

University Library

Author/Filing Title TAYLOR - PHILLIPS

Class Mark . T

Please note that fines are charged on ALL
overdue items.

0403820391



The presentation of prior mammograms in the transition to digital mammography

by

Sian Taylor-Phillips

Doctoral Thesis


Submitted in partial fulfillment of the requirements

for the award of

Doctor of Philosophy of Loughborough University

21/9/2009

© by Sian Taylor-Phillips 2009

 Loughborough University Pilkington Library	
Date	9/7/10
Class	T
Assoc No.	0403860391

Abstract

The NHS Breast Screening Programme (NHSBSP) is changing from using film screen to digital mammography. The aim of this research was to determine the optimum display medium for the prior mammograms (from the previous screening round) for the transition period.

Three options for the display of the prior mammograms were investigated: film display adjacent to the digital workstation, digitised display on the digital workstation, or not displaying them at all. The effect of this choice in terms of workstation ergonomics, participant behaviour, and cancer detection performance were all investigated. Eight participants were videotaped reading digital mammograms with either film or digitised prior mammograms for four 45 minute sessions as part of the NHS Breast Screening Programme. Workstation ergonomics was assessed using event based Rapid Upper Limb Assessment (RULA), and body part discomfort charts. Behaviour of participants was measured from the video recordings, with start time, end time, and the number of times the participant looked at the prior mammograms recorded for every case. Workload was measured using NASA RTLX questionnaires. Cancer detection performance was measured using a set of 160 difficult test cases of which 41% were malignant. Eight participants (all qualified to read mammograms in the NHSBSP) read the cases in three conditions: with film; with digitised; and without prior mammograms. Both Jackknife Free Response Receiver Operating Characteristic (JAFROC), and recall rate analyses were conducted. This study was also video-taped and use of prior mammograms and time taken per case measured.

There was no significant difference between using film or digitised prior mammograms for body part discomfort scores or RULA risk scores. In breast screening practice prior mammograms were used for 19% more cases when displayed in digitised rather than film format ($p=.04$). Reading with digitised prior mammograms was 18% faster ($p=.04$), and was associated with lower workload ($p=.03$) than reading with film prior mammograms. The JAFROC results showed that performance differed between conditions ($p=.006$), with performance superior with prior mammograms than without. No difference in performance was found between using film or digitised prior mammograms, but the greater use of digitised than film prior mammograms in screening practice was not mirrored in the experiment. After weighting for case type, the number of false positives (normal cases recalled) was 26% higher without than with prior mammograms. Ceasing to use prior mammograms in the transition to digital mammography may cause an increase in recall rate from 3.9% to 4.6% at the study hospital with no associated increase in cancer detection performance.

In the transition to digital mammography the prior mammograms should be presented for every case, and where possible in digitised format.

Contents

Glossary of Terms and Abbreviations .. .	1
1 Introduction	7
1.1 Breast Cancer Screening in the UK	7
1.2 The Introduction of Digital Mammography	10
1.3 Aims and Objectives	19
1.4 Methodological Approach.....	20
1.5 Thesis Outline	21
1.6 Subject Matter Immersion	23
1.7 Literature Review	25
1.8 Contribution to Knowledge	29
2 Physical Comfort in the transition to Digital Mammography	31
2.1 Introduction	31
2.2 Aims	35
2.3 Choice of Methods	36
2.3.1 Design approach vs ergonomic assessment.....	36
2.3.2 Workstation Evaluation without Participants	36
2.3.3 Participant Observation and Measurement Methods	37
2.3.4 Subjective Comfort Reports	39
2.4 Method	40
2.4.1 Workstations	40
2.4.2 Participants	44
2.4.3 Research methods.....	45
2.5 Results	49
2.5.1 Workstation Dimensions	49
2.5.2 Body Part Discomfort	52
2.5.3 Postural Analysis	59

