reduce variability amongst trainees to a minimum. The work presented in this section was thus strongly motivated by applied issues of reader selection and minimisation of individual differences. On a more theoretical level, it was considered desirable to identify individual differences at the level of cognitive processes, which were formally described in the model of mammogram interpretation in the preceding part. The first chapter in this section is concerned with outlining the applied issues arising from individual differences in mammogram interpretation. It also presents a review of previous research in individual differences in radiologic and mammographic performance. Methodological issues in investigating individual differences in cognitive skills are discussed with particular reference to cognitive processes. The empirical approach taken to individual differences in the present research was essentially a psychometric one. Chapter 9 is specifically concerned with attempts to identify a psychometric aptitude domain for mammogram interpretation. The subsequent chapter reports the development of a cognitive test battery, which was to measure specific cognitive processes involved in mammogram interpretation as outlined in the task model. Measurement of separate cognitive process proved elusive, however, and chapter 11 reports prediction studies in which both, psychometric and cognitive measures are employed psychometrically as predictors of mammogram interpretation performance.
Chapter 8
Identifying Sources of Individual Differences in Mammogram Interpretation Skill

Observer variability has been noted in most medical domains (see Feinstein (1985) for a review), but has presented a particular problem in radiological diagnosis (Garland, 1959). Radiological diagnosis is subject to both inter- and intra-observer variability even in the presence of established criteria for the classification of X-rays (Yerushalmy, 1969, Herman, Gerson, Hessel, Mayer, Watnick, Blesser and Ozonoff, 1975, Gale et al, 1989). In addition, a number of studies comparing observer performance in screening mammography have revealed considerable variability in interpretation between individual film readers (Baines, McFarlane and Wall, 1986, Vineis, Sinistrero, Temporelli et al, 1988, Gale et al, 1989, Swart, in press).

Observer variability in mammography has been specifically discussed by Gale et al (1989) and by the German Mammographie Studie (Swart, in press). In both cases, it was concluded that a case may be made for continued performance monitoring after initial training (Gale et al, 1992). Swart (in press) even advocates specific competence testing as a prerequisite for film reading practice in order to ensure quality screening. In addition, there is evidence that measurable individual differences in mammogram interpretation skills may be independent of the individual level of expertise in the performance of the task. Gale et al's (1992) work with a large number of experienced breast screening radiologists, for example, has shown that there are, indeed, some individuals who can benefit from specific intervention and further training, although the overall standard of mammogram classification performance was good.

Individual differences were also evident in the strategies employed by the two radiographers in the study reported in chapter 4. Although the small sample size
precluded the identification of significant performance differences in the post-training period, it is nevertheless conceivable that such strategy differences have an impact on the continued development of expertise. Further evidence for individual differences in mammogram interpretation performance was provided by the student training studies. In the training methods study (chapter 6), it was shown that feature training is particularly suitable for homogenising individuals' performance after training. Nevertheless, considerable individual differences in mammogram interpretation performance remained even after training. In addition, there was some indication that the most able trainees did not profit from feature training, which was generally considered the most useful method in lay reader training. In the context of selecting lay observers for breast screening, it is therefore important to investigate the nature of individual differences that may affect performance in mammogram interpretation skill after training. Training for lay readers may be optimized by selecting individuals who most benefit from training on account of certain identifiable individual ability and/or personality characteristics or specific cognitive skills they possess. Alternatively, it may be possible to match individuals to training methods that suit their particular skills and abilities.

In this chapter, the practical applications of taking an individual differences approach in the context of selecting and training radiographers for breast cancer screening are outlined. This is followed by a more detailed review of previous empirical studies of individual differences in radiological skill, which are examined with respect to their relevance to screening mammography. Methodological issues in studying individual differences in cognitive ability and skill acquisition are considered next. In the last section of this chapter, the present approach to individual differences in screening mammography is outlined.

8.1 Applications of individual differences in screening performance

Most complex tasks are subject to individual differences in performance, and, as indicated above, screening mammography is no exception. In such tasks, individual
differences are likely to operate not only in terms of aptitude for the skills involved, but also more generally with respect to the availability of task-appropriate strategies, motivation to acquire the skill, cognitive style and personality (Messick, 1987). Identification of those differences that influence the acquisition and performance of mammogram interpretation skills is important for a number of practical reasons. These concern issues of training design and trainee selection which will be discussed in detail in this section.

8.1.1 Training Design and Individual Differences

A number of issues in training lay readers for mammogram interpretation have already been dealt with in some detail in previous chapters. The present section is concerned with considering the use of knowledge about measurable individual differences to solve some pertinent training problems. Stammers (1987) has pointed out that training activities should not be considered in isolation from other aspects of occupational psychology, in particular personnel selection. It is important to consider the characteristics of the population from which people are selected for training when designing a training programme. Training that has been designed to take account of the kind of people it caters for, is much more efficient than inappropriately designed training. With respect to mammography, it can be said that the tutorial approach to training was originally designed for radiologists rather than lay readers. It has already been shown in chapter 6, that a more task appropriate training design may have some beneficial effects on overall training outcome.

Identification of specific individual differences within a group of potentially diverse lay readers, that contribute to greater efficiency in acquiring mammogram interpretation skills, may further benefit training design. Training could, for example be designed to support specific individual weaknesses or to suit particularly able trainees. This approach is particularly useful in situations where the participants are self-selected on the basis of personal interest, as has been the
case with radiologists in screening mammography. This technique is already being implemented on a performance level in Gale et al's (1992) work who provide special training for individual radiologists, who have been identified as having difficulties in detecting particular types of abnormality. In a similar vein, psychometric characteristics that are important predictors of learning a complex perceptual skill such as mammogram interpretation might be used to assess an individual's readiness for learning prior to the commencement of training. Any individual weaknesses that become apparent at this stage could be supported by training before problems in skilled performance even arise.

In the previous chapter, an attempt was made to specify the cognitive processes involved in mammogram interpretation on the basis of performance of individuals learning the task. The model of mammogram interpretation provides some pointers to specific cognitive processes involved in such skilled performance. This is a useful starting point for identifying cognitive strengths and weaknesses which may be supported by training. In the context of training design it is thus particularly important to have some hypotheses about the nature of the task to be learned. Much of the work reported in the following chapters aims to identify individual learner characteristics and their relationship with task characteristics as they affect skilled performance. Although training design is not directly investigated any further in the present thesis, it remains to say that the identification of individual differences in skilled performance may, in the long term, result in the development of a more efficient approach to training than is achieved by either the tutorial or the feature identification method.

8.1.2 Selection and Aptitude for Mammogram Interpretation

An alternative approach to matching people to jobs is that of selection. It is possible to select individuals who are particularly capable of either performing a skill or learning to perform the skill. This is the main concern from an applied point of view of the present, third part of this thesis on individual differences. One
prerequisite of selection procedures is that the characteristics and personality traits which are responsible for success or failure at the skill in question can be unequivocally identified. Such aptitudes for particular task domains have been psychometrically identified for a number of skills, one example being Carroll's (1958, 1985) work on aptitude for second language learning.

Drenth and Algera (1987) have distinguished four aspects of selection procedures, one oriented towards identification of predictors, one directed towards criterion validation, one concerned with the prediction model and decision oriented ones. They argue that proper personnel selection must take account of all four aspects. Prediction of mammogram interpretation, thus, involves the identification of ability and personality characteristics which are predictive of either mammogram interpretation performance after training (aptitude for the skill) or improvement in mammogram interpretation as a function of training (readiness to learn the skill) or both. Selection decisions on the basis of tests measuring these characteristics have to be properly based on a clear definition of the task to be performed by the trainees, i.e. whether they are to perform a pre-screening or a 'second pair of eyes' function.

The application of such procedures to identifying individuals with particular aptitude for mammogram interpretation is naturally dependent on the availability of a large enough population of suitable candidates. If trainees continue to self-select into film reading as has been the case with radiologists in screening mammography, there is not much scope for positive selection, although it might still be possible to discourage trainees who have been identified as lacking prerequisite characteristics from pursuing mammogram interpretation training. If, however, lay film reading becomes an accepted alternative to consultant screening, there is much potential for selecting radiographers with particular aptitude for mammogram interpretation on the basis of ability, personality and cognitive skill characteristics. Such procedures would, indeed, be highly desirable as Saxton (1992) has argued.
Testing for selection depends on the availability of valid and reliable predictors of performance in mammogram interpretation. The nature of such predictors is to some extent dependent on the purpose of the test. In some situations, it may be desirable to select individuals with particular aptitudes from a large pool of candidates. In such a situation a psychometric approach, which is specifically suited to selecting a small number of people who are most likely to succeed at the task in question from a large pool, is probably most appropriate. In the screening situation the problem is, at present, slightly different in that the pool of available radiographers is relatively limited. In such a situation it may be more desirable to identify each individual who is likely to succeed. Lesgold, Lajoie, Logan and Eggan (1990) have argued that it is more appropriate in such situations to use cognitive task analysis (see discussion below) in the development of predictor tests as such a procedure leads to tests which are capable of 'deciding who has the prerequisite thinking and learning skills, the domain-specific cognitive procedures, and the understanding to do a job or to be trained to do a job' (p.329).

8.2 Individual differences in radiological interpretation

There is evidence in mammography that expert screeners vary not only with regard to overall performance but also with respect to the features they use in identifying abnormality (Gale et al, 1989). This may be the result of highly individual visual schemata which have been developed during training and as a function of experience. However, it may be possible to identify some specific predictors of the accuracy with which such schemata are developed and applied in mammogram interpretation. Studies of individual differences in radiological performance have been conducted by Smoker, Berbaum and their co-workers. These are discussed in the following section. This work is essentially psychometric in nature and, therefore, has little concern with the processes involved in radiological interpretation. It shows, however, that there are stable, psychometrically measurable individual differences in radiological performance as well as differences in responsiveness to training.
8.2.1 Aptitude for radiological interpretation

Berbaum, Smoker and Smith (1985) conducted a longitudinal study of radiology residents' radiographic interpretation performance during training as rated by their supervisors. They found that there were stable individual differences in performance across five consecutive half-year units of residency which were unaffected by training. Although correlations between more distant half years were lower than those of adjacent half years, this provides evidence that performance across half years is predictable. Such individual differences were thought to be due to differences in the "... perceptual and cognitive capacities required for effective diagnosis" (p.1307).

The potential for measurement of such capacities was investigated by Smoker et al (1984), who developed a psychometric test of three-dimensional Visual Form Reconstruction in an attempt to predict training outcome of radiology residents. This test was designed to measure spatial ability in a manner that they hypothesized to be directly related to skills required in interpreting radiographic images. In particular, it contained a strong element of making inferences about three-dimensional objects from two-dimensional images. In essence, the test involved reconstructing shapes with Lego pieces from two-dimensional drawings. The test was validated as a test of spatial ability as it yielded a high correlation with the Thurstone Surface Development Test. Smoker et al found that the three-dimensional form test was a better predictor of ratings of residents' performance than the two-dimensional paper-and-pencil Thurstone test. Error scores were more closely related to performance ratings than the time taken to complete the test. This study seemed to support Smoker et al's fundamental assumption that there is an aptitude relating to three-dimensional perception from two-dimensional images which is essential for good performance in making inferences from radiographs.

The Visual Form Reconstruction Test was employed in a second study reported by Berbaum et al (1985). In this study, the test was shown to be highly predictive of
performance from the third half-year onwards indicating the long term benefit of using such a test for selecting radiology residents. In addition, performance ratings in eleven radiology subfields, ranging from neuro-radiology and body cat scan to chest radiology but not including mammography, were correlated with performance on the form test. It was found that the test predicted performance in radiological domains using cross-sectional imaging but was only of moderate or little use in predicting performance in two-dimensional radiology. Berbaum et al suggested that three-dimensional spatial ability is also important in radiological subfields, in which structures important for diagnosis are overlaid or obscured by soft-tissue structures, despite the fact that they found only moderate correlations between test scores and performance in such domain.

8.2.2 Aptitude for mammogram interpretation

The notion that there may be a particular perceptual aptitude for interpreting mammograms derives directly from Berbaum et al's work on aptitude in other radiological disciplines. Exploratory investigations of aptitude in mammogram interpretation have been reported by Walker, Gale, Roebuck and Worthington (1989). They administered a battery of perceptual and cognitive psychometric tests to medical and radiography students who had also taken part in mammography training experiments. They found that spatial aptitude and closure flexibility were predictive of mammogram interpretation performance. For experienced radiologists, however, the ability to disembed figures from background and a scanning task were more highly correlated with performance.

Unpublished work at the University of Surrey has further shown that performance in screening mammogram interpretation is affected by personality variables. This work is based on the assumption that reading screening mammograms is not only a complex cognitive skill, but that it involves task aspects which are more appropriately investigated in terms of social phenomena. Ansell et al (1990a) reports sex differences in the production of false positives, but not in overall
performance, for example. The indication is that women radiologists are more cautious, adopting less stringent decision criteria, which results in more cancers identified but at the cost of a higher false positive rate. This finding can be explained in terms of higher levels of empathy with breast cancer victims by women radiologists compared with their male counterparts. Similarly, there is some evidence that extravert readers may recall more women, a finding which has been explained in terms more impulsive decision making in extravert individuals (Unwin, 1988, Ansell et al, 1990a). The conclusion from this previous work is that mammogram interpretation may be subject to measurable individual differences in both ability and personality.

8.3 Methodological issues in investigating individual differences in complex skills

Previous work on individual differences in radiological performance has been essentially psychometric in nature. This may be entirely appropriate if the pertinent individual differences can be reliably identified as stable personality characteristics. However, in chapter 7, mammogram interpretation was characterised as a complex cognitive skill involving cognitive processes such as pattern recognition, categorisation, decision making and memory processes. Individual differences in task performance are thus likely to result at least partly from the execution of cognitive processes, although personality characteristics are almost certainly involved when a wider definition of the screening task is adopted (see above). The question is whether a purely psychometric approach is the most efficient way of identifying those individuals who are going to be successful at the mammogram interpretation task after training. There may be more utility in taking a cognitive approach to selection in the context of recruiting lay film readers (see discussion above). From a theoretical point of view, the mere measurement of individual differences is not particularly satisfactory either. Cognitive task analysis provides an opportunity to locate and investigate individual differences in terms of specific cognitive processes. There have been some recent approaches to the study of
individual differences from an information processing perspective which are worth considering in this context. In this section, current methodological approaches to individual differences in complex cognitive skill performance and acquisition are discussed.

8.3.1 Individual differences in skilled performance and skill acquisition

Referring to psychometrics and experimental psychology, Cronbach (1957) has emphasized the need for 'a true federation of the disciplines [of scientific psychology]' (p. 673), in order to do justice to individual differences in cognitive processes. Although cognitive theory provides the basis for much rigorous, methodologically well-established research, experimental approaches have in the past had little concern for individual differences. There has been a tendency to assign any variance attributable to such differences to the error term in the analysis of experimental data. Psychometrics, on the other hand, has only been concerned with the measurement of individual differences rather than their explanation in terms of cognitive processes.

Although applied educational and instructional researchers have largely heeded Cronbach's call for closer attention to individual differences in abilities and skill acquisition, there have been fewer such attempts in the theoretical analysis of cognitive skills processes. In this section, the treatment of individual differences from an applied educational and instructional perspective is compared with more theoretical approaches to individual differences in cognitive ability and cognitive skills research.

8.3.1.1 Aptitude-Treatment Interactions

The most direct impact of Cronbach's work has been the large body of aptitude treatment interaction (ATI) research (Tobias, 1987). Such research has been motivated by applied classroom teaching and instructional problems and has been directed at the question to what extent learning can be improved by accommodating
an individual's preferred style of learning when designing instructional programmes. Most of the work has focused on the cognitive processes involved in task performance in relation to the cognitive processes available to the individual learner (learner characteristics), the emphasis being on the latter. The main premise in this approach is that no one instructional design is equally beneficial for all learners. Instruction needs to be matched to individual learning styles and take into account the characteristics of the individual learner (Reigeluth and Curtis, 1987).

The method has recently been reviewed by Tobias (1987), who has emphasised the importance of a detailed analysis and specification of cognitive processes. One major problem of ATI research has been the content-specificity of tasks and learner characteristics. Some interactions between aptitudes and instruction which have been discovered to operate in a particular task domain have been difficult to replicate in other task domains. This is thought to be due to different cognitive characteristics and different task demands of the different tasks in question. Tobias (1987) has suggested that detailed cognitive process analyses of the task domain are required in order to specify the precise interactions between learner characteristics, instruction and task demands.

8.3.1.2 Individual differences in ability and skilled performance

Efforts to understand individual differences in cognitive ability/skilled performance have been less frequent than in the instructional context. Notable exceptions have been the correlational approach developed by Hunt and his co-workers (Hunt, Frost and Lunneborg, 1973, Hunt, Lunneborg and Lewis, 1975, Mcleod, Hunt and Matthews, 1978) and Sternberg's cognitive components (1977, 1979). The basic premise of this early work was that individual differences arose from differences in the processing speed or efficiency of cognitive task components as engendered in a performance model of a task or ability. Hunt et al (1973, 1975) correlated individuals' verbal ability measures with performance on basic information processing tasks, which they hypothesized to measure basic elements of verbal ability. They found differences in a number of basic information processing
parameters such as speed of processing and short term memory which differentially characterised individuals of high and low verbal ability.

Sternberg's information processing approach to human ability (1985) involves the decomposition of tasks into subtasks. Componential models of task performance are then validated internally and externally. Individual differences in task performance can be traced to differences in subtasks which represent information processing steps of the overall task. Sternberg's procedure is similar to that of Hunt et al in that it aims to identify basic information processing stages in complex cognitive tasks which are used to characterise individual differences in overall task performance. However, whereas Hunt's 'mechanistic' processes are defined operationally in terms of performance on simple information processing tasks, Sternberg's components are based on a theoretical information processing model of the task. Measurement of performance on the subtasks is thus task-specific in Sternberg's approach.

Criticism of information processing approaches to individual differences in cognitive abilities has largely come from an ATI perspective. Snow (1987), for example, has argued that any information processing model of complex cognitive performance which assumes that all individuals perform a task in one particular way is inadequate. His objection is that no single information processing model for a task fits all persons and nor does a single model fit any given person on all items of the task. Although, both Sternberg and Hunt have shown that individuals can be categorized as using qualitatively different routes, i.e. that they follow different strategies in performing a task, the problem of within-person differences has not been addressed. A similar point regarding intra-individual differences has been made by Battig (1979). He has argued that cognitive flexibility or the individual's ability to adopt alternative strategies is a more important individual difference variable than inter-individual differences. More recently this problem has been addressed by a number of researchers in particular with respect to cognitive analyses of individual differences in spatial ability. Lohman and Kylloinen (1983),
for example, found that there can be strategy shifts within an individual's performance of a spatial task. Cooper and Mumaw (1985) observed that 'one component of high spatial aptitude may be a high degree of flexibility in strategy selection' (p.89). Just and Carpenter (1985) suggested a theoretical information processing framework which could account for such within- and between-subject strategy differences. According to this, strategic differences in spatial tasks are explained in terms of different coordinate systems that people adopt for reasoning in such tasks.

8.3.1.3 Cognitive skill acquisition and individual differences

Unlike the theoretical work on individual differences in cognitive abilities, approaches to individual differences in cognitive skill acquisition have often been motivated by a desire to achieve better prediction for job selection (Lesgold et al, 1990, Egan and Gomez, 1985, Ackerman and Schneider, 1985). Nevertheless, the methods employed to study individual differences in skill acquisition rely heavily on those developed in the context of cognitive ability differences. Thus, most studies employ some kind of task analysis, although real-life tasks tend to be broken down at a more macroscopic level than abilities for the purpose of identifying individual differences (Egan and Gomez, 1985, Green and Gilhooly, 1990a).

A number of studies have been conducted in the domain of human-computer interaction. Egan and Gomez (1985), for example, have studied individual differences in acquiring skills in computer text editing. Their approach involves the 'assaying of individual differences', involving the break down of the complex task so that the sources of individual differences can be understood, 'isolating individual differences' which consists of identifying the effect of learner characteristics on particular task components, and 'accommodation of individual differences' by changing the task, by providing performance aids or by specific training intervention. The methodology employed in this study is an example of how matching people to jobs can be achieved without the usual problems inherent in
conventional selection procedures. Green and Gilhooly (1990a) have investigated differences between fast and slow learners in acquiring statistical computing skills. In addition to protocol analysis, they employed a more macroscopic information processing based breakdown of the task in order to study differences at both, tactical and strategic levels. They found that differences in learning between novices were largely due to executive processes. Similar findings have been reported by Thorndyke and Stasz (1980) in a study comparing expert and novice knowledge acquisition from maps. Success in this task largely depended on the use of task-appropriate strategies and not experience with map reading. Like in Green and Gilhooly's study good learners differed from poor ones in the execution of attentionally controlled processes such as accurate evaluation of the learning process. These findings lead to the question to what extent differences in efficiency of controlled processing is the main determinant of individual performance, at least at the novice stage.

This question has been addressed on a theoretical level by Ackerman's work (Ackerman and Schneider, 1985, Ackerman, 1987, 1988). The aim of this work has been to develop a comprehensive theory which relates individual differences in cognitive and intellectual abilities to individual differences in task performance over the course of skill acquisition within the automatic/controlled processing framework developed by Schneider and Shiffrin (1977). In this framework a distinction is made between consistent and varied practice as learning mechanisms. Whilst the former leads to automaticity in performance after skill learning, controlled processing is required even after extensive experience of the latter. The theory predicts that initial task performance is subject to controlled processes as Green and Gilhooly (1990) found in the case of statistical analysis skills acquisition and Thorndyke and Stasz (1980) in the case of knowledge acquisition from maps.

With respect to performance prediction after training, Ackerman and Schneider (1985) argue that task analysis can be used to identify those task components that are likely to become automatic after practice. Separate assessment of these
components for predictive purposes may lead to more accurate prediction with regard to job performance after extensive training or practice since initial task performance is only likely to be predictive of those task aspects that remain under conscious control. Highly skilled complex task performance may, however, consist of many automatic as well as a number of controlled task components. This approach is of considerable interest in the context of mammography. In chapter 7, it was argued that visual processing aspects of the tasks are highly automatised in the skilled radiologist (see also Myles-Worsley et al, 1988). Interpretation aspects of the task remain under conscious control though. Cognitive performance assessment should thus aim to assess components relating to visual processes separately from those relating to diagnostic decision making.

8.3.2 Psychometric versus information processing approaches

At first sight, it appears that information processing approaches to individual differences are superior to psychometric approaches from a practical and a theoretical point of view, where it is desirable to identify individual differences at the level of cognitive processes. Psychometric measurement fails to explain the cause of individual differences, whereas cognitive task analysis conducted on the basis of information processing theory can at least provide a pointer to the potential causes of observed individual differences. There are other reasons for preferring an information processing method to psychometric measurement, too. Lesgold et al (1990) have argued that cognitive task analysis based testing might lead to better prediction if the sample of applicants is small. Similar points have been made by Sternberg (1981), Ackerman and Schneider (1985) and Egan and Gomez (1985). Carroll (1988) has argued that information processing approaches to individual differences are always preferable because measurement of cognitive task performance involves absolute measurement (e.g. reaction times or error rates) rather than normative or criterion referenced measurement and, therefore, is more reliable.
The main disadvantage of cognitive testing is that tests take longer to develop and are laborious to administer. Compared with traditional psychometric group testing procedures, cognitive testing usually involves individual testing using computers or other apparatus. Such elaborate procedures may yield better predictability but are not very practical if candidates are to be selected from a large pool of applicants. The question of whether to adopt a purely psychometric approach or an information processing approach in personnel selection is, to a large extent, dependent on the purpose to which findings about individual differences in a task are to be put. If the aim is simple prediction of task performance, measurement of individual difference variables at the level of psychometric paper and pencil tests may be entirely appropriate, if personality and ability variables can be shown to be valid and reliable predictors of the task in question. If, on the other hand, the aim is to identify individuals from a relatively small group of people, who are most likely to succeed at the task, cognitive testing may be preferable (Lesgold et al, 1990). The cognitive task analysis approach is essential, however, if the aim is to investigate causes of individual differences in the task at a cognitive processing level.

The main strength of information processing methods, with respect to individual differences, is that theoretical conclusions about the processes that are the source of individual differences in complex tasks can be drawn. Carroll (1980) has put the case that componential analysis in Sternberg's sense could be usefully combined with psychometric methods such as factor analysis to analyse the interrelation between component scores and psychometric tests of ability. Thus, he has shown that Sternberg's preparation and response component as well as the encoding component in analogical reasoning tasks (Sternberg, 1977) are intrinsically related to reasoning speed as measured by psychometric tests (Carroll, 1988). Such psychometric factors can then be interpreted as source traits of individual differences which enter into the parameters for information processing.
8.4 Psychometric and cognitive task analysis approaches to individual differences in mammography

The studies reported in the following chapters were motivated by a desire to investigate applied issues in film reader selection for screening mammography, on the one hand, and to uncover clues as to what kind of processes might be the source of individual differences in such complex perceptual tasks, on the other. The main premise was that predictable individual differences in the performance of perceptual reasoning tasks such as mammogram interpretation can be traced to psychometrically measurable differences in personality, visual ability and cognitive skill variables, which in turn can be explained in terms of differences in cognitive processing during task performance.

Three studies are reported in this context. The first addresses the question of whether aptitude for mammography can be defined psychometrically in terms of personality and ability variables alone. The second study investigates whether performance on psychometric predictors tests is related in a meaningful way to performance on cognitive tests devised on the basis of cognitive task analysis. The final study is concerned with a number of issues in prediction of mammogram interpretation performance from psychometric and cognitive predictor tests. The specific approaches to individual differences taken in the three studies is outlined in more detail below.

8.4.1 Aptitude for screening mammography

The first study reported in this part is concerned with the possibility of identifying psychometrically what constitutes aptitude for mammography. Some of the studies discussed above (in particular Walker et al, 1989 and Ansell et al, 1990a) have shown that there is some evidence that mammogram interpretation performance can be predicted with psychometric tests of ability and personality. In the present study, an attempt was made to identify broad aspects of ability and personality that
contribute to good interpretation performance in mammography, the aim being the
development of a paper-and-pencil test set which is predictive of screening
mammogram performance. This work was conducted strictly in the tradition of
previous psychometric work on individual differences in radiology (Smoker et al,

Messick (1987) has argued that the distinction between ability on the one hand, and
personality on the other is an artificial one. Modern theories of personality regard
human personality as a 'system in the technical sense of something that functions as
a whole by virtue of interdependence of its parts.' (p.36). The 'interdependent parts'
of personality in this view are cognition and non-cognitive aspects of personality
such as affect and motivation. Messick (1987) has outlined the structural
relationships that exist between cognitive and non-cognitive aspects of personality,
e.g. ways in which personality influences the organisation of cognition and
determines how stable structures are. The utility of such an integrative approach to
cognition and personality becomes clear when considering individual differences in
a complex cognitive task such as interpreting mammograms. Different levels
represented by different task aspects are associated with such tasks, which are
likely to be differentially determined by cognitive ability and non-cognitive
personality structures. In mammography, for example, there are easily identifiable
cognitive, strategic and social task aspects as summarised in Table 8.1. Task
performance can, thus, be seen to be dependent on and influenced by a multitude
of interdependent ability and personality variables each of which could potentially
facilitate or depress performance. Table 8.1 presents a description of different task
aspects on cognitive, strategic and social levels of the screening task with an
outline of those psychometric variables, which are hypothesized to be predictive of
the task on the respective levels.
Table 8.1: Specific task aspects of screening mammography and possible predictors of performance after training

<table>
<thead>
<tr>
<th>Task level</th>
<th>Task nature</th>
<th>Possible Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>cognitive</td>
<td>feature detection in degraded images classification into ill-defined categories</td>
<td>ability: visual acuity visual aptitude</td>
</tr>
<tr>
<td>strategic</td>
<td>extracting ambiguous targets making decisions under conditions of uncertainty</td>
<td>cognitive style: field-dependence creative style</td>
</tr>
<tr>
<td>social</td>
<td>social impact of making a mistake positive-negative imbalance time pressure when reading 80-100 mammograms per session</td>
<td>personality: impulsivity vigilance stress-proneness</td>
</tr>
</tbody>
</table>

The idea that a comprehensive psychometric test battery based on different task levels can be instrumental in the development of a set of test items measuring aptitude for screening mammogram interpretation is based on previous psychometric work in cognitive skill domains. One example is Carroll's (1958, 1985) development of a foreign language aptitude battery. Like screening mammography, foreign language learning is essentially a complex cognitive skill involving memory processes and an ability for sensory discrimination, auditory discrimination in the case of language acquisition. In addition, though, foreign language acquisition, like screening mammography, is subject to social pressures (see e.g. Tarone, 1988 for more detail). Carroll's work was based on the measurement of certain cognitive abilities and personality variables involved in second language acquisition. Thus, the aim is to identify a domain of personality and ability which predicts performance in a skill, taking into account general personality variables bearing on the performance of the skill as well as specific ability. In chapter 9, an attempt to identify such a psychometric aptitude domain for screening mammography is made.
8.4.2 Cognitive task analysis

It has been argued previously that more task-relevant tests may be better predictors of mammogram interpretation than psychometric paper-and-pencil tests (Lesgold et al., 1990). In psychometric approaches, skilled performance is not necessarily measured directly, but is often equated with cognitive ability which is hypothesized to be predictive of skilled performance. Direct measurement of aspects of skilled performance may thus be preferable for more accurate prediction, because it is not based on assumptions of abilities underlying cognitive performance (Carroll, 1988). Apart from these applied considerations regarding accuracy of prediction, there also are theoretical reasons for preferring a cognitive skills approach to individual differences. Identification of individual differences at a level of cognitive task components allows the determination of causes for individual differences in overall performance at the level of cognitive task components.

The cognitive task model of screening mammogram interpretation, which was presented in chapter 7, provides a starting point for cognitive task analysis. It is clear from the model that mammogram interpretation involves visual processes, on the one hand, and reasoning and decision making processes on the other. Memory is involved at both visual assessment and interpretation levels. Previous work on complex cognitive skills has attempted to break the task down into intrinsic processing components. One method of achieving this is to use verbal protocols obtained from subjects during the performance of the task to derive different cognitive processes involved in task performance (see, for example, Green and Gilhooly, 1990a). In the present research, verbal protocols proved useful only in the elicitation of relatively late stages of information processing in mammogram interpretation. Specifically, it was possible to identify clinical reasoning and decision making processes in the radiographer case studies (chapter 4). This was thought to be due to the explicitly perceptual nature of radiograph interpretation skill. As much of the early information processing in mammogram interpretation is perceptually coded, trainees were unable to gain conscious, verbal access to the
processes involved in acquiring and processing critical visual information. Since the radiological task could not readily be analysed in terms of verbal data, it was necessary to separate the task into appropriate components on the basis of processes identified in the task model.

Mammogram interpretation appears to proceed largely in terms of consciously inaccessible processes even at the novice level. Because of this, it is difficult to break down the task in terms of actual task components, although such components undoubtedly exist as outlined in the task model in chapter 7. Information processing steps may be derived directly from the task model. Thus the task model clearly predicts that information processing during mammogram interpretation involves a visual search component, followed by a target recognition component. A further cognitive component involves decision making processes. Skill acquisition further relies heavily on the appropriate execution of memory processes. On the whole, this represents a somewhat macroscopic task analysis of mammogram interpretation performance, which may nevertheless be useful in investigating broad information processing parameters involved in the task. Since information processing is not directly observable in mammogram interpretation, external cognitive tasks were developed, which were designed to model closely certain aspects of the processes involved in skilled performance. Performance on these external tasks can then be used as a predictor of overall performance. Moreover, it may be possible to identify individual differences as resulting from relatively specific differences in cognitive performance. The development of such cognitive tests developed as predictors of mammogram interpretation performance, which are based on the task analysis according to the model in chapter 7, is discussed in detail in chapter 10.

8.4.3 Prediction of performance and individual differences

The final chapter in this part is concerned with issues in the prediction of mammogram interpretation performance. Three aspects of prediction are addressed
in particular. The first question is whether mammogram performance after training can be predicted psychometrically. From the point of view of trainee selection it would be very useful to be able to specify those characteristics of people that would predict whether they will be good mammogram screeners after training independent of pre-training performance. Alternatively it might be possible to identify those individuals who are most likely to benefit from training, i.e. those who improve most as a function of training. A related issue in this context is the extent to which tests based on cognitive task analysis actually achieve better prediction in the context of recruiting lay film readers for mammography. Although cognitive testing procedures are more elaborate than traditional psychometric testing, gains in predictive value in the latter may outweigh the procedural inconvenience implied in the former (Lesgold et al, 1990). Lastly, an attempt to identify the sources of individual differences will be made. For this purpose, an analysis of differential performance on predictor tests of good and poor learners is conducted. Such a procedure has proved useful in a number of studies investigating individual differences in complex cognitive tasks (Green and Gilhooly, 1990, Thorndyke and Stasz, 1980). Such studies have typically succeeded in identifying some of the processes responsible for individual differences in overall task performance. Knowledge about the nature of these processes may be utilised in designing training to support specific individual weaknesses.

8.4.4 Summary

The work on individual differences presented in this part is largely motivated by applied issues in the selection of lay readers for screening mammography. The problem of observer variability needs to be addressed by selecting those individuals who are most likely to succeed at the complex interpretation task. Previous work has indicated that a psychometric approach to individual differences in screening mammography may be successful in identifying ability and personality variables which are predictive of interpretation performance. In addition, it may be possible to identify a psychometric aptitude domain for the radiological task. Theoretical
and methodological considerations led to the conclusion, however, that in the long term, a cognitive processes based approach to individual differences may be more appropriate both in terms of accuracy of selection procedures and possibilities in terms of training design. Both, psychometric and cognitive process approaches to individual differences can be examined in terms of their respective predictive power for mammogram interpretation skill after training.
Chapter 9
Searching for a psychometric perceptual personality domain

The identification of a psychometric aptitude domain which is predictive of mammogram interpretation performance is potentially very useful in selecting lay film readers for training. As discussed in the previous chapter, this is particularly true if there is a large pool of potential trainees and the requirement is to select those most likely to succeed at the task. The main pre-requisite for such a procedure is that it is possible to identify reliable predictors of task performance from a range of personality and ability characteristics. These can be measured with psychometric paper-and-pencil tests which are easy to administer to relatively large groups of candidates. With respect to selecting lay film readers for mammography it is conceivable that such a procedure may be useful if lay-screening becomes an accepted alternative to consultant screening. Under these circumstances it may become necessary to implement valid selection procedures (Saxton, 1991).

In the previous chapter, it was argued that mammogram interpretation is a multilevel task. The search for reliable and valid predictors of such a task needs to take into account cognitive, strategic and social task aspects. Such an integrative approach promises better prediction, because more of the variance due to individual differences can be accounted for. In addition, such a framework based on modern theories of personality permits the generating of hypotheses about the structural interrelationships of the different aspects of personality involved (Messick, 1987). Thus, an aptitude domain for mammography may be largely defined in terms of ability for visual tasks as has been indicated by Walker et al (1989). However, there are also strategic and social task aspects which influence task performance such as disembedding of ambiguous targets and time pressure during viewing. Messick's (1987) view that 'human functioning may sometimes represent a defensive compromise amongst contending forces ....' (p.36) clearly invokes the
possibility that inaptitude at the strategic or social level may depress performance
despite aptitude at the cognitive level having been established.