2.6	Discussion.....	74
2.6.1	Workstation Dimensions and weights	74
2.6.2	Body Part Discomfort	75
2.6.3	RULA – The transition to digital mammography.....	78
2.6.4	RULA – Digital or Film Display of Prior Mammograms.....	80
2.7	Conclusions.....	86
3	Workload and Productivity in the Transition to Digital Mammography.....	88
3.1	Introduction	88
3.2	Aims	92
3.3	Choice of Methods	92
3.4	Method	98
3.4.1	Workload Method	98
3.4.2	Speed of Reading Method	99
3.4.3	Statistical Analysis	100
3.5	Results	102
3.5.1	Workload Results.....	102
3.5.2	Speed of Reading	107
3.6	Discussion.....	110
3.6.1	Speed of reading.....	110
3.6.2	Workload.....	111
3.6.3	Conclusions	116
4	Behavioural Use of Prior Mammograms	118
4.1	Introduction	118
4.2	Aims	120
4.3	Choice of Methods	120
4.4	Method	122
4.4.1	Pilot Studies.....	122
4.4.2	Main Study.....	124
4.4.3	Statistical analysis.....	126

4.5	Results	127
4.6	Discussion.....	134
4.7	Conclusions.....	138
5	Chapter 5 – The use of Prior Mammograms and Performance	140
5.1	Introduction	140
5.2	Aims	142
5.3	Choice of Methods	142
5.4	Method	151
5.4.1	Ambient Lighting conditions	151
5.4.2	Case Selection.....	153
5.4.3	Participants and methods.....	156
5.4.4	Cost Analysis	162
5.4.5	Equipment.....	164
5.5	Results	165
5.5.1	Ambient lighting levels	165
5.5.2	Performance	166
5.5.3	Cost Analysis	173
5.6	Discussion.....	175
5.7	Conclusions.....	182
6	Comparison of Behaviour in Experimental Setting and Screening Practice	184
6.1	Introduction	184
6.2	Aims.....	184
6.3	Method	185
6.4	Results	187
6.4.1	Behaviour in the Experimental Setting	187
6.4.2	Comparison of Behaviour in the Experimental and Live Screening Settings	195

6.5	Discussion.....	204
6.6	Conclusions.....	211
7	Discussion	215
7.1	Progression of the Research Direction.....	216
7.2	Aims and Objectives	218
7.3	Choice of Methods	223
7.4	Limitations of the Study and Further Research	226
8	References	231
	Appendix 1 – Task Analysis of Screening at the Study Hospital	255
	Appendix 2 - Participant Information Sheet and Informed Consent Form	263
	Appendix 3 - Body Part Discomfort Chart.....	265
	Appendix 4 - NASA TLX Workload Questionnaire.....	266
	Appendix 5 – Normality Tests for Workload and Time Taken per Case.....	267
	Appendix 6 – Participant Information Sheet and Informed Consent Form	288
	Appendix 7 – Participant Instructions	290
	Appendix 8 – Examples of Data Recording Sheets for Performance Experiment.....	292
	Appendix 9 - Publications by Sian Taylor-Phillips	298

Acknowledgements

I would like to thank my supervisor Prof. Alastair Gale for his support and vast knowledge, and Dr Matthew Wallis and Dr Iain Darker for their time and expertise. My sincere thanks to the people at the Coventry, Solihull and Warwickshire Breast Screening Programme for the hours they put in as participants, and for the support and encouragement they gave.

Thank you to Lis for thinking it was a good idea to do a PhD rather than get a proper job, and most especially thanks to all of my friends and family for their support throughout, in particular MumDad and Amy. And finally thank you to Jilly Peggle, for not giving up! Peace to the momonga!

List of Figures

FIGURE 1.1 - THE DECISION MAKING PROCESS IN UK SCREENING ..	9
FIGURE 1.2 - TWO WORKSTATIONS FOR READING MAMMOGRAMS IN THE BREAST SCREENING PROGRAMME THE FILM WORKSTATION (ABOVE) HAS BEEN USED SINCE THE NHS BREAST SCREENING PROGRAMME WAS SET UP IN 1988, THE DIGITAL WORKSTATION (BELOW) WILL REPLACE SUCH FILM WORKSTATIONS OVER TIME BOTH HAVE CURRENT MAMMOGRAMS ON THE UPPER ROW AND PRIOR MAMMOGRAMS ON THE LOWER ROW, AND A SMALLER SCREEN TO THE LEFT TO SHOW THE DETAILS OF THE WOMEN SCREENED AND ALLOW THE READER TO INPUT RESULTS	12
FIGURE 1.3 - DIGITAL MAMMOGRAMS WITH IMAGE ENHANCEMENT TOOLS. MAGNIFICATION WITH THE ELECTRONIC MAGNIFYING TOOL (ABOVE LEFT), AND WITH FULL IMAGE MAGNIFICATION (ABOVE RIGHT). CONTRAST ADJUSTMENT FOR THE SAME MAMMOGRAM IS SHOWN ON THE LOWER ROW	13
FIGURE 1.4 - SCHEMATIC VIEW OF (A) DIGITISING THE PRIOR MAMMOGRAMS AND DISPLAYING ONSCREEN ALONGSIDE THE DIGITAL CURRENT MAMMOGRAMS, AND (B) DISPLAY IN FILM FORMAT ON AN ADJACENT BACKLIT MULTI-VIEWER...	15
FIGURE 2.1 - THE FILM WORKSTATION WITH FILM CURRENT MAMMOGRAMS ON THE TOP ROW AND FILM PRIOR MAMMOGRAMS ON THE LOWER ROW.	41
FIGURE 2.2 - THE HYBRID WORKSTATION WITH CURRENT MAMMOGRAMS VIEWED ON LCD SCREENS TO THE LEFT, AND PRIOR MAMMOGRAMS VIEWED ON A MULTI-VIEWER TO THE RIGHT.	42
FIGURE 2.3 - THE DIGITAL WORKSTATION WITH THE FIRST HANGING DISPLAYED; CURRENT MAMMOGRAMS ON THE UPPER ROW AND PRIOR MAMMOGRAMS ON THE LOWER ROW	43
FIGURE 2.4 - BOXPLOT OF DISCOMFORT SCORES FOR DIFFERENT BODY PARTS BEFORE THE SESSIONS BEGAN. A SCORE OF 1=NO DISCOMFORT, 2=VERY MILD DISCOMFORT, 3=MILD DISCOMFORT, 4=MODERATE DISCOMFORT, 5=SEVERE DISCOMFORT	53
FIGURE 2.5 - DISCOMFORT SCORES AFTER A 45 MINUTE SESSION AT THE FILM (TOP), HYBRID (MIDDLE) AND DIGITAL (BOTTOM) WORKSTATION	54
FIGURE 2.6 - BOXPLOT FOR DISCOMFORT SCORES AFTER A 45 MINUTE SESSION AT ONE OF THE WORKSTATIONS DATA FOR SESSIONS 1 AND 2 AND FOR THE FILM, HYBRID, AND DIGITAL WORKSTATIONS ARE COMBINED HERE.	56
FIGURE 2.7 - BOXPLOT FOR DISCOMFORT SCORES AFTER A 45 MINUTE SESSION AT ONE OF THE WORKSTATIONS DATA FOR SESSIONS 1 AND 2 AND FOR THE FILM, HYBRID, AND DIGITAL WORKSTATIONS IS COMBINED HERE, BUT SCORES FOR RADIOLOGISTS AND RADIOGRAPHY ADVANCED PRACTITIONERS DISPLAYED SEPARATELY.	57
FIGURE 2.8 - BOXPLOT FOR THE CHANGE IN DISCOMFORT OVER A 45 MINUTE SESSION AT ONE OF THE WORKSTATIONS. DATA FOR SESSIONS 1 AND 2 AND FOR THE FILM, HYBRID, AND DIGITAL WORKSTATIONS IS COMBINED HERE, BUT SCORES FOR RADIOLOGISTS AND RADIOGRAPHY ADVANCED PRACTITIONERS DISPLAYED SEPARATELY.	58
FIGURE 2.9 - RESULTS OF RULA POSTURAL ANALYSIS. A SCORE OF 1 OR 2 INDICATES THAT POSTURE IS ACCEPTABLE IF IT IS NOT MAINTAINED OR REPEATED FOR LONG PERIODS, 3 OR 4 INDICATES THAT FURTHER INVESTIGATION IS NEEDED AND CHANGES MAY BE REQUIRED, 5 OR 6 INDICATES THAT INVESTIGATION AND CHANGES ARE REQUIRED SOON. A SCORE OF 7 INDICATES THAT INVESTIGATION AND CHANGES ARE REQUIRED IMMEDIATELY.	60
FIGURE 2.10 - RESULTS OF RULA POSTURAL ANALYSIS DIVIDED BY PARTICIPANT TYPE (RADIOLOGIST OR RADIOGRAPHY ADVANCED PRACTITIONER) ..	61
FIGURE 2.11 - TWO POSTURES ADOPTED WHEN PICKING UP AND PUTTING DOWN THE SCREENING BAGS	62
FIGURE 2.12 - TWO EXAMPLES OF A PARTICIPANT LOOKING AT THE CURRENT MAMMOGRAMS AT THE FILM WORKSTATION	64
FIGURE 2.13 - RECORDING THE DECISION AT THE FILM WORKSTATION (ABOVE) AND THE DIGITAL WORKSTATION (BELOW)	66
FIGURE 2.14 - A MAMMOGRAPHY READER TURNING OVER A SCREENING BAG TO MOVE TO THE NEXT CASE AT THE FILM WORKSTATION (ABOVE) AND THE DIGITAL WORKSTATION (BELOW).	68
FIGURE 2.15 - TWO POSTURES ADOPTED AT THE HYBRID WORKSTATION WHEN LOOKING AT THE PRIOR MAMMOGRAMS. TOP IMAGE. THE PARTICIPANT LEANS OVER BOTH TWISTING AND SIDE BENDING THE TORSO TO GET CLOSER TO THE IMAGES. BOTTOM IMAGE: PARTICIPANT	