In this chapter, a psychometric study is reported that attempts to identify a multi-
level visual perceptual aptitude domain, which might be relevant to predicting
mammogram interpretation performance. For this purpose, a battery of five ability
and personality tests was administered to a large sample of students. The primary
task facet of mammogram interpretation is visual perception. Perceptual
performance is influenced by both ability and personality structures. The nature of
these relationships and how they might bear on mammogram interpretation
performance are discussed in more detail in the following two sections.

9.1 Perceptual Ability and Mammography

Investigations into aptitude for radiological interpretation have almost exclusively
focused on ability. Smoker et al (1984), for example, concentrated on one aspect of
visual ability, drawing inferences about three-dimensional structures from two-
dimensional data. They claimed to be tapping an underlying aptitude for
radiological interpretation with their form perception test (Berbaum et al, 1985).
Although approaches to mammographic aptitude have been somewhat broader in
the inclusion of tests of perceptual strategy (Walker et al, 1989), visual ability was
again the primary focus. Both studies found visual ability measures to be predictive
of radiological and mammographic performance, respectively, lending support to
the notion that it might be possible to define what constitutes aptitude for
mammogram interpretation psychometrically.

Ability is undoubtedly a major factor in the performance of most complex tasks.
General ability is predictive of learning speed and flexibility of strategy use in
many tasks. It is well established that ability in a particular domain can predict
skill acquisition and subsequent performance in that domain. Specific abilities
which are particularly relevant to mammogram interpretation are visual-spatial
abilities because they are concerned with perceptual evaluation. In general, visual perception factors summarise abilities in perceiving and mentally manipulating visual forms and patterns and dealing with the spatial layout of the environment. Some of the factors included in such ability have not yet been well defined (Carroll, 1988), but general visual ability is thought to be comprised of essentially three aspects, space relations, visualisation and spatial orientation. The processes involved in these factors have, however, been difficult to separate (Lohman, 1979) reported in Carroll, 1988).

For the present study, both general and spatial ability tests were included in the battery. Spatial ability tests are often contaminated by sex differences, a problem that does not apply to general ability tests. At the lower end of the intelligence distribution such tests are expected to correlate highly. However, this is not necessarily the case when testing a population drawn from the upper end of the distribution. As the acquisition of mammogram interpretation skill involves not only perceptual evaluation but also strategic decision making, it was thought that both kinds of ability are likely to be involved.

9.1.1 Perception, Personality and Mammogram Interpretation

The relationship between personality variables and mammogram interpretation performance are less obvious than those with perceptual ability. The influence of these variables are more likely to operate at the level of strategic and social task aspects. Nevertheless, there is some evidence that personality has a direct impact on cognitive functioning with respect to strategic decisions made during task performance as outlined below. On the other hand, interpreting mammograms involves some completely non-cognitive aspects. Ways in which personality may influence mammogram interpretation performance via non-cognitive task aspects are discussed below.
9.1.1.1 Personality and Perception

Personality has traditionally been closely implicated with perception. In clinical psychology, for example, the use of perceptual tests such as the Rorschach test were firmly rooted in the belief that patients' perceptions can somehow reveal their personalities. In the 40's and 50's perception was also frequently linked with personality in experimental psychology. Early attempts were made to explain perceptual processes through personality as, for instance, in Bruner's work (1951). Although much of this work is of little consequence for modern approaches to perception, there remains a psychometric interest in the link between perception and personality as evidenced in the work of Witkin and his colleagues on the field-dependence trait (Witkin, Lewis, Hertzman, Machover, Bretnall-Meissner and Wapner, 1954, Witkin, Lewis, Hertzman, Machover, Meissner and Wapner, 1972, Witkin, Dyk, Faterson, Goodenough and Karp, 1974).

Field dependence - independence (FD-I) is a dimension of individual variation which characterises an aspect of information processing. Field-dependent people are readily influenced by environmental cues and processes information non-selectively. The field-independent individual is more reliant on internal cues and is more discriminating in the processing of environmental information. The trait is often measured in terms of the individual's ability to locate simple figures in a complex field. Two modes of perceiving can thus be identified, a holistic one (FD) and an analytic one (FI). The FD-I dimension is conceptualised as a cognitive style, i.e. a preferred mode of functioning. Although there is a notion that the individual differences engendered in the FD-I dimension are differences in information processing and, therefore, closely linked to the cognitive perception of information from the environment, it is claimed that the dimension also has a more social aspect. FD-I, thus, also has an influence on how social information is processed by individuals, i.e. holistically or more selectively.

Given that mammogram interpretation involves disembedding ambiguous features from ambiguous backgrounds the relevance of the field-dependence dimension to
predicting mammography performance becomes clear. Individuals who habitually adopt an analytic perceptual strategy are more likely to succeed in extracting important features from difficult mammograms. That such target detection is, in fact, associated with perceptual style has been shown by Thornton, Barrett and Davis (1968) with respect to target identification performance in aerial photographs. Field-independence is, thus, likely to be an important perceptual characteristic of a good film reader.

9.1.1.2 Personality and decision making

Kirton (1978) has argued that Witkin's field-dependence theory has theoretical implications beyond the perceptual level into conceptual and cognitive realms. Perceptual details which have been identified have to be interpreted at the conceptual level. In mammography, decisions about whether to recall or not to recall on the basis of a set of ambiguous features identified in a mammogram represents an integral part of task performance. Kirton (1976) has suggested that such decision making is subject to individual variation in a way similar to perception being determined by field-dependent or independent strategies. Individual differences in decision making can be conceptualised along a dimension of creative style according to Kirton (1976), the extremes of which are labelled 'adaptation' and 'innovation'. 'Adaptors' tend to approach problems within the given terms of reference, whilst 'innovators' are more likely to search for novel solution. The differences on this dimension are not, however, differences of level of creativity but purely of style.

Kirton (1978) hypothesized that there would be a relationship between field-independence and innovation on the basis of the theoretical overlap between the two constructs. This was confirmed in significant correlations between adaptation-innovation and FD-I. This finding has recently been replicated by Robertson, Fournet, Zelhart and Estes (1987) with alcoholic subjects. It, thus, seems that both adaptation-innovation and FD-I theories attempt to classify people in terms of how they deal with new perceptual or cognitive information, either by accommodating it
into an existing paradigm (FD, adaptors) or by bringing about the destruction and replacement of an existing paradigm (FI, innovators). Both dimensions can be seen as relevant to mammogram interpretation, field-dependence by virtue of its close association with perceptual processes and adaptation-innovation with respect to information processing style in complex decision making.

9.1.1.3 Personality and non-cognitive aspects of screening mammography

It was argued in the previous chapter that there are aspects of mammogram interpretation, which are independent of visual ability, strategic decision and level of interpretation skill, but which may nevertheless have a significant impact on interpretation performance. These non-cognitive aspects of screening mammography are comprised of those that are stress inducing for the film reader on the one hand, and those that demand screener vigilance on the other. As these task aspects have been discussed in some detail in chapter 3, they are only briefly dealt with here.

One aspect of screening mammography, that sets it apart more obviously from clinical radiology than any other, is the imbalance between negative and positive instances. In viewing a set of mammograms a screener would only expect to classify approximately 10% as abnormal and would make a decision to recall only about half of those. This illustrates that vigilance performance, i.e. the ability to maintain concentration in view of very few positive instances, is an important aspect of screening mammography. In addition, there are a number of stress-inducing factors implied in screening mammography. Mammograms are typically read in sets of up to 100 per session. Apart from maintaining vigilance over a considerable period of time, observers may worry about making relatively rapid decisions in view of ambiguous information. Most film readers are very aware that wrong decisions on their part might cause another person's premature death. At the same time they are under pressure to keep down recalls in order to reduce the costs of the screening programme as their job depends on its cost effectiveness. It can be seen that screeners may suffer a permanent conflict between being efficient and making mistakes.
The extent to which film readers can cope with the stresses implied in the job and with maintaining vigilance performance is likely to be largely influenced by personality factors. Eysenck's theory of personality, for example, assumes that differences in coping with stress and vigilance situations are, in fact, 'hard-wired' at a physiological level (Eysenck, 1967). These differences lead to different behavioural manifestations which can be measured at a trait level. Eysenck's theory comprises three dimensions, extraversion-introversion, neuroticism and, more recently, psychoticism. Personality influences on mammography performance can be largely addressed within Eysenck's comprehensive framework.

There has been extensive work on vigilance and personality as defined by Eysenck's framework, for example. Most experiments show that extraverts perform worse on vigilance task than introverts (Eysenck, 1967, p.87-89, Eysenck and Eysenck, 1985). This has been conceptualised at the biological level in terms of differential cortical arousal between extraverts and introverts. A similar argument has been put forward as explanation for a second experimental finding, that introverts are more prone to fear of punishment than extraverts as evidenced in human conditioning experiments (Eysenck, 1967). This has implications with respect to stress proneness in the face of making mistakes. Neuroticism, which is implicated with high emotional reactivity, is another dimension which might be important in determining how stressful an individual perceives the screening situation to be. In particular in combination with introversion, neuroticism implies high levels of anxiety which may cause the screener to be overcautious when making recall decisions thus leading to low specificity. The third dimension of personality in Eysenck's theory, psychoticism, is claimed to address an individual's disposition to psychotic breakdown. Psychoticism is associated with a lack of concern for others which may be a beneficial trait to have as a screener because high specificity may be compromised by undue concern for others' lives as indicated above. Previous work has tentatively indicated that there might be an association between mammography performance and psychoticism scores.
9.2 **Method**

The procedure employed in order to identify an aptitude domain for mammography involved the selection of psychometric personality and ability tests, which were likely to be predictive of aspects of screening performance. Individual characteristics that might be responsible for differences in interpretation performance in the screening situation have been discussed in some detail above. The study was essentially correlational in nature. The factor structure underlying test battery scores was investigated in search on an over-arching perceptual personality-ability factor, which might be indicative of a perceptual aptitude domain predictive of screening mammography performance.

9.2.1 **Subjects**

Participation in the psychometric study was limited to undergraduate students in their first year at Surrey University. There were 235 students who volunteered to take part in a two-hour testing session. Their ages ranged from seventeen to forty-four, the median age being nineteen. There were 165 female volunteers and 70 male ones. Sample composition in terms of subjects studied is shown in Figure 9.1, which represents a breakdown according to the faculty membership of participants.

**Figure 9.1**: Percentage breakdown of student subjects by faculty
9.2.2 The Test Battery

Five tests were included in the psychometric battery, two of which were ability tests and the remaining three were personality tests. All these tests included are well-established, commercially available measures. Test selection was to some extent restricted by practical considerations with respect to administration. Thus only group tests or tests that could easily be administered to groups were included. In addition, tests were selected so that as broad a range of ability and personality characteristics of relevance to mammography could be established in as brief a time period as possible. Since all subjects were volunteers, an upper time limit of two hours for testing was imposed.

9.2.2.1 Ability measures

Ability assessment covered both, general ability and spatial ability. In both cases, high level ability measures were employed in order to achieve discrimination amongst participants, who as university students represent a sample highly selected for intelligence.

MD 5 Mental Ability Test (The Test Agency, 1972)
The MD 5 Mental Ability Test is a test of general ability assessing verbal vocabulary and arithmetic skills. Although the test measures educational attainment to some degree, it is primarily concerned with participants' ability to deduct relationships and to apply rules that govern them. Test items are arranged in order of ascending difficulty. As the test has a time limit of 15 minutes it measures both power and speed. The test claims to be culture-fair, i.e. appropriate for all English speakers. Norms are available for graduates and people with A level, both of which produce good score distributions.

Shapes Analysis Test (Heim, Watts and Simmonds, 1972)
The Shapes Analysis Test (SAT) is a test of spatial perception. It assesses the ability of the participant to manipulate mentally different shapes and sizes, to visualize how geometric figures will appear when turned over or round, to estimate
area and to assess spatial relations. In short, the SAT is a test of 'k', general visual ability. An important feature of the SAT is that it contains both two-dimensional and three-dimensional items so that both 2D and 3D scores can be derived separately. The test has a time limit of 25 minutes.

9.2.2.2 **Personality measures**

Personality was assessed with three tests, two of which represented cognitive style measures. Specifically, field-dependence-independence, adaptation-innovation, extraversion, neuroticism and psychoticism were assessed.

**Group Embedded Figures Test (Witkin, Oltman, Raskin and Karp, 1971)**

The Group Embedded Figures Test distinguishes between two modes of perceiving, an analytic one (FI) and a holistic one (FD). Participants are required to trace the outline of a specified simple shape, displayed on the backpage of the test booklet in a more complex drawing. A high score, i.e. successful tracing of many simple shapes, indicates field-independence. Low scorers are assumed to be more field-dependent. Again testing is time-controlled at five minutes for the completion of each of the two main test sections.

**Kirton Adaptation-Innovation Inventory (Kirton, 1985)**

The Kirton Adaptation-Innovation Inventory distinguishes two kinds of decision making or problem solving style, an adaptive one, which is characterised by seeking 'better' solutions to problems, and an innovative one, which is associated with attempting to find 'different' solutions. The inventory divides into three subscales. Originality (O), which is about creativity, has been defined by Kirton (1976) as indicative of the 'creative person' described by, for example, Rogers as the 'creative loner'. Efficiency (E) is about precision, reliability, and discipline and has been described by Kirton (1976) as 'Methodological Weberianism'. The third subscale contains items about rule/group conformity, 'Mertonian Conformism' being Kirton's description of the person with proper respect for authority and rules. The Kirton Inventory is a self-report measure which requires participants to rate on a
five point scale the extent to which they find it easy or hard to present images (as described in the inventory items) of themselves as particular kinds of persons over a sustained period of time. There is no time limit for the completion of the test. People who score below the theoretical mean score of the inventory are classified as adaptors, whilst those scoring above are innovators.

Eysenck Personality Questionnaire (Eysenck, Eysenck and Barrett, 1985)
The Eysenck Personality Questionnaire is a self-report questionnaire requiring subjects to give yes/no answers to a range of questions about themselves. In this study, the Short-scale EPQ-R with a revised Psychoticism scale was used as published by Eysenck, Eysenck and Barrett (1985). Four scales are measured with the EPQ, extraversion (E), neuroticism (N), psychoticism (P) and a lie scale (L). Each scale is comprised of twelve items in the short EPQ. There is no time limit for this test.

9.2.3 Procedure

The test battery was administered to participants in groups of between five and thirty subjects. The order of the tests within the battery was counter-balanced across the different group sessions. All tests were administered according to the instructions in the respective manuals, with the exception of the Group Embedded Figures Test the administration of which was adjusted as described below. The Kirton Adaptation-Innovation was not part of the initial battery, but was added at a later stage. There are, thus, a number of cases with missing data on the KAI. These were excluded from correlational and factorial analyses of the data.

9.2.3.1 Group Embedded Figures Test - Adjusted Procedure
Preliminary analysis of the data was conducted after administering the test battery to the first 62 subjects. This revealed that there was a marked ceiling effect with respect to GEFT scores. Time limits for completion of sections 2 and 3 of the test were, therefore, reduced from five minutes to three minutes each, according to
suggestions in the test manual. The adjusted testing procedure led to much greater differentiation between individuals in subsequent testing. For the correlational and factorial analyses GEFT scores of all subjects were transformed into z-scores so as to make pre- and post-adjustment scores comparable.

9.3 Results

The purpose of collecting psychometric test data was to investigate the potential existence of a perceptual aptitude domain relevant to mammography screening performance. For this purpose, a correlational analysis of the data was performed. In addition, a principal components analysis was conducted in order to examine the factor structure underlying the test battery.

Table 9.1 shows the means and standard deviations of the scores of all five tests including test subscales in comparison with the relevant norms provided in the respective test manuals. It can be seen that the mean test scores achieved by the present sample are similar to the published normative scores. On the whole, male test scores deviate more from the norms, particularly in the case of psychoticism, extraversion and neuroticism as well as on the Kirton Adaptation-Innovation scale. This may be due to the fact that the male sample is smaller than the female one and, therefore, less reliable. The differences between sample test scores and norms are not very large, though. In the case of the Shapes Analysis Test norms for University Science students were used. As the students in the present sample were drawn from all faculties, it is not surprising that, overall, lower mean scores were recorded for both men and women, when compared with the norms. The means and standard deviations for the Group Embedded Figures Test in Table 9.1 are those from subjects tested after the adjustment in testing procedure described above. It can be seen that almost halving the standardized testing time resulted in comparable means and standard deviations to those of liberal arts students under normal testing procedures. Adjusting the testing time thus resulted in a more discriminative score distribution, avoiding the ceiling effects encountered under normal testing conditions.
Table 9.1: Means, Standard Deviations and Norms of five Personality and Ability Tests

<table>
<thead>
<tr>
<th></th>
<th>PRESENT STUDY</th>
<th>NORMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TOTAL</td>
<td>MALE</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>M</td>
</tr>
<tr>
<td>EPQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>234</td>
<td>2.791</td>
</tr>
<tr>
<td>KAI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>204</td>
<td>18.975</td>
</tr>
<tr>
<td>O</td>
<td></td>
<td>40.632</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td>34.990</td>
</tr>
<tr>
<td>(Separate norms for Male and Female not available)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEFT</td>
<td>173</td>
<td>11.324</td>
</tr>
<tr>
<td>SAT</td>
<td>235</td>
<td>8.004</td>
</tr>
<tr>
<td>2D</td>
<td></td>
<td>7.098</td>
</tr>
<tr>
<td>3D</td>
<td></td>
<td>15.102</td>
</tr>
</tbody>
</table>

* GEFT norms are based on liberal arts college students at a standardized testing time of 5 minutes for the second and third section. In the present study testing time for these sections was reduced to 3 minutes each in order to achieve the same degree of differentiation between individuals.
9.3.1 Test Reliability

Reliability of the tests was established by calculating Cronbach’s alpha coefficient for subscale and total test scores. These are presented in Table 9.2. Although the majority of the tests and their subscales can be seen to be reliable above the cut-off point of .7 (Nunnally, 1978), there are some exceptions. In particular, the SAT subscales and the P scale of the EPQ seem to have rather low internal reliabilities. The Shapes Analysis Test yields a satisfactory reliability coefficient, if the whole scale of 36 items is assessed, but reliability on both subscales is low. Problems of reliability of the P-scale on the EPQ have been well-documented in the past (Eysenck, Eysenck and Barrett, 1985). The reliability coefficient of the psychoticism scale in the present sample compares with those reported by Eysenck et al (1985) for the short-scale EPQ though (.63 compared with reliabilities of between .51 and .68 for different samples).

Table 9.2: Evaluation of test reliability with Cronbach’s alpha coefficient

<table>
<thead>
<tr>
<th>EPQ-P</th>
<th>EPQ-E</th>
<th>EPQ-N</th>
<th>KAI-O</th>
<th>KAI-E</th>
<th>KAI-R</th>
<th>GIPT</th>
<th>MD5</th>
<th>SAT2D</th>
<th>SAT3D</th>
<th>SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>N CASE</td>
<td>234</td>
<td>234</td>
<td>234</td>
<td>204</td>
<td>204</td>
<td>204</td>
<td>173</td>
<td>235</td>
<td>235</td>
<td>235</td>
</tr>
<tr>
<td>N ITEM</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>7</td>
<td>12</td>
<td>32</td>
<td>18</td>
<td>57</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>ALPHA</td>
<td>.627</td>
<td>.845</td>
<td>.791</td>
<td>.844</td>
<td>.702</td>
<td>.905</td>
<td>.877</td>
<td>.862</td>
<td>.578</td>
<td>.638</td>
</tr>
</tbody>
</table>

9.3.2 Correlational Analysis

There were 201 complete data sets for all psychometric tests in the battery. These test and subscale scores were entered into a correlational analysis. Table 9.3 shows the correlations between tests and test subscales. It can be seen that there were high correlations between general (MD5) and spatial ability (SAT). The Group Embedded Figures Test also correlated highly with both spatial, and general ability tests, in particular the three-dimensional subscale of the Shapes Analysis test.
Table 9.3: Correlations between tests and subscales

<table>
<thead>
<tr>
<th></th>
<th>EPQ-P</th>
<th>EPQ-E</th>
<th>EPQ-N</th>
<th>KAI-O</th>
<th>KAI-E</th>
<th>KAI-R</th>
<th>KAI</th>
<th>MD5</th>
<th>GEFT (z)</th>
<th>SAT2D</th>
<th>SAT3D</th>
<th>SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPQ-P</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPQ-E</td>
<td>-1.62</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPQ-N</td>
<td>-0.031</td>
<td>-0.245**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAI-O</td>
<td>0.332**</td>
<td>0.303**</td>
<td>-0.196*</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAI-E</td>
<td>0.149</td>
<td>0.029</td>
<td>-0.149</td>
<td>0.202*</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAI-R</td>
<td>0.315**</td>
<td>0.231**</td>
<td>-0.099</td>
<td>0.464**</td>
<td>0.462**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAI</td>
<td>0.368**</td>
<td>0.260**</td>
<td>-0.195*</td>
<td>0.807**</td>
<td>0.585**</td>
<td>0.810**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD5</td>
<td>0.041</td>
<td>-0.084</td>
<td>-0.063</td>
<td>0.031</td>
<td>-0.003</td>
<td>0.022</td>
<td>0.035</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEFT (z)</td>
<td>0.098</td>
<td>-0.114</td>
<td>0.001</td>
<td>0.024</td>
<td>-0.000</td>
<td>0.063</td>
<td>0.042</td>
<td>0.471**</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT2D</td>
<td>0.084</td>
<td>-0.067</td>
<td>-0.138</td>
<td>0.027</td>
<td>0.182*</td>
<td>0.123</td>
<td>0.115</td>
<td>0.341**</td>
<td>0.582 **</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT3D</td>
<td>0.011</td>
<td>-0.132</td>
<td>-0.019</td>
<td>0.042</td>
<td>0.047</td>
<td>-0.011</td>
<td>0.021</td>
<td>0.371**</td>
<td>0.560**</td>
<td>0.571**</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>SAT</td>
<td>0.011</td>
<td>-0.115</td>
<td>-0.084</td>
<td>0.059</td>
<td>0.125</td>
<td>0.058</td>
<td>0.073</td>
<td>0.402**</td>
<td>0.537**</td>
<td>0.871**</td>
<td>0.900**</td>
<td>1.000</td>
</tr>
</tbody>
</table>

* significant at $p < .01$

** significant at $p < .001$
Amongst the personality measures there were high correlations between all EPQ subscales and the Kirton Adaptation-Innovation scores. Psychoticism and extraversion were positively related to innovation, whereas neuroticism was associated with adaptation. These overall relationships arise specifically from correlations between the EPQ scales and the KAI-Originality subscale and, in the case of psychoticism and extraversion, also from significant associations with the KAI-Rule conformity subscale. There were no significant correlations between EPQ scales and KAI-Efficiency scores. Although the traits measured by the EPQ are claimed to be orthogonal (Eysenck and Eysenck, 1984), a substantial negative correlation between extraversion and neuroticism was observed in the present data. With respect to overlap between personality and ability measures it can be said that no correlations of any significance were observed between field-independence and innovation. There was, however, a moderate but significant association between the two-dimensional scale of the Shapes test and the KAI-Efficiency scale.

9.3.3 Principal Components Analysis

In order to examine the factor structure underlying the observed correlations, a principal components analysis was performed. For this purpose, only subscale scores were included for the KAI and the SAT. Total test scores were not entered into the analysis, because subscale scores were expected to be more informative than the overall scores. Inclusion of both would have resulted in problems with a poorly conditioned matrix. Although some of the subscales of the tests in question are intercorrelated, these correlations are much lower than any of the subscale/total test score associations. Subjects with missing variables were excluded from this analysis, resulting in the inclusion of 204 complete data sets in the analysis.

The principal components analysis was conducted using the SPSS-PC analysis package. Examination of the scree-plot led to the conclusion that a two or three factor solution is most appropriate for the present data set. There were two factors with eigenvalues greater than two, accounting for 46% of the variance between
them. A third factor had an eigenvalue exceeding one, which added a further 12.3% to the variance accounted for. In total, the three-factor solution accounted for 58.3% of the observed variance. An oblique rotation was performed using the oblimin method. This converged in 8 iterations to reveal the factor structure shown in Table 9.4.

Table 9.4: Three factor solution of principal components analysis with oblique rotation

<table>
<thead>
<tr>
<th></th>
<th>Factor 1 Ability</th>
<th>Factor 2 Personality 'Psychotic Innovator'</th>
<th>Factor 3 Personality 'Anxiety'</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD5</td>
<td>.7014</td>
<td>-.0314</td>
<td>.0063</td>
</tr>
<tr>
<td>SAT 2D</td>
<td>.7582</td>
<td>.0307</td>
<td>-.1250</td>
</tr>
<tr>
<td>SAT 3D</td>
<td>.8327</td>
<td>-.0397</td>
<td>.0237</td>
</tr>
<tr>
<td>GEFT</td>
<td>.7826</td>
<td>.0428</td>
<td>.0112</td>
</tr>
<tr>
<td>P</td>
<td>-.0169</td>
<td>.7574</td>
<td>.4434</td>
</tr>
<tr>
<td>KAI-E</td>
<td>.0384</td>
<td>.5561</td>
<td>-.0684</td>
</tr>
<tr>
<td>KAI-R</td>
<td>-.0121</td>
<td>.8008</td>
<td>-.1267</td>
</tr>
<tr>
<td>KAI-O</td>
<td>-.0121</td>
<td>.6847</td>
<td>-.2527</td>
</tr>
<tr>
<td>E</td>
<td>-.1608</td>
<td>.0340</td>
<td>-.8122</td>
</tr>
<tr>
<td>N</td>
<td>-.1161</td>
<td>-.1145</td>
<td>.6382</td>
</tr>
</tbody>
</table>

It can be seen that there was one ability factor and two personality factors. All variables were well defined by this solution with communality values exceeding .55 in all cases. The first factor, the ability factor, had high positive loadings on all ability measures and the Group Embedded Figures scores. The second factor loads positively on psychoticism and the three KAI subscales. This clearly represents a personality factor, which could be labelled 'psychotic innovation'. The third factor accounts for less variance than the first two. It loads positively on neuroticism and negatively on extraversion. This is also a pure personality factor, which has been
labelled 'anxiety'. The reasoning behind the labels of the two personality factors are discussed in detail below. Table 9.5 shows that intercorrelations between the three factors were negligible.

Table 9.5: Factor correlation matrix for the principal components analysis

<table>
<thead>
<tr>
<th></th>
<th>FACTOR 1</th>
<th>FACTOR 2</th>
<th>FACTOR 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACTOR 1</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FACTOR 2</td>
<td>0.0899</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>FACTOR 3</td>
<td>0.0198</td>
<td>-0.1508</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

9.4 Discussion

The primary aim of the present study was to establish whether it was possible to identify a perceptual/cognitive aptitude domain predictive of mammography performance with a set of commercially available, psychometric paper-and-pencil tests. For this purpose, a test battery consisting of two ability and three personality tests, each of which had been argued to be relevant to aspects of mammogram interpretation, was administered to a large sample of undergraduate students. Correlational and principal components analysis of these data showed that there was little evidence for any overlap between ability and personality in the perceptual domain under investigation.

9.4.1 Correlational and factor structures of the test battery

The correlation matrix (Table 9.3) shows that there were high correlations between the ability tests and the KAI and EPQ. Contrary to expectations no significant associations were found between field-independence as measured on the Group
Embedded Figures Test and innovation scores on the KAI. The GEFT did not, in fact, associate with any of the personality measures but correlated very highly with the all ability measures, particularly the three-dimensional SAT scale. The case of the GEFT is discussed in more detail below. There was only one significant correlation between a personality and an ability scale, that between KAI-Efficiency and the two-dimensional SAT scale ($r^2 = .182$, $p < .01$). However, the moderate size of this correlation raises doubts with respect to the importance it should be assigned. It is a clear possibility that this is a spurious association, particularly in view of the large number of correlation coefficients computed.

9.4.1.1 Personality and ability as separate components
The picture emerging from the principal components analysis supports the impressions gained from the examination of the correlation matrix. There is no genuine overlap between the ability and personality dimensions measured in the present test battery. Three factors were extracted each of which was well-defined in terms of a number of variables (Table 9.4). The first factor loaded highly on all ability measures as well as GEFT scores (see discussion below). This is clearly an ability factor as indicated by the negligible loadings of all personality variables on this factor. The second factor is a personality factor as none of the ability variables load on it. The factor is largely defined in terms of the KAI subscales and psychoticism all of which have high positive loadings on this factor. The association between innovation and psychoticism, which represents the essence of this factor, can be explained in terms of behavioural similarities between Eysenck's trait and Kirton's cognitive style. Thus, both have been said to be about disregard of social norms and emphasis of the self as well as a like for the unusual (Eysenck and Eysenck, 1976, Kirton, 1976). The third factor is defined in terms of positive loadings of neuroticism and negative extraversion. Although extraversion and neuroticism are orthogonal factors in the general population, it is not uncommon to find that they are correlated in student populations. The combination of introversion and neuroticism on an individual level has been interpreted in terms of anxiety. The emergence of this third factor, which is defined in terms of unstable introversion,
i.e. anxiety, can possibly be explained in terms of the demanding testing situation. This factor may, thus, be indicative of state anxiety as induced by task demands.

9.4.1.2 GEFT and perceptual style
The close alliance of field-independence and ability found in the present study raises some questions about the validity of the GEFT as a measure of perceptual style. FD-I has frequently found to be correlated with measures of intelligence and Vernon (1972), for instance, has argued that the construct is indistinguishable from intelligence, in particular from spatial ability, the "k" factor of intelligence. Similarly, McKenna (1983, 1984) has claimed that measures of FD-I are more appropriately viewed as measures of cognitive ability rather than cognitive style. The Group Embedded Figures Test certainly has some characteristics in common with ability tests such as the fact that there is a time limit. This specifically emphasizes speeded performance of the disembedding task, an aspect of the task which was of particular importance in the present test battery because testing time was reduced in order to avoid ceiling effects. This may also be one of the reasons why Kirton's (1978) and Robertson et al's (1987) findings of correlations between KAI and EFT/GEFT were not replicated in the present study.

A further problem with the GEFT is that it measures only one side of a theoretically bipolar dimension. Thus, the score on the GEFT is a field-independence score, field-dependence being inferred from a low score rather than direct measurement. It has been argued that differences in FD-I may well be differences in strategy, despite correlations with intelligence (Hampson, 1982, Clark and Roof, 1988). The analytic strategy associated with field-independence happens to be the better strategy for achieving both, high scores on the GEFT and high intelligence test scores, particularly "k" factor intelligence. If tests were developed that measured field-dependence in a direct manner, i.e. tests on which field-dependent strategies result in better performance, similar relationships between field-dependence and intelligence might be discovered.
9.4.2 Psychometric prediction of mammography performance

It can be concluded that the present study has failed to define an aptitude domain which encompasses cognitive and non-cognitive elements in mammography performance. However, although the identification of a perceptual personality type germane to mammogram interpretation has proved elusive in this study it is, nevertheless, possible that some or all of the psychometric paper-and-pencil tests are individually predictive of mammogram interpretation performance. These issues will be further investigated in chapter 11.

It has been argued previously, that better prediction may be achieved with more task-relevant tests. Aspects of mammogram interpretation may be better captured by tests which are directly based on a cognitive task analysis than by fairly remote paper-and-pencil tests of ability and personality. Such cognitively derived tests can still be used psychometrically for the prediction of post-training performance, for instance. At the same time, however, such tests are based on a theoretical analysis of the cognitive processes involved in the task under investigation leading not only to more validity than psychometrically derived tests but also to greater flexibility with respect to studying individual differences at the process level. Following on from the failure to identify a psychometric aptitude domain for mammography, it was attempted to develop cognitive tests predictive of mammography based on an analysis of the interpretation task (chapter 10).
Chapter 10
Development of a cognitive test battery

The attempt to identify a perceptual aptitude domain reported in the previous chapter was intended to indicate possible predictors of mammography performance. However, it was not possible to identify a unidimensional personality/aptitude domain involving the hypothesized psychometric constructs. The underlying skills involved in a complex perceptual task such as mammogram interpretation are thus not well captured by general psychometric paper-and-pencil tests of ability and personality. Lesgold et al (1990) have argued that the prediction of complex cognitive task performance may be achieved more appropriately by the use of cognitive tests, which are based on thorough task analysis. Similarly, Carroll (1988) has indicated that the absolute measurement involved in assessing task components is preferable to criterion referenced measurement, because it is a more direct.

In this chapter, the development of a cognitive test battery based on skills prerequisite for mammogram interpretation is described. The aim was to develop a set of tests, which directly measure the cognitive skills required for screening mammography. In chapter 8, the possibility of breaking the task into intrinsic information processing components was discussed. The indication was that mammogram interpretation does not lend itself to the traditional methods of achieving this, since verbal protocols, for example, do not capture important task aspects such as visual processing. Use of the information processing model in chapter 7, thus, represents one way of informing the search for appropriate task aspects to be measured in tests, which are representative of intrinsic task components. Three aspects of task performance of particular interest were visual perceptual search, target analysis and strategic decision making, each of which is discussed in more detail below.
10.1 Derivation of cognitive predictor tests

The main purpose of developing cognitive tests which measure performance on hypothesized elements of mammogram interpretation skill is another, more specific attempt at predicting mammogram performance as a function of training. This issue is dealt with in some detail in chapter 11. In the present chapter, the aim is to report on the development and validation of cognitive tests derived from cognitive task analysis. In addition, the relationship between the psychometric predictor variables and cognitive tests are explored. The question arises whether the psychometric tests are related in a meaningful way to performance on cognitive tests devised on the basis of cognitive task analysis, thus establishing a rudimentary form of concurrent validity for both test sets as predictors of mammogram interpretation performance.

The identification of cognitive task components can be conducted on a number of different levels. Previous work on individual differences in cognitive abilities, for example, has been directed at attempts to identify basic information processing parameters of specific abilities (Sternberg, 1977, Hunt et al, 1973). Research on complex cognitive skills has invariably resulted in more macroscopic analyses of the tasks, however (Thorndyke and Stasz, 1980, Egan and Gomez, 1985, Green and Gilhooly, 1990, Lesgold et al, 1990). The approach in the present research has been to follow the cognitive skills tradition and select relatively macroscopic task elements for prediction, the reason being that more comprehensive task elements are more likely to capture the essence of overall task performance. Cognitive task were developed as representations of intrinsic information processing components, because there is no obvious way of investigating information processing in mammogram interpretation directly.

Examination of the simplified process model of mammogram interpretation in chapter 7 (Figure 7.3) reveals that, on a macroscopic level, there are two kinds of processes associated with mammography skill, each of which broadly consists of
two information processing components. The visual search stage consists of perceptual evaluation of the radiograph and classification based on perceptual evidence. The outcome of these two components is the identification of potential target abnormalities. If such an abnormality is identified, the second processing stage is initiated. This consists of the interpretative analysis of target abnormalities and results in a further classification depending on the perceived significance of the target abnormality. The production of a true positive, i.e. the correct identification of an abnormality is dependent on the execution of both processing stages without error.