SIMPLY TURNS HIS HEAD RESULTING IN LESS TWISTING AND SIDE BENDING OF THE TORSO, BUT A GREATER EYE TO IMAGE DISTANCE REDUCING DETAIL PERCEPTION.	70
FIGURE 2 16 – USING THE MAGNIFYING GLASS ON THE PRIOR MAMMOGRAMS AT THE HYBRID WORKSTATION, BY LEANING OVER WHILST MAINTAINING CHAIR POSITION (ABOVE) AND BY MOVING THE CHAIR (BELOW)	72
FIGURE 2 17 - EXAMPLES OF USING MAGNIFICATION AT THE DIGITAL WORKSTATION. UPPER ROW SHOWS USE OF FULL SCREEN MAGNIFICATION (ABOVE) AND USE OF THE MAGNIFICATION TOOL (BELOW) OF HANGING THREE, IN BOTH CASES THE PARTICIPANT IS LEANING FORWARD TO GET CLOSER TO THE SCREEN.	73
FIGURE 3.1 – MODEL OF VARIATION OF WORKLOAD AS A RESULT OF TASK DIFFICULTY WITH PRIMARY TASK PERFORMANCE; ADAPTED FROM O'DONNELL AND EGGEMEIER (1986)....	94
FIGURE 3 2 – MEAN NASA RTLX WORKLOAD SCORES FOR EACH OF THE WORKSTATIONS. ERROR BARS REPRESENT ± 1 STANDARD ERROR.....	103
FIGURE 3.3 - PLOT OF THE INTERACTION BETWEEN WORKSTATION TYPE AND PARTICIPANT TYPE FOR WORKLOAD SCORE. THERE IS A TREND TOWARDS AN INTERACTION ($P=0.09$)... ..	104
FIGURE 3.4 – Q-Q PLOT OF WORKLOAD SCORES WITH PARTICIPANT 7 INCLUDED (ABOVE) AND EXCLUDED (BELOW) INCLUSION OF PARTICIPANT 7 RESULTS IN OUTLIERS AND SKEWNESS.	106
FIGURE 3.5 – MEAN TIME TAKEN PER CASE AT EACH WORKSTATION. ERROR BARS REPRESENT \pm ONE STANDARD ERROR FROM THE MEAN.	108
FIGURE 3 6 – PLOT OF THE INTERACTION BETWEEN WORKSTATION TYPE AND PARTICIPANT TYPE. BOTH WORKSTATION TYPE AND PARTICIPANT TYPE ARE SIGNIFICANT MAIN EFFECTS BUT THERE IS NO INTERACTION BETWEEN THEM.. ..	109
FIGURE 3 7 – HYPOTHETICAL RELATIONSHIP BETWEEN PERFORMANCE AND RESOURCE ALLOCATION READING MAMMOGRAMS WITH PRIOR MAMMOGRAMS IN DIGITISED AND FILM FORMATS. RELATIONSHIP IS BASED ON THE THEORY THAT THE HIGHER PERCEIVED WORKLOAD AT THE HYBRID WORKSTATION IS DUE AT LEAST IN PART TO THE GREATER MENTAL RESOURCES REQUIRED TO MAKE COMPARISONS TO THE FILM PRIOR MAMMOGRAMS	113
FIGURE 3.8 – HYPOTHETICAL RELATIONSHIP BETWEEN EFFORT AND PERFORMANCE IN READING MAMMOGRAMS WITH OR WITHOUT THE PRIOR MAMMOGRAMS. AT VERY LOW LEVELS OF EFFORT PERFORMANCE MAY BE THE SAME WITHOUT PRIOR MAMMOGRAMS, BUT AT HIGHER LEVELS OF EFFORT WITHOUT THE PRIOR MAMMOGRAMS PERFORMANCE BECOMES DATA LIMITED.....	114
FIGURE 4.1 – EXPERIMENTAL SET UP FOR CAMERA PLACEMENT TRIALS NUMBERS 1 TO 4 INDICATE THE FOUR CAMERA PLACEMENTS TRIALLED. THE INSET SHOWS THE CAMERA USED	124
FIGURE 4 2 – THE HYBRID WORKSTATION WITH MINIATURE VIDEO CAMERAS HIGHLIGHTED BY THE BLUE ARROWS, AND VIDEO RECORDING DISPLAY BELOW THE WORKSTATION.	125
FIGURE 4 3 – THE PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS WERE USED (LOOKED AT ONCE OR MORE) AT THE FILM, HYBRID AND DIGITAL WORKSTATIONS. ERROR BARS REPRESENT \pm ONE STANDARD ERROR.	130
FIGURE 4.4 - PLOT OF THE INTERACTION BETWEEN WORKSTATION TYPE AND PARTICIPANT TYPE FOR PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS ARE USED. TYPE 1 IS RADIOLOGISTS AND TYPE 2 IS RADIOGRAPHY ADVANCED PRACTITIONERS THERE IS A SIGNIFICANT INTERACTION ($P=0.008$)	131
FIGURE 4.5 – MEAN NUMBER OF TIMES PER CASE THAT THE PARTICIPANTS LOOKED AT THE PRIOR MAMMOGRAMS. ERROR BARS REPRESENT \pm ONE STANDARD ERROR.	132
FIGURE 4.6 - PLOT OF THE INTERACTION BETWEEN WORKSTATION TYPE AND PARTICIPANT TYPE FOR MEAN NUMBER OF COMPARISONS PER CASE TO THE PRIOR MAMMOGRAMS. TYPE 1 IS RADIOLOGISTS AND TYPE 2 IS RADIOGRAPHY ADVANCED PRACTITIONERS THERE IS NOT A SIGNIFICANT INTERACTION ($P=0.2$).....	133
FIGURE 5.1 – THE FORMATION OF ROC CURVES. AT EACH POSSIBLE THRESHOLD VALUE THE NUMBER OF TRUE POSITIVES (TP), FALSE POSITIVES (FP), TRUE NEGATIVES (TN), AND FALSE NEGATIVES (FN) ARE CALCULATED AND PLOTTED IN TERMS OF SENSITIVITY AND 1-SPECIFICITY ON THE ROC CURVE. THE LARGER THE AREA UNDER THE ROC CURVE THE BETTER THE PERFORMANCE. THE DOTTED DIAGONAL LINE REPRESENTS CHANCE PERFORMANCE.....	143