The aim of the cognitive test battery was to present as comprehensive an analysis of different aspects of mammography skill as possible. Four tasks were included, each primarily but not uniquely concerned with one aspect of mammogram interpretation performance. Visual search performance was assessed with two tasks. The first one was a simple measure of contrast sensitivity, which was included in order to determine the importance of very basic physiological perceptual processes in target detection. The second task involved the detection of cognitively simple targets, which were difficult to discriminate perceptually. Target analysis performance was measured with a computer-administered task, which employed probabilistically defined targets. The aim of this task was to assess the degree to which ill-defined targets could be correctly identified and classified. The task demanded that appropriate schemata of probabilistic targets were developed and applied in rapid two-way classification. A fourth task aimed to identify individuals' preferred decision making/classification style. It has been argued previously that making decisions about mammograms involves a strategic element (chapter 5). Target interpretation involves classification processes, which may be influenced by individual differences in decision strategies. The four cognitive tests and the justification for their inclusion in the battery are discussed in more detail below.
10.1.1 Target detection

There is clear evidence that radiological skill involves a substantial visual perceptual component as is evident from the visual search literature concerned with radiological performance (see chapter 2). The importance of this was recognised in the model of mammogram interpretation (Figure 7.1), in which the entire first stage of processing is concerned with vision. Visual processing may be even more important in screening mammography than in diagnostic radiology, because it involves the detection of pre-clinical abnormality. Such abnormalities are typically small and difficult to detect (see chapter 3). One aspect of visual perception which appears to be of particular importance in this context is contrast sensitivity, as the detection of minimal abnormality in radiographs is often associated with distinguishing low contrast targets from fuzzy backgrounds (Revesz et al, 1974, Kundel and Revesz, 1980). This is the case in screening mammography when detecting small masses, an early indication of cancer, for example.

The importance of visual acuity for detecting small white specks indicating calcification and contrast sensitivity for detecting low contrast masses in screening mammography has been indicated by previous, unpublished work on differences between expert mammography screeners and students (Ansell et al, 1990a). It appears that expert screeners perform significantly better when making judgements about low contrast dots displayed on X-rays than the control group. Mammography screeners were not only more accurate in their judgement of the presence or absence of dots, but also made their decisions with a higher degree of confidence. With respect to individual differences, the question is whether mammography screeners' superior performance on the low contrast decision task was due to self-selection into screening because of this specific aptitude or to perceptual learning whilst performing screening duties. Although it was impossible to distinguish between these two possibilities on the basis of the above study, it can be said that there is some evidence for perceptual learning in tasks concerned with visual acuity (Fiorentini and Berardi, 1981). Such learning proved to be very specific, though.
Nevertheless, the possibility that contrast sensitivity is an important individual difference variable in mammogram interpretation can not be ruled out.

Two aspects of target detection performance were examined as part of the cognitive test battery. Simple contrast sensitivity was assessed with a commercially available gratings test, the Cambridge Low Contrast Gratings (Wilkins, Della Sala, Somazzi and Nimmo-Smith, 1988). The test administration follows a psychophysical two-alternative forced choice procedure which allows for the distinction between observer criterion and observer sensitivity. The test, thus, represents a relatively pure measure of contrast sensitivity rather than contrast detection performance. The second aspect of target detection performance was concerned with complex visual search. The task involved the detection of simple targets which were at the limit of perceptibility. A test X-ray displaying a grid, which contained barely perceptible low contrast dots, was generated by X-raying a perspex 'phantom' object with holes drilled into it. On the X-ray the air-filled holes show up as low contrast dark dots. Phantom objects such as the one used to generate the test X-ray are routinely used to check the sensitivity of X-ray equipment in screening mammography. Subjects were required to search for twenty-five dark dots, each located in one of five possible locations. The forced multiple choice nature of the task ensured a high level of task complexity, involving both visual search and decision making processes. In this manner, the 'grid' task was expected to simulate the main aspects of visual search and target detection processes involved in mammogram interpretation.

10.1.2 Target analysis

Interpretation of abnormalities identified during visual search of the radiograph involves the assessment of targets in terms of relevant target and background features. In screening mammography minimal abnormalities present particular problems in that they may not be well defined. In addition, features of abnormality are often fuzzy and probabilistic. Interpreting mammograms, thus, involves the
analysis and evaluation of the area containing the abnormality in terms of the presence or absence of several probabilistically defined features. These assessments have to be merged into a binary decision in the most appropriate manner in order to avoid misclassification (Getty et al, 1988). As in target detection both, evaluation and decision making processes are involved in assessing detected abnormalities.

The task of identifying and analysing uncertain targets was simulated in a computer-administered test involving probabilistically defined targets representing low contrast 'masses' on complex backgrounds. Of all the tests in the cognitive battery the 'target' task is the closest simulation of mammogram interpretation as a whole, as it contains elements of visual search, target analysis and binary classification. The test involved several levels of target probability and two levels of background complexity. This facilitated the measurement of subjects' accuracy and speed of response at different combinations of target and background complexity. In addition, it was possible to measure the degree of learning in such a task, as the first half of the trials, which involved simple targets, was followed by another set of trials with more complex targets. If learning occurred, accuracy and response time should be better in the transfer task than in the initial target analysis session. Several accuracy and speed measures can, be derived from the 'target' task, which were hypothesized to be related to overall mammogram interpretation performance. In terms of the model (Figure 7.1), the 'target' task is still largely concerned with visual processes rather than interpretation of complex information. Thus, the initial task is designed to measure performance in making decisions about the presence or absence of 'abnormal areas' (simple targets) in a complex visual display. The processes involved in interpretative decision making are more closely modelled in the transfer task, since this requires the conjunctive integration of two target features to arrive at the right decision.
10.1.3 Decision making and classification

Strategies in decision making and classification were identified as a major issue in mammogram interpretation in chapter 5. The way in which classification decisions are made involves the evaluation of information from mammograms. This evaluation is conducted with close reference to schemata stored in memory which guide visual exploration on the one hand (visual schemata), and interpretative processes on the other (disease schemata). In mammography, learning to make appropriate classifications on the basis of available information has been identified as a prerequisite for good interpretation performance (Gale et al, 1987, Getty et al, 1988). Although the acquisition of appropriate disease schemata, which inform the interpretation and classification process on the second level, can be aided by specific instruction and training, it was argued in chapter 5, that such learning was at least in part inductive in nature. The mechanisms involved in acquiring mammogram interpretation schemata are, therefore, likely to involve the testing of hypotheses which are either confirmed or disconfirmed on further investigation.

Individual differences may, thus, arise from strategy differences in acquiring and evaluating information contained in mammograms. There may also be differences disease schemata, which in turn result from different strategies in knowledge acquisition. Strategy differences of this kind were investigated as part of the cognitive test battery with a concept learning task based on Bruner et al's (1956) original work on concept learning. Bruner et al were able to distinguish two different strategies that are employed in learning well-defined concepts under cognitive strain conditions, scanning and focusing. In the present battery, a computer-administered concept learning procedure developed by Johnson (1971, 1978) was used. Johnson (1971) adapted Bruner's procedure for identifying individual differences in concept learning in order to overcome problems with the interpretability of certain strategy groups. The 'concept' task requires subjects to identify artificial concepts by trying out instances or testing hypotheses. Previous applications of this procedure have used participant's responses to classify them as
predominantly following one particular strategy (Johnson, 1978, Morrison and Duncan, 1988). In the present application of the procedure, subjects were assigned scores for each strategy type.

One problem with employing a concept learning task based on Bruner's work in order to identify strategy differences in the acquisition of information is that the concepts to be acquired in the task are artificial and well-defined. Strategies identified in this manner may not be representative of classification strategies employed in mammogram interpretation which typically involves probabilistic cues. Bruner et al (1956) have discussed the case of probabilistic concepts and have indicated that the way individuals acquire information about such concepts is similar to well-defined concepts (pp. 182 - 230). Although this has been put into question by subsequent research on natural concepts (Collins and Quillian, 1969, Rosch, 1973, 1978), there is some evidence for continuity between artificial and natural concept acquisition as far as information processing strategies are concerned. Morrison and Duncan (1988) have shown that the strategies identified by Johnson's procedure are predictive of complex fault diagnosis performance, a task which can involve probabilistic reasoning. In terms of the mammogram interpretation model, this kind of task may thus represent processes involved in schema acquisition required in the second, interpretative stage.

A further reason for including this task in the cognitive battery is its hypothesized relationship with field-dependence. The significance of this cognitive style for predicting mammogram interpretation performance has already been discussed (see also Walker et al, 1989). Johnson (1978) has argued that perceptually based field-dependence dimensions could be the developmental forerunners of cognitively based problem solving style. The strategies identified by the 'concept' task may thus be representative of a general cognitive style brought to bear in all problem solving behaviour by an individual. In so far as mammogram interpretation actually involves 'problem solving', i.e. the gathering and evaluating evidence for or against a particular decision, attempts to resolve ambiguity by reference to prior
knowledge, and deciding whether perceptual evidence is sufficient to warrant further investigation, strategies identified by means of the 'concept' task may be predictive of performance.

10.1.4 The nature of the cognitive tasks and their relationships with psychometric predictor tests

The four tasks included in the cognitive test battery can thus be seen to be representative of different aspects of mammogram interpretation performance as modelled in chapter 7. The tasks are not independent, however. Although each one aims to simulate a particular aspect of mammogram interpretation, there is considerable overlap between the tasks in terms of the processes accounted for. Thus, the contrast sensitivity measure, the 'grid' task and the 'target' task are all concerned with visual aspects of mammogram interpretation. All three tasks comprise elements of visual search and decision making. However, the contrast sensitivity measure is primarily concerned with very basic visual processes, whilst the 'grid' task is a measure of complex visual search. 'Target' is more concerned with complex target identification and evaluation. Since the processes modelled in Figure 7.5 (chapter 7) do not claim to be independent of each other, but are highly recursive in nature, it would be expected that overlap between the cognitive tasks would increase their predictive power with regard to mammogram interpretation performance.

As far as the psychometric predictors discussed in the previous chapter are concerned, further overlap between the two test batteries can be expected, since both, the psychometric tests and the cognitive tasks were intended as predictors of mammogram interpretation performance. The theoretical links between problem solving styles assessed in the 'concept' task and perceptual style as measured by the Group Embedded Figures Test (Johnson, 1978) have already been discussed. In addition, cognitive predictors may overlap with the more visually orientated cognitive style and ability measures. Thus, field-independence may be related to
certain visual search parameters, and spatial ability may be implicated in target analysis. The aim of this chapter is to investigate the nature of relationships between psychometric and cognitive predictor variables in order to derive a set of relatively independent predictor variables, which can be employed in the attempts to predict mammogram interpretation performance in the next chapter.

10.2 Method

Tasks for the cognitive test battery were selected to measure performance on simplified aspects of mammography skill. The relationship between the cognitive tests and mammogram interpretation performance has been discussed above. As the psychometric study in the previous chapter, the present study was essentially correlational. The aim was to validate the cognitive tasks as measures and to examine the structure of underlying patterns of performance. In addition, relationships between cognitive task performance and psychometric tests were investigated.

10.2.1 Subjects

Ninety-nine subjects for whom full sets of psychometric data were available took part in the cognitive task study. All 235 students from the psychometric study were contacted, and sixty-nine female and thirty male subjects agreed to participate in this second study. Their ages ranged from eighteen to forty-four, the median age being nineteen. Seventy-two of the participants were on psychology or psychology related degree course. The remaining twenty-seven students came from other disciplines (see chapter 9).

10.2.2 Procedure

Participants completed the four tests in one testing session lasting approximately two hours. Tasks were administered individually or in pairs of two subjects at a
time. Two of the tasks were presented on an Archimedes 310 computer. All tasks were completed in a windowless experimental laboratory with artificial lighting. Lights were dimmed for the 'grid' task and the 'target' task. Administration procedures for each task are detailed below. As with the psychometric test battery, task order was counterbalanced across subjects.

10.2.3 The test battery - Task designs and administration procedures

10.2.3.1 Cambridge Low Contrast Gratings
The Cambridge Low Contrast Gratings (Wilkins and Robson, 1987, Wilkins et al, 1988) represent a psychophysical measure of contrast sensitivity at a spatial frequency of 4 c/deg. The test, which is commercially available, is designed to assess contrast sensitivity in individuals who have normal visual acuity as measured by letter acuity tests. The spatial frequency of 4 c/deg was selected for assessment in the test because this is the frequency at which the normal human visual system is close to maximally sensitive.

The gratings are presented in an A4 size book which is hung on a wall. Two pages, one appearing above the other, are presented monocularly at a viewing distance of six meters. One page in each pair contains a grating while the other is blank. Subjects are required to indicate whether the lines are at the top or at the bottom. After completion of a practice set of pages, gratings are shown in order of decreasing contrast until an error is made. The test consists of four descending series being shown to each eye, each terminating if an error occurs or when the end is reached. The second, third and fourth series commence four gratings prior to the one on which an error occurred on the previous series. The Cambridge Low Contrast Gratings were administered according to the procedure and instructions suggested by the authors and described by Wilkins et al (1988).

10.2.3.2 The 'grid' task
The grid test is a task specifically developed to measure complex visual search
with low contrast targets. The grid results from X-raying a perspex base into which twenty-five individual one inch square perspex pieces had been placed. Each of these had a hole drilled into it in one of five possible locations, in the middle or in one of the four corners. The holes in the X-ray grid employed in the present task were drilled to a depth of .5 mm with a .35 mm diameter drill. The computer-controlled drilling procedure ensured that they were located either in the centre of each one inch square perspex piece or half way between the centre and one of the corners. Five pieces of perspex had a centre hole and the remaining twenty pieces a corner hole. The corner holes could be rotated so as to display the hole in either corner. The hole size was calibrated on the basis of test development experiments (see 10.3.1) to be maximally sensitive to individual differences.

The test X-ray was displayed on a backlit light box such as those commonly used for viewing radiographs. Any light from the box extraneous to viewing the practice or test X-rays was blocked out with black paper. Prior to the test each subject was shown an X-ray displaying an example dot and were asked to complete a small practice grid consisting of five squares. As in the test proper, the practice grid contained one dot in each square and served to demonstrate the five possible locations. The main task required subjects to locate each of the twenty-five dots in the X-ray grid and to record their degree of confidence with regard to the dot being in the indicated location. A record of '1' signified that the subject was 'not sure that the dot is in the indicated position' whilst a '3' meant that the participant was 'very sure that the dot is in the indicated position'. Participants were instructed to work as fast as possible and to 'guess' the location of the dot if they could not see one. Task performance was timed. An instruction and recording sheet for this task can be found in Appendix IIId.

10.2.3.3 The 'target' task
'Target' is a computer-administered target search task (see Appendix IIIa for a program listing) which is designed to measure individuals' ability to correctly identify probabilistically defined targets on backgrounds of varying complexity.
Backgrounds and targets were computer-generated using the 256 colour palette available on the Archimedes 310 computer. Backgrounds were generated using either two colours or eight colours. The two colour backgrounds consisted of four-pixel blue and grey blocks generated at random by the computer. The eight colour backgrounds were generated the same way but utilising the four tints available for each of the two colours, thus producing eight shades with subtle variation in colour. Targets of varying discriminatory difficulty were also computer-generated at random. They were defined as 15 x 15 pixel blocks which contained 30%, 40% or 50% of the blue base colour, with the remaining pixels in the background colours. Half the trials involved backgrounds superimposed with one such target, in the other half the background did not contain a target.

Subjects received computer-administered instructions (see program listing, Appendix IIIa) which included a demonstration of backgrounds and targets. The task was to decide on each trial whether a target was present in the display or not. Participants responded by pressing specified keyboard keys for 'yes' and 'no' responses. They were instructed to 'keep their fingers on the relevant buttons and press them as soon as they had made up their minds', as performance was timed. Prior to commencing the task subjects were able to complete a six trial practice session as often as they felt was necessary. Each of these trials comprised one of the six target displays. There were no no-target practice trials. The purpose of the practice session was to ensure that participants were aware of the kind of target they were looking for. Each practice trial ended in feedback to the subject which indicated the location of the target in the display.

The task proper commenced when the subject indicated that he or she was ready. Each trial was prefixed by a fixation screen which displayed a black cross in the middle of the screen. Timing started as soon as the display appeared on the screen and terminated on pressing of one of the response keys. Subjects received feedback only in terms of whether they had been correct or not after each display had been terminated. There were five trials with each type of target/background combination
and a matched number of no-target trials resulting in sixty trials altogether which were presented in random order based on a computerised randomisation procedure. After the first sixty trials, the task was changed slightly. Subjects were instructed to 'look only for large targets from now on'. Targets were now defined as 30 x 30 pixel blocks which were generated in the same manner as above. In addition, backgrounds contained one, two or three small (15 x 15 pixel) distracters. As before, instructions included a full demonstration of different target/background combinations. The new targets were considered more complex in terms of the features that needed to be identified for correct classification and, in combination with the distractors, presented a more complex target identification task.

10.2.3.4 The 'concept' task

'Concept' is a computerised concept identification task (see Appendix IIIb for a program listing), which is based on the Zaps-Duds task described and validated by Johnson (1971, 1978). The purpose is to identify the extent to which a particular subject follows a particular strategy in obtaining the information required to solve the task. This involves the identification of a specific rule, which is used to classify patterns consisting of a sequence of the letters X and O. Rules specify two positions in the pattern which are relevant and must contain an X. Two kinds of rules are possible, either a conjunctive one, which indicates that Xs must be present in both specified locations, or a disjunctive one, in which case the X is located in either of the two specified positions. Instructions to participants (Appendix IIIc) were slightly amended from Johnson's (1971) to accommodate small changes in procedure (see below). Subjects could either 'test hypotheses' by typing in potential solutions (e.g. type in '1 and 3', if they think this is the relevant rule) or 'try out patterns' by typing in any combination of Xs and Os (e.g. XOXOO or XOOOO). Each pattern which has been tried out was classified as a positive or a negative instance by the computer (yes or no). Hypotheses were similarly evaluated in terms of accuracy (true or false). All patterns and hypotheses tried in a particular trial and their outcomes remained displayed on the computer screen throughout the trial.
The task and its administration procedure largely followed Johnson's (1971) design. However, instead of using sequences consisting of six letters as Johnson (1971, 1978) did, five letter sequences were employed in the present implementation of the task. This resulted in a reduction of possible solutions from 30 to 20 while preserving the essence of the task. This amendment was made in order to reduce the time taken to complete the task. As subjects were required to complete five runs, a practice run plus four runs of the task proper, the saving in terms of time to solution was considerable for the majority of subjects. It was essential to keep the task as brief as possible in view of the fact that it was completed as part of a whole battery of cognitive tasks. Apart from this, all features of Johnson's procedure (see Johnson (1971, 1978)) for detailed descriptions of the task and procedure were retained, including the minimum-information-feedback technique. This involves making each trial as long as possible by not having a predetermined solution which could be guessed. The computer gives feedback on each trial so that the maximum number of hypotheses remain tenable. This ensures easier identification of relevant strategies.

10.3 Results

The present study is concerned with validating cognitive task performance as measures of cognitive processes underlying mammogram interpretation, examining the structure of the cognitive test battery and investigating the relationships between cognitive and psychometric variables. Tasks specifically developed for the test battery, i.e. the 'grid' task and the 'target' task, were subject of pilot studies in order to assess the validity of derived measures. Scores on the other two tasks were examined in terms of their reliability compared to published norms. Data from the cognitive test battery were analysed using principal components analysis examine the underlying factor structure. In order to investigate the relationships between the cognitive and the psychometric batteries a canonical correlation was performed.
10.3.1 Test development experiments and score derivation of the 'grid' task

Previous work employing a similar X-ray based task found that breast screening radiologists were better at identifying low contrast dots than a control group. The research involved the use of an X-ray grid based on a phantom object with holes of varying sizes (both depth and diameter). The larger holes produced massive ceiling effects in that experiment, though. As the purpose of the task in the present battery was to discriminate between individuals, several initial studies were conducted in order to define a hole size which was maximally sensitive to individual differences.

10.3.1.1 Exploring depth and diameter of hole sizes as variables in the grid task

The purpose of the test development experiments was to identify those hole sizes which combined depth and diameter in such a manner as to produce dots on the verge of perceptibility. One of the questions arising in the context of developing a new phantom object for use in the grid task was whether the perceptibility of the low contrast dots in the X-ray were influenced by the depth of the air-filled holes (contrast) or by the diameter of the holes (size) or both. Three small, informal experiments were conducted in the course of the test development, firstly to address the question of maximal discriminability between individuals and, secondly, in order to investigate perceptibility as a function of hole depth and diameter.

First test development experiment

Two slabs of perspex were drilled with holes ranging from .35 mm to 1 mm in diameter and .1 mm to 2 mm in depth. There were thirty-two holes on each slab, each combining one of the four diameters (.35 mm, .6 mm, .8 mm and 1 mm) with one of eight levels of depth (.1 mm, .5 mm and a further six depths in steps of .25 mm up to 2 mm). The slabs were X-rayed and the X-ray presented to subjects displayed on a light box. Seven subjects were asked to record on a piece of paper, which showed outlines of the slab, the places in which they saw a black dot. The experiment showed that most misses occurred with the smallest holes, in particular
those of .35 mm diameter up to a depth of 1 mm and those of .1 mm depth across all diameters.

**Second test development experiment**
The second pilot experiment aimed to investigate the discriminatory power of the smallest hole sizes in a more formal manner. For this experiment an X-ray grid consisting of 50 squares each with a hole drilled into the middle or one of the corners was produced. These included forty target squares, twenty with .35 diameter holes half of which were drilled to a depth of .75 mm and the other half to 1 mm depth, and twenty .6 mm diameter holes, .1 mm and .5 mm deep. The other ten squares contained clearly visible dots at 1 mm diameter and 2 mm depth for example and practice purposes. Instructions to the ten subjects were the same as in the first experiment. With ten subjects and ten holes of each size there were one hundred observations per target dot size. Descriptive analysis revealed that the mean number of misses on the target holes was 35.6 (ranging from 28 to 41) with a standard deviation of 3.68. It also become clear that the two large hole sizes (.35 mm dia, 1 mm depth and .6 mm dia, .5 mm depth) still produced ceiling effects with only 26% and 5% misses respectively compared with around 50% for the smaller holes. The 6 mm dia, .1 mm depth hole size appeared to be the most difficult with 59% misses. The extent to which depth or diameter contributed to the difficulty in perception could not be established in this pilot study.

**Third test development experiment**
The third pilot experiment was set up specifically to investigate whether hole depth, diameter or both contributed most to the perceptibility of dots on the X-ray. A new grid containing forty-five squares was produced. The target squares were thirty squares each with a hole of the following dimensions, a diameter of .35 mm, .45 mm and .6 mm combined with a depth of .1 mm, .25 mm, .5 mm, .75 mm or 1 mm. There were two squares with holes of each size. The remaining fifteen squares were taken from the previous experiments. They had holes in one corner of varying sizes and were not included in the analysis. Eleven subjects took part in this
experiment. Analysis of variance was performed on the number of hits each hole size produced. The summary of the Anova is shown in Table 10.1.

Table 10.1: Summary of analysis of variance investigating the role of depth and diameter of holes in phantom objects in the perceptibility of dots in X-ray grids.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>91.803</td>
<td>4</td>
<td>22.95</td>
<td>4.72</td>
<td>p &lt; .025</td>
</tr>
<tr>
<td>Diameter</td>
<td>7.27</td>
<td>2</td>
<td>3.635</td>
<td>.74</td>
<td>p &gt; .05</td>
</tr>
<tr>
<td>Depth x Dia</td>
<td>121.397</td>
<td>8</td>
<td>15.175</td>
<td>3.12</td>
<td>p &lt; .05</td>
</tr>
<tr>
<td>Residual</td>
<td>73</td>
<td>15</td>
<td>4.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>293.47</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The analysis shows that hole depth appears to be more important for making dots perceptible on the X-ray, but that there is also a significant interaction between hole size and hole diameter. Although the larger hole sizes again produced ceiling effects, both size and contrast appear to be important in defining a hole size which produces individual differences. Contrast is more important as would be expected in the case of X-rays, since they are complex grey-scale images.

10.3.1.2 Development of the 'grid' task phantom and scoring of the test

As indicated above the squares contained in the X-ray grid used for the 'grid' task in the test battery were .35 mm in diameter and .5 mm in depth. This particular size was chosen because it produced dots on the verge of perceptibility. All three pilot studies showed that smaller holes were hardly ever seen, whilst larger holes tended to produce a ceiling effect. Only one hole size was used in order to produce a reliable score which was based on a large number of readings under the same conditions.

The test was scored in terms of correct identification of dots and in terms of the confidence with which judgements of location had been made. Two scores were derived from the test, the number of dots correctly identified (G-corr) and the
confidence with which errors were made (G-error). Although the task was timed, this was not included as a variable because there were too many cases with missing data. Descriptive statistics of the measures derived from the grid task are shown in Table 10.2.

Table 10.2: Means, standard deviations and ranges of scores derived from the grid task.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>S.D.</th>
<th>Actual Range</th>
<th>Theoretical Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-corr</td>
<td>11.87</td>
<td>2.78</td>
<td>3 to 17</td>
<td>0 to 25</td>
</tr>
<tr>
<td>G-error</td>
<td>1.65</td>
<td>.36</td>
<td>1 to 2.64</td>
<td>1 to 3</td>
</tr>
</tbody>
</table>

10.3.2 Task validation and score derivation for the 'target' task

'Target' was designed to simulate major aspects of mammogram interpretation. Targets were probabilistic and had to be interpreted in relation to background features. In addition, the task consisted of an initial 'simple target' session followed by a more complex 'multiple feature target' session. Whereas targets in the former were simply concentrations of a particular shade within a particular area, in the second part of the task, targets had to be identified in terms of clusters of a particular shade and of a particular size. This part of the task can thus be seen to require the transfer of search and interpretation skills learned in the first part to a more complex variation of the initial task. The transfer task was designed to assess in how far practised performance on the first task is relevant to performance in the second one.

Performance on the target task was measured in terms of two quantities, accuracy as indicated by the number of correctly classified trials and response speed which was measured in terms of reaction time from onset of the stimulus display to the pressing of one of the two response keys clearly marked on the computer keyboard. The overall mean accuracy and response scores are shown in Table 10.3.
Table 10.3: Mean overall accuracy scores and response times with standard deviations and ranges for the 'target' task.

<table>
<thead>
<tr>
<th></th>
<th>no. correct trials</th>
<th>response time (msec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>89.778</td>
<td>2554.7</td>
</tr>
<tr>
<td>s.d.</td>
<td>9.804</td>
<td>1050</td>
</tr>
<tr>
<td>range</td>
<td>61 to 108</td>
<td>1175 to 7165</td>
</tr>
</tbody>
</table>

10.3.2.1 Reliability and validity of the 'target' task

In order to assess the reliability and validity of the task as a simulation of mammogram interpretation, three factors may be examined in terms of their contribution to variation in performance. These are background properties (two colour or eight colour), target properties (30%, 40%, 50% or no target) and skill transfer (single feature target trials versus multiple feature target trials). Previous research has shown that both, background complexity and target discriminability, contribute to the difficulty of target detection in complex visual displays (Kundel et al, 1985). In addition, it was anticipated that there would be a learning effect as subjects build up appropriate visual schemata of targets and backgrounds as a function of practice. The task was set up as a three-factor within-subjects design with all subjects completing trials at all levels of background, target and transfer factors. Analysis of variance can be employed to explore the relative contribution of each factor to variation in performance. Summaries of analyses of variance for both, accuracy data and response time data can be found in Tables 10.4 and 10.5. It can be seen that all three main effects are significant for both, accuracy and response time data (transfer from single to multiple target task, $F = 163.7$ and $F = 147.3$ respectively, $p << .001$, background $F = 11.27$ and $F = 4.25$ respectively, $p < .05$, and target type $F = 156.77$ and $F = 124.37$ respectively, $p << .001$).
Table 10.4: Summary of analysis of variance for the number of correct trials on the 'target' task.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer</td>
<td>139.46</td>
<td>1</td>
<td>139.46</td>
<td>163.7</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>83.49</td>
<td>98</td>
<td>.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background</td>
<td>10.78</td>
<td>1</td>
<td>10.78</td>
<td>11.27</td>
<td>p &lt;.005</td>
</tr>
<tr>
<td>Residual</td>
<td>93.69</td>
<td>98</td>
<td>.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>686.92</td>
<td>3</td>
<td>228.97</td>
<td>156.77</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>429.41</td>
<td>294</td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trf x Back</td>
<td>.22</td>
<td>1</td>
<td>.22</td>
<td>.37</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>59.03</td>
<td>98</td>
<td>.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transf x Target</td>
<td>58.98</td>
<td>3</td>
<td>19.66</td>
<td>20.57</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>281.02</td>
<td>294</td>
<td>.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back x Target</td>
<td>107.94</td>
<td>3</td>
<td>35.98</td>
<td>35.94</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>294.31</td>
<td>294</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trf x B x Tar</td>
<td>.93</td>
<td>3</td>
<td>.31</td>
<td>.44</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>208.32</td>
<td>294</td>
<td>.71</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10.5: Summary of analysis of variance of response times (correct trials) on the 'target' task.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer</td>
<td>10508776.74</td>
<td>1</td>
<td>10508776.74</td>
<td>147.3</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>6991776.32</td>
<td>98</td>
<td>71344.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background</td>
<td>83476.64</td>
<td>1</td>
<td>83476.64</td>
<td>4.25</td>
<td>p &lt;.05</td>
</tr>
<tr>
<td>Residual</td>
<td>1926784.92</td>
<td>98</td>
<td>9819661.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>8135102.19</td>
<td>3</td>
<td>2711700.7</td>
<td>124.37</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>6410443.24</td>
<td>294</td>
<td>21804.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trf x Back</td>
<td>139237.50</td>
<td>1</td>
<td>139237.5</td>
<td>6.38</td>
<td>p &lt;.05</td>
</tr>
<tr>
<td>Residual</td>
<td>2138115.31</td>
<td>98</td>
<td>21817.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transf x Target</td>
<td>906058.6</td>
<td>3</td>
<td>302019.53</td>
<td>21.32</td>
<td>p &lt;&lt;.001</td>
</tr>
<tr>
<td>Residual</td>
<td>4165471.09</td>
<td>294</td>
<td>14168.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back x Target</td>
<td>3758347.24</td>
<td>3</td>
<td>12783.49</td>
<td>.84</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>32118.95</td>
<td>294</td>
<td>10706.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trf x B x Tar</td>
<td>33395.82</td>
<td>3</td>
<td>11131.94</td>
<td>1.00</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>3273726.12</td>
<td>294</td>
<td>11135.12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Means of the number of correct trials and response time of correct trials are shown in Table 10.6. Both sets of data indicate a strong learning effect since performance is better in the more difficult transfer task than in the initial task as shown by a higher mean accuracy and a lower mean response time in the transfer task. The background main effect is evidenced in a lower mean accuracy score and a larger mean reaction time for two-colour backgrounds as compared with eight colour ones. This is contrary to initial expectations since it can be argued that two colour backgrounds are simpler and less noisy than eight colour background. Kundel et al (1985) found that the level of random noise in an X-ray display has a negative effect on the detection of low contrast targets. However, it might be the case that two colour backgrounds yield poorer detection results because target discriminability in such backgrounds is lower than in eight colour backgrounds (Kundel and Revesz, 1980). Target conspicuity has been defined as the ratio of feature contrast to surround complexity (Revesz et al, 1974). Contrast between the target and surround strongly influences detection (Kundel and Revesz, 1980). In the 'target' task eight colour backgrounds provide more opportunity for target/background boundary contrast, since concentrations of the target colour contrast more with backgrounds consisting of target colour plus seven other shades than with backgrounds of target colour plus one other colour. The third main effect, the target effect, is borne out in a linear effect of target discriminability in trials with targets present. The higher the percentage of the target colour the more accurate and the faster they are classified. Trials with no targets present yielded an accuracy score better than 30% targets but worse than 40% targets. These displays produced the longest response times.

Table 10.6: Mean accuracy scores (n. corr) and response times (rt in msec) for task transfer, background and target effects. Better performance is indicated by a higher accuracy score and a lower response time.
10.3.2.2 Derivation of scores and structure underlying the 'target' task

It is possible to derive a large number of speed and accuracy scores from the 'target' task. The above analysis of task parameters indicates that performance in any of the eight conditions as well as in combinations thereof may uniquely contribute to individual differences in task performance. In the first instance it was, therefore, decided to compute a large number of performance measures for each individual measuring performance on all levels of the target, background and transfer factors. Descriptive statistics for most of these scores can be found in Table 10.3 and Table 10.6. Two additional scores were computed using the single and multiple target task scores which represent direct measures of learning. These are measures of difference between accuracy in the transfer task and the initial task (learn acc) and between the response time in the initial task and the transfer task (learn speed).

The structure underlying the scores derived from the target task was examined with a principal components analysis. Given the large number of intercorrelated scores derived from the task, the presence of substantial redundancy is very likely. Such multicollinearity does not present a problem in principal components analysis, however. Twenty-one variables were entered into the analysis. These included the total number of correctly classified trials (Tot corr), logs of the mean response time and its standard deviation (Tot rt and Tot sd), accuracy and response time measures of initial single target trials and transfer task multiple target trials (Sing acc, Sing rt, Mult acc, Mult rt), measures of learning (L acc, L rt), accuracy and speed scores for two and eight colour background trials (2col acc, 2col rt, 8col acc, 8col rt), and for each type of target (30% acc, 30% rt, 40% acc, 40% rt, 50% acc, 50% rt) as well as non-target trials (None acc, None rt).

In order to establish construct validity for the target task, a principal components analysis with oblique rotation was performed to examine the component structure underlying these scores. The analysis can also be used as the basis for reducing the large number of intercorrelated variables derived from the target task in terms of
factor scores. The use of principal components analysis in preference to factor analysis, in this context, mitigates the problem of factor indeterminacy (Schonneman and Wang, 1972). Four components with eigenvalues larger than one were extracted, but only two of these accounted for more than 10% of the variance. Examination of the scree plot supported the notion that a two factor solution was the most appropriate. The two factors accounted for 67.8% of the variance, and the two factor solution is shown in Table 10.7.

**Table 10.7:** Two factor solution for principal components analysis of scores derived from the 'target' task.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1: Speed</th>
<th>Factor 2: Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tot corr</td>
<td>-.11802</td>
<td>.88311</td>
</tr>
<tr>
<td>Tot rt</td>
<td>.91264</td>
<td>.04804</td>
</tr>
<tr>
<td>Tot sd</td>
<td>.82460</td>
<td>.04343</td>
</tr>
<tr>
<td>Sing acc</td>
<td>-.01587</td>
<td>.92707</td>
</tr>
<tr>
<td>Sing rt</td>
<td>.98037</td>
<td>.01299</td>
</tr>
<tr>
<td>Mult acc</td>
<td>-.00807</td>
<td>.74404</td>
</tr>
<tr>
<td>Mult rt</td>
<td>.71997</td>
<td>-.21737</td>
</tr>
<tr>
<td>L acc</td>
<td>.01131</td>
<td>-.45009</td>
</tr>
<tr>
<td>L rt</td>
<td>.86127</td>
<td>.11955</td>
</tr>
<tr>
<td>2col acc</td>
<td>-.16182</td>
<td>.78835</td>
</tr>
<tr>
<td>2col rt</td>
<td>.93355</td>
<td>-.08567</td>
</tr>
<tr>
<td>8col acc</td>
<td>.14192</td>
<td>.86291</td>
</tr>
<tr>
<td>8col rt</td>
<td>.94338</td>
<td>-.01346</td>
</tr>
<tr>
<td>30% acc</td>
<td>.22370</td>
<td>.43224</td>
</tr>
<tr>
<td>30% rt</td>
<td>.85603</td>
<td>.06265</td>
</tr>
<tr>
<td>40% acc</td>
<td>.03949</td>
<td>.76270</td>
</tr>
<tr>
<td>40% rt</td>
<td>.85866</td>
<td>-.09298</td>
</tr>
<tr>
<td>50% acc</td>
<td>-.18253</td>
<td>.72689</td>
</tr>
<tr>
<td>50% rt</td>
<td>.70096</td>
<td>-.34192</td>
</tr>
<tr>
<td>None acc</td>
<td>-.19493</td>
<td>.50140</td>
</tr>
<tr>
<td>None rt</td>
<td>.94663</td>
<td>.07857</td>
</tr>
</tbody>
</table>

249
It can be seen that the first factor is a speed factor. All measures involving response time have positive loadings of above .65 on this component. The second factor represents accuracy and is defined by high positive loadings from all accuracy measures and a negative loading on the measure of difference between initial and transfer task (L acc). An informal examination of the alternative three and four factor solutions revealed that they maintain the essential structure of the two factor solution with speed and accuracy components, but accuracy is spread out over two or three factors in a way which is not easily interpretable.

Table 10.8 shows the factor correlation matrix. It can be seen that the two factors are negatively intercorrelated. Since the two factors represent speed and accuracy respectively, this would indicate that there is a degree of speed-accuracy trade-off. The relationship is not statistically significant, though. Since speed and accuracy represent the two aspects which best summarise performance on the target task, factor scores for both components were computed for each individual. These were used as the two scores representing performance on the 'target' task in all further analysis.

Table 10.8: Factor correlation matrix for two components solution.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Factor 2</td>
<td>-.19</td>
<td>1.00</td>
</tr>
</tbody>
</table>

10.3.3 Validity of the contrast sensitivity and concept learning strategy scores

Both the Cambridge Low Contrast Gratings and the 'concept' task described by Johnson (1971) represent fully validated tests (Wilkins et al, 1988, Johnson, 1978) and, therefore, do not require separate validation. Contrast sensitivity scores can be directly compared to published norms. Strategy scores can be compared to those obtained by Johnson (1978). The purpose of this is to establish the comparability of the data collected as part of the present battery.
10.3.3.1 Comparability of contrast sensitivity scores

The Cambridge Low Contrast Gratings were scored according to the procedure described by Wilkins et al. (1988). The score is calculated by summing up the page numbers on which each of the four series was terminated and ranges between zero and forty-four (if no errors were made at all) for each eye. For the purpose of the current battery, a composite score for both eyes was computed by taking the mean performance of the two eyes (C.S.). Table 10.4 shows the age norms for mean contrast sensitivity for both eyes and mean contrast sensitivity for the poorer eye (Wilkins et al., 1988) in comparison with the mean scores obtained by subjects in the present battery.

Table 10.9: Age norms and mean contrast sensitivity scores obtained in the present sample for both eyes and the poorer eye.

<table>
<thead>
<tr>
<th>Age</th>
<th>n</th>
<th>mean and s.d. both eyes</th>
<th>mean and s.d. poorer eye</th>
<th>n</th>
<th>mean and s.d. both eyes</th>
<th>mean and s.d. poorer eye</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-19</td>
<td>13</td>
<td>30.38 5.47</td>
<td>27.54 5.92</td>
<td>64</td>
<td>31.15 6.1</td>
<td>28.09 6.91</td>
</tr>
<tr>
<td>20-29</td>
<td>25</td>
<td>35.08 5.06</td>
<td>33.76 5.12</td>
<td>24</td>
<td>30.60 6.7</td>
<td>27.04 7.13</td>
</tr>
<tr>
<td>30-39</td>
<td>20</td>
<td>32.90 3.09</td>
<td>31.25 3.10</td>
<td>6</td>
<td>31.08 5.7</td>
<td>27.5 2.02</td>
</tr>
<tr>
<td>40-49</td>
<td>13</td>
<td>30.46 3.79</td>
<td>28.54 4.75</td>
<td>2</td>
<td>39.25 2.47</td>
<td>37.5 2.12</td>
</tr>
</tbody>
</table>

It can be seen that the present means and standard deviations are similar to the norms. Exceptions are a slightly lower mean score in the 20 to 29-year-olds and a much higher mean in the over 40's. The latter may be a result of the very small sample representing this age group in the present battery. The differences are not significant though.

10.3.3.2 Validity of strategy scores

The abbreviated 'concept' task was designed and administered according to the procedures described by Johnson (1971, 1978) with the amendments described above. Subjects completed five runs of the task, the first of which was a practice
run and was not included in the scoring. All trials in the other four runs were computer-scored as representing one of four possible strategies, tactical, focusing, scanning and blundering. The scoring program was developed to implement the strategy classification rules outlined by Johnson (1978, p. 266). According to Johnson's procedures all trials were scored until at least two elements of the final solution had been determined. The remaining trials were not classified as belonging to a particular strategy. Johnson also outlined rules for classifying a particular run as belonging predominantly to one of the four strategies and for classifying individuals as users of a particular strategy. In the present application of the task, subjects were each assigned a score for each of the four strategies based on the proportion of trials classified as belonging to one of the four strategies. However, for the purpose of validation it was necessary to categorise subjects temporarily in terms of strategy use.

One of Johnson's (1978) internal validation procedures involved demonstrating that there was a relationship between strategy classification and other task variables which were independent of the classifications. One such variable was the number of trials taken to reach a solution. Johnson argued that discontinuities in the distribution of this variable would provide evidence that different strategies existed and affected performance differentially. Table 10.10 shows the mean number of trials to solution for the different strategy groups for Johnson's procedure compared with those obtained in the present abbreviated implementation of the task. None of the participants in the present research received an overall classification as a tactician. The remaining strategy groups yielded mean trials to solution of about half the magnitude of those found by Johnson. This is a result of reducing the letter sequence from six elements to five which limited the number of potential solutions to 20. This change to the procedure was made to make the task less difficult and faster to implement as part of the cognitive battery. The trials to solution data indicate that the reduction in sequence elements by one effectively halved the effort required to find the solution. This is within the range of expectation, since the changes reduced the number of possible solutions by one third.
Table 10.10: Trials to solution for different strategy groups in Johnson's Zaps and Duds task compared to those obtained in the abbreviated 'concept' task.

<table>
<thead>
<tr>
<th></th>
<th>Zaps and Duds task (Johnson, 1978)</th>
<th>'Concept' task (present battery)</th>
<th>n</th>
<th>trials to solution</th>
<th>n</th>
<th>trials to solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacticians</td>
<td>38</td>
<td>10.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Focusers</td>
<td>51</td>
<td>12.1</td>
<td>65</td>
<td>6.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scanners</td>
<td>37</td>
<td>20.8</td>
<td>30</td>
<td>10.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blunderers</td>
<td>4</td>
<td>26.4</td>
<td>1</td>
<td>9.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to Johnson, only those trials that occur before two solution elements are known should be scored since this renders the solution determinable within the next one or two trials. This results in relatively few cases being classified as blundering, four out of 130 in Johnson's data set and one out of 98 in the present set. However, inspection of the task protocols revealed that a significant number of subjects try redundant patterns or make invalid hypotheses after two elements of the solution have been determined. These trials represent blundering trials which would not normally be classified in Johnson's scoring procedure. Subjects are clearly not able to draw the correct conclusions from the evidence before them. The extent to which such blundering trials occur is noteworthy. Sixty-six subjects produced at least one post-scoring blundering trial. Whilst it may be possible to regard a small number of such trials as temporary slips in attention, there is evidence in the present data set that a considerable proportion of subjects classified as focusers produced more than ten such blunders, i.e. substantially more than the average number of trials classified as focusing. Descriptive data for post-solution blunderers are shown in Table 10.11.

Table 10.11: Prevalence of blundering after the determination of the solution in the 'concept' task

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>mean no. of blunders</th>
<th>standard deviation</th>
<th>range of number of blunders</th>
</tr>
</thead>
<tbody>
<tr>
<td>post-solution blunderers (more than ten blunders)</td>
<td>13</td>
<td>28.61</td>
<td>17.72</td>
<td>14 to 83</td>
</tr>
</tbody>
</table>

253
Laughlin (1978) has characterised blundering subjects as those subjects for whom Johnson's task proved too difficult. Blundering thus does not represent a type of strategy as such but a measure of efficiency on the concept task. In this respect the post-solution blunderers are of as much interest in the present research as the other strategy types. For this reason task protocols were scored so that they yielded not only scores on the three main strategies, tactical (Strat-T), focusing (Strat-F) and scanning (Strat-S), but also included two blundering scores, one representing blunders before the solution was determined (Strat-BB) and one after at least two elements of the solution had been determined (Strat-BA). The two blundering score represent subjects' efficiency in obtaining information which leads to the solution and efficient use of such information, respectively. High blundering scores represent lower efficiency.

10.3.4 Construct validity of the cognitive task battery

Tasks included in the cognitive test battery were thought to be predominantly but not uniquely concerned with measuring specific aspects of mammogram interpretation performance. In particular, the mammogram interpretation model (Figure 7.1) predicts that there would be some overlap between the contrast sensitivity measure and the grid task, since both are concerned with basic visual skills implemented in the first stage of the model. In addition, some overlap between the 'grid' and the 'target' task was anticipated as both tasks appear to involve target detection, target analysis and decision-making elements, the 'target' task presenting the more complex case. Table 10.12 shows the correlation matrix of the ten variables derived from the cognitive battery.
Table 10.12: Correlations between measures derived from the cognitive test battery.

<table>
<thead>
<tr>
<th></th>
<th>C.S.</th>
<th>G-error</th>
<th>G-corr</th>
<th>T-speed</th>
<th>T-acc</th>
<th>Strat-T</th>
<th>Strat-F</th>
<th>Strat-S</th>
<th>Strat-BA</th>
<th>Strat-BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.S.</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-error</td>
<td>-.13</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-corr</td>
<td>-.07</td>
<td>.07</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-speed</td>
<td>-.00</td>
<td>-.06</td>
<td>.10</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-acc</td>
<td>.05</td>
<td>-.06</td>
<td>.23*</td>
<td>.19</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strat-T</td>
<td>.07</td>
<td>.18</td>
<td>.11</td>
<td>-.01</td>
<td>.11</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strat-F</td>
<td>-.04</td>
<td>-.08</td>
<td>.19</td>
<td>.00</td>
<td>-.05</td>
<td>.08</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strat-S</td>
<td>.02</td>
<td>.03</td>
<td>-.14</td>
<td>.03</td>
<td>.14</td>
<td>-.25*</td>
<td>-.83**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strat-BB</td>
<td>.10</td>
<td>.05</td>
<td>-.08</td>
<td>.05</td>
<td>.05</td>
<td>.06</td>
<td>-.38**</td>
<td>.08</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Strat-BA</td>
<td>-.03</td>
<td>.00</td>
<td>-.08</td>
<td>-.10</td>
<td>-.27*</td>
<td>.00</td>
<td>-.12</td>
<td>-.37**</td>
<td>.11</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* significant at p < .01  
** significant at p < .001

It can be seen that there are relatively few significant correlations between the cognitive variables. The correlation between G-corr, the number of dots correctly identified on the X-ray grid, and T-acc, the factor score derived from the target accuracy factor, confirms the hypothesis that there is some overlap between the 'grid' and 'target' tasks. The size of the correlation is relatively small, though, indicating that the two tasks are measuring different aspects of target detection. There are no correlations between contrast sensitivity and any of the other variables. Performance on the 'grid' task or the 'target' task, thus, does not appear to be related to visual sensitivity in a direct manner. However, there is a significant negative relationship between T-acc and Strat-BA, the blundering score after determination of the solution in the 'concept' task. This indicates that the efficiency with which subjects are able to use available information (a low Strat-BA score) is related to how accurate they are at deciding whether a target is present or not in the 'target' task.
All other significant correlations are between the strategy scores derived from the 'concept' task. Most notably, there is a very high negative correlation between the focusing and scanning strategy scores. This indicates that people who predominantly follow a focusing strategy yield low scanning scores and vice versa. The high correlation between the two strategy scores is evidence for an ipsative relationship between the two variables, as might be expected given that Johnson's task measures information processing style. It should be noted, however, that this ipsative relationship is not an artefact of scoring, but is inherent in the actual way in which strategies are employed by subjects. The negative relationship between tactical scores and scanning can be explained in the same manner. The observed correlation is lower than that between focusing and scanning, which is probably due to the lesser prevalence of the tactical strategy. Focusing and scanning also yield significant negative correlations with blundering. Focusing is associated with efficiency in information acquisition (low Strat-BB) while scanning correlates with efficient use of available information (low Strat-BA).

10.3.4.1 Principal components analysis

Principal components analysis was conducted to examine the component structure of the test battery. Components were extracted by principal components extraction and oblique rotation. Five factors accounted for more than 10% of the variance each. The five factor solution, which accounted for 71.7% of the variance, is shown in Table 10.13. Interpretation of variables with a loading of more than .4 reveals that the first component reflects strategy choice. Both style and efficiency are captured by this component as is borne out by very high loadings of focusing and scanning scores and a moderate loading on the blundering score. The ipsative nature of the two main strategy types is reflected in this factor as focusing loads negatively and scanning loads positively. Blundering is also positive, indicating that as far as efficiency is concerned the focusing strategy fares better. The second factor has high positive loadings on three variables, the two 'target' variables and G-corr. This component can be interpreted as a 'target detection' component as it combines all variables which directly measure visual search performance.
Table 10.13: Five factor solution of principal components analysis of all cognitive tasks. Highest loadings for each variable are printed in bold.

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>-.0053</td>
<td>-.1576</td>
<td>-.0366</td>
<td>.0102</td>
<td>-.9090</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-error</td>
<td>.1672</td>
<td>-.2045</td>
<td>-.0493</td>
<td>-.7101</td>
<td>.4044</td>
</tr>
<tr>
<td>G-corr</td>
<td>-.2707</td>
<td>.6230</td>
<td>-.0662</td>
<td>-.0762</td>
<td>.1279</td>
</tr>
<tr>
<td>T-speed</td>
<td>.0736</td>
<td>.6874</td>
<td>.1212</td>
<td>.1783</td>
<td>.0878</td>
</tr>
<tr>
<td>T-acc</td>
<td>.1562</td>
<td>.6285</td>
<td>-.2883</td>
<td>-.1653</td>
<td>-.1717</td>
</tr>
<tr>
<td>Strat-T</td>
<td>-.2209</td>
<td>.1013</td>
<td>.0571</td>
<td>-.7776</td>
<td>-.2182</td>
</tr>
<tr>
<td>Strat-F</td>
<td>-.9278</td>
<td>.0373</td>
<td>-.1910</td>
<td>.0090</td>
<td>-.0123</td>
</tr>
<tr>
<td>Strat-S</td>
<td>.9117</td>
<td>-.0233</td>
<td>-.3477</td>
<td>.1252</td>
<td>.0387</td>
</tr>
<tr>
<td>Strat-BA</td>
<td>-.1629</td>
<td>-.1535</td>
<td>.8345</td>
<td>.0641</td>
<td>-.1956</td>
</tr>
</tbody>
</table>

The third component is defined by positive loadings of both blundering scores. This factor thus appears to reflect an inefficiency in gathering and applying available information. The fourth factor has high negative loadings on tactical strategy scores and G-error, which is a measure of confidence with which errors are made on the 'grid' task. This component results directly from the positive correlation between the two variables (Table 10.13). The indication is that readiness to make mistakes on the perceptual 'grid' task is related to the use of a deep processing strategy when acquiring information. The component is probably best explained in terms of the degree of confidence a subject is prepared to invest in the strategy they choose to perform a particular task. The final factor is characterised by a very high negative loading of contrast sensitivity which is combined with a moderate positive loading on G-error. Whilst the previous component indicated a strategic link with the degree of confidence with which perceptual decisions are made, the present factor establishes the link with basic visual processes. Visual sensitivity thus appears to play a role in complex target detection as represented by the 'grid' task, but it operates in terms of the confidence with which subjects get it wrong rather than the degree to which their decisions are accurate.
10.3.5 Concurrent validity of the batteries: Relationships between cognitive tasks and psychometric tests

The relationship between cognitive tasks and psychometric tests is of particular interest in terms of their respective roles as predictor variables for mammogram interpretation, because it can establish a rudimentary form of concurrent validity for the two batteries. Some of the cognitive tasks were included in the cognitive battery because of their theoretical links with the psychometric variables under investigation. In order to investigate whether these relationships are confirmed in the present study, the correlations between the individual variables on the two test batteries can be examined. Table 10.14 shows that there are significant correlations between a large number of psychometric variables and the two blundering scores. In addition, focusing correlates with all psychometric ability measures as well as field-independence and introversion. The perceptual tasks correlate less frequently with the psychometric predictors, but significant relationships can be observed between G-error and the KAI, reconfirming the strategy connection of the former, and between spatial ability and G-corr. Sizable but not significant associations between field-independence and 'target' speed and Psychoticism and 'target' accuracy were also present.

Table 10.14: Correlations between cognitive tasks and psychometric tests.

<table>
<thead>
<tr>
<th></th>
<th>C.S.</th>
<th>G-error</th>
<th>G-corr</th>
<th>T-speed</th>
<th>T-acc</th>
<th>Strat-T</th>
<th>Strat-F</th>
<th>Strat-S</th>
<th>Strat-BB</th>
<th>Strat-BA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEFT</td>
<td>.06</td>
<td>-.08</td>
<td>.03</td>
<td>-.16</td>
<td>.05</td>
<td>-.01</td>
<td>.36**</td>
<td>-.21</td>
<td>-.30*</td>
<td>-.14</td>
</tr>
<tr>
<td>SAT</td>
<td>.06</td>
<td>.01</td>
<td>.23*</td>
<td>-.08</td>
<td>.10</td>
<td>-.03</td>
<td>.28*</td>
<td>-.16</td>
<td>-.24*</td>
<td>-.12</td>
</tr>
<tr>
<td>MD5</td>
<td>-.05</td>
<td>-.01</td>
<td>.08</td>
<td>-.02</td>
<td>.12</td>
<td>.00</td>
<td>.27*</td>
<td>-.09</td>
<td>-.23*</td>
<td>-.26*</td>
</tr>
<tr>
<td>KAI</td>
<td>-.04</td>
<td>.24*</td>
<td>-.02</td>
<td>-.03</td>
<td>.13</td>
<td>.06</td>
<td>.04</td>
<td>.11</td>
<td>-.28*</td>
<td>-.23*</td>
</tr>
<tr>
<td>P</td>
<td>.02</td>
<td>-.14</td>
<td>.03</td>
<td>-.02</td>
<td>.18</td>
<td>.09</td>
<td>.15</td>
<td>-.01</td>
<td>-.27*</td>
<td>-.20</td>
</tr>
<tr>
<td>E</td>
<td>.14</td>
<td>.19</td>
<td>-.18</td>
<td>-.09</td>
<td>-.06</td>
<td>.14</td>
<td>.31**</td>
<td>.20</td>
<td>.14</td>
<td>.10</td>
</tr>
<tr>
<td>N</td>
<td>.10</td>
<td>.09</td>
<td>.04</td>
<td>-.02</td>
<td>.08</td>
<td>-.11</td>
<td>-.04</td>
<td>.00</td>
<td>.16</td>
<td>.05</td>
</tr>
</tbody>
</table>

* significant at p < .01
** significant at p< .001
10.3.5.1 Canonical correlation

In order to investigate the relationship between cognitive tasks and psychometric predictors further, a canonical correlation was performed between the two sets of variables. All cognitive variables discussed in the preceding pages were included, whilst all psychometric tests were represented by total scores on each test. The analysis was conducted using the PAP analysis package (Hammond, 1990). The procedure identified seven canonical variates. The first canonical correlation was .66 and accounted for 44% of the variance ($\chi^2 = 232.64$, $df = 70$, $p < .0001$). The second canonical correlation was .47 ($\chi^2 = 116.72$, $df = 54$, $p < .0001$), accounting for a further 22% of the variance. The next two canonical correlations accounted for 12% and 10% of the remaining variance and represented correlations of .36 ($\chi^2 = 67.33$, $df = 40$, $p < .005$) and .33 ($\chi^2 = 39.24$, $df = 28$, $p = .07$). Although the last canonical correlation failed to reach statistical significance at the 5% level, it is considered in the present interpretation because it accounts for some interesting relationships. The remaining three canonical correlations yielded $\chi^2$ tests involving much higher chance levels and were not included. Table 10.15 shows the item-function correlations for the first four canonical variates. Using a cut-off correlation of .4 for inclusion in the interpretation (Hammond, 1990), it can be seen that the variables in the cognitive set that were correlated with the first canonical variate were focusing (.66) and the two blundering scores (-.73 for Strat-BB and -.54 for Strat-BA). Among the psychometric variables there were high positive correlations for the Group Embedded Figures Test (.60), the MD5 Ability Test (.57), the Shapes Analysis Test (.53), the Kirton Adaptation-Innovation Inventory (.48), and Psychoticism (.47). This indicates that subjects who prefer a focusing strategy and who are efficient in both the acquisition and use of information with this strategy, perform better on both spatial and general ability tests. They also tend to be field-independent innovators with high scores on psychoticism. The second canonical correlation combined G-error (-.83) with two psychometric variables, extraversion (-.59) and KAI (-.54). The negative correlations indicate that subjects who are reluctant to make mistakes on the 'grid' task (low G-error) tend to be introverts who prefer to adopt an adaptive problem solving style. The third canonical variate was
Table 10.15: Item-function correlations between cognitive task and psychometric variables for the first four canonical variates.

<table>
<thead>
<tr>
<th>cognitive tasks</th>
<th>first canonical correlation</th>
<th>second canonical correlation</th>
<th>third canonical correlation</th>
<th>fourth canonical correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast S.</td>
<td>-.08</td>
<td>-.23</td>
<td>-.01</td>
<td>.71</td>
</tr>
<tr>
<td>G-error</td>
<td>.04</td>
<td>-.83</td>
<td>.09</td>
<td>-.33</td>
</tr>
<tr>
<td>G-corr</td>
<td>.22</td>
<td>.10</td>
<td>.52</td>
<td>.03</td>
</tr>
<tr>
<td>T-speed</td>
<td>-.10</td>
<td>.18</td>
<td>.05</td>
<td>-.32</td>
</tr>
<tr>
<td>T-acc</td>
<td>.30</td>
<td>-.15</td>
<td>-.02</td>
<td>.10</td>
</tr>
<tr>
<td>Strat-T</td>
<td>-.01</td>
<td>-.04</td>
<td>-.49</td>
<td>.71</td>
</tr>
<tr>
<td>Strat-F</td>
<td>.66</td>
<td>.37</td>
<td>.31</td>
<td>.08</td>
</tr>
<tr>
<td>Strat-S</td>
<td>-.54</td>
<td>-.37</td>
<td>-.37</td>
<td>-.24</td>
</tr>
<tr>
<td>Strat-BB</td>
<td>-.73</td>
<td>-.07</td>
<td>.22</td>
<td>-.00</td>
</tr>
<tr>
<td>Strat-BA</td>
<td>-.54</td>
<td>.07</td>
<td>.21</td>
<td>.27</td>
</tr>
<tr>
<td>psychometric tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD5</td>
<td>.57</td>
<td>.09</td>
<td>-.00</td>
<td>-.08</td>
</tr>
<tr>
<td>SAT</td>
<td>.53</td>
<td>-.04</td>
<td>.47</td>
<td>.36</td>
</tr>
<tr>
<td>GEFT</td>
<td>.60</td>
<td>.16</td>
<td>.04</td>
<td>.51</td>
</tr>
<tr>
<td>KAI</td>
<td>.48</td>
<td>-.54</td>
<td>.44</td>
<td>-.29</td>
</tr>
<tr>
<td>P</td>
<td>.47</td>
<td>.20</td>
<td>-.51</td>
<td>.26</td>
</tr>
<tr>
<td>E</td>
<td>-.37</td>
<td>-.59</td>
<td>-.57</td>
<td>.23</td>
</tr>
<tr>
<td>N</td>
<td>-.11</td>
<td>-.31</td>
<td>.58</td>
<td>.22</td>
</tr>
</tbody>
</table>

composed of the accuracy score on the 'grid' task and the tactical strategy score on the cognitive tasks, and Psychoticism, Extraversion, Neuroticism, spatial ability and Adaption-Innovation on the psychometric tests. The indication is that those individuals who yield a high accuracy score on the 'grid' task (.52) and who are disinclined to use a tactical information processing strategy (-.49) tend to yield high neuroticism scores (.58), are introvert (-.57 with extraversion) and have a high degree of impulse control (-.51 with psychoticism). In addition, these subjects yield high spatial ability scores (.47) and tend to adopt an adaptive decision making style (-.44). The last canonical correlation combines speed on the 'target' task (-.52) and contrast sensitivity (.71) with field independence (.51). Although this canonical
correlation was not statistically significant \( (p = .07) \) and therefore does not account for a significant relationship between the variables, it can be seen that it gives an indication of a direct perceptually based link between the cognitive and psychometric variables.

10.4 Discussion

In this chapter, the aim was to discuss the development of cognitive tasks which were representative of aspects of mammography performance. Four tasks were introduced each of which was designed to measure specific but not unique aspects of information processing required for mammogram interpretation. Attempts were made to establish each of the tasks as valid and reliable by examining relevant task parameters in the case of newly developed tasks or by comparing performance data obtained in the current study with previously published data. The main concern in this chapter was to examine the factor structure of skills underlying performance on all cognitive tasks as well as to establish in how far the cognitive tasks are related to performance on the psychometric tests discussed in the previous chapter.

10.4.1 Validity of cognitive task measures

The purpose of developing cognitive tasks as part of this research was to examine the extent to which measures derived from such tasks could predict performance in mammogram interpretation. It is therefore clear that the most important criterion in establishing validity of the tasks is in terms of their predictive power. This is an issue which will be dealt with in detail in the next chapter. The analysis conducted in the present chapter was concerned with the internal validity of task measures in terms of task parameters or previously published performance data.

10.4.1.1 Validity of 'grid' and 'target' measures

The test development studies for the 'grid' task showed that the perceptibility of dots on the X-ray was influenced not only by contrast as is evident from the
significant main effect regarding depth of holes, but also showed that hole diameter played a significant role in determining the perceptibility of low contrast dots. This is important in determining the hole size for the task that would produce maximal individual differentiation. Descriptive analysis of the two measures derived from the task (Table 10.2) showed that there was a good range of scores on both measures. The two measures are complementary in assessing two major aspects of the task, contrast sensitivity in complex visual search (G-corr) and preparedness to make errors as measured by the confidence with which erroneous judgements are made (G-error).

The 'target' task was shown to be reliable in terms of yielding expected results in relation to specific task parameters. Thus target discriminability and conspicuity in relation to background features were found to have a significant effect on the extent to which targets were accurately identified and the speed with which decisions were made. Principal component analysis of a large number of scores derived from the 'target' task revealed that it was possible to reduce all potential measures to two underlying parameters, accuracy and response time. Although it would appear obvious that there should be these two underlying components, since accuracy and speed scores were the two types of scores derived from the task, the unitary nature of each of these components is by no means given. Thus separate components might have been found for other aspects of the task such as a dichotomy between scores on target versus non-target trials, or for target conspicuity. The lack of evidence for such further divisions permitted the adoption of two factor scores derived from the target task which represented the two interlinked aspects of performance on the target task, speed and accuracy. The trade-off between the two scores was not significant, though.

10.4.1.2 Validity of contrast sensitivity and strategy scores
Reliability for scores on the Cambridge Low Contrast Gratings was established by comparison with the published age norms (Wilkins et al, 1988). Mean scores for both overall performance and poorer eye performance were comparable to those
published as age norms. The contrast sensitivity scores obtained in the present battery can thus be considered reliable and valid measures of true contrast sensitivity.

Reliability of the concept learning strategy scores was assessed by comparing data from the present battery to internal validation data presented by Johnson (1978). The data presented in Table 10.10 show that, on the whole, the results obtained in the present study are comparable to those reported by Johnson (1978), when taking into account the changes that had been made to the task for the present purpose. One notable difference between Johnson's data and the present data was that none of the subjects received an overall tactical classification. The tactical strategy has been described as a form of 'simultaneous scanning', an ideal strategy, which is rarely employed in practice because it requires too much mental effort (Bruner et al, 1956). Johnson explained the occurrence of this strategy in terms of procedural variables which reduced cognitive strain imposed by his task as compared with Bruner's task. These included the emphasis on minimizing of the number of trials in the instructions to subjects and a lack of time pressure in the administration procedure. In the current implementation the task was administered as part of a battery which resulted in increased time pressures. Subjects were also told that the task was timed. The lack of tactical solutions in the present task may, therefore, be a result of slight procedural variations resulting in increased cognitive strain which in turn prevented the use of this ideal strategy.

10.4.2 Overlap between cognitive tasks

Principal components analysis of the cognitive task battery revealed that there were overlaps between the four tasks which were largely in accordance with expectations. It had been argued that there might be overlaps between the 'grid' and 'target' tasks, since both involved target detection and analysis. The correlation between G-corr and T-acc (Table 10.12), which is reaffirmed by the second factor in the principal components analysis (Table 10.13), is a clear indication of
communality between the two tasks. Similarly, it was anticipated that there might be some overlap between basic visual mechanisms and target detection. The fifth factor indicates that performance in the 'grid' task is indeed related to visual acuity, although the association operates in terms of the degree of confidence observers have in their decisions rather than actual detection performance. As far as the 'concept' task is concerned there is clear evidence for two factors being involved in performance. The first is about information processing style (Factor 1 in Table 10.13) and results from the ipsative relationship between focusing and scanning strategies. If Laughlin's (1987) interpretation of blundering as a lack of efficiency on the task is accepted, the second factor associated with strategies can clearly be explained as an information processing efficiency factor, which summarises the two blundering scores (Factor 3 in Table 10.13). There is also some evidence for overlap between perceptual tasks and strategy as is borne out in the fourth factor (Table 10.13). This factor combines the tactical decision making strategy with preparedness to make errors in the 'grid' task indicating that performance on the perceptual tasks is actually susceptible to being influenced by the choice of a high risk strategy.

On the whole, it can be said that although there is a degree of overlap between the tasks there is no substantial redundancy amongst the variables. The correlations between the variables are quite low, the only exception being the focusing/scanning correlation which is due to the mutual exclusiveness of the two scores, as already discussed. In addition, a relatively large number of components was required to adequately account for the ten variables derived from the battery. This indicates that there is minimal redundancy in the cognitive task battery, and each of the variables contributes uniquely to measuring performance on the tasks. It can therefore be concluded that each of the variables derived from the cognitive battery can be regarded as a potentially important predictor of mammogram interpretation performance in its own right.
10.4.3 Overlap between cognitive and psychometric measures

Since both, psychometric tests and cognitive tasks, were selected as possible predictors for mammogram interpretation performance, it can be expected that there is some communality between the two batteries. In particular, theoretical links have been established between information processing strategy and field-independence (Johnson, 1978). The correlation matrix (Table 10.14) showed there was a significant association between focusing and the GEFT lending support to Johnson's argument. However, focusing also correlated significantly with the two ability measures, albeit to a lesser extent.

The question to what extent cognitive styles such as field-dependence and information processing strategy are actually independent of ability thus remains open. It can be said that the focusing information processing style is not entirely explicable in terms of ability, since the correlation with field-independence was much higher, and there was a significant association with introversion. In addition, there were no similar negative correlations with the scanning style which would be expected, if ability was the main factor in producing high focusing scores. The pattern of correlations between strategic efficiency, the two blundering scores, and ability and personality measures is more likely to reflect general ability, though. Blundering is not a style in its own right but reflects the inability to adopt a task-appropriate strategy. This was reflected in negative correlations with the ability measures and field-independence.

The more perceptually based cognitive tasks yielded fewer associations with psychometric variables (Table 10.14). Exceptions are significant association between spatial ability and the correct identification of dots on the 'grid' task and innovative style and confidence on error trials in the 'grid' task. The former indicates that relatively low-level visual processes in radiographic interpretation is associated with general spatial ability. The correlation between the other two variables is more indicative of strategy in that innovators appear to be more
prepared to adopt a high-risk strategy in an uncertain situation. Some evidence for
the predicted overlap between target detection and field-dependence can be derived
from the association between the speed variable derived from the target task and
the GEFT. There were no significant correlations between target detection variables
and field-dependence, however.

The canonical correlation (Table 10.15) largely supports the picture presented by
the patterns of correlation between psychometric and cognitive variables. The first
canonical correlation reaffirms the association between information processing
strategy and cognitive style on the one hand and ability on the other. Both style
and efficiency of information processing are addressed by this correlation, giving
the overall impression that those subjects who are able to use a focusing strategy
efficiently, are highly able field-independent innovators who also have high
psychoticism scores. The second canonical correlation addresses style from a
different angle. Introversion and adaptive problem solving style are associated with
a cautious approach to making perceptual decisions on the grid task. The overall
interpretation of this correlation is that it represents subjects, who prefer a low-risk
approach. The third canonical correlation links target detection performance with
spatial ability and personality. Anxious subjects (low on psychoticism, high on
introversion and neuroticism, adaptors) with high spatial ability appear to perform
well on the 'grid' task. These subjects prefer not to adopt high risk tactical
information processing strategies. The final canonical correlation combines
perceptual variables from the cognitive battery with field-independence, reaffirming
the position of the GEFT as a measure of perceptual style. Field-independence is
associated with better contrast sensitivity and faster response times in target
analysis.

In summary, the overlap between psychometric and cognitive variables appears to
be more about information processing style and strategy than about specific
perceptual abilities. The only canonical variate to include predominantly perceptual
variables in both batteries was the last one which does not, in fact, account for a
10.4.4 Representativeness of derived measures of cognitive processes in mammogram interpretation

On a theoretical level, the main purpose of developing the cognitive tasks was to establish measures representative of cognitive processes involved in mammogram interpretation, which were formally identified in the task model in chapter 7 (Figure 7.1). On the face of it, the 'grid' task was designed to measure basic visual processes, the 'target' task was concerned primarily with target analysis and binary categorisation, and the 'concept' task involved the identification of information processing strategies in complex concept learning. These three tasks thus broadly match the main information processing components involved in mammogram interpretation as modelled in chapter 7. In order to investigate individual differences at the level of cognitive processes, it is necessary to derive measures from the task representative of such processes. This is particularly true for the 'grid' and 'target' tasks, since the 'concept' task was concerned with identifying strategies rather than processes, and the Cambridge Low Contrast gratings presented a simple visual sensitivity measure, which was included to control for sensitivity effects in the more complex visual processing tasks.

The model of mammogram interpretation in chapter 7 assumed that different processes were involved in visual processing of mammograms on the one hand, and in diagnostic evaluation of target abnormality on the other. The former involves the detection and recognition of abnormal features, whilst the latter is concerned with the interpretative evaluation of these features in terms of prior knowledge of breast anatomy, pathology and radiological features of malignancy. The 'grid' task was specifically developed to represent processes in basic visual analysis, whereas the
'target' task primarily simulated target analysis aspects of the task. The two tasks thus appear to fit well in terms of representing cognitive processes. However, an examination of the measures derived from the two tasks shows that they do not directly measure any hypothesized cognitive processes underlying mammogram interpretation, but represent fairly global measures. Thus, the accuracy measure derived from the 'grid' task is a global measure representing performance in visual processing of simple targets rather than representing detection or recognition processes. Similarly, the measures derived from the 'target' task are global accuracy and speed performance measures.