FIGURE 5.2 – COMPARISON OF ROC AND LROC, FROC AND AFROC CURVES. THE RED LINE DENOTES THE TRAPEZOIDAL APPROXIMATION TO THE AFROC CURVE MADE BY JAFROC ANALYSIS	148
FIGURE 5.3 – TWO IDENTICAL AFROC CURVES WITH DIFFERENT TRAPEZOIDAL APPROXIMATIONS. THE LESS ACCURATE APPROXIMATION ON THE LEFT WOULD RESULT FROM PARTICIPANTS NOT USING THE LOWER MALIGNANCY RATINGS.....	150
FIGURE 5.4 – EXPERIMENTAL APPARATUS FOR MEASUREMENT OF COEFFICIENT OF DIFFUSE REFLECTION.	153
FIGURE 5.5 – FREE RESPONSE RECEIVER OPERATING CHARACTERISTIC (FROC) CURVES FOR THE CONDITIONS: NO PRIOR MAMMOGRAMS, DIGITISED PRIOR MAMMOGRAMS; AND FILM PRIOR MAMMOGRAMS. LESION LOCALISED FRACTION IS THE PROPORTION OF LESIONS CORRECTLY LOCALISED AT A THRESHOLD, AND NON-LESION LOCALISED FRACTION IS THE NUMBER OF NON-LESIONS LOCALISED PER IMAGE AT THAT THRESHOLD	168
FIGURE 5.6 – PLOT OF THE INTERACTION BETWEEN PRESENTATION OF THE PRIOR MAMMOGRAMS AND PARTICIPANT TYPE FOR JAFROC FIGURE OF MERIT. THERE IS A TREND TOWARDS AN INTERACTION ($P= .09$). ..	169
FIGURE 6.1 – BOXPLOTS OF THE MEAN NUMBER OF COMPARISONS TO THE FILM AND DIGITISED PRIOR MAMMOGRAMS INCLUDING PARTICIPANT 8 (ABOVE) AND EXCLUDING PARTICIPANT 8 (BELOW).	189
FIGURE 6.2 – BOXPLOTS OF THE MEAN TIME TAKEN PER CASE USING FILM AND DIGITISED PRIOR MAMMOGRAMS, INCLUDING PARTICIPANT 8 (ABOVE) AND EXCLUDING PARTICIPANT 8 (BELOW)	191
FIGURE 6.3 – THE INTERACTION BETWEEN PARTICIPANT TYPE AND PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS FOR THE PERCENTAGE OF CASES FOR WHICH THE PRIOR MAMMOGRAM WAS USED ($F(1,6)=11.6, P=.01$).	193
FIGURE 6.4 – MEAN TIME TAKEN PER CASE BY PARTICIPANT TYPE AND PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS. THERE IS NO SIGNIFICANT INTERACTION	194
FIGURE 6.5 – THE INTERACTION BETWEEN PARTICIPANT TYPE AND PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS ($F(1,5)=18.7, P=.008$) FOR PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS WERE USED.....	197
FIGURE 6.6 – THE INTERACTION BETWEEN THE SETTING (EXPERIMENTAL OR SCREENING PRACTICE) AND THE PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS FOR PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS WERE USED ($F(1,5)=13.8, P=.01$)	198
FIGURE 6.7 – THE INTERACTION BETWEEN SETTING, PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS AND PARTICIPANT TYPE ($F(1,5)=9.8, P=.03$) FOR PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS WERE USED. THE RELATIONSHIP BETWEEN SETTING AND PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS IS SHOWN FOR BOTH RADIOLOGISTS AND RADIOGRAPHY ADVANCED PRACTITIONERS. ..	200
FIGURE 6.8 – PLOTS OF THE VARIATION OF THE MEAN NUMBER OF COMPARISONS TO THE PRIOR MAMMOGRAMS WITH SETTING (EXPERIMENTAL OR SCREENING PRACTICE), STRATIFIED BY PARTICIPANT TYPE AND PRESENTATION MEDIUM OF THE PRIOR MAMMOGRAMS. THERE WERE NO SIGNIFICANT INTERACTIONS.	202
FIGURE 6.9 – THE RELATIONSHIP BETWEEN TIME TAKEN PER CASE IN THE EXPERIMENT AND IN SCREENING PRACTICE FOR BOTH FILM AND DIGITISED PRIOR MAMMOGRAMS. THE INTERACTION WAS NOT SIGNIFICANT ($F(1,5)=4.6, P=.08$).	203
FIGURE A5.1 – A Q-Q PLOT OF THE DIFFERENCE BETWEEN THE TWO SCORES FOR EACH PARTICIPANT AT THE HYBRID AND DIGITAL WORKSTATIONS.	268
FIGURE A5.2 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF OVERALL WORKLOAD SCORES	270
FIGURE A5.3 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF MENTAL DEMAND SCORES	271
FIGURE A5.4 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF PHYSICAL DEMAND SCORES..	272
FIGURE A5.5 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF TEMPORAL DEMAND SCORES	273
FIGURE A5.6 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF PERFORMANCE SCORES.....	274