In terms of investigating individual differences at the cognitive process level, the measures derived from the cognitive tasks are thus not very useful. One of the problems may be that the tasks, and particularly the 'target' task, are too macroscopic. It might have been possible to develop tasks which were better able to distinguish between visual and interpretative processes. On the other hand, it may be the case that it is simply not possible to develop a task which closely models the processes involved in mammogram interpretation and simultaneously distinguishes between cognitive processes. The indication is that mammogram interpretation, at least in the context of lay film reading, may be inseparably dependent on visual processes. Such processes are more easily examined experimentally as in Kundel et al's (1978) work on error analysis based on eye movement data.

On an applied level, these theoretical considerations do not, however, invalidate the cognitive task approach. The principal purpose of developing cognitive tasks was to achieve a higher degree of predictive value in the context of selection. The idea was that on the whole the direct measurement of cognitive processes involved in a task are more likely to predict performance in that task than more traditional psychometric measures based on normative measurement (Carroll, 1988, Lesgold et al, 1990). Failure to identify specific cognitive processes in term of task parameters does not preclude the predictive potential of the global task measures which
represent more complex clusters of cognitive processes involved in mammogram interpretation. The macroscopic nature of the tasks, in particular of the target task, may even result in better prediction than more component orientated tasks, because it is more representative of mammogram interpretation as a whole. The potential for prediction on the basis of cognitive tasks is investigated further in the next chapter, in which the cognitive measures derived from the present battery are used psychometrically along side the ability and personality measures discussed in the preceding chapter.

10.4.5 Summary

The purpose of the present chapter has been to report on the development of a battery of cognitive tasks based on an analysis of mammogram interpretation skill (chapter 7). The internal validity of individual measures was established and the structure of the cognitive task battery examined. It was found that, although the measures derived from the battery were too global to be representative of specific cognitive processes in mammogram interpretation, they all represent potentially valuable predictors of mammogram interpretation. Their relationship with psychometric predictors was largely limited to overlap on the strategic level, indicating that different aspects of mammogram interpretation are covered in the two batteries. The predictive power of both, psychometric and cognitive variables can only be established in relation to actual mammogram interpretation performance, though. This is attempted on an exploratory level in the following chapter. Judgement as to the actual value of psychometric tests and cognitive tasks as predictors of film reading performance must, therefore, be reserved for the next chapter.
Chapter 11
Predicting mammogram interpretation performance

11.1 Introduction

The last two chapters have been concerned with the development and validation of two sets of variables which were hypothesized to have utility in predicting film reading performance in mammography. The first set of variables was derived from psychometric paper-and-pencil tests. The tests were selected on the basis of theoretical relationships between ability, cognitive style and personality and performance in mammography. The second set of predictor variables originated from cognitive tasks which had been designed to measure the cognitive skills involved in mammogram interpretation. These variables were more specifically based on an analysis of the skills inherent in X-ray film reading performance. The task model, which was presented in chapter 7, formed the basis for the development of these tasks, which are representative of hypothesized cognitive processes in film reading.

In this chapter, the issue of the predictive value of both psychometric and cognitive variables is addressed. In chapter 9, the possibility of developing a psychometric battery of paper and pencil tests which was capable of measuring 'aptitude for screening film reading' was investigated. It was shown that there was no evidence for a unitary perceptually-based aptitude domain amongst the particular set of psychometric tests employed. Nevertheless, there is a strong possibility that individual psychometrically derived variables may be good predictors of mammogram interpretation (Walker et al, 1989). It is likely, however, that the cognitive variables, which are more directly related to film reading, have more utility for predicting mammogram interpretation (Lesgold et al, 1990).

In order to assess the predictive value of the psychometric and cognitive variables in a lay-screening situation, it is possible to use performance in the training studies
reported in chapters 6 and 7 as criteria. Three issues involving prediction of performance are addressed on an exploratory level. The first question is in how far the two sets of predictor variables are predictive of mammogram interpretation performance after training. A second concern is to determine whether it is possible to identify those individuals who are likely to benefit most from training, i.e. people whose performance improves as a function of training. Thirdly, it may be possible to identify those variables which are associated with either good or bad overall performance after training, thus permitting the diagnosis of particular individual weaknesses by identifying poor performance on specific aspects of the skill involved. The three approaches to performance prediction are explored in more detail below.

11.1.1 Selection of film readers for breast cancer screening

Although, at present, the breast screening service in Britain is to a large extent staffed by consultant radiologists who are self-selected, it is clear that there is scope for the employment of lay people to assist with routine film reading duties. There is some evidence that in a number of screening centres, non-specialist doctors and some medico-technical staff are already involved in film interpretation procedures (see chapter 3). The case for lay film reading in NHS breast screening has been made by Saxton (1992), who also emphasized the need for careful selection of individuals for training.

In order to establish proper selection procedures, it is necessary to identify those predictor variables that are unequivocally associated with success or failure in mammogram interpretation. In the present research, this can only be done on an exploratory level, since a much larger subject sample would be required to develop reliable and valid selection procedures. Nevertheless, it is possible to examine what type of psychometric and cognitive variables are associated with mammogram interpretation performance on an exploratory level. In addition, the relative importance of the different predictor variables for criterion prediction can be

271
examined. The criterion in the present study was the mammogram interpretation performance of novices after very brief training. Conclusions about the predictability of mammogram interpretation performance thus address only the lay screening situation.

In summary, this study aims to explore the value of the psychometric and cognitive predictors in terms of performance on the mammogram test. So far all predictor variables have been included on the basis of face validity alone. In this study, correlations between both sets of predictor variables and performance are examined. The establishment of predictive validity of psychometric and cognitive variables is effected by an investigation of their associations with mammogram interpretation performance after training.

11.1.2 Assessing readiness to learn in lay film readers

Training intervention can be expected to change individual performance differentially depending on ability and personality. In chapter 7, it was shown that individual differences persist even after training which appears to homogenise overall performance. Although there was an overall improvement of performance as a function of brief training, some subjects did not improve or even deteriorated in their mammogram interpretation performance. In particular, participants who yielded high initial performance scores, failed to profit from brief feature training. Selection of individuals to participate in a particular training programme should consider the trainees ability to profit from the training provided, i.e. their readiness to learn in a particular training context.

Berbaum et al (1985) have shown the value of predicting improvement as a function of training in the context of radiology residency programmes. However, although they achieved a high level of prediction in radiological domains involving cross-sectional imaging, two-dimensional radiology was not predicted well by the three-dimensional Visual Form Reconstruction Test (Smoker et al, 1984). The
second aim of the present exploratory study was to evaluate the value of the psychometric and cognitive predictors for assessing an individual's ability to profit from brief feature training in mammogram interpretation. For this purpose, improvement measures were computed from performance on the mammogram interpretation test for use as criteria. These were measures of difference between pre- and post-training performance and pre-training and follow-up performance, respectively. The former thus represents a measure of improvement as a function of training (improvement), whilst the latter is a measure of enduring improvement without further training intervention (retention).

11.1.3 Diagnosis of weaknesses in performance

Although selection of trainees in terms of predicted post-training performance or in terms of predicted readiness to profit from a particular training programme represents an efficient way of matching people to job requirements, the method is only applicable when there is a relatively large pool of applicants. In a situation where most trainees are self-selected, knowledge of relationships between personality characteristics and task performance may be more profitably employed to identify individual strengths and weaknesses. It may, for example, be possible to match individuals to training schemes from which they are most likely to benefit. Alternatively, specific weaknesses could be supported with additional training. Identification of personal characteristics which are associated with good or poor mammogram interpretation performance can thus be seen to be useful in a number of different contexts.

The third aim of the current investigation was to present a qualitative analysis of variables associated with good and poor performance in mammogram interpretation after brief training. In addition, variables important in determining whether an individual is likely to benefit from brief feature training were examined. Since the subject sample was quite small, the most appropriate way of exploring relationships the between predictor variables and good or poor performance and improvement
was in terms of a qualitative, descriptive analysis.

11.2 Method

The studies reported in the current chapter are exploratory analyses of the predictability of mammogram interpretation skills. For this purpose, performance data from the training studies reported in chapters 6 and 7 were employed as prediction criteria. Predictors were the psychometric and cognitive variables discussed in chapters 9 and 10.

11.2.1 Subjects

Each of the thirty-two participants in the training studies (chapter 6) had also completed the psychometric test battery and the cognitive tasks. These subjects thus represented individuals for whom both, predictor and performance data, were available. There were twenty-one female and eleven male subjects for whom complete data sets had been collected. Their ages ranged from eighteen to forty-two with a median age of nineteen.

11.3 Results

The first issue to be investigated was the predictive validity of the psychometric and cognitive variables. Performance criteria for the assessment of concurrent and predictive validity of psychometric predictors were the performance on the mammogram test measured before and after training respectively. The mammogram test was completed by all students prior to the commencement of training to provide a baseline measure of performance. This performance prior to training in a student sample without previous experience of breast screening is presumably indicative of a natural ability to cope with X-ray interpretation, since only four participants performed worse than chance, and mean performance was significantly better than chance ($t = 6.5, p < .001, \bar{x} = .269, sd = .233$).
Performance after training on the mammogram test is an indicator of individual potential as a function of training. This is the performance criterion, the prediction of which is most critical for the purpose of selecting lay screeners. Two further performance criteria were improvement measures, difference measures of post-training and follow-up performance with pre-training performance. These were computed as outlined above and represent improvement as a function of training and long term skill retention after training respectively. Prediction achieved by the psychometric and cognitive variables in post-training performance, improvement and retention were examined using exploratory multiple regression.

11.3.1 Predictive validity of psychometric and cognitive variables

Table 11.1 shows the correlations between performance criteria and predictor variables. It can be seen that the highest correlations are between performance and the cognitive predictor variables. Psychometric variables, on the whole, do not show significant relationships with performance before training or with either of the two improvement measures. The only statistically significant relationship between a psychometric predictor variable and post-training mammogram interpretation performance is observed with the originality scale on the Kirton Adaptation-Innovation inventory (r = -.35, p < .05), indicating that an 'original' problem solving style is detrimental to trained film reading performance.

Some of the psychometric variables appear to have a degree of association with mammogram interpretation, albeit not a statistically significant one. Spatial ability is related to performance before training. Apart from the negative correlation with the KAI Originality subscale, post-training performance also yields a positive association with the KAI Efficiency subscale. The relationship of spatial ability with post-training performance is much weaker than that with pre-training performance. As far as the improvement measures are concerned, the main relationships, both negative, are with KAI-O in the short term and KAI-E in the longer term (retention). Spatial ability appears to play a role only in the long term
and is mediated by three-dimensional spatial ability. Both improvement measures also show a weak negative relationship with field-independence.

Table 11.1: Correlations between predictor variables and mammogram interpretation performance before and after training and improvement and retention as a function of training.

<table>
<thead>
<tr>
<th></th>
<th>G before training</th>
<th>G after training</th>
<th>improvement</th>
<th>retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPQ-P</td>
<td>.17</td>
<td>-.01</td>
<td>-.18</td>
<td>-.19</td>
</tr>
<tr>
<td>EPQ-E</td>
<td>-.05</td>
<td>-.03</td>
<td>.04</td>
<td>.18</td>
</tr>
<tr>
<td>EPQ-N</td>
<td>.09</td>
<td>.16</td>
<td>.02</td>
<td>-.01</td>
</tr>
<tr>
<td>KAI-O</td>
<td>.00</td>
<td>-.35*</td>
<td>-.26</td>
<td>-.11</td>
</tr>
<tr>
<td>KAI-E</td>
<td>.11</td>
<td>.25</td>
<td>.07</td>
<td>-.23</td>
</tr>
<tr>
<td>KAI-R</td>
<td>.17</td>
<td>.07</td>
<td>-.12</td>
<td>-.09</td>
</tr>
<tr>
<td>KAI</td>
<td>.12</td>
<td>-.06</td>
<td>-.17</td>
<td>-.16</td>
</tr>
<tr>
<td>MD5</td>
<td>-.09</td>
<td>-.03</td>
<td>.08</td>
<td>-.08</td>
</tr>
<tr>
<td>GEFT</td>
<td>.14</td>
<td>.08</td>
<td>-.09</td>
<td>.07</td>
</tr>
<tr>
<td>SAT-2D</td>
<td>.26</td>
<td>.19</td>
<td>-.14</td>
<td>-.13</td>
</tr>
<tr>
<td>SAT-3D</td>
<td>.19</td>
<td>.13</td>
<td>-.10</td>
<td>-.21</td>
</tr>
<tr>
<td>SAT</td>
<td>.25</td>
<td>.17</td>
<td>-.13</td>
<td>-.19</td>
</tr>
<tr>
<td>Contrast S.</td>
<td>.10</td>
<td>.36*</td>
<td>.17</td>
<td>-.14</td>
</tr>
<tr>
<td>G-error</td>
<td>.29</td>
<td>-.28</td>
<td>-.51**</td>
<td>-.44**</td>
</tr>
<tr>
<td>G-corr</td>
<td>.38*</td>
<td>.00</td>
<td>-.40*</td>
<td>-.42**</td>
</tr>
<tr>
<td>T-speed</td>
<td>-.40*</td>
<td>.01</td>
<td>.43**</td>
<td>.45**</td>
</tr>
<tr>
<td>T-acc</td>
<td>-.16</td>
<td>-.09</td>
<td>.11</td>
<td>.04</td>
</tr>
<tr>
<td>Strat-T</td>
<td>-.01</td>
<td>-.06</td>
<td>-.03</td>
<td>.01</td>
</tr>
<tr>
<td>Strat-F</td>
<td>.36*</td>
<td>.17</td>
<td>-.25</td>
<td>-.31*</td>
</tr>
<tr>
<td>Strat-S</td>
<td>-.29</td>
<td>-.07</td>
<td>.25</td>
<td>.24</td>
</tr>
<tr>
<td>Strat-BB</td>
<td>-.48**</td>
<td>-.43**</td>
<td>.19</td>
<td>.26</td>
</tr>
<tr>
<td>Strat-BA</td>
<td>.05</td>
<td>-.02</td>
<td>-.06</td>
<td>.06</td>
</tr>
</tbody>
</table>

* p < .05        ** p < .01
The majority of cognitive predictor variables yielded a strong correlation with pre-training mammogram interpretation, exceptions being contrast sensitivity, tactical strategy and efficiency in dealing with available information. Two of the correlations, that with G-error (.29) and that with accuracy on the target task (.16), were in the opposite direction to that expected. Neither of these were statistically significant, though. Efficiency in acquiring information (Strat-BB) was the variable showing the highest correlation with pre-training ($r = -.48, p < .01$). Further significant relationships with pre-training performance were yielded by accuracy in the grid task ($r = .38, p < .05$), speed on the target task ($r = -.40, p < .05$) and use of a focusing strategy ($r = .36, p < .05$). Post-training performance is less well predicted by the cognitive variables. Although efficiency in acquiring information retains its position as the variable with the highest correlation ($r = -.43, p < .01$), most other previously important variables yield no correlations with post-training performance. There is, however, a significant association with contrast sensitivity ($r = .36, p < .05$) and a degree of association with preparedness to make mistakes on the grid task ($r = -.28, n.s.$).

Prediction of improvement in performance as a function of training appears to be more easily achieved. Both improvement and retention are highly associated with the same variables as pre-training performance but with opposite relationships. This is not surprising, since improvement measures are difference measures between post-training and follow-up performance and pre-training performance. Individuals performing well on the pre-training assessment thus yield low scores on the two improvement measures, since there is little scope for improvement as a function of training for these individuals.

11.3.2 Prediction of performance after training

The examination of correlations between predictor variables and pre-training performance above demonstrated that the more numerous associations were with the cognitive predictors. Post-training performance, however, showed some equally
high correlations with personality variables. In order to assess the relative importance of variables correlated with post-training performance a standard multiple regression was performed. This analysis is suitable for exploratory purposes, even if the sample size is small (Tabachnick and Fidell, 1989). The main issue was to investigate the degree to which the variables identified as associated with post-training performance achieved prediction of performance.

All predictor variables with any degree of association (more than .15) with post-training performance were included in the analysis. The analysis was performed with the SPSS-PC data analysis package using the regression command. In standard multiple regression all variables are entered into the equation simultaneously and each variable is assessed as if it had entered the regression after all other independent variables. This method thus allows an assessment of the unique prediction achieved by an independent variable, which is different from the predictability achieved by all other independent variables. Table 11.2 shows the standardized regression coefficients, associated t statistics and levels of significance. R for the regression was .76 which was significantly different from zero ($F = 3.85, df = 8, 23, p = .005$). The predictor variables accounted for 42% of the variance when adjusted for overestimation due to the small sample size. The unadjusted variance estimate was 57%.

Four of the eight variables included in the analysis contributed significantly to the prediction of post-training performance. The only cognitive variable to contribute significantly was contrast sensitivity. Cognitive style as evidenced in the relationships with innovative efficiency, adaptive originality and high efficiency in information acquisition was the single most important contributor to prediction of post-training mammogram interpretation performance. Other personality variables such as spatial ability and neuroticism appeared to be less important for predictive purposes, as did the strategic variables, readiness to make mistakes in the grid task (G-error) and use of a focusing strategy in information acquisition.
Table 11.2: Standard multiple regression of predictor variables associated with performance after training on post-training G.

<table>
<thead>
<tr>
<th></th>
<th>beta</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAI-E</td>
<td>.42797</td>
<td>2.843</td>
<td>.01</td>
</tr>
<tr>
<td>KAI-O</td>
<td>-.40382</td>
<td>-2.591</td>
<td>.02</td>
</tr>
<tr>
<td>Strat-BB</td>
<td>-.37056</td>
<td>-2.376</td>
<td>.03</td>
</tr>
<tr>
<td>Contrast S.</td>
<td>.29770</td>
<td>2.021</td>
<td>.05</td>
</tr>
<tr>
<td>EPQ-N</td>
<td>.14714</td>
<td>.931</td>
<td>.36</td>
</tr>
<tr>
<td>SAT-2D</td>
<td>.10624</td>
<td>.738</td>
<td>.47</td>
</tr>
<tr>
<td>G-error</td>
<td>-.10584</td>
<td>-.635</td>
<td>.53</td>
</tr>
<tr>
<td>Strat-F</td>
<td>-.08628</td>
<td>-.552</td>
<td>.58</td>
</tr>
</tbody>
</table>

R = .76
F = 3.85 df = 8, 23 p = .0052
R² = .57 Adjusted R² = .42

11.3.3 Prediction of improvement as a function of training

Both, improvement and retention of mammogram interpretation performance yielded high associations with a number of predictor variables, in particular cognitive predictors (Table 11.1). The two measures which represent difference measures between pre-training performance and post-training and follow-up performance respectively, are highly intercorrelated (r = .74) and would therefore be expected to yield similar relationships with predictor variables. In the present analysis a set of predictor variables yielding associations with improvement as a function of training was assessed in terms of their predictability of improvement and retention. The relative importance of these cognitive and psychometric variables for predicting improvement and retention of mammogram interpretation skill was examined in two standard multiple regressions conducted in the same manner as the analysis for the prediction of post-training performance above. Table 11.3 shows the standardized regression coefficients and t statistics for both multiple
regressions. The multiple regression with improvement as the dependent variable yielded a multiple R of .74, while that with retention as the dependent variable was .76. Both regression coefficients were significantly different from zero (F = 3, p = .017 and F = 3.7, p = .006, respectively). The predictor variables accounted for 37% and 44% of the variance respectively when adjusted for overestimation due to the small sample. In both analyses, only G-error, the readiness to make mistakes on the grid task, contributed significantly to the prediction of the dependent variables. The order of importance of the remaining predictor variables was similar for improvement and retention. However, contrast sensitivity appeared to play a relatively more important role in predicting retention, whereas speed on the target task might be more salient for the prediction of improvement.

Table 11.3: Standard multiple regressions of predictor variables on improvement and retention of mammogram performance as a function of training.

<table>
<thead>
<tr>
<th>predictor</th>
<th>beta</th>
<th>t</th>
<th>predictor</th>
<th>beta</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-error</td>
<td>-.55062</td>
<td>-2.583*</td>
<td>G-error</td>
<td>-.67924</td>
<td>-3.383**</td>
</tr>
<tr>
<td>Strat-F</td>
<td>-.25079</td>
<td>-1.313</td>
<td>Strat-F</td>
<td>-.33796</td>
<td>-1.879</td>
</tr>
<tr>
<td>EPQ-P</td>
<td>-.24149</td>
<td>-1.291</td>
<td>Contrast S.</td>
<td>-.27695</td>
<td>-1.806</td>
</tr>
<tr>
<td>T-speed</td>
<td>.17479</td>
<td>.916</td>
<td>EPQ-P</td>
<td>-.18951</td>
<td>-1.076</td>
</tr>
<tr>
<td>G-corr</td>
<td>-.14184</td>
<td>-.804</td>
<td>Strat-BB</td>
<td>.14203</td>
<td>.898</td>
</tr>
<tr>
<td>Strat-BB</td>
<td>.10497</td>
<td>.625</td>
<td>G-corr</td>
<td>-.10769</td>
<td>-.648</td>
</tr>
<tr>
<td>Contrast S.</td>
<td>.07306</td>
<td>.449</td>
<td>SAT</td>
<td>-.09389</td>
<td>-.598</td>
</tr>
<tr>
<td>SAT</td>
<td>-.04036</td>
<td>-.242</td>
<td>T-speed</td>
<td>.06553</td>
<td>.364</td>
</tr>
<tr>
<td>KAI</td>
<td>.03326</td>
<td>.178</td>
<td>KAI</td>
<td>.03162</td>
<td>.179</td>
</tr>
</tbody>
</table>

R = .74
F = 3.00  df = 9, 22  p = .017
R² = .55  Adjusted R² = .37

R = .76
F = 3.70  df = 9, 22  p = .006
R² = .60  Adjusted R² = .44

* p < .02  ** p < .003
11.3.4 Prediction of good and poor performance

Although variables may be predictive of performance or improvement as a function of training on the whole, it may be desirable to identify those variables specifically predictive of poor or good performance. This involves identifying those variables which maximise the differences between good and poor performance after training. This can be achieved in a number of ways, but in the present study prediction of individual differences after training was conducted by means of correspondence analysis (Weller and Romney, 1990, Hammond, 1987). This analysis presents a descriptive, non-parametric way of exploring associations between specific predictor variables and good or poor performance after training.

For the purpose of this analysis the thirty-two participants in the training studies were divided into three groups. Good performance after training was indicated by a G index of more than .49 (n = 12), and poor performers after training were those participants who scored less than .32 (n = 10). The remaining ten subjects received a neutral classification. Good and poor classifications are broadly equivalent to performance at half a standard deviation above or below the mean of .39. Table 11.4 shows the three groups' mean scores on the best predictors identified for post-training performance (see Table 11.2).

Correspondence analysis (Hammond, 1987) was employed to investigate the associations between the predictor variables and performance level graphically. The analysis accounted for 90.38% of inertia in one principal component. The relationship between the variates is shown graphically in Figure 11.1 (see Table Ic.3, Appendix Ic for corresponding coordinates). The plot can be interpreted across one dimension, since one component accounted for most of the variance. The locations of the three performance levels are indicated by star markers, and predictor variables are shown as triangles. The three performance levels are ordered from right to left with poor performance being rightmost, followed by neutral
It can be seen that poor post-training performance is specifically associated with
high scores on the Originality scale of the KAI. Preparedness to make errors on the grid task (G-error) is also associated with poor performance but to a lesser degree. This variable is equally associated with neutral performance after training. Good post-training performance is most closely associated with higher two-dimensional spatial ability scores, higher contrast sensitivity scores and higher neuroticism scores. Use of a focusing strategy (Strat-F) and Efficiency on the KAI (KAI-E) distinguish less well between good and neutral learners, and appear to be equally associated with both groups. Only one predictor variable, efficiency in acquiring new information (Strat-BB), did not associate with any of the three performance groups. This variable is, therefore, of little use in predicting performance level after training.

Two further correspondence analyses were conducted in an analogous manner for improvement after training and retention, respectively. Although the analyses accounted for 82.64% (improvement) and 91.78% (retention) of inertia in one principal component respectively, inspection of the plots revealed that the variables predictive of the differences between post-training and follow-up performance and pre-training performance did not differentiate well between disimprovers (negative difference score), slight improvers and good improvers. In particular, none of the predictor variables associated closely with good improvement or retention of mammogram interpretation skill. Prediction of individual differences in improvement was, thus, not possible with the set of variables which were most useful in the overall prediction of performance.

### 11.4 Discussion

The main concern of the present chapter was to investigate the extent to which psychometric and cognitive variables, which were selected on the basis of theoretical relationships as predictors of mammogram interpretation performance, fulfilled their potential in achieving prediction of mammography performance criteria. Examination of Table 11.1 revealed that the associations between the
psychometric personality and ability variables and the four performance criteria are quite small. Although individual predictors yield some correlations with criterion variables, there is no overall strong pattern of associations. None of the observed correlations are high enough to indicate beyond reasonable doubt that they have not arisen by chance, given the large number of correlation coefficients computed. On the whole, this confirms the conclusions drawn in chapter 9, that there is no general pattern of relationships among these particular psychometric tests that might be predictive of mammogram interpretation performance.

Patterns of correlation between cognitively derived variables and performance criteria (Table 11.1) are more consistent than those with psychometric variables. The validity of the cognitive variables as predictors of mammogram interpretation is supported by high correlations between most cognitive tasks and pre-training performance. This criterion is indicative of natural ability to cope with X-ray interpretation, since none of the participants had any prior experience with X-rays or any prior medical knowledge. Thus, accuracy on the grid task, speed on the target task, and a focusing strategy combined with efficiency in acquiring information as identified by the concept task all contribute to the identification of such natural ability in the predicted manner, the only exception being the contrast sensitivity measure. Post-training performance, however, appears to be less well predicted by cognitive variables alone. The issue of post-training performance is considered further below.

The other two criteria, improvement and retention, yield high correlations with the same variables as pre-training performance, but in the opposite direction. This indicates that there is a direct relationship between initial performance and improvement. It is clear that participants with high initial performance scores, i.e. those with some natural ability for X-ray interpretation did not have much scope for improvement as a function of training. In chapter 7, the possibility was discussed that these participants did not profit from brief feature training at all. This appears indeed to be the case as is evidenced in the inverse correlation patterns
between pre-training performance and cognitive predictor variables on the one hand, and the improvement measures and cognitive predictors on the other. This does not, however, present a problem because the purpose for including the improvement measures as prediction criteria was to identify those subjects likely to profit from brief feature training in a self-selection situation. The argument has already been made that able individuals, i.e. those with high pre-training mammogram interpretation accuracy may not benefit from such training.

11.4.1 Personality and post-training performance

Although pre-training performance appears to be predicted better by cognitive variables, training intervention results in close associations between personality variables as well as some cognitive variables. Table 11.1 shows that post-training performance is correlated with personality variables, in particular with the Originality scale of the KAI. On the cognitive side, there are relative few correlations between predictors and the criterion, exceptions being contrast sensitivity and efficiency in information acquisition (Strat-BB). Exploration of these relationships in terms of multiple regression confirms the impression that personality variables are relatively more important in the prediction of post-training performance.

Table 11.2 shows that of the four predictors which contribute significantly to the prediction of the criterion, two were psychometric personality variables. The two subsales of the KAI contribute differentially to performance prediction. Thus, it is not overall innovation or adaptation style that is predictive of mammogram interpretation after training, but a combination of low Originality and high Efficiency. The indication is that the proliferation of original ideas is detrimental to performance whereas a degree of discontinuity in information processing actually aids performance. However, discontinuity does not imply inefficiency as is indicated by the significant contribution of one of the cognitive predictors. Efficiency in acquiring new information as measured by the number of blundering
trials in the concept task is the third most important predictor of post-training mammogram interpretation. Lastly, contrast sensitivity contributes significantly to criterion prediction. Contrast vision, which contributes little to explaining pre-training variance, gains importance after training because lesions in screening mammography may be very subtle in terms of target and background contrast. Knowledge of the visual features which are indicative of mammographic abnormality appears to facilitate the application of low level visual skills.

In summary, a stereotype of the kind of person who is likely to make a good film reader after initial brief training is someone who is not too imaginative, is efficient but not too structured in acquiring information from the environment and possesses good contrast vision. Whilst contrast vision and efficiency are attributes that might have been expected to be predictive of X-ray interpretation, the other two attributes are less intuitively appealing. However, consideration of the literature on search strategies in X-ray interpretation may give some clues as to the role of discontinuity in information acquisition in reading mammograms. Gale and Worthington (1983), for example, found that too much concern with precision in search patterns or the imposition of such very orderly patterns may disrupt performance in film interpretation. There is some anecdotal evidence from radiologists and radiographers involved in mammogram interpretation that abnormality should 'jump out', i.e. it is often the case that first impressions on viewing a mammogram are more accurate than decisions made after long deliberation. The role of Originality in inhibiting good interpretation performance can be explained in terms of the detrimental effects of an overactive imagination. The visual differences between normality and abnormality are sometimes very small, and, with an active imagination, it is possible to 'see' abnormalities in most mammograms (see photos in chapter 3). Individuals with high scores on the KAI-O scale, thus, might be expected to produce large numbers of false positives.
11.4.2 Readiness to learn from feature instruction as a function of pre-training performance

Improvement as a function of training is a measure of the extent to which individuals with no prior aptitude for X-ray interpretation benefit from brief feature instruction in mammography. Table 11.3 shows that both improvement and retention as a function of brief feature training are predictable in terms of cognitive and psychometric variables. Cognitive predictors are evidently more important in predicting improvement than in predicting post-training performance. Only one variable contributed significantly to predicting both criteria. The degree of confidence with which mistakes were made on the grid task was negatively correlated with both, improvement and retention, indicating that resistance to errors in complex visual search is the most important predictor of readiness to learn from brief feature instruction.

The remaining variables present largely reversed relationships compared with pre-training performance. Improvement and retention of mammogram performance appear to be predicted by low levels of focusing, low Psychoticism scores and low scores on contrast vision (retention only). Similarly all other predictor variables yield associations with improvement in the opposite direction as those predicted for good performance. This appears to be a direct result of the improvement measures being high for those subjects who performed poorly before training. Considering the relative importance of the predictor variables in terms of the relative magnitude of their contribution to overall prediction, it can be said that, with the exception of Psychoticism, cognitive variables appear to play a more significant role in predicting improvement. Both spatial ability and adaptation-innovation play only a marginal role in predicting improvement and retention.

In conclusion, there is some evidence that cognitive strategy, in particular preparedness to make errors in complex visual tasks, may be predictive of the extent to which individuals with no prior experience in X-ray interpretation can
profit from brief feature instruction. On the whole, though, it appears to be the case that those individuals who do not possess natural ability to cope with X-ray interpretation are most likely to benefit from such instruction. Natural (pre-training) aptitude for reading mammograms is perhaps most accurately assessed by a pre-training mammogram test rather than tests involving external predictor variables.

11.4.3 Strengths and weaknesses in post-training performance

The third aim of this study was to examine the extent to which predictor variables could be employed to identify specific individual strengths and weaknesses in mammogram interpretation performance. Figure 11.1 shows that poor performance in post-training is closely associated with high scores on the Originality scale of the KAI and on G-error (Table 11.4). The correspondence analysis plot further shows that good performance is most closely associated with two-dimensional spatial ability and contrast sensitivity, closely followed by Neuroticism and use of a focusing strategy. The correspondence analysis, on the whole, supports the findings of the multiple regression analysis involving prediction of post-training performance. The main discrepancy is that efficiency in information acquisition (Strat-BB) did not associate with any of the three performance levels in the current analysis, despite the fact that it contributed significantly to the overall prediction of post-training performance (Table 11.2).

The specific association of poor post-training performance with KAI-O and G-error indicates that the main reasons for poor performance may be of a strategic nature. Both variables are representative of particular information processing strategies which may be related. The role of Originality as an indicator of an active imagination and its likely effects on false positive rates has already been discussed above. Similarly, the preparedness to make errors at a high level of confidence in the grid task denotes an inclination for over-interpretation of minimal visual signals.
One of the reasons for attempting to identify variables indicative of specific weaknesses in mammogram interpretation after training was that it might be possible to rectify individual weaknesses with specific training and support in a self-selection situation. The strategic nature of the weaknesses associated with poor post-training film reading performance, which were identified in the present study, indicates that it may be possible to assist individuals in adopting more task-appropriate strategies, thus improving their overall performance. Such specific intervention would probably aim to clarify the visual features of mammographic abnormality further, discouraging the over-interpretation of minimal deviant appearances. It should be noted, however, that the specific link between poor performance and visual over-interpretation is likely to be associated with the brevity of feature training, which results in students performing the skill on a visual level (see chapter 7). It is possible that a study involving longer training might have found closer links between decision making strategies involved in the classification of mammograms as normal or abnormal and performance.

11.4.4 Conclusions

It can be concluded that improvement and retention of mammogram interpretation performance is not well predicted by external variables because the two improvement criteria are confounded with pre-training performance. The latter appears to represent a natural ability to deal with the visual information contained in X-rays. Such natural ability can be satisfactorily assessed with an X-ray interpretation task such as the mammogram test itself. However, training intervention influences the predictability of mammogram interpretation performance. Post-training performance is more difficult to predict with the task-derived cognitive predictors. This appears to indicate that variables other than the match between specific task parameters and cognitive functioning contribute to performance after training.

In chapter 8, it was hypothesized that mammogram interpretation in the context of
breast cancer screening is influenced by variables on social, cognitive and strategic levels due to the unique nature of the screening film interpretation in radiological and medical diagnosis. Support for this contention can be derived from the wide variety of variables, which significantly contribute to predicting performance after training. In particular, strategic variables have been identified as good overall predictors of post-training performance and are associated with poor performance after training. In terms of predicting post-training performance, spatial ability and good contrast sensitivity appear to be important prerequisites of individuals who read mammograms successfully after training. The involvement of cognitive style and personality in predicting film reading performance leads to the conclusion that mammogram interpretation is one complex cognitive task which is not necessarily better predicted by very task-specific tests. This finding stands in contrast to Lesgold et al's (1990) contention that the best assessment of cognitive performance in complex tasks is in terms of cognitive task parameters.
Discussion

Chapter 12
Complex visual skill acquisition, training and individual differences

This chapter presents a summary and final discussion of the main empirical findings reported in this thesis. The research is discussed in terms of its implications for theoretical and applied questions raised in the context of learning to interpret mammograms. In particular, issues of skill acquisition and the nature of mammogram interpretation skill, training and individual differences are addressed. The chapter is further concerned with an analysis of the main limitations of the studies and theoretical analysis reported in this thesis. This is followed by an outline of how this applied line of research might be advanced in the future to address further applied and theoretical issues of relevance to breast screening and radiological skill acquisition.

12.1 Theoretical issues in mammogram interpretation

Three main empirical approaches were employed to investigate pertinent skill issues in breast cancer screening. These involved firstly, a longitudinal case study of two radiographers learning to read mammograms, secondly, some training experiments with students who had no prior knowledge of mammography, and, thirdly, a psychometric study which aimed to identify ways of predicting mammogram performance as a function of training. Theoretical underpinnings of these studies, which were designed to answer the applied questions raised in the context of this thesis, related to the issue of the kind of cognitive mechanisms involved in skilled performance. The interpretation of mammograms was initially described as a complex visual skill involving perceptual and interpretative task elements. This notion was formalised in a model of the cognitive processes involved in mammogram interpretation. The integrity of the model was examined
in terms of its ability to account for empirical findings resulting from the mammogram training studies in particular, and investigations of radiological skill in general. The task model also proved useful in deriving tests measuring performance in terms of actual cognitive skills implied in mammogram interpretation. Although the issue of individual differences was approached from a psychometric perspective, the cognitive processes identified as part of the process model played a major role in addressing observer variability and its prediction in mammogram interpretation. The model of mammogram interpretation was thus central to both, questions about training and approaches to individual differences. In this section, implications of the empirical studies for theoretical considerations in this thesis are discussed.

12.1.1 The nature of mammogram interpretation

12.1.1.1 Summary of main issues

Previous research on X-ray interpretation has approached the topic from two different perspectives. In the first instance, radiological skill has been investigated in terms of visual processes, in particular those involving the detection and recognition of significant visual features. Secondly, radiological interpretation has been studied from a cognitive skills point of view, an approach which emphasizes the medical diagnosis aspects of radiological skill. It was argued in this thesis, that full mammogram interpretation in the context of screening mammography involves both, visual and interpretative processes. This view was formalised in the model of mammogram interpretation in chapter 7.

Support for the distinction between perceptual and interpretative processes in the model could be derived primarily from the fact that it was possible to distinguish between two types of errors which correspond to the two levels of processing. People clearly make errors which are due to not seeing or not recognising abnormality, which represents a failure of visual processes. The other type of error, the interpretation error, occurs when observers recognise an abnormality and clearly
mark it, but decide that it is insignificant. Both types of error could be clearly distinguished in the longitudinal radiographer study and in the student training experiments. However, interpretation errors were much more common in radiographer performance than amongst students, and it was concluded that this represents some cross-sectional evidence for a developmental sequence in skill acquisition, in which perceptual skill is a prerequisite of interpretation skill (Lesgold et al, 1988). This conclusion was based on the assumption that students have less developed skills in mammogram interpretation on account of their very short training and prior inexperience with mammograms.

The model also accounted well for some general findings in the skill acquisition studies. One of the most persistent findings in radiological skill acquisition is that people appear to learn what represents normality in X-rays, i.e. they improve at recognizing true negatives but not necessarily true positives. This was confirmed in the present training studies, all of which showed that improvement could be accounted for in terms of reduction in false positives, whilst false negatives appeared unaffected by training. The main reason for questioning the notion that this finding is due to trainees learning to recognise normality, was that the same pattern of post-training performance was found for individuals who had almost exclusively been trained with abnormal mammograms. Taking into consideration the processing paths required for true positives and true negatives in the model, an alternative explanation for improved true negative classification can be found. The production of true positives requires a longer processing path, in which both perceptual processes and interpretative processes need to be executed without error. The reduction of false positives can thus be either due to improved visual discrimination or to better interpretation or both. False negatives, on the other hand, can occur independently either at the visual processing stage or at the interpretation stage. In terms of processing considerations, true positives are, therefore, more difficult to produce because they require more complex processing.
12.1.1.2 Implications for breast screening

The main implication for breast screening regards the conclusions that can be drawn about the nature of the film reading task. The indication is that full interpretation of screening mammograms involves both, visual search and diagnostic interpretation processes. It has been shown that the task can be performed at a pattern recognition level by lay people (see discussion below). Although the evidence in the present research is indirect, it would appear that discrimination between normal and abnormal mammograms can be learned relatively rapidly by people without prior experience with mammograms. Appropriate interpretation of abnormality remains a difficult task, however, and may require specialist medical knowledge. It can be concluded that lay film readers can readily acquire the visual skills required for discrimination between normal and abnormal mammograms, an aspect of mammogram interpretation which is very rapid and fully automatic in expert radiologists. Search for and recognition of abnormality in mammograms depends on the availability of appropriate visual schemata, which may be acquired most easily through specific feature training.

12.1.2 Individual differences and cognitive processes in mammogram interpretation

12.1.2.1 Summary of main issues

Another theoretical concern in this thesis was the identification of sources of individual differences in task performance at the level of cognitive processes. One attempt to investigate knowledge representation and reasoning processes as a function of skill acquisition was the collection of verbal data in the radiographer study. It was clear in that study that the two radiographers used pertinent concepts differently when categorising mammograms. There were some problems, however, with the detail of the verbal protocols obtained. Mammogram interpretation involves dominant perceptual components which were not well captured by the 'thinking aloud' method, because perceptual processes may not be consciously accessible. In addition, even if they were accessible, they would require processing into a verbal format which might affect task performance. The verbal protocols
obtained from the radiographers thus demonstrated that training changes the knowledge structures used in the interpretation phase which involves verbally coded reasoning processes, and that there are individual differences in how such knowledge is applied.

The identification of individual differences in terms of cognitive processes was also considered important in the context of selection of lay people for mammogram interpretation. It was argued that from a theoretical point of view the unequivocal identification of strength and weaknesses in cognitive processes associated with specific aspects of mammogram interpretation is preferable to the pure measurement of general differences. This is because knowledge about differences at a process level leads to a better understanding of what the task entails on a theoretical level, as well as the potential for much more specific training intervention at the applied level. Theoretical ideas about the nature of the task were formalised in the task model of mammogram interpretation, and this information was used in the development of cognitive tests representing specific skill aspects of mammogram interpretation. It was anticipated that simulating aspects of the task would provide measures representative of macroscopically defined cognitive processes involved in mammogram interpretation. However, it was only possible to derive very broad information processing scores from the tasks, which could nevertheless be employed statistically in predicting mammogram performance.

12.1.2.2 Implications of theoretical considerations concerning individual differences

Cognitive processes in the visual evaluation and interpretation of radiological features appear to be difficult to investigate because they are closely interlinked with visual processes, which are not consciously accessible. This is compounded by the fact that the present research was largely concerned with novice performance on what has to be considered a complex perceptual medical skill. In the present context, the two methods for identifying individual differences have not been very successful in distinguishing between individuals at a cognitive process level. This
issue will be discussed further below. The indication from the studies on individual differences in lay mammogram interpretation is that the pertinent differences are strategic and stylistic rather than cognitive. This is noteworthy, because other research involving complex cognitive skills with perceptual elements has also identified information processing strategies as important individual difference variables, an example being spatial reasoning strategies (Lohman and Kyllonen, 1983, Cooper and Mumaw, 1985).

12.2 Empirical findings and implications for breast screening

The studies reported in this thesis were designed to answer three applied questions, whether lay film readers can be trained to perform the complex skills involved in mammogram interpretation, how training may be designed to support the acquisition of appropriate skills and how the effects of observer variability may be minimised. The present section presents a summary of the main empirical findings and an analysis of their implications for breast cancer screening with specific reference to the issue of lay film reading.

12.2.1 Lay film reading

Issues in lay film reading and training were investigated in the first two parts of this thesis. In particular, chapter 4 was concerned with the acquisition of film reading skills of two radiographers in the context of breast screening, and chapter 6 reported experimental studies of undergraduate students receiving brief training in mammogram interpretation. Both, the radiographer study and the experimental work with students, demonstrated that lay people could be trained to read mammograms. Performance in reading mammograms improved significantly after training in three groups of students. The radiographers, who received two weeks of training, achieved performance levels comparable with those of radiologists trained in the same way. This was supported by similarly good performance in actual screening. Compared to a radiologist, the radiographers regularly achieved performance levels
in excess of .7 as measured by the G-index (see Figure 4.9). In terms of screening outcome, a more objective screening performance criterion, they achieved even higher agreement scores. It can thus be concluded that the data presented in this thesis support the notion that lay film reading in screening mammography is feasible.

The implications of this for breast screening in Britain are twofold. Firstly, it may provide an opportunity for addressing the problem of observer error, which is inherent in all imaging based screening programmes, in terms of dual reading of mammograms. Data presented in chapter 4, indicate that there may well be some benefit of dual film reading with radiographers in terms of increased true positive rates. Although the 6.25% increase in true positives found in the present study is based on very small numbers, it may provide a reasonable estimate of the impact of dual screening on the true positive rate. Kirkpatrick (1993) estimated the increase at a minimum of 2.5% in Scotland, whereas in Sweden dual reading has improved sensitivity by as much as 15%. The second implication is that it highlights the need to re-assess the role of radiographers in breast screening. The purely technical role radiographers are currently playing is in conflict with attempts to professionalize nursing and paramedical occupations. The possibility of extending the radiographers role by involving them in the performance of routine radiological tasks such as film reading has gained wider acceptance in recent years. Thus, some radiological hospital departments operate a 'red dot system', whereby radiographers 'pre-screen' all radiographs and mark abnormal ones with a red sticker. The idea is that this cuts the routine film reporting task of radiologists, who can direct most of their effort towards diagnosing X-rays marked as abnormal. The issue of radiographers' role in mammogram film reading is discussed in more detail below.

12.2.2 Issues in training lay film readers for screening mammography

12.2.2.1 Summary of main empirical findings

Issues in training were discussed both, with reference to the radiographer study and
in terms of experiments conducted with students. The radiographer study showed that the main effect of two weeks of training was to reduce false positives whilst true positives remained relatively variable. The screening data supported this finding. It was concluded in that study that the current training does not equip radiographers for a pre-screening function, because this should entail picking up the large majority of true positives. Instead, training based on tutorials with radiologists encourages them to emulate radiologist performance. It was argued, though, that the way in which radiographers approach the film reading task may be influenced by training. In particular, it was thought that training aimed at the visual aspects of mammogram interpretation might succeed in producing pre-screeners, which is a more appropriate role for lay readers in the medico-legal context of NHS breast cancer screening.

The role of training in producing differences in how the mammogram interpretation task is performed was investigated more specifically in the training experiments involving students. This study, comparing two different training methods, one concerned with visual features and the other modelled on the tutorial approach, produced no differences in overall performance between the two groups. It was argued that this finding may be the result of procedural shortcomings during the administration of training. Specifically, the contention was that the two methods were not differentiated enough to produce observable differences in performance. An alternative explanation was that training in both methods was too brief to produce such differences. The latter explanation was based on the idea that complete novices must acquire visual skills prior to being able to interpret mammograms fully. Although the feature training method was specifically aimed at developing visual schemata, trainees in the tutorial methods probably acquired similar discrimination skills on the basis of practice with normal and abnormal examples.

The utility of feature training in lay reader training became the main focus of the training studies, because a chance difference in variance between the two training
methods groups led to the discovery of a homogenising effect of feature based training. This was replicated in a larger sample in a second training study. It was argued that the apparent homogenising properties of feature training can be usefully exploited in lay reader training, because it can serve to reduce individual differences. However, the indication was that the more able trainees did not profit from feature training. It was thought that able individuals did not improve with feature training because they developed appropriate visual schemata very rapidly and did not require detailed explanation of visual features.

Another issue addressed in the context of the training studies was whether acquiring X-ray interpretation skills proceeds in terms of a developmental sequences, requiring perceptual skills to be in place before interpretation skills are acquired (Lesgold et al, 1988). The training studies appeared to indicate that brief training results in better discrimination performance regardless of training method. In terms of the model of mammogram interpretation this indicates that trainees acquired visual schemata which enabled them to distinguish abnormal mammograms from normal ones. There was no evidence for improvement with regard to interpretation of abnormality though. In fact, subjects appeared to perform at a visual level, i.e. preceding interpretation. Error analysis of false negatives showed that the majority of errors were perceptual ones, and there were very few interpretation errors. In the radiographer study, which involved longer training and individuals who had prior experience with mammograms, the majority of false negative errors were interpretation errors. Cross-sectionally, this represents evidence for a developmental sequence in learning to read mammograms, in which perceptual discrimination precedes interpretative decision making.

12.2.2.2 Implications of training studies for breast screening
The training studies have demonstrated that specific instruction in mammogram interpretation results in improving performance in lay people, whether they are complete novices as in the student training studies or whether they have prior experience of mammography as in the case of the radiographers. The value of
training in learning to read mammograms is thus uncontroversial. The longitudinal performance data in the radiographer study appear to indicate, however, that training may need to be more continuous than it currently is. Both, performance in test mammograms and in screening performance declined after initial increases immediately following training. The indication is that exposure to screening mammograms alone does not result in further increases in performance and that it is not even sufficient to maintain maximum performance. Reasons for this might be the lack of direct feedback on performance in the screening situation (Brehmer, 1980). The beneficial effects of initial training indicate, however, that periodic further training sessions may be a useful medium for achieving more continuous performance in mammography film reading (Fajardo et al, 1987).

The student training studies showed that brief feature training has properties that may make it particularly suitable for training lay film readers. However, it was also noted that such training is not appropriate for a proportion of trainees, mainly those who already perform at a high level. This raises the question whether such training can be useful in the context of training radiographers to read mammograms, for example. Radiographers in breast screening have prior experience with mammograms, albeit from a technical point of view. They may, therefore, be much more aware of what constitutes abnormality in mammograms, and feature training designed to improve visual schemata of abnormality may not be appropriate for them. However, preliminary findings from recent work with five more radiographers (see section 12.4), who received the same feature training as students as part of their two week course has shown that such specific visually based training has got beneficial effects in more experienced trainees, too. In comparison with the two radiographers in the present research, the five new trainees yielded a significantly higher true positive rate in the post-training assessment ($t = -3.15$, $df = 5$, $p < .025$). The indication is that brief feature training may be a useful element of training for most lay film readers, even those with some prior experience of mammograms.
12.2.3 Selection of lay readers for mammogram interpretation

12.2.3.1 Summary of main empirical findings
The issue of individual differences in mammogram interpretation skill was approached from two perspectives. In the first instance, an attempt was made to define a psychometric aptitude domain in terms of ability and personality variables. The argument was that mammogram interpretation performance has to be seen in the context of social aspects of the task as well as in terms of strategic variables and cognitive ability. A battery of tests consisting of personality and ability measures, which were hypothesized to be associated with performance on the social, cognitive or strategic task level, failed to define a general aptitude domain. The second approach involved the development of cognitive tasks representative of specific cognitive processes involved in mammogram interpretation performance. It was argued that measurement of cognitive skills required for mammogram interpretation represents a more appropriate way of predicting performance, because such skills are more directly related to task performance than cognitive ability (Lesgold et al, 1990). Although task development for the cognitive battery was conducted with reference to the cognitive task model in chapter 7, the performance measures derived from the cognitive tests were too global to be representative of specific cognitive processes in mammogram interpretation as implemented in the model. Nevertheless, cognitive skill measures yielded little overlap with the psychometric tests, leading to the conclusion that the two test batteries may independently contribute to the prediction of mammogram interpretation performance.

A further issue under investigation was the extent to which personality, ability and cognitive skills variables from the two batteries were predictive of actual mammogram interpretation skill. This was investigated in an exploratory prediction study, in which mammogram interpretation performance in the student training studies was used as a performance criterion. It was shown that performance after training can be predicted quite well in terms of both cognitive style and basic
cognitive variables. Specific cognitive skill variables, on the other hand, did not predict performance after training. Good performance was specifically associated with good contrast vision, a basic visual processing variable, and two-dimensional spatial ability, and to a lesser extent by strategy and personality variables. Poor post-training performance can be predicted in terms of certain information processing strategies associated with over-interpretation of minimal abnormality.

12.2.3.2 Implications of psychometric studies for lay reader selection

Although the number of subjects involved in the prediction study is small, it can be said that it has been shown on an exploratory level that individual differences in mammogram interpretation after training are predictable in terms of strategic and cognitive variables. This implies that it should be possible to develop selection procedures which aim to identify lay people who are likely to succeed at the mammogram interpretation task after training. It has also been shown that such attempts in prediction have to approach the problem on more than one level. The variables which proved the most useful in the present study originated from both, the psychometric battery and the cognitive task battery, but neither the psychometric nor the cognitive skills battery on their own accounted for mammogram interpretation skill in a convincing manner. As far as cognitive skills are concerned, it was shown, that it was the more basic elements concerned with visual processing, in particular contrast sensitivity which proved to be most important for good performance in the present study. This supports the notion that lay film readers, at least in the initial stages of training, perform the task as a visual discrimination task rather than on a cognitive interpretation level. As visual discrimination ability is a prerequisite of the full interpretation task and is, in fact, the only skill required in a pre-screening context, it may be appropriate to base any selection procedures on this important task aspect.

12.2.4 Implications for lay film reading: Roles of lay readers

The importance of taking into account the role lay film readers are to play when
considering issues in the selection and training of individuals for screening mammogram interpretation has been emphasized throughout this thesis. In this section, the implications of the present research and theoretical considerations for lay film reading are discussed in terms of four possible roles the lay reader could fill. Which role a film reading radiographer adopts depends to a large degree on what function she is to perform. Throughout this thesis, a distinction has been made between pre-screening and 'second pair of eyes' roles. The former implies that the second reader has knowledge of pre-screening decisions and not all films are necessarily seen by two observers. The latter involves independent evaluation of all films by two observers. In addition, radiographers could learn to perform a pattern recognition function (Swinburne, 1971) or they could operate at a level consistent with full X-ray interpretation. It is possible to define the different roles for lay film readers by cross-classifying different performance levels with different implementations of dual screening. This way of conceptualising potential roles for lay readers in X-ray interpretation is shown in Table 12.1 in combination with the expected performance in terms of true and false positives.

Table 12.1: Potential roles and characteristics of performance of lay film readers in terms of task level and role definition.

<table>
<thead>
<tr>
<th>Pattern Recognition</th>
<th>Pre-screening (second reader knows of pre-screening decisions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting abnormality</td>
<td>high TP/high FP rate</td>
</tr>
<tr>
<td>'Red dot system'</td>
<td>high TP/high FP rate</td>
</tr>
<tr>
<td>Full Interpretation</td>
<td>Emulating radiologist</td>
</tr>
<tr>
<td></td>
<td>balanced TP/FP rate</td>
</tr>
<tr>
<td></td>
<td>Cervical cancer screening</td>
</tr>
<tr>
<td></td>
<td>no FNs</td>
</tr>
</tbody>
</table>

In the present radiographer study (chapter 4), the two individuals concerned clearly performed as 'second pairs of eyes' in a full interpretation function. The question
arises whether this is the most suitable role for lay readers in breast screening, however. The training studies have shown that lay readers can be trained to discriminate between normal and abnormal appearances on a visual level. Emulation of radiologist performance involves making decisions about probable benignity and malignancy, i.e. diagnostic decisions. The presently used tutorial method of training breast screeners does not include any specific instruction with regard to medical, anatomical or pathological issues in evaluating radiological evidence of breast disease. This is because most trainees to date have been fully trained medical practitioners who were expected to possess the prerequisite knowledge for diagnosing breast disease. The successful deployment of paramedical lay readers may depend on either defining their role in breast screening as 'pattern recognizers' with no responsibility for any diagnostic decisions or on providing them with appropriate training in order to enable them to make informed diagnostic decisions.

Pre-screening solutions appear unacceptable in breast cancer screening, because of the lack of stable performance, particularly with regard to true positives. Pre-screening as implemented in cervical cancer screening, for example, would imply that the second reader only views those films which are marked as abnormal plus a small proportion of normal ones. In this context, the occurrence of any false negatives is highly undesirable, but the training studies appeared to indicate that it is not possible to train people not to miss abnormality. Even in the 'red dot' pre-screening scheme, in which the radiologist views all films, false negatives are not desirable. Red dots serve to direct the observer's attention and create expectations. In screening mammography, subtle lesions are more likely to be missed by the second reader, if they were missed by the pre-screener. The absence of a red dot in this instance may create false expectations of normality in the radiologist, thus compounding the error of the pre-screener.

In summary, a 'second pair of eyes' pattern recognition role appears to be the most appropriate film reading role for radiographers in breast cancer screening. Properly
implemented, this dual reading scheme also represents the most profitable way of increasing true positives without trade-off in terms of increased recall rates. The training studies did not support the notion that feature training is intrinsically better at producing film readers who excel in the identification of abnormality as far as complete novices are concerned, but recent work with more radiographers, which was mentioned above (see also section 12.4.1), has shown that there may be some benefit in terms of recognition of abnormality if feature training is integrated into the two week training period. On the whole, the homogenising effects of feature training may be beneficial in reducing individual differences in lay film readers. Similarly, selection procedures can be developed to select those individuals most likely to be successful film readers after training. The indication is that in the pattern recognition context, attention should be paid to basic visual ability such as contrast sensitivity on the one hand, and strategic and cognitive style variables associated with over-interpretation and impulsive decision making on the other.

12.3 Limitations of the present research

The main limitations of the research and the theoretical considerations presented in this thesis arise from the pressures and restrictions inherent in most research concerned with applied issues. The problems inherent in applied cognitive research but also the contributions that cognitive psychology can make to applied issues and vice versa have been discussed by Richardson (1989). In the present research an attempt was made to address applied issues in skill learning in screening mammography from a cognitive psychology perspective. Applied issues concerned in particular the selection of individuals who would succeed at the task and training that would equip them to perform mammogram interpretation in the screening context. In the present section, the main applied and theoretical limitations arising from this research are discussed.
12.3.1 Practical limitations

On a practical level the main limitations of the present research are threefold and relate to the sample size, the use of student subjects in the training studies and the brevity of training in the training studies. Although the longitudinal study of radiographers presents a rich source of qualitative data about skill acquisition in screening mammography, it is essentially a case study, and there is little scope for statistical inference and generalisation of findings in such a study. Since radiographers are the most likely candidates for lay film reading, a study with a larger sample of radiographers would have been desirable. This was not possible, however, because the constraints imposed by the applied setting of the research did not permit the involvement of radiographers in this research beyond the initial case study.

Since radiographers were not available to take part in any further studies, all the other research in this thesis was conducted with undergraduate students. In this context, the question arises whether students are representative of a radiographer population. On the whole, it can be said that students are probably not too dissimilar from radiographers in that both groups are educated to 'A' level standard, and professionalisation of the paramedical occupations involves most radiographers attending part-time degree courses. One of the characteristics of students which matches those of radiographers less well is their younger age. However, the inclusion of several mature students in the training studies may serve to mitigate this problem. Another issue is whether the inclusion of male participants in the training studies was justified. In breast screening, the large majority of radiographers are female. However, there are some male radiologists in breast screening, a fact which renders the complete exclusion of male lay film readers unrealistic.

A further practical issue of concern in the present research is the question whether the brevity of training in the student training studies compromises the conclusions
that can be drawn about skill acquisition in screening mammography. It is clear
that the student training studies do not address the entire process of skill acquisition
in mammography. Rather the intention was to contrast the effects of alternative
training methods on the nature of performance. Thus, although it would be
desirable to study skill acquisition on a longer term basis, this was not possible
with student subjects who participated on a voluntary basis. Nevertheless, it could
be shown that brief feature training may have a place in general training for lay
film readers, especially in view of the fact that all training for breast screeners in
the NHS is, by necessity, relatively brief.

Another weakness, as far as the selection issues were concerned, is the sample size
in the training studies, which was constrained by practical limitations, e.g. the
number of people volunteering to take part and the cost of the £15 payment for
participation. This resulted in rather a small sample being available for the
prediction study in chapter 11, on account of the limited number of people for
whom post-training performance criteria were available. However, even with a
relatively restricted sample it was possible to examine the predictive value of
ability, personality and cognitive skills variables on an exploratory level.

12.3.2 Limitations of theoretical considerations

The limitations of the present research from a theoretical point of view are, on the
whole, more profound than the practical limitations. The main problem appears to
be a trade-off between the demands imposed by applied questions and theoretical
aspirations, respectively. Since the applied questions represented the main
motivation for this research, issues of training and selection have taken precedence
over theoretical issues in skill acquisition. The results of this have been two-fold.
Firstly, theoretical discussions of skill in screening mammography go beyond the
empirical data, and secondly, theoretical aspirations with respect to identifying
individual differences at a cognitive process level had to be abandoned in favour of
a psychometric approach.
The cognitive task model in chapter 7 was preceded by a detailed discussion of cognitive processes involved in mammogram interpretation in chapter 5. The latter was important for an understanding of the former, but both go beyond the data presented in the training studies. These do not deal with skill acquisition in mammography at the same level of detail as the theoretical expositions might suggest. In fact, it was shown that the training data do not even distinguish effectively between learning of visual processes and the acquisition of interpretation skills. Conclusions about different processes in mammogram interpretation and developmental sequences in their acquisition can only be drawn by implication in comparison with data from the radiographer study. Nevertheless, the theoretical analysis of mammogram interpretation skill is useful in attempting to understand the processes underlying mammogram interpretation performance and must be considered entirely justified for this reason alone.

A further need for the development of a detailed cognitive task model of screening mammogram interpretation was presented by the attempt to identify individual differences in film reading at the cognitive process level. It was intended to bridge the gap between psychometric testing and cognitive analysis by developing a set of cognitive tests which could measure performance on aspects of mammogram interpretation at a process level. However, the tasks which were developed to represent specific aspects of mammogram interpretation failed to measure performance at a cognitive process level. One reason for this may be that the tasks were too macroscopic, attempting to measure performance in both, visual and decision making processes simultaneously. It might have been possible to develop alternative tasks which are better able to separate out different cognitive processes. Although the cognitive tasks could not be used to differentiate between different cognitive processes, it was possible to use cognitive performance measures psychometrically for purposes of prediction. This does not satisfy the theoretical aspirations of this part of the thesis, but it presented a contribution to answering applied questions regarding the utility of cognitive skills measures in predicting film reading performance after training.
12.4 Suggestions for further work

The work reported in this thesis has been concerned with the acquisition of film reading skills in screening mammography with particular emphasis on issues in selection and training of lay film readers. On the whole, the empirical work presented in this thesis was conducted on an exploratory level, and further research is required to replicate and confirm the findings of the present research. On an applied level, this involves further investigations of the suitability for training of radiographers for a film reading role and the development of sound selection procedures. In addition, there are some specific avenues along which the present work can profitably be extended on a theoretical level. For example, the analysis of mammography film reading as a complex cognitive skill involving perceptual and interpretative processes, which was formalised in the model of mammogram interpretation, provides a starting point for detailed examination of radiological skill in screening mammography. Both, applied extensions and theoretical advances of the present research are discussed in more detail below.

12.4.1 Future work concerning lay film reader selection and training

Further research on radiographer film reading in breast cancer screening has recently started as a direct result of the pilot work with two radiographers reported in the present thesis. Issues raised in this thesis are being addressed in this work. The study involves the training of five further radiographers from three screening centres in the South-West Thames Health Authority region. Training for these radiographers is the same as for the two radiographers who took part in the present research, but a feature training module very similar to that used in the student training studies has been added. Preliminary results of this study appear to confirm the suitability for training of radiographers for breast screening, and, as mentioned above, there is some evidence that feature training has beneficial effects in such training. The numbers involved in this study are still very small, though, and it is hoped that a national study involving a substantial number of radiographers may
become possible, when and if dual lay/consultant film reading becomes an acceptable alternative to single consultant reading.

This would also provide an opportunity for developing appropriate selection procedures for lay film readers with a much larger sample. The present work can only provide a pointer at the most likely predictors of mammogram interpretation performance in a lay population. The indication is that information processing strategies and cognitive styles relating to lack of imagination may have better predictive potential than ability or even cognitive skills measures. However, one personality characteristic which has not been investigated at all in the present study but which is likely to be highly predictive of film reading performance is motivation. In the present research, both radiographers were highly motivated to succeed at the film reading task, because of the special status this entailed. Similarly, students partaking in the training study possessed a high degree of intrinsic motivation, which is indicated by the zero drop-out rate despite the demanding and time consuming nature of the training. However, it is certainly true to say that not all radiographers are interested in learning to read mammograms or would be equally motivated to do well. It is for this reason that special attention should be paid to intrinsic motivational factors in selecting film readers for breast screening.

12.4.2 Future work on mammogram interpretation skill and its acquisition

There is much scope for further investigation of screening mammogram interpretation as a complex cognitive skill. A most important first step would be to confirm the two stage nature of the skill, which was postulated in the model of screening mammography. Although the performance data in the present research and in particular the data concerning decision outcomes and false negative errors fit the model well, independent corroboration of the structure of the model is required. One way of achieving this might be to demonstrate that there are indeed two separate types of knowledge involved in mammogram interpretation, which are
relatively independently stored in memory. These should consist of visual schemata on the one hand and knowledge about 'symptoms' or features of malignant abnormality in mammograms on the other. Such investigations would need to involve expert participants with fully developed knowledge for obvious reasons.

One problem with investigating conceptual structure and reasoning processes in screening mammography in the present research has been that cognitive processes were difficult to separate out. An example in terms of observing the development of conceptual structure were the verbal reports elicited from the radiographers. These only captured those aspects of performance which were consciously accessible to the participants, i.e. verbally coded diagnostic reasoning processes, which were not necessarily well developed in the radiographers because of their lay status. Important perceptual processes and visual concepts were not accessed by this method. Failure of the cognitive tests to distinguish individual differences at a cognitive process level were probably similarly due to practical issues of separating rapid perceptual processing from memory and decision making processes. It is thus clear that more suitable methods are required to access underlying knowledge structures, in particular at the visual perceptual processing level. Reitman Olson and Rueter (1987) have reviewed a number of methods for knowledge elicitation from experts. One that might be particularly useful in eliciting tacit feature knowledge in mammogram interpretation is the indirect repertory grid method (see Reitman Olson and Rueter (1987) for a more detailed discussion).

Acquisition of mammogram interpretation performance may be a further subject of future research. Lay film readers are perhaps not ideal participants in such an endeavour, because their lack of medical, anatomical and pathological knowledge may lead to performance of the task at a perceptual discrimination level rather than the full interpretation task. It is certainly the case that the radiographers in the present research were performing full mammogram interpretation emulating a radiologist. However, the question is whether radiographers should be encouraged to perform the task at this level without the availability of appropriate training (see
discussion above). Research on skill acquisition should ideally be conducted in a longitudinal framework. The current research has indicated however that skill acquisition in screening mammography proceeds very slowly after initial training. It might, therefore, be better to investigate it in an expert/novice cross-sectional study.

12.4.3 Conclusion

The empirical work presented in this thesis has shown that training lay film readers is feasible. In addition, it has been demonstrated that there is scope for developing selection procedures and for improving training in order to ensure that lay film readers interpret mammograms to the highest standard possible, so that dual lay/consultant film reading can yield the maximum benefit in terms of increased cancer detection. Much of the work conducted for the present research has already led to changes in the implementation of training for lay film readers at the Jarvis Breast Cancer Screening Training Centre. The present research is thus of genuine applied value.

Theoretical issues in mammogram interpretation have been discussed with regard to the nature of the skills involved in reading mammograms. In this context, many questions remain open, leaving scope for further investigation. The notion that mammogram interpretation is a two stage process involving separate visual and interpretative processes has proved a useful concept in explaining some of the more pertinent empirical findings in research on X-ray interpretation. Nevertheless, further research verifying the postulated cognitive processes at a more detailed level is required to give greater credence to the model of mammogram interpretation.
References


316


The Test Agency (1972). MD5 Mental Ability Test. High Wycombe.