FIGURE A5. 7 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
EFFORT SCORES275

FIGURE A5. 8 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
FRUSTRATION SCORES ...276

FIGURE A5. 9 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
OVERALL WORKLOAD SCORES WITH PARTICIPANT 7 REMOVED.278

FIGURE A5. 10 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
MENTAL DEMAND SCORES WITH PARTICIPANT 7 REMOVED279

FIGURE A5. 11 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
PHYSICAL DEMAND SCORES WITH PARTICIPANT 7 REMOVED. ...280

FIGURE A5. 12 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
TEMPORAL DEMAND SCORES WITH PARTICIPANT 7 REMOVED281

FIGURE A5. 13 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
PERFORMANCE SCORES WITH PARTICIPANT 7 REMOVED.....282

FIGURE A5. 14 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
EFFORT SCORES WITH PARTICIPANT 7 REMOVED.283

FIGURE A5. 15 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DISTRIBUTION OF
FRUSTRATION SCORES WITH PARTICIPANT 7 REMOVED.....284

FIGURE A5. 16 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DIFFERENCES
BETWEEN TIME TAKEN PER CASE AT THE DIGITAL AND HYBRID WORKSTATIONS286

FIGURE A5. 17 – Q-Q PLOT AND BOXPLOTS TO ASSESS THE NORMALITY OF THE DIFFERENCES
BETWEEN TIME TAKEN PER CASE (EXCLUDING RECALLED CASES) AT THE DIGITAL AND
HYBRID WORKSTATIONS287

List of Tables

TABLE 1.1 – SUMMARY OF WORKSTATIONS INVESTIGATED IN THIS THESIS.....	20
TABLE 2.1 – DIMENSIONS OF THE FILM, HYBRID AND DIGITAL WORKSTATIONS IN COMPARISON TO RECOMMENDATIONS ...	50
TABLE 2.2 – EXISTING MUSCULOSKELETAL DISORDERS IN PARTICIPANTS....	52
TABLE 2.3 – CHANGE IN DISCOMFORT SCORES AFTER THE 45 MINUTE SESSIONS FOR THOSE COMBINATIONS OF WORKSTATION AND BODY PART FOR WHICH THE INTERQUARTILE RANGE EXTENDS BEYOND A SCORE OF 1.....	55
TABLE 3.1 – CORRELATIONS BETWEEN SUBSCALES OF WORKLOAD AND OVERALL WORKLOAD, WITH PARTICIPANT 7 REMOVED FROM THE ANALYSIS, *DENOTES CORRELATION IS SIGNIFICANT AT THE 0.05 LEVEL (2-TAILED), ** DENOTES CORRELATION IS SIGNIFICANT AT THE 0.01 LEVEL (2-TAILED)..	107
TABLE 5.1 – BREAKDOWN OF THE 160 CASES USED IN THE EXPERIMENT BY CASE TYPE, DIFFICULTY RATING AS ASSIGNED BY AN EXPERT RADIOLOGIST, AND SUSPICIOUS PATTERN TYPE. CALC IS AN ABBREVIATION OF CALCIFICATIONS.	155
TABLE 5.2 – COUNTERBALANCING APPLIED BETWEEN THE CONDITIONS OF FILM OR DIGITISED PRIOR MAMMOGRAMS. PRIORITY WAS GIVEN FOR COUNTERBALANCING BY PARTICIPANT TYPE AND EXPERIENCE, AND WHETHER CASES WERE READ WITH FILM OR DIGITISED PRIOR MAMMOGRAMS FIRST WITHIN EACH SESSION (TO AMELIORATE THE EFFECTS OF FATIGUE)	159
TABLE 5.3 – THE PROPORTION OF THREE TYPES OF NORMAL CASES PRESENT IN THE BREAST SCREENING PROGRAMME IN COMPARISON TO THE STUDY..	170
TABLE 5.4 – FOUR MEASURES OF PERFORMANCE USING FILM PRIOR MAMMOGRAMS, DIGITISED PRIOR MAMMOGRAMS AND NO PRIOR MAMMOGRAMS.	172
TABLE 5.5 – RESULTS OF THE MODEL CONVERTING SINGLE READER RESULTS TO DOUBLE READER WITH ARBITRATION .	173
TABLE 5.6 – PROJECTED COSTS PER 10,000 WOMEN AT A UK BREAST SCREENING CENTRE FOR IMPLEMENTING THREE DIFFERENT APPROACHES TO THE DISPLAY OF PRIOR FILM MAMMOGRAMS ...	175
TABLE 6.1 – TESTS FOR NORMALITY FOR THE NUMBER OF COMPARISONS TO THE PRIOR MAMMOGRAMS AT THE HYBRID AND DIGITAL WORKSTATIONS. ^ DENOTES LILLIEFORS SIGNIFICANCE CORRECTION *. DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE..	188
TABLE 6.2 – TESTS FOR NORMALITY FOR THE TIME TAKEN PER CASE AT THE HYBRID AND DIGITAL WORKSTATIONS. ^ DENOTES LILLIEFORS SIGNIFICANCE CORRECTION * DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE. ..	190
TABLE 6.3 – MEAN PROPORTION OF CASES FOR WHICH THE PRIOR MAMMOGRAMS WERE USED, NUMBER OF COMPARISONS AND TIME TAKEN PER CASE, FOR BOTH SCREENING AND EXPERIMENTAL SETTING WITH FILM AND DIGITISED PRIOR MAMMOGRAMS. DATA IS PROVIDED FOR ALL CASES, AND FOR JUST THE NORMAL SCREENING CASES (A NORMAL SCREENING CASE IS ONE WHICH WAS NOT RECALLED IN BREAST SCREENING PRACTICE BY EITHER READER). DATA IS FOR THE SEVEN PARTICIPANTS WHO TOOK PART IN BOTH THE OBSERVATIONS OF SCREENING PRACTICE AND THE EXPERIMENT. SIGNIFICANCE TESTS ARE WITHIN SUBJECTS T TESTS BETWEEN USING DIGITISED AND FILM PRIOR MAMMOGRAMS, A BLANK FIELD REPRESENTS NO SIGNIFICANT DIFFERENCE.....	196
TABLE A5.1 – TESTS FOR NORMALITY FOR THE COMPARISONS BETWEEN WORKLOAD SCORES AT THE HYBRID AND DIGITAL WORKSTATIONS. ^ DENOTES LILLIEFORS SIGNIFICANCE CORRECTION *. DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE.	267
TABLE A5.2 – TESTS FOR NORMALITY FOR THE CORRELATIONS BETWEEN SUBSCALES OF WORKLOAD AND OVERALL WORKLOAD. ^ DENOTES LILLIEFORS SIGNIFICANCE CORRECTION *. DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE.....	269
TABLE A5.3 – TESTS FOR NORMALITY FOR THE CORRELATIONS BETWEEN SUBSCALES OF WORKLOAD AND OVERALL WORKLOAD WITH PARTICIPANT 7 REMOVED. ^ DENOTES LILLIEFORS SIGNIFICANCE CORRECTION * DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE .	277