Appendix I: Descriptive and Inferential Statistics

Appendix Ia: Means and Standard Deviations for Performance on the Mammogram Test (Radiographers, Radiologists and Students)

Table Ia.1: G-index

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.S.</td>
<td>0.114</td>
<td>0.62</td>
<td>0.696</td>
<td>0.671</td>
<td>0.646</td>
<td>0.57</td>
</tr>
<tr>
<td>F.A.</td>
<td>0.418</td>
<td>0.62</td>
<td>0.646</td>
<td>0.722</td>
<td>0.696</td>
<td>0.646</td>
</tr>
<tr>
<td>Radiologists (n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{x}$</td>
<td>0.544</td>
<td>0.644</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>n/a</td>
<td>n/a</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Tutorial) (n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{x}$</td>
<td>0.21</td>
<td>0.38</td>
<td>0.47</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.15</td>
<td>0.12</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 1) (n=8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{x}$</td>
<td>0.26</td>
<td>0.33</td>
<td>0.45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.39</td>
<td>0.25</td>
<td>0.16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 2) (n=15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{x}$</td>
<td>0.31</td>
<td>0.43</td>
<td>0.45</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.166</td>
<td>0.128</td>
<td>0.166</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Data from Ansell et al (1990b)
Table Ia.2: Probability of True Positives (TP)

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.S.</td>
<td>0.56</td>
<td>0.75</td>
<td>0.94</td>
<td>0.88</td>
<td>0.75</td>
<td>0.94</td>
</tr>
<tr>
<td>F.A.</td>
<td>0.81</td>
<td>0.69</td>
<td>0.75</td>
<td>0.94</td>
<td>0.88</td>
<td>0.63</td>
</tr>
<tr>
<td>Radiologists*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.82</td>
<td>0.83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>n/a</td>
<td>n/a</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Tutorial)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.54</td>
<td>0.61</td>
<td>0.53</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>0.19</td>
<td>0.15</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.57</td>
<td>0.6</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>0.24</td>
<td>0.11</td>
<td>0.19</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Data from Ansell et al (1990b)

Table Ia.3: Probability of False Positives (FP)

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.S.</td>
<td>0.44</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.16</td>
<td>0.25</td>
</tr>
<tr>
<td>F.A.</td>
<td>0.32</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.13</td>
</tr>
<tr>
<td>Radiologists*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.24</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>n/a</td>
<td>n/a</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Tutorial)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.37</td>
<td>0.29</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>0.12</td>
<td>0.09</td>
<td>0.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>0.35</td>
<td>0.32</td>
<td>0.24</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sd</td>
<td>0.28</td>
<td>0.16</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table Ia.4 Perceptions Errors

<table>
<thead>
<tr>
<th>A.S.</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F.A.</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>All Students (n=32)</td>
<td>$\bar{x}$</td>
<td>7.38</td>
<td>6.53</td>
<td>6.72</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>2.56</td>
<td>2.22</td>
<td>2.74</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Tutorial) (n=9)</td>
<td>$\bar{x}$</td>
<td>6.89</td>
<td>6.44</td>
<td>7.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>1.96</td>
<td>2.07</td>
<td>3.61</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 1) (n=8)</td>
<td>$\bar{x}$</td>
<td>6.63</td>
<td>7.25</td>
<td>6.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>2.2</td>
<td>1.67</td>
<td>2.14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 2) (n=15)</td>
<td>$\bar{x}$</td>
<td>8.07</td>
<td>6.20</td>
<td>6.93</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>2.07</td>
<td>2.82</td>
<td>2.35</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table Ia.5: Interpretation Errors

<table>
<thead>
<tr>
<th>A.S.</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
<th>Test 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>F.A.</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>All Students (n=32)</td>
<td>$\bar{x}$</td>
<td>1.38</td>
<td>1.38</td>
<td>1.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>1.15</td>
<td>1.32</td>
<td>1.54</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Tutorial) (n=9)</td>
<td>$\bar{x}$</td>
<td>1.56</td>
<td>1.44</td>
<td>2.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>2.88</td>
<td>1.42</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 1) (n=8)</td>
<td>$\bar{x}$</td>
<td>1.38</td>
<td>1.25</td>
<td>1.88</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>1.6</td>
<td>1.39</td>
<td>2.23</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Students (Feature 2) (n=15)</td>
<td>$\bar{x}$</td>
<td>1.27</td>
<td>1.4</td>
<td>1.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.71</td>
<td>1.8</td>
<td>1.59</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Appendix Ib:
Screening Data

Table Ib.1: Screening Data for Radiographers

<table>
<thead>
<tr>
<th>Screening Interval</th>
<th>A.S.</th>
<th></th>
<th>F.S.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>G</td>
<td>TP</td>
</tr>
<tr>
<td>1</td>
<td>0.39</td>
<td>0.11</td>
<td>0.72</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>0.64</td>
<td>0.12</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>3</td>
<td>0.47</td>
<td>0.08</td>
<td>0.76</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.04</td>
<td>0.83</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>0.43</td>
<td>0.07</td>
<td>0.80</td>
<td>0.43</td>
</tr>
<tr>
<td>6</td>
<td>0.19</td>
<td>0.06</td>
<td>0.81</td>
<td>0.73</td>
</tr>
<tr>
<td>7</td>
<td>0.45</td>
<td>0.04</td>
<td>0.86</td>
<td>0.46</td>
</tr>
<tr>
<td>8</td>
<td>0.64</td>
<td>0.04</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>0.46</td>
<td>0.05</td>
<td>0.82</td>
<td>0.53</td>
</tr>
<tr>
<td>10</td>
<td>0.41</td>
<td>0.05</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>11</td>
<td>0.23</td>
<td>0.07</td>
<td>0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>12</td>
<td>0.37</td>
<td>0.10</td>
<td>0.73</td>
<td>0.31</td>
</tr>
<tr>
<td>13</td>
<td>0.45</td>
<td>0.09</td>
<td>0.72</td>
<td>0.55</td>
</tr>
<tr>
<td>14</td>
<td>0.33</td>
<td>0.12</td>
<td>0.67</td>
<td>0.39</td>
</tr>
<tr>
<td>15</td>
<td>0.54</td>
<td>0.05</td>
<td>0.79</td>
<td>0.50</td>
</tr>
<tr>
<td>16</td>
<td>0.64</td>
<td>0.06</td>
<td>0.84</td>
<td>0.55</td>
</tr>
</tbody>
</table>
### Appendix Ic:
Coordinates of Correspondence Analysis Plots

**Table Ic.1:** Correspondence Analysis - A.S.'s Verbal Protocol Data

<table>
<thead>
<tr>
<th>Category</th>
<th>Coordinate 1</th>
<th>Coordinate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN</td>
<td>-51.62</td>
<td>18.66</td>
</tr>
<tr>
<td>FP</td>
<td>17.61</td>
<td>-15.95</td>
</tr>
<tr>
<td>TP</td>
<td>46.04</td>
<td>49.70</td>
</tr>
<tr>
<td>TN</td>
<td>3.12</td>
<td>-24.72</td>
</tr>
<tr>
<td>Spec</td>
<td>-1.71</td>
<td>-21.15</td>
</tr>
<tr>
<td>Unspec</td>
<td>33.02</td>
<td>-27.62</td>
</tr>
<tr>
<td>Hist</td>
<td>-55.07</td>
<td>21.02</td>
</tr>
<tr>
<td>Normal</td>
<td>139.83</td>
<td>184.76</td>
</tr>
<tr>
<td>Back</td>
<td>86.67</td>
<td>50.59</td>
</tr>
<tr>
<td>Technical</td>
<td>77.92</td>
<td>35.99</td>
</tr>
<tr>
<td>Causal</td>
<td>-13.16</td>
<td>12.18</td>
</tr>
<tr>
<td>Spec 2</td>
<td>-32.26</td>
<td>17.95</td>
</tr>
<tr>
<td>Unspec 2</td>
<td>13.17</td>
<td>-26.08</td>
</tr>
<tr>
<td>Hist 2</td>
<td>-7.04</td>
<td>18.52</td>
</tr>
<tr>
<td>Normal 2</td>
<td>-58.63</td>
<td>35.68</td>
</tr>
<tr>
<td>Back 2</td>
<td>90.69</td>
<td>60.74</td>
</tr>
<tr>
<td>Tech 2</td>
<td>-11.59</td>
<td>42.98</td>
</tr>
<tr>
<td>Causal 2</td>
<td>-4.53</td>
<td>-14.00</td>
</tr>
</tbody>
</table>
Table lc.2: Correspondence Analysis : F.A.'s Verbal Protocol Data

<table>
<thead>
<tr>
<th></th>
<th>Coordinate 1</th>
<th>Coordinate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN</td>
<td>13.10</td>
<td>-20.58</td>
</tr>
<tr>
<td>FP</td>
<td>-9.35</td>
<td>-2.21</td>
</tr>
<tr>
<td>TP</td>
<td>46.37</td>
<td>19.82</td>
</tr>
<tr>
<td>TN</td>
<td>-33.01</td>
<td>10.04</td>
</tr>
<tr>
<td>Spec</td>
<td>-33.64</td>
<td>10.70</td>
</tr>
<tr>
<td>Unspec</td>
<td>4.25</td>
<td>-29.85</td>
</tr>
<tr>
<td>Hist</td>
<td>-41.78</td>
<td>23.90</td>
</tr>
<tr>
<td>Normal</td>
<td>39.05</td>
<td>6.31</td>
</tr>
<tr>
<td>Back</td>
<td>10.09</td>
<td>11.84</td>
</tr>
<tr>
<td>Tech</td>
<td>11.31</td>
<td>-2.30</td>
</tr>
<tr>
<td>Causal</td>
<td>-26.45</td>
<td>-4.22</td>
</tr>
<tr>
<td>Spec 2</td>
<td>-34.06</td>
<td>12.38</td>
</tr>
<tr>
<td>Unspec 2</td>
<td>-3.99</td>
<td>-23.56</td>
</tr>
<tr>
<td>Hist 2</td>
<td>8.76</td>
<td>-57.88</td>
</tr>
<tr>
<td>Normal 2</td>
<td>32.19</td>
<td>-1.40</td>
</tr>
<tr>
<td>Back 2</td>
<td>37.58</td>
<td>13.97</td>
</tr>
<tr>
<td>Tech 2</td>
<td>-13.05</td>
<td>16.24</td>
</tr>
<tr>
<td>Causal 2</td>
<td>-14.80</td>
<td>-22.07</td>
</tr>
</tbody>
</table>
Table Ic.3: Correspondence Analysis- Predictors of Individual Differences in Interpretation of Mammograms

<table>
<thead>
<tr>
<th></th>
<th>Coordinate 1</th>
<th>Coordinate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>9.72</td>
<td>-2.11</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.79</td>
<td>3.85</td>
</tr>
<tr>
<td>Good</td>
<td>-10.49</td>
<td>-1.67</td>
</tr>
<tr>
<td>SAT-2D</td>
<td>-4.23</td>
<td>-1.79</td>
</tr>
<tr>
<td>KAI-E</td>
<td>-1.72</td>
<td>3.08</td>
</tr>
<tr>
<td>KAI-O</td>
<td>9.26</td>
<td>-1.46</td>
</tr>
<tr>
<td>EPQ-N</td>
<td>-18.47</td>
<td>-7.56</td>
</tr>
<tr>
<td>CS</td>
<td>-6.73</td>
<td>1.80</td>
</tr>
<tr>
<td>G-error</td>
<td>6.59</td>
<td>0.69</td>
</tr>
<tr>
<td>Strat-F</td>
<td>-3.89</td>
<td>0.50</td>
</tr>
<tr>
<td>Strat-BB</td>
<td>34.53</td>
<td>17.91</td>
</tr>
</tbody>
</table>
Appendix Id:  
Analysis of Variance of Student Training Data with Arcsine Transformation

**Table Id.1:** Analysis of Variance table for overall film reading performance measured with the G-index, which was normalised by arcsine transformation.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>18.64</td>
<td>1</td>
<td>18.64</td>
<td>1.02</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>273.62</td>
<td>15</td>
<td>18.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>54.56</td>
<td>2</td>
<td>27.28</td>
<td>6.95</td>
<td>&lt;.005</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>11.48</td>
<td>2</td>
<td>5.74</td>
<td>1.46</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>117.75</td>
<td>30</td>
<td>3.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table Id.2:** Analysis of Variance table for true positive (hit) classification performance, which was normalised by arcsine transformation.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>12.52</td>
<td>1</td>
<td>12.52</td>
<td>0.95</td>
<td>n.s</td>
</tr>
<tr>
<td>Residual</td>
<td>197.71</td>
<td>15</td>
<td>13.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>8.86</td>
<td>2</td>
<td>4.43</td>
<td>1.11</td>
<td>n.s</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>7.34</td>
<td>2</td>
<td>3.67</td>
<td>0.92</td>
<td>n.s</td>
</tr>
<tr>
<td>Residual</td>
<td>119.75</td>
<td>30</td>
<td>3.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table Id.3: Analysis of Variance table for false positive (false recall) classification performance, which was normalised by arcsine transformation.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>11.64</td>
<td>1</td>
<td>11.64</td>
<td>0.91</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>191.85</td>
<td>15</td>
<td>12.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>74.76</td>
<td>2</td>
<td>37.38</td>
<td>6.21</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Train x Ass</td>
<td>12.76</td>
<td>2</td>
<td>6.38</td>
<td>1.06</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>180.62</td>
<td>30</td>
<td>6.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table Id.4: Three oneway analyses of variance for overall performance (G-index), false positives and true negatives as a function of feature training (n = 23) over the three assessments. Normalised by arcsine transformation.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-index</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>110.56</td>
<td>2</td>
<td>55.28</td>
<td>6.81</td>
<td>&lt; .005</td>
</tr>
<tr>
<td>Residual</td>
<td>357.73</td>
<td>44</td>
<td>8.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>52.52</td>
<td>2</td>
<td>26.26</td>
<td>5.21</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Residual</td>
<td>225.72</td>
<td>44</td>
<td>5.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td>43.98</td>
<td>2</td>
<td>21.99</td>
<td>2.64</td>
<td>n.s.</td>
</tr>
<tr>
<td>Residual</td>
<td>366.96</td>
<td>44</td>
<td>8.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix II:
Student Training Protocol

Appendix IIa:
Introductory Lecture

Introductory Lecture attended by all trainees (n = 32) on the first training day prior to completing the first mammogram test (pre-training test)

In Britain, there has been a commitment to providing mammographic breast screening for all women between the ages of 50 and 65 in recent years (Forrest, 1986). The means that these women receive X-rays of both breasts which are then evaluated for abnormality. If an abnormality is identified the woman in question is recalled and further investigations are undertaken, i.e. further mammograms are done, clinical examination for lumps, ultrasound examination and possibly fine needle biopsies. The incidence of breast cancer in the screening population is about 0.7%, i.e. 140 cases in the Jarvis screening centre which has a screening population of 20000 women per year. The recall rate for mammographic abnormality varies between 5% and 10% in different screening centres. So many more women are assessed for mammographic abnormality on the initial screening mammograms than turn out to have a carcinoma. In order to ensure the feasibility of the screening programme, though, the aim is to detect as many abnormalities as early as possible while keeping false recalls to a minimum, because of both costs to the NHS and costs to the women in terms of unnecessary anxiety.

A mammogram is a soft tissue X-ray, i.e. there are no bone structures and in general very little 'fixed anatomy'. It is also a two-dimensional representation of a three-dimensional object which means that structures can overlap. Sometimes abnormalities are very obvious, other are very subtle. This depends partly on the breast background, partly on the type of abnormality and partly on technical (X-ray-specific) factors. In the remainder of this talk we will therefore look at basic breast anatomy, how this is translated into X-ray images, basic types of abnormality and mammogram viewing techniques.

[The section on breast anatomy is taken in its entirety from Tabar's Teaching Atlas of Mammography (Tabar and Dean, 1983) and illustrated with schematic drawings taken from the atlas (Figures 1 and 2)]

Anatomically the breast can be subdivided into the following structural entities:

LOBE (Fig 1): The human breast contains 15-18 lobes. Each lobe has a main duct opening into the nipple.
TERMINAL DUCT LOBULAR UNIT (Fig 2): The main duct branches and eventually forms the TDLU consisting of the extralobular terminal duct and the lobule.

LOBULE: The intralobular terminal duct and ductules surrounded by a special, loose intralobular connective tissue form a lobule (Fig 2) [...].

Screening mammograms at the Jarvis screening centre are taken at an oblique projection. For this the breast is compressed at a 45° angle [Figure 12, showing mammogram being taken, from Andersson (1986) is used for illustration]. This results in an image which shows the pectoral muscle lying diagonally at the back of the breast to about mid breast area. The inframammary skin fold can be seen underneath the breast. This projection best images the ancillary tail. Usually the image contains a denser (whitish area) in the central part of the breast. This constitutes connective tissues. In older women this type of tissue is progressively replaced by fatty tissue which is radiolucent (greyish and see-through). This is one of the reasons for offering screening to women over 50 only. Cancer is more frequent in that age group but also there is a greater chance of detecting abnormality as the breasts tend to be more fatty. There are individual differences though and one generally distinguishes fatty, mixed fatty and dense and dense breasts [examples are shown on slides]. Fatty breasts have an overall grey appearance with many curved white lines that represent supporting framework of the breast. Dense breasts contain ill-defined areas of increased density which appear whiter than the general background. This may be patchy (mixed dense and fatty) or cover the whole breast. It is often more marked behind the nipple and in the axillary tail. It may appear in small 'nodules' 2-3 mm or larger which are often merging into each other (cotton wool effect). The structures seen in mammograms are lines which tend to converge towards the nipple. Thin white crisp ones are trabeculae, medium grey, less well-defined worm-like ones (1 mm diameter) are ducts. Blood vessels can sometimes be seen especially when calcified (whitish thicker lines).

Abnormalities in mammograms can be classified broadly into three types:

MASSES are discrete areas of density greater than the background breast pattern. These may have smooth or irregular edges. If a mass of more than 5 mm is seen it needs further analysis. In general, irregular masses tend to be very indicative of cancer, whereas rounded masses tend represent benign processes such as cysts.

CALCIFICATIONS are small white flecks sometimes seen on mammograms. They can be macrocalcifications, which are dense mainly spherical, usually well-defined with a crisp outline. Microcalcifications are less dense and sometimes irregular in outline. You may need a magnifying glass to identify them. It is only microcalcifications that are indicative of malignant changes.
Other abnormalities are indicated by MINIMAL features, e.g. distortion, a disturbance in the architectural pattern of the breast, or by asymmetry, an unequal distribution of 'white fluff' between the two breasts.

When viewing mammograms it may be worthwhile to adopt a structured method. Calcifications can often only be seen with a magnifying glass. Small masses and asymmetries may be detected more easily if the right and left breasts are viewed systematically e.g. by sequentially viewing restricted areas of each side. Also make sure you look at both breasts as they may both contain an abnormality.

The Test roller (instructions for completing the mammograms test)

The mammogram test represents the task a mammography screener faces on a day to day basis. The screening mammograms come in sets of 80 to 100 in one session. The tests replicate this situation with one exception. In a normal screening roller you would expect to find 3 or 4 abnormalities worth recalling, in this test there are about 20% abnormal mammograms, although you may wish to recall more so as to not miss any. When you look at the test mammograms, please, do not spend too long on one mammogram. There are 79 sets to get through, and you should spend about 1 or 2 minutes on each, including writing down your decision. First make a decision on whether to recall or not. You have a choice between definitely recalling (1), maybe recalling (2) (e.g. checking with another screener first), maybe not recalling (3) and definitely not recalling (4). The record sheet requires you to report on any abnormalities detected. Please draw into the schematic breasts everything that you think represents an abnormality. Also tick what categories it fits into (mass, calcification or other (specify)). If there is more than one, please, indicate in the comments section which of the abnormalities noted you are recalling for and whether it is the right or left breast you are recalling.
Appendix IIb:
Feature Training Lectures

Training protocol for students in the feature training groups (1st training study n=8, 2nd training study n=15)

Day 2 Lecture: Masses

A mass is a discrete area of density appearing whiter than the general background. It may simply be more noticeable than the rest of the breast tissue and stand out in some way. If such a mass is 5 mm or over it needs further feature analysis.

Primary features

1. Shape

Masses can be
   a) rounded
   b) irregular (tends to be stellate)

2. Contour

This can be
   a) sharp (encapsulated, well-defined or circumscribed by a fine radiolucent 'halo')
   b) lobulated or fuzzy-edged for rounded masses
   or c) spiky for irregular masses
      - sharp dense fine lines of variable length radiating in all directions
      [Fig from Tabar and Dean, 1983]

As a general rule, completely sharply outlined round masses tend to represent benign processes such as cysts, lymph nodes etc. (don't need recalling - see secondary features, though). Masses with radiating spicules are very likely to be carcinoma. The larger the central tumour mass, the longer the radiating spicules.

Secondary features for rounded masses

3. Density

Masses can be
   a) radiolucent
   b) low density (surrounding structures (e.g. veins trabeculae etc) can be seen through the lesion)
   c) high density (more dense than surrounding breast tissue).

The denser an abnormality is the more likely that it is cancerous.
4. Location

a) Axilla
Oval circumscribed masses are often found in the axilla. These represent lymph nodes. Abnormal lymph nodes are larger (> 1cm) and round. These may represent pathology anywhere in the body, not necessarily in the breast.

b) Subcutaneous tissue
Tumours with an epicentre in this region tend to be benign.

c) Subareolar region
This is where the bulk of the fibroglandular tissue is situated. Rounded lesions within this area tend to be benign. Rounded lesions outside this area (in the more fatty regions) are suspicious even if well circumscribed.

5. Number

a) multiple circumscribed masses

b) single circumscribed mass

More than one rounded mass tends to be benign processes such as cysts or lymph nodes.
Calcifications are small white flecks seen on mammograms. Macro calcification is dense and usually has a smooth crisp outline. It is well-defined and is mostly spherical. It is larger than 1 mm. Microcalcification is less dense and may have an irregular contour. It is smaller than 1 mm and you may need a magnifying glass to identify it. They often appear in clusters.

**Primary features**

1. **Form**

Calcifications can have

a) a granular form
   tiny dot like or elongated, innumerable and irregularly grouped very closely together in an area of the breast, resembling fine grains of salt.
   irregular in form and size and density grouped very closely in one area of the breast. Resembles granulated sugar.

b) casting
   casts of segments of the duct (elongated), sometimes filling in the branch. They are irregular in contour.

c) arising within a cavity
   homogenous, solid, sharply outlined, spherical pearl-like density or crescent-shaped.

   d) punctate
      very fine powder-like, barely perceptible

The first two and the last are likely to be malignant while the third type is benign.

2. **Size**

This can be

a) variable

b) homogenous

Calcifications uniform in size tend to be benign.
3. Density

This can be a) variable between particles

b) uniformly dense

The former is an indication of granular or casting type calcification, whereas the later is the case for small pearl-like densities. If they fill a larger cavity they can be more variable though.

4. Number

This can be a) innumerable

b) few particles

Casting type calcification can sometimes consist of very few particles. More often malignant type calcification (especially granular) appears in innumerable particles.

5. Distribution

Particles can be a) clustered within one area

b) scattered

Clustered calcification especially if associated with a mass or soft tissue component is likely to be malignant. If the calcifications are distributed throughout most or much of the breast it is likely to be benign.
Day 4 Lecture: Other (minimal) Features

Sometimes a mammographic abnormality is neither indicated by a definite mass or a calcification. These abnormalities are of two types:

a) Asymmetry
This is indicated by an unequal distribution of 'white fluff' between the two breasts. Usually both breasts are similar but not quite mirror images of each other.

b) Distortion
This is indicated by a disturbance of the orientation of normal breast pattern lines. These lines seem to converge towards a central point or density.

Features of asymmetry

Asymmetry may be due to

a) a cancerous abnormality
In this case, the same features apply as to masses, i.e. density in relation to the general background, shape, contour etc.

b) faulty positioning of the breast when the X-ray is taken.
Check that both breast are positioned in the same way. If the pectoral muscles are at different angles, or there are skin creases, this may be the cause of especially slight asymmetry.

c) uneven involution of the two breasts
When the breast tissue becomes progressively replaced by fatty 'see through' tissue this process may not be entirely even in both breasts, so that there is more 'white fluff' in one breast than the other. Again look for mass-like features to confirm an abnormality.

d) positional shadow
This is when 2 or 3 linear structures apparently merge or converge towards a point. The lines need to be followed around fatty lobules through the area of concern. There should not be any soft tissue density associated with a positional shadow.
Features of distortion

a) a pattern disturbance in the curvy lines that go towards the nipple.

b) nipple or skin retraction
if you note retraction in the nipple or skin area (indrawing of tissue), this could support a finding of distortion.

c) positional shadow
(see asymmetry above)
Appendix III:
The Cognitive Task Battery

Appendix IIIa:
Program Listing for 'Target'
A-20
PROCIntroScr6  PROCWaitKey
ENDPROC

REM Give Instructions for three Targets

CLS VDU 5
REPEAT PR0CDemoScr2L GCOL 63 TINT 192
MOVE 650,100:PRINT "Press C to Continue, 
MOVE 650,50:PRINT "R to Repeat"
REPEAT PROCWaitKey UNTIL (A$ = "C" OR A$ = "R" OR A$ = "r") UNTIL (A$ = "C" OR A$ = "c")
REPEATPROCDemoScr8L

GCOL 63 TINT 192
MOVE 650,100:PRINT "Press C to Continue, 
MOVE 650,50:PRINT "R to Repeat"
REPEAT PROCWaitKey UNTIL (A$ = "C" OR A$ = "r") UNTIL (A$ = "C" OR A$ = "c")

PROCIntroScr7 PROCWaitKey
ENDPROC

REM Set up the array of percentages for the trial
FOR J = 1 TO NoTrial% / 2: PercArray% (J) = 0: NEXT J
FOR J = 1 TO NoPerc% FOR K = 1 TO NoPerc% Trial
REPEAT L = RND (NoTrials% / 2) UNTIL PercArray% (L) = 0
PercArray% (L) = J
NEXT K
NEXT J

ENDPROC

DEFPROCFixate (FixTime%)
REM Display the Fixation Screen
GCOL 128 + BasoCol% COLOUR 128 + BasoCol% CLSCOLOUR 60 GCOL 0
RECTANGLE FILL 590,507,100,10 RECTANGLE FILL 635,462,10,100 PROCWait (FixTime%)
ENDPROC

DEFPROCDisplayBack (BackCond%, BackNo%)
REM Display the Background Sprite
CLS IF BackCond% = 1 OR BackCond% = 3 THEN
BACKCol% = 2 ELSE
BACKCol% = 8 ENDIF
REM Tell the subject the result and record reaction times
REM Switch off Cursor
VDU 5
COLOUR BaseCol%
CLS
REM Switch on Graphic Printing
IF SubjNo% MOD 2 = 0 THEN Yes$ = "Z": No$ = "/" ELSE Yes$ = "/": No$ = "Z" ENDIF
GCOL 40: RECTANGLE FILL 39,30,1200,100 GCOL 0: RECTANGLE 39,30,1200,100
MOVE 500, 90: PRINT "WRONG"
TrialArray (Trial% + 6) = 0 SOUND 2,-15,100,50 ELSE
GCOL 7: RECTANGLE FILL 39,30,1200,100 GCOL 0: RECTANGLE 39,30,1200,100
MOVE 500, 90: PRINT "CORRECT"
ENDPROC
REM **************************************************
REM read a system variable
REM **************************************************
SYS "OS_ReadVarVal",SysVarName%,SysVarValue%,100,,3 TO ,,ValueLen%S
IF ValueLen%>0 THEN SysVarValue%=ValueLen%:SysVarValue%=ValueLen%
ENDPROC

REM **************************************************
REM Display the Background Screen
REM **************************************************
OSCLI( •  *SLOAD <Obey$Dir>.  Sprites.  "  +ScrName$)OSCLI( •  *SCHOOSE "  +ScrName$)

REM First Instruction Screen
REM **************************************************
DEFPROCIntroScr1

PROCScreenWrite(2):PROCScreenDisplay(1)VDU 5
MODE 13
CLS

PROCSScreenWrite(2):PROCSScreenDisplay(1)

FOR J = 3 TO 1 STEP -1
MOVE PlotX%, DemoPlotY% (J)
MOVE PlotX%+((1280/320)*30),DemoPlotY%(J)+((1024/256)*30)
OSCLI(  "*SGET SPRTEMP"+STR$ (J) )SName$(J)="TS8P"+STR$ (PercValue% (J))="N1"
OSCLI(  "*SMERGE <Obey$Dir>.Sprites."  +SName$ (J) )
OSCLI(  "  *SCHOOSE "  +SName$ (J) )
OSCLI(  "  *SCHOOSE "  +SName$ (J) )
PLOT AND,Pix2%,DemoPlotY%(J)
NEXT J NEXT I

REM **************************************************
REM Demo Screen - 2 Colour Sprites Small
REM **************************************************
DEFPROCDemoScr2

PROCSScreenWrite(2):PROCSScreenDisplay(1)

FOR J = 3 TO 1 STEP -1
NEXT J

REM **************************************************
REM Demo Screen - 8 Colour Sprites Small
REM **************************************************
DEFPROMODemoScr8

REM **************************************************
REM Demo Screen - 4 Colour Sprites Small
REM **************************************************
DEFPROMODemoScr4

REM **************************************************
REM Demo Screen - 1 Colour Sprite Small
REM **************************************************
DEFPROMODemoScr1
FOR I = 1 TO 3
FOR J = 1 TO 3
OSCLI("*SCHOOSE SPRTEMP*"+STR$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)
NEXT J
PROCWait(50)
FOR J = 1 TO 3
OSCLI("*SCHOOSE*"+SNorae$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)
NEXT J
PROCWait(50)
NEXT I

♦SNOW
ENDFROM

ENDSCREEN

DEFPROCDemoScr2L
PROCScreenWrite(2):PROCScreenDisplay(1) PROCScrLoad("TBK21")
GCOL 0 TINT 0
RECTANGLE FILL 640,0,640,1024
DemoPlotY%(1)=200 DemoPlotY%(2)=450 DemoPlotY%(3)=700 PlotX%=210
GCOL 63 TINT 192
MOVE 700,100:PRINT "You are now required to identify trials where a large target is present"
MOVE 700,95:PRINT "A small target is no longor valid"
MOVE 700,90:PRINT "Targets on a two shade background can look like any of those three examples"
PROCScreenWrite(2):PROCScreenDisplay(2)
PROCWait(200)
FOR J = 3 TO 1 STEP -1
MOVE PlotX%,DemoPlotY%(J)
MOVE PlotX%+((1280/320)*40),DemoPlotY%(J)+((1024/256)*48) OSCLI("*SGET SPRTEMP*"+STR$ (J)) SName?(J)="TL2P"+STR$(PercValue%(J))="N1"
OSCLI("*SMERGE <Obey$Dir>.Sprites."+SName$ (J)) OSCLI("*SCHOOSE*"+SName$ (J))
GCOL 8,0
PLOT &ED,PlotX%,DemoPlotY%(J)PROCWait(100)
FOR I = 1 TO 3
OSCLI("*SCHOOSE SPRTEMP*"+STR$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)PROCWait(50)
OSCLI("*SCHOOSE*"+SName$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)PROCWait(50)
NEXT I
NEXT J
PROCWait(100)
FOR I = 1 TO 3
FOR J = 1 TO 3
OSCLI("*SCHOOSE SPRTEMP*"+STR$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)
NEXT J
PROCWait(50)
FOR J = 1 TO 3
OSCLI("*SCHOOSE*"+SNorae$ (J)) PLOT &ED,PlotX%,DemoPlotY%(J)
NEXT J
PROCWait(50)
NEXT I
♦SNOW
ENDFROM

DEFPROCIntroScr3
CLS
GCOL 63 TINT 192
MOVE 700,500:PRINT "During the task you are required to press a key indicating whether or not you thought the target was present"
PROCWait(600)
MOVE 150,450:PRINT "A box at the bottom of the screen will detail which keys to press"
PROCShowFrompt
PROCWait(800)
ENDFROM

DEFPROCIntroScr4
CLS
GCOL 63 TINT 192
MOVE 150,950:PRINT "During the task you are required to press a key indicating whether or not you thought the target was present"
PROCWait(600)
MOVE 150,450:PRINT "A box at the bottom of the screen will detail which keys to press"
PROCShowFrompt
PROCWait(800)
ENDFROM

FUNCTION

REM * Demo Screen - 2 Colour Sprites Large

REM ******************************************************
REM *  Demo Screen - 8 Colour Sprites Large

REM ******************************************************
REM * Introduction Screen Three - Response and Feedback *

REM ******************************************************
REM * Introduction Screen Four - Proceed to trial? 

REM ******************************************************
REM * Screen - 8 Colour Sprites Large *

REM ******************************************************
You can now practice soma trials before starting or repeat the instructions. Each of the practice trials has a target present. Press P to Practice or R to Repeat the Instructions.

You are about to resume the main task. Press any key to continue.

You are now ready to start the trials. Press S to Start or P to Practice again or R to Repeat the Instructions.

You are about to start the main task. Press any key to continue.

You are now about to start the main task. Press any key to start.

You have finished all trials. Thankyou for taking part.

You are now about to start the main task. Press any key to continue.
Appendix IIIb:
Program Listing for 'Concept'

```vbnet
REM >>> YesNo4
REM ********************************************************** REM
REM Global Variables
REM ********************************************************** EntryNo%=0: NoTrialsft=99: Para eOK = FALSE TestNo%=0
DIM AnsTest (5)
DIM RuleArray (80)
DIM AnoArray (200)
DIM SysVarName% 100, SysVarValue% 100
PROCInitialise
PROCConclude
END

REM **********************************************************REM * Initialise: Set up for trial *REM ********************************************************** DEFPROCInitialise
MODE 15
ResultX%=0: ResultY%=0
FOR N=1 TO 2000 -.AnsArray (N) =0 :NEXT N
COLOUR 128+57: COLOUR 46 TINT 192
CLS
ON ERROR PROCError: END
ResultDir$=FNrGadvar("YesNo$RooDir")
REPEAT
PRINT "  INPUT " Subject Name ", SubjName? UNTIL SubjName? <> "*
REPEAT
PRINT "  INPUT " Subject Number ", SubjNo% UNTIL SubjNo% > 0
ResFile?=ResultDir? + "YSUBJ-+STR?(SubjNo%)
test=OPENIN(RoaFile$)
IF test<>0 THEN CLOSE#0: PRINT " Subject already tested 
UNTIL test=0
CLS
PROC LoadArray
ENDPROC

REM **********************************************************REM * Process: Administer the Task *REM ********************************************************** DEFPROCProcess
PROCRunTrial
FOR TestNo%= 1 TO 4
PRINT VAR (27,4) " PRACTICE RUN"
PRINT VAR (27,5) "------------------------
PRINT VAR (15,14) " This is a practice run prior to the main task."
PRINT VAR (15,15) " Please make sure that you have read the instructions."
PRINT VAR (15,15) " before continuing."
PRINT VAR (25,25) " PRESS ANY KEY TO START"
PROCWaitKey
CLS
PROCConclude
ENDFOR

REM ************************************************************************REM * RunTrial: Run the program trials *REM ************************************************************************DEFPROCRunTrial
TrialNo%=0 RuleTrue=FALSE ResultX%=0: ResultY%=0
PROCProportFirst
REPEAT
TrialNo%+=1 ParseOK=FALSE TIME=0
REPEAT
PROCInputBox
PROCInput ParseOK=TRUE 
UNTIL ParseOK = TRUE
PRINT VAR (27,7) " THANK YOU"
PRINT VAR (27,8) "-------
PRINT VAR (15,15) " Thankyou for taking part in this experiment."
PRINT VAR (25,25) " PRESS ANY KEY TO EXIT"
PROCWaitKey
ENDPROC
```

A-26
REM ********************************************************** REM *  Paraelnput :  Decode tho input string *REM **********************************************************
DEFPROC Paro oInput

REM Strip out leading spaces
WHILE LEFT? (Answer?,1) = " 
Answer? = RIGHT? (Answer?, LEN (Answer?) -1 )
ENDWHILE

CASE LEFT? (Answer?,1) OF
WHEN "1" : PROCDecodeRule
WHEN "2" : PROCDecodeRule
WHEN "3" : PROCDecodeRule
WHEN "4" : PROCDecodeRule
WHEN "5" : PROCDecodeRule
WHEN "0" , "o" : PROCDecodeTest
WHEN "X" , "x" : PROCDecodeTest
OTHERWISE ParseOK = FALSE:PROCInvalidBox { "Invalid Input")
ENDCASE

ENDPROC

REM ********************************************************** REM *  DecodeRule :  Decode the input rule *REM **********************************************************
DEFPROC DecodeRule

Rule=TRUE:Teat=FALSE

Nurnl?=LEFT? (Answer? ,1)
Nural = VAL (Numl?)
Answer? =RIGHT?(Answer?,LEN(Answer?)-1)

WHILE LEFT? (Answer?, 1) = " 
Answer?=RIGHT?(Answer?, LEN(Answer?) -1 )
ENDWHILE

A2? = LEFT?(Answer?,2)
A3? = LEFT?(Answer?,3)

IF ( A2? <> "or") AND ( A2? <> "OR") AND ( A3? <> "and") AND ( A3? <> "AND") THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ENDIF

IF ( A2? = "or" )  OR ( A2? = "OR" )  THEN
OrRule = TRUE AndRule = FALSE
Answer?=RIGHT? (Answer?, LEN (Answer?)  - 3)
ELSE
OrRule = FALSE AndRule = TRUE
Answer? =RIGHT? (Answer?, LEN (Answer?)  - 3)
ENDIF

WHILE LEFT? (Answer?,1) = " 
Answer?=RIGHT? (Answer?, LEN (Answer?)  - 1)
ENDWHILE

Num2?=LEFT? (Answer?,1)
N

UIb

2  = VAL(Num2?)

IF Num2 < 1 OR Num2 > 5 THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ELSE ParseOK = TRUE
ENDIF

IF Numl = Num2 THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ELSE ParseOK = TRUE
ENDIF

ENDIF

ENDPROC

REM ********************************************************** REM *  Paro oInput :  Decode the input string *REM **********************************************************
DEFPROC Paro oInput

REM Strip out leading spaces
WHILE LEFT? (Answer?,1) = " 
Answer? = RIGHT? (Answer?, LEN (Answer?) -1 )
ENDWHILE

CASE LEFT? (Answer?,1) OF
WHEN "1" : PROCDecodeRule
WHEN "2" : PROCDecodeRule
WHEN "3" : PROCDecodeRule
WHEN "4" : PROCDecodeRule
WHEN "5" : PROCDecodeRule
WHEN "0" , "o" : PROCDecodeTest
WHEN "X" , "x" : PROCDecodeTest
OTHERWISE ParseOK = FALSE:PROCInvalidBox { "Invalid Input")
ENDCASE

ENDPROC

REM ********************************************************** REM *  DecodeRule :  Decode the input rule *REM **********************************************************
DEFPROC DecodeRule

Rule=TRUE:Teat=FALSE

Nurnl?=LEFT? (Answer? ,1)
Nural = VAL (Numl?)
Answer? =RIGHT?(Answer?,LEN(Answer?)-1)

WHILE LEFT? (Answer?, 1) = " 
Answer?=RIGHT?(Answer?, LEN(Answer?) -1 )
ENDWHILE

A2? = LEFT?(Answer?,2)
A3? = LEFT?(Answer?,3)

IF ( A2? <> "or") AND ( A2? <> "OR") AND ( A3? <> "and") AND ( A3? <> "AND") THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ENDIF

IF ( A2? = "or" )  OR ( A2? = "OR" )  THEN
OrRule = TRUE AndRule = FALSE
Answer?=RIGHT? (Answer?, LEN (Answer?)  - 3)
ELSE
OrRule = FALSE AndRule = TRUE
Answer? =RIGHT? (Answer?, LEN (Answer?)  - 3)
ENDIF

WHILE LEFT? (Answer?,1) = " 
Answer?=RIGHT? (Answer?, LEN (Answer?)  - 1)
ENDWHILE

Num2?=LEFT? (Answer?,1)
N

UIb

2  = VAL(Num2?)

IF Num2 < 1 OR Num2 > 5 THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ELSE ParseOK = TRUE
ENDIF

IF Numl = Num2 THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ELSE ParseOK = TRUE
ENDIF

ENDIF

ENDPROC

REM ********************************************************** REM *  DecodeRule :  Decode the input test *REM **********************************************************
DEFPROC DecodeTest

Rule=FALSE:Teat=TRUE

IF LEN (Answer?) <> 5 THEN ParseOK = FALSE:PROCInvalidBox ("Invalid Input")
ENDIF

FOR J s 1 TO 5  CASE MID? ( Answer?,J,1) OF
WHEN "0","o" :  AnoTest ( J) =0
WHEN "X" ,  "x" :  AnoTest ( J) =1
OTHERWISE ParseOK = FALSE:PROCInvalidBox("Invalid Input")
ENDCASE

ParseOK = TRUE

ENDPROC

REM ********************************************************** REM *  WaitKey :  Wait for a key to be got *REM **********************************************************
DEFPROC WaitKey

REPEAT:  A?=GET?:  UNTIL A? <> " 

ENDPROC

REM ********************************************************** REM *  EvaluateTost :  Opt for yes or no *REM **********************************************************
DEFPROC EvaluateTest

PROCCountYes PROCCountNo

IF YesCount = NoCount THEN Rnum=RND(2)REM PRINT "Rand "  ;Rnum IF Rnuro=1 THEN PROCSetYes ELSE PROCSetNo ENDIF
ENDIF

IF YesCount > NoCount THEN PROCSetYes ELSE PROCSetNo ENDIF

ENDPROC

REM ********************************************************** REM *  EvaluateRule :  Opt for yes or no *REM **********************************************************
DEFPROC EvaluateRule

PROCCountYes PROCCountNo

IF YesCount = NoCount THEN Rnum=RND(2)REM PRINT "Rand "  ;Rnum IF Rnuro=1 THEN PROCSetYes ELSE PROCSetNo ENDIF
ENDIF

IF YesCount > NoCount THEN PROCSetYes ELSE PROCSetNo ENDIF

ENDPROC

A-27
DEFPROC EvaluateRule
NoRulesCount = 0
MatchRule = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 3 THEN
IF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
AndRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
ELSE
IF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
NEXT J
REM PRINT NoRulesCount, MatchRule
IF (MatchRule > 0) AND (NoRulesCount = 1) AND (RuleArray(MatchRule+3) = 0) THEN
PROCValidBox("TRUE") RuleTrue = TRUE
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
DONPROC
REM *********************************************************
REM * LoadArray : Load the array of possible rules *
REM *********************************************************
DEFPROC LoadArray
RESTORE
FOR J = 1 TO 80 STEP 4
READ RuleArray(J)
READ RuleArray(J+1)
READ RuleArray(J+2)
RuleArray(J+3) = 0
NEXT J
REM Number 1, Number 2, AND/OR AND = 1 OR = 2
DATA 1,2,1
DATA 1,2,2
DATA 1,3,1
DATA 1,3,2
DATA 1,4,1
DATA 1,4,2
DATA 2,1,1
DATA 2,1,2
DATA 2,2,1
DATA 2,2,2
DATA 3,1,1
DATA 3,1,2
DATA 3,2,1
DATA 3,2,2
DATA 4,1,1
DATA 4,1,2
DATA 4,2,1
DATA 4,2,2
ENDPROC
REM *********************************************************
REM * CountYes : Count the Number for yes *
REM *********************************************************
DEFPROC CountYes
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
IF AnsTest(RuleArray(J)) = 1 AND AnsTest(RuleArray(J+1)) = 1
THEN
RuleArray(J+3) = 0 THEN
YesCount = 1
ENDIF
ENDIF
ELSE
IF AnsTest(RuleArray(J)) = 0 OR AnsTest(RuleArray(J+1)) = 0
THEN
IF RuleArray(J+3) = 0 THEN
YesCount = 1
ENDIF
ENDIF
NEXT J
REM PRINT "YesCount = ";YesCount
ENDPROC
REM *********************************************************
REM * CountNo : Count the Number for no *
REM *********************************************************
DEFPROC CountNo
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
IF AnsTest(RuleArray(J)) = 0 OR AnsTest(RuleArray(J+1)) = 0
THEN
IF RuleArray(J+3) = 0 THEN
NoCount = 1
ENDIF
ENDIF
ELSE
IF AnsTest(RuleArray(J)) = 1 OR AnsTest(RuleArray(J+1)) = 1
THEN
IF RuleArray(J+3) = 0 THEN
NoCount = 1
ENDIF
ENDIF
ELSE
IF AnsTest(RuleArray(J)) = 0 AND AnsTest(RuleArray(J+1)) = 0
THEN
IF RuleArray(J+3) = 0 THEN
NoCount = 1
ENDIF
ENDIF
NEXT J
REM PRINT "NoCount = ";NoCount
ENDPROC
REM *********************************************************
REM * SetYes : set the Number for yes *
REM *********************************************************
DEFPROC SetYes
PROCValidBox("NO")
NOTest = TRUE: Yestest = FALSE
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
IF AnsTest(RuleArray(J)) = 1 AND AnsTest(RuleArray(J+1)) = 1
THEN
RuleArray(J+3) = 0 THEN
ENDIF
ENDIF
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
NEXT J
REM PRINT "Yestest = ";Yestest
ENDPROC
REM *********************************************************
REM * SetNo : set the Number for no *
REM *********************************************************
DEFPROC SetNo
PROCValidBox("YES")
Yestest = TRUE: NOTest = FALSE
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
IF AnsTest(RuleArray(J)) = 1 AND AnsTest(RuleArray(J+1)) = 1
THEN
IF RuleArray(J+3) = 0 THEN
RuleArray(J+3) = 1
ENDIF
ENDIF
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
ELSEIF (Num1 = RuleArray(J)) AND (Num2 = RuleArray(J+1)) AND
OrRule = TRUE THEN
MatchRule = J
ENDIF
ENDIF
NEXT J
REM PRINT "Notest = ";Notest
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for yes *
REM *********************************************************
DEFPROC CountTheNumber
YesCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
YES
ENDIF
ELSE
NO
ENDIF
ENDPROC
REM *********************************************************
REM * CountTheNumber : Count the Number for no *
REM *********************************************************
DEFPROC CountTheNumber
NoCount = 0
FOR J = 1 TO 80 STEP 4
IF RuleArray(J+2) = 1 THEN
NO
ENDIF
ELSE
YES
ENDIF
ENDPROC
DEFPROCReportFirst
PRINT TAB(ResultX%+5, ResultY%+10) "X X XXX"
PRINT TAB(ResultX%+11, ResultY%+10) "YES"
ResultY%+=1
PRINT TAB(ResultX%+5, ResultY%+10) "0 0 0 0 0"
PRINT TAB(ResultX%+11, ResultY%+10) "NO"
ResultY%+=1
GCOL 4 6 TINT 192
RECTANGLE FILL 40,770,1200,100
VDU 5
MOVE 100,830
GCOL 0
PRINT "Patterns: XXXX = YES OOOOO = NO"
VDU 4
ENDPROC

DEFPROCReportTest
PROCStoreTest
Try? = STR?(TrialNo%)
PRINT TAB(ResultX%+1, ResultY%+10) Try?
PRINT TAB(ResultX%+5, ResultY%+10) Answer?
IF YesToTest = TRUE THEN
PRINT TAB(ResultX%+11, ResultY%+10) "YES"
ELSE
PRINT TAB(ResultX%+11, ResultY%+10) "NO"
ENDIF
ResultY%+=1
IF ResultY% > 20 THEN
ResultY% = 0
ResultX%+=25
ENDIF
ENDPROC

DEFPROCReportRule
PROCStoreRule
Try? = STR?(TrialNo%)
PRINT TAB(ResultX%+1, ResultY%+10) Try?
PRINT TAB(ResultX%+5, ResultY%+10) Rules?
IF RulesToTest = TRUE THEN
PRINT TAB(ResultX%+11, ResultY%+10) "TRUE"
ELSE
PRINT TAB(ResultX%+11, ResultY%+10) "FALSE"
ENDIF
ResultY%+=1
IF ResultY% > 20 THEN
ResultY% = 0
ResultX%+=20
ENDIF
ENDPROC

DEFPROCInputBox
RECTANGLE FILL 40,900,1200,100
VDU 5
MOVE 100,957
GCOL 0
COLOUR 128 46 TINT 192: COLOUR 0
Answer? = ""
PRINT "Enter Pattern or Hypothesis?"
VDU 4
INPUT TAB(37,2) Answer?
COLOUR 128 46 TINT 0: COLOUR 46 TINT 192
ENDPROC

DEFPROCContinueBox
GCOL 6 0 TINT 192
RECTANGLE FILL 40,900,1200,100
VDU 5
MOVE 90,957
GCOL 0
COLOUR 128 46 TINT 192: COLOUR 0
IF TestNo% > 0 THEN
PRINT "Correct, Test " + STR?(TestNo%) + " completed. Press a key to continue"
ELSE
PRINT "Correct, Practice completed. Press a key to continue"
ENDIF
VDU 4
ENDPROC

DEFPROCInvalidBox(Invalid?)
GCOL 39 TINT 192
RECTANGLE FILL 40,770,1200,100
MOVE 100,830
Out? = "Trial " + STR?(TrialNo%) + " : " + Answer? + Invalid?
PRINT Out?
VDU 4
ENDPROC

DEFPROCValidBox(Valid?)
GCOL 46 TINT 192
RECTANGLE FILL 40,900,1200,100
MOVE 100,830
GCOL 0
COLOUR 128 46 TINT 192: COLOUR 0
IF ValidBox = 0 THEN
PRINT "Correct, Practice completed. Press a key to continue"
ELSE
PRINT "Correct, Test " + STR?(TestNo%) + " completed. Press a key to continue"
ENDIF
GCOL 46 TINT 192
RECTANGLE FILL 40,770,1200,100
MOVE 100,930
GCOL 0
A-29
Out$="";STR$(TrialNo%)+" : "+Answer$+" Valid?"
PRINT Out$

VDU 4
ENDPROC

REM * Store Rule : record a rule result*REM *************************************
DEFPROCS toroRule
EntryNo%+=1
StoreSub = ( EntryNo% - 1 ) * 12 ) + 1
AnsArray(StoreSub)+1 = TrialNo%
AnsArray(StoreSub)+2 = 2
AnsArray(StoreSub)+3 = Num1
IF AndRule=TRUE THEN AnsArray(StoreSub+ 3) = 1 ELSE AnsArray(StoreSub+3) =2 ENDIF
AnsArray(StoreSub+4 ) = Num2
IF Rule True =  TRUE THEN AnsArray(StoreSub+5 ) = 1 ELSE AnsArray(StoreSub+5) =0 ENDIF
IF SomeRule =  TRUE THEN AnsArray(StoreSub+6) = 1 ELSE AnsArray(StoreSub+6) =0 ENDIF
IF InconsistentRule = TRUE THEN AnsArray(StoreSub+7) = 1 ELSE AnsArray(StoreSub+7) =0 ENDIF
AnsArray(StoreSub+10 ) = InputTime
AnsArray(StoreSub+11)=TestNo%
ENDPROC

REM *************************************REM *  Store Test : record a test result*REM *************************************
DEFPROCSto
tTest
EntryNo%+=1
AnsSub = ( EntryNo%-1 ) * 12 ) + 1
AnsArray(Assub)+1 = TrialNo%
AnsArray(Assub)+1 = 1
FOR J =  1  TO 5 AnsArray(Assub+J+1) = AnsTest(J) NEXT J
IF YesTest = TRUE THEN AnsArray(Assub+6) = YesCount ELSE AnsArray(Assub+6) =0 ENDIF
AnsArray(Assub+7) = YesCount
AnsArray(Assub+8) = NoCount
AnsArray(Assub+10) = InputLines
AnsArray(Assub+11)=TestNo%
ENDPROC

REM * Check Rule : Check a ruleREM *************************************
DEFPROCCheckRule
SomeRule = FALSE
InconsistentRule = FALSE
InconsistNo = 0
FOR O =  1  TO EntryNo%
N = ( 0 - 1 ) * 12 ) + 1
IF AnsArray(N +11) = TestNo% THEN
IF AnsArray(N+1) = 2 THEN IF ( Num1 = AnsArray(N+2)) AND ( Num2 = AnsArray(N+4)) THEN
InconsistentRule = TRUE
ENDIF IF ( Num1 = AnsArray(N+2)) OR ( Num2 = AnsArray(N+4)) THEN
InconsistentRule = TRUE
ENDIF
ENDIF
ENDIF
NEXTO
FOR H =  1  TO EntryNo%
J = ( 0 - 1 ) * 12 ) + 1
IF AnsArray(J +11) = TestNo% THEN
IF AnsArray(J+1) = 1 THEN IF ( AnsArray(J+1+Num1) = 1 ) AND ( AnsArray(J+1+Num2) = 0 ) AND (AnsArray(J+7) =0) THEN
InconsistentRule = TRUE
ENDIF IF ( AnsArray(J+1+Num1) = 0 ) OR ( AnsArray(J+1+Num2) = 0 ) AND (AnsArray(J+7) =0) THEN
InconsistentRule = TRUE
ENDIF
ENDIF
ENDIF
NEXTH
ENDPROC

REM General Error TrapREM *************************************
DEFPROCError
CLS
ENDPROC

REM Check system variableREM *************************************
DEFFNroadvar($SysVarNamG%)

END
Appendix IIIc: Instructions for 'Concept'

INSTRUCTIONS

You are about to participate in a computer run concept formation experiment. As in all concept formation problems, your task is to examine two sets of objects and figure out why the sets are formed as they are. Suppose that you find out that:

These 3 patterns are instances of the concept

| XXXXX  |
| XXXXO  |
| OXXO   |

These 3 patterns are not instances of the concept

| OOOOO  |
| XXOOX  |
| OXXO   |

Your task is to find the rule which makes a pattern an "instance" as opposed to "not an instance". This is an example of the kind of problem you are to solve in this experiment. Look again at the patterns. The six given above come from a set of 32 patterns each one consisting of a unique arrangement of 5 X's and O's.

In the example above the rule is "3 AND 4" which means that patterns with X's in both positions 3 and 4 are instances (e.g. XXXXX and XXXXO) while patterns with X's in neither position (e.g. OXOOO) and those with an X in just one of these positions (e.g. OXOXO) are not instances. There are also "OR" rules. Suppose you find out that:

These are instances

| XXXXX  |
| OXOXO  |
| OOXO   |

These are not instances

| OOOOO  |
| XXOXO  |
| OXXO   |

In this example, the rule is "3 or 4" which means that any pattern with an X in either position 3 or 4 (or both) are instances (e.g. XXXOO, OOOXX) while patterns with X's in neither position 3 or 4 are not instances (e.g. OXOXO, XOXOX).

The computer which administers the experiment is going to give you four of these problems. For each problem there is a rule which determines which patterns are instances. You are to discover this rule. Discovering such a rule involves, as you saw above, involves finding out three things: the two key positions and the kind of rule (AND, OR) it is. There are only 20 rules from which the computer can select an answer. A complete list is on the next page.

This concept problem resembles the old "20 Questions" game in that you start off knowing practically nothing about the answer. You make progress by asking questions and carefully evaluating the replies. There are two kinds of question you may ask: "How
does the unknown rule classify such-and-such?" and "Is rule such-and-such the correct answer to this problem?" Thus you may either ask about a pattern or try a hypothesis.

The computer starts you off on a problem by telling you that "XXXXX" is an instance (YES) and "OOOOO" is not an instance (NO). This is all the information you have to begin with. It is now up to you to ask questions which will help you determine which one of the 20 rules is the answer to this problem. You may ask about a pattern or you may try a hypothesis.

The computer will type "Enter Pattern or Hypothesis" and wait for you to request more information. If you want to know about a pattern you would simply type it in and wait for an answer. Suppose that you type XXOOX. The computer will reply by telling you whether the unknown rule classifies it as an instance or not. If it is an instance you know (amongst other things) that the answer might be "1 AND 2" but that it couldn't be "3 AND 4". In this case the trial would have turned out looking like this (the part you type is underlined):

TRIAL 1: XXOOX YES

Every time you learn how the unknown rule classifies a pattern you have a clue as to what the rule must be. If you want to try a hypothesis you would type it in and wait for an answer. If you type in "1 and 2", for example, the computer will tell you whether or not it is the answer. If it is the computer types "TRUE" and you go on to the next problem. If your hypothesis isn't the right answer it will tell you that it is "FALSE" and goes on to the next trial. In this case the trial would have turned out looking like this (again the part you typed is underlined):

TRIAL 1: 1 and 2 False

To review: You will solve five problems. The answer to a problem is drawn from a pool of 20 possibilities. A rule states which two positions are the key positions in a pattern and whether they are linked by AND or OR. You discover the rule by asking questions of the computer - one question per trial. On each trial you may either ask about a pattern (any pattern you think will be helpful in your search) or try a rule. The problem ends when you have typed in a rule to which the computer answers "TRUE".

Finally: You should always strive to do these problems in as few trials as possible. Time is of little importance. So take as long as you wish and solve the problems in as few trials as you can.

The possible AND rules:

1 AND 2  2 AND 3  3 AND 4
1 AND 3  2 AND 4  3 AND 5
1 AND 4  2 AND 5  4 AND 5
1 AND 5

The possible OR rules:

1 OR 2  2 OR 3  3 OR 4
1 OR 3  2 OR 4  3 OR 5
1 OR 4  2 OR 5  4 OR 5
1 OR 5
Appendix IIId:
Instructions for the 'Grid Test'

The Grid Test

This test looks at your ability to see very faint spots on an X-ray image.

Look at Example 1 on the light box. In the centre of the small square in the middle you can see a dark spot. In the test you will be required to decide whereabouts in the square the spot is located. Looking at Example 2, you can see the spot can be in one of five places: in one of the four corners or in the middle. Each square in the test contains one and only one such dark spot.

The grid on the answer sheet represents the grid you can see on the test X-ray. There are 25 squares in the test. Your task, therefore, is to find the locations of the 25 spots, on in each square. The five possible positions are marked on the answer sheet. The spots in the test tend to be much smaller than in the examples. You may even have to guess. Please indicate which section you think contains the dark spot by writing the number representing the level of confidence with which you make your decision in the appropriate section.

| UNSURE  | 1 |
| CONFIDENT | 2 |
| VERY CONFIDENT | 3 |

So if you can see a spot in the top left corner of the square, and you are completely sure about this, your answer would look like this:

```
3
```

If, however, you think you can't see anything much, but you decide that the spot might be in the middle, your answer is meant to look like this:

```
1
```

Beware of dusty flecks or other artefacts and remember that each large square contains one and only one area with a black spot!
ANSWER SHEET

Please mark only one section in each square with one of these numbers.

1 - Unsure
2 - Confident
3 - Very Confident

Name: ________________________________

Dept: ________________________________

THANK YOU FOR PARTICIPATING

A-34
Appendix IV: Description of the Test Roller of 79 Mammograms by a Consultant Radiologist.

<table>
<thead>
<tr>
<th>Cancers:</th>
<th>8,18,24,35,55,63,65,70,73,78 &amp; 71</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referrals</td>
<td>16,29</td>
</tr>
<tr>
<td>Benign Biopsies</td>
<td>6,41,49</td>
</tr>
</tbody>
</table>

1. The breasts are generally radiolucent and fatty in consistency. In the left breast there is an ovoid 1.5 cm low density opacity with a well defined margin which contains coarse calcification. The margins of the calcification are sharply defined and knobby and it has a bilobed configuration and this is a characteristic of a fibroadenoma.

2. There is a dense glandular type of pattern in both breasts. In the lower half of the left breast there is a spidery type of density present. It appears to have curvilinear shadow pulled into it. There seems to be an overlapping blood vessel and I think that this is probably a summation shadow involving trabeculae and vascular shadows. Nevertheless the patient should be recalled for a slightly different view to demonstrate this.

3. The breasts are radiolucent and fatty in consistency with complete atrophy of the glandular tissue. Some minute filigree rings of benign calcification are present on the left and these are probably in skin pores. On the right sub-areolar fusiform density may well be a dilated duct. None of the radiological criteria of malignancy are seen.

4. There is a dense inhomogeneous type of background pattern to the breasts. Some very faint rounded densities can be seen up to 10 mm in diameter and these could indicate the presence of cystic change. Some large but normal lymph nodes are seen in the axillae. There is no radiological evidence of neoplasia in the breasts.

5. Involutionary changes are occurring with some shrinkage of the breast disc indicating involution. The pattern of the breast is somewhat nodular but the appearances are within normal limits.
6. Involution throughout both breasts appears to be uneven and there is some asymmetry present which is caused by a curvilinear broad density in the lower half of the right breast. There is no true distortion present in this region and as the angle of the pectoralis major muscle varies on both sides I think this is probably a positional shadow. On the left breast however there is a radiating low density shadow in the upper half of the breast with one long linear 1.5 cm bright line extending upwards and anteriorly but there are also some faint radial shadows extending from a fairly dense tiny central core with some very faint flecks of calcification shown in its vicinity. These are almost at the limits of perceptibility. The lesion should be regarded as a carcinoma until proved otherwise.

7. The breasts are radiolucent and fatty in consistency with involution limits and there is no radiological evidence of neoplasia.

8. Quite noticeable asymmetry is present. The presence of a very low density shadow in the right axillary tail. This has a very diffuse and shaded margin and measures over 1.5 cm in diameter. When compared with the texture of similar shadows bi-centrally in the breast it is perhaps a little denser and a little more inhomogeneous. Two parallel linear shadows are seen extending posteriorly from the density and this lends added support for a suspicion of malignancy.

9. There is slightly increased nodularity in both breasts but the appearances are within normal limits and there is no radiological evidence of neoplasia.

10. Breasts are generally radiolucent but there is some asymmetry present with fluffy and slightly lobulating densities in the upper half of the right breast but the asymmetry is mainly due to uneven involution. Some finger-like projections in the right sub-areolar region are probably due to duct ectasia. Benign ring calcification is also noted throughout both breasts. None of the radiological criteria of malignancy are seen.

11. Involutionary changes are present in the breasts which show a normal glandular type of breast disc. The appearances are within normal limits and none of the criteria of malignancy are seen.

12. There is increased nodularity in both breasts. Some of the larger nodules are probably tiny cysts. Posteriorly in the right breast there is some dense calcification measuring 1 cm in diameter. This appears as a conglomeration of rounded particles of calcification, small rounded buds extending off it and this is characteristic of calcification in a fibroadenoma.
13. The breasts are generally radiolucent. In the upper half of the right breast there is a 10mm rounded homogeneous density with a very clearly defined margin with a fatty halo which has all the characteristics of a benign lesion and is quite innocuous and can be safely ignored. A similar rounded density lying posteriorly in the left breast at the base of the pectoral muscle is again of low density and has a faint fatty corona around it. This has the characteristics and position of a small lymph node.

14. There is a normal glandular type of pattern in both breasts. Calcification is noted in the walls of the arteries on both sides and some benign ring calcification is noted on the right. The appearances are within normal limits and there is no radiological evidence of neoplasia.

15. Rounded and fluffy densities are scattered throughout both breasts. This is probably more marked on the right than the left but the difference in the two breasts is probably due to uneven compression. Faint rounded 12 mm density lying centrally in the right breast is almost certainly a cyst and there are several smaller very low density rounded shadows which again are probably tiny cysts. There is no radiological evidence of neoplasia.

16. There is considerable distortion in the upper half of the right breast. The breast tissue appears to be condensed and retracted. As this is at the site of previous surgery it is probably of little consequence but one really needs another view to demonstrate this adequately. A small circular 6 mm round opacity of calcium density is seen just behind it and this is calcification in the wall of a cyst and is a small breast stone. Below this there is conglomerate calcification in a tiny fibroadenoma.

On the left side in the region of the axillary tail and overlying the margin of the pectoral muscle there is a smudge shadow in that there is an ill-defined density which seems to be structureless. This could be due to uneven involution but I think should be recalled for further evaluation.

17. The apparent asymmetry between the breasts is I think due to slightly faulty positioning as the posterior fat line cannot be identified on the left. Benign ring calcification is noted in both breasts and there is no radiological evidence of neoplasia.

18. There is a small knobbly well defined mass fairly homogenous in the left axillary tail. There is a small tail extending into the breast disc. This should be regarded as a carcinoma until proved otherwise.

19. Streaky nodular type of pattern in both breasts. The appearances are within normal limits and there is no radiological evidence of disease.
20. Generally nodular pattern with early involutionary changes and fatty replacement creeping in from the posterior portion of the breast. An island of glandular tissue in the upper half of the posterior fat band on the left seems to be due to uneven involution and causes some asymmetry.

21. Both breasts are generally radiolucent with involution of the glandular tissue and shrinkage of the breast discs. Odd particles of calcification in both breasts are due to either duct ectasia or secretory disease and there is also some benign ring calcification in tiny areas of fat necrosis.

22. There is a very dense glandular type of pattern in both breasts with quite marked calcification in the arteries but no radiological abnormality is demonstrated.

23. The breasts are radiolucent and fatty in consistency with involution of the glandular tissue. Some retraction of the right nipple and there is a small lymph node against the margin of the pectoralis major on the left. No criteria of malignancy are seen.

24. There is quite marked asymmetry present. Widening of the subcutaneous space in the upper half of the right breast can be identified. In this region there are also possibly abnormal vessels. There is quite marked distortion of the breast disc which is pulled inwards to an area of increased density rather stiff looking trabeculae with some linear fibrosis. Calcification is present but this is fairly generalised and I think is due to microcystic disease. The distortion however strongly suggests the presence of a carcinoma.

25. No abnormality seen in the breasts.

26. The breasts are radiolucent and fatty in consistency. Benign ring calcification is noted on the right. There is a small lymph node in the upper half on the left and posteriorly on the right. No evidence of neoplasia is seen.

27. Involutionary changes have resulted in marked contraction of the breast discs but appearances are entirely within normal limits.

28. No abnormality is demonstrated in the breasts.

29. There is a generally granular appearance in both breasts. On the lower half of the right breast there is movement blurring in the film and the whole of the breast is not adequately covered. The patient should be recalled for this alone. There is some faint stippled calcification within some tiny nodules in the upper half of the right breast which is almost certainly due to microcystic change.
30. Nice, normal radiolucent breasts.

31. There is a rather streaky type of pattern in both breasts with a little nodularity. The appearances are otherwise normal.

32. The appearance of the breasts is normal. No radiological evidence of pathology.

33. No abnormality is shown in the breasts.

34. Generally nodular breasts but no abnormality is seen.

35. There is a 6 mm mass in the axillary tail of the right breast. This has some very fine whiskers radiating from it and there are bright linear shadows also extending from it. The linear shadows are up to 1.5 cm in length. The architectural pattern is broken up within the mass. There is a second ovoid mass 1.5 cm diameter with ill-defined but distinct margins centrally in this breast. Again there is a break in the architectural pattern within and adjacent to the mass and this should be regarded as a multifocal breast cancer until proved otherwise.

36. There is a very dense glandular almost homogeneous pattern in the breasts but the appearances are within normal limits.

37. The breast is involuting and the discs are contracting, No abnormality is demonstrated.

38. The breasts are fatty and radiolucent but no abnormality is seen.

39. There is a normal glandular type of pattern in the breasts. The appearances are within normal limits and none of the criteria of malignancy is seen.

40. No abnormality is seen in the breasts.

41. There is very marked asymmetry between the breasts in that there is a low density fairly sharply defined ovoid mass in the lower half of the left breast and this contains fine speckled calcification. The sharp lower margin of the mass and the gradual shading of the upper margins makes it unlikely that this is an invasive type of carcinoma but an intraductal or an ectatic duct with lipisated and calcified secretion is possible. This should be recalled for further investigation.

42. Fine scattered calcification which appears to be intra-lobular is present in the right breast. This is fairly characteristic of microcystic change. No evidence of neoplasia is seen in the breasts. The slightly serpiginous calcification seen centrally in the upper half of the right breast might well merit recall for magnification views to provide a base line for screening.
43. Very radiolucent normal breasts.

44. There are fluffy and nodular densities scattered throughout both breasts. There is very faint, fine calcification scattered throughout the breasts consistent with microcystic change. Some of the larger nodules may also be tiny cysts.

45. Slight increase in nodularity and streakiness in both breasts but no abnormality is seen.

46. Radiolucent breasts. Tubular calcification on the left is in duct ectasia. No other abnormality seen.

47. Apparent positioning fault on the left. This appears due to the congenital absence of the pectoralis muscle on the left and that's not fair.

48. No abnormality seen in the breasts.

49. There is some very fine punctate calcification in the right sub-areolar region. This is not obviously malignant but nevertheless it should be recalled for further investigation.

50. Normal glandular pattern and benign ring calcification scattered throughout the right breast. Posteriorly in the lower half of the left breast there are some linear shadows extending over the edge of the film but these are not distorted and not extrapolated into the Cooper's ligaments. I think this is due to inadequate positioning but there is no radiological evidence of neoplasia.

51. There is a normal glandular pattern in both breasts but no abnormality is seen.

52. Generally nodular type of breasts. Ring calcification in the upper half of the right breast is in an end-on blood vessel. The other ring calcification is due to fat necrosis. There is no evidence of neoplasia.

53. There is a very dense, glandular type of breast but no abnormality is seen.

54. There is a streaky and nodular pattern in both breasts. Some benign ring calcification is also noted but the appearances are otherwise within normal limits with no radiological evidence of neoplasia.
55. Rather streaky background pattern in both breasts. In the right breast there are some scattered punctate flecks of calcification which are probably nodular in type but rather more centrally in the upper half of the breast there is a small cluster of calcification which could be intraductal in type and therefore requires recall for magnification views but the chances of this being malignant are very low.

In the left breast there is a very ill defined 1.5 cm mass in the lower half of the breast. Its margin is irregular and the adjacent trabeculae are distorted and some are pulled into the mass and there is a break-up of the pattern and a lot of homogeneity within the mass. This should be regarded as a carcinoma until proved otherwise.

56. There is a slight increase in nodularity in both breasts. Some very faint calcification of the type usually associated with burnt out micro-cystic change seen in the right sub-areolar region and there is also some benign ring calcification around. The appearances are within normal limits. No abnormality is demonstrated in the breasts.

57. Some involution of the glandular tissue is shown with shrinkage of the breast. Exuberant lymph nodes are shown in the axillae and the blood vessels are more than usually conspicuous but I think that this is due to the compression in the axillae. No abnormality is seen in the breasts.

58. There is marked involution of the glandular tissue and the breasts are mainly fatty in consistency. Some uneven involution is present which does cause some asymmetry with rather more breast tissue in the right sub-areolar region than on the left. None of the radiological criteria of malignancy are seen.

59. Calcification in an oil cyst is shown in the lower half of the right breast. An island of conglomerate tissue is shown centrally in the upper half of the right breast which does give an impression of asymmetry but on analysing this one can see that it consists of several low density discrete nodules. None of the radiological criteria of malignancy are seen in either breast.

60. No involuotary changes are present. There is a streakiness to the glandular tissue, particularly in the upper half of the breast. This streakiness appears to be caused by radiolucent material within the ducts so that there could well be some evidence of duct stagnation. There is no radiological evidence of neoplasia in the breasts.
61. Breasts have a generally streaky background pattern. There is very slight asymmetry present with faint ill-defined shadowing in the upper half of the right breast but examination of the angles of the pectoral muscle would suggest that this is due to positioning. The texture of the asymmetric tissue again is the same as glandular tissue in the breast and there is no distortion or disruption of breast pattern within it. The appearances are within normal limits and there is no radiological evidence of neoplasia.

62. The breasts are radiolucent with atrophy of the glandular tissue and no abnormality is seen.

63. The breasts are generally radiolucent with atrophy of the glandular tissue. Numerous lymph nodes are seen in the region of the axillae and these are of normal texture and configurations. In the lower half of the right breast there is a .5 cm mass. The margins of this mass are slightly fuzzy but fairly well delineated. There is no pattern and no homogeneity within the mass. An overlapping vein is noted and there are retraction phenomena present with tapering trabeculae pulled in towards the mass anteriorly and just below the mass there are smaller trabeculae pulled into it with a slight curve giving a whiskery appearance. This should be regarded as carcinoma until proved otherwise.

The small .75 cm mass in the upper half of the right breast is very homogeneous has a well defined margin with a clear zone around it and this again is almost certainly a small intramammary lymph node.

64. There is a normal glandular type of pattern in both breasts. No abnormality is seen.

65. Involutorial changes are seen and the breasts are generally fatty. In the left breast there is an 8 mm spiculated mass lying centrally towards the lower half of the breast with slightly curved trabeculae curved inwards towards the mass anteriorly and there are also whiskers present and sharply defined sharp bright linear shadows radiating from the mass. Within the mass these lines break up, some of them bend and the pattern is generally chaotic.

66. These are generally nodular but very normal breasts.

67. There is a fairly dense glandular pattern in both breasts but the appearances are within normal limits.

68. There is a rather active nodular type of pattern in both breasts but the appearances are within normal limits and there is no radiological evidence of neoplasia.
69. Glandular tissue in both breasts have atrophied and the breasts are generally fatty in consistency. No abnormality is seen.

70. There is a dense nodular pattern in both breasts. In the right breast there is a very diffuse ill-defined shadow in the lower half of the breast lying mainly posteriorly. The pattern breaks up within this density and posteriorly the margin appears to be shaggy with some linear shadows that give the impression of being pulled into the suspect area. The linear shadows tend to be bright and two are in parallel and this in itself is a very suspicious sign. The shagginess and distortion extend well beyond the breast disc and into the posterior fat line and when compared with the left breast there is very significant asymmetry present.

71. The breasts are very nodular and dense. In the left breast there is a stellate 4 mm mass lying posteriorly and just outside the breast disc. The 4 mm refers to the central core and from this central core there are long linear slightly wavering and bright shadows. This is most probably a carcinoma but it is just possible that it is a radial scar.

72. Some asymmetry is present due to uneven involution and there is more fatty tissue on the left than the right. Positioning is also slightly askew and the increased density in the upper half of the left breast is probably due to inappropriate compression.

73. There is a small cluster of calcification centrally in the right breast. This has a linear morse code type of pattern and one of the rods has got a bud extending from it. The appearances of this is therefore very likely to be due to intraluminal intraductal calcification indicating an increase in the turnover within the ducts and arouses strong suspicion of an in situ carcinoma.

74. There is extensive calcification in both breasts. The average size of the particles is less than 1 mm. Some of these particles have a tubular type of appearance thus I think are almost certainly associated with duct ectasia. The ring calcification tends to have a radiolucent centre and is therefore in tiny oil cysts associated with fat necrosis. None of the radiological criteria of malignancy are seen.

75. The breasts are fatty in consistency with a granular type of background pattern but the very conspicuous skin pores seem to contribute to this.Appearances are within normal limits.

76. Streaky and nodular background pattern and some of the larger nodules are probably tiny cysts but there is no evidence of neoplasia.

77. There is a normal glandular type of background pattern in both breasts and the appearances are entirely within normal limits.

A-43
78. Extensive ring and tubular calcification seen in both breasts, some associated with duct ectasia and some with fat necrosis. There is calcification in the vessel walls. There is some asymmetry present and in the right breast there is an area of increased density in the upper half of the breast with surface of the breast disc being pulled inwards. There is quite a lot of distortion in this area together with some faint stippled calcification. Subcutaneous changes are noted with abnormal blood vessels, stiff looking trabeculae and faint whiskers extending upwards in a brush-like fashion. This should be regarded as a carcinoma until proved otherwise.

79. Normal mammograms with generally nodular breasts.