TABLE A5. 4 – TESTS FOR NORMALITY FOR DIFFERENCE BETWEEN THE TIME TAKEN PER CASE AT
THE DIGITAL AND HYBRID WORKSTATIONS ^ DENOTES LILLIEFORS SIGNIFICANCE
CORRECTION * DENOTES A LOWER BOUND OF THE TRUE SIGNIFICANCE.285

Glossary of Terms and Abbreviations

Abnormal case - A set of mammograms (from the same woman) which contain a malignant lesion which requires further treatment.

Batch – A set of screening bags to be read together typically consisting of the screening bags for one screening van for one day.

Benign case - A set of mammograms from a woman who was recalled from the breast screening programme, and had a biopsy with negative results (i.e. it was not malignant).

Biopsy – Removing cells from a breast for testing under a microscope for indications of malignancy

Breast Screening – Inviting healthy women to have mammograms (x-ray images) taken of their breast in order to detect breast cancer at its early stages.

Craniocaudal mammogram – An x-ray view of the breast where the x-ray beam enters at the cranial side (from the direction of the head) and exits at the caudal side.

Current mammograms – The mammograms most recently taken in the breast screening programme.

Digital mammography – Process of taking mammograms using digital acquisition and display, and thereby removing the need for film screen.

Digital workstation – Mammography workstation with digital current and digitised prior mammograms.

Film workstation – mammography workstation with film digital and film prior mammograms

Hybrid workstation – Mammography workstation with digital current and film prior mammograms

JAFROC analysis – Jackknife free response receiver operating characteristic analysis. A method of measuring signal detection performance using data of both confidence level that the signal is present/absent, and the location of the signal, which allows for more than one lesion per case. In breast screening the confidence level and signal location refers to the probability of malignancy and lesion location.

JAFROC figure of merit – Measure of cancer detection performance obtained using JAFROC analysis. Equals the probability that lesions on abnormal images are rated higher than false positive marks on normal images. Analogous to ROC area under the curve

Film screen mammography – Process of taking mammograms using x-rays incident on photographic film. The film is developed and viewed to search for indications of malignant growth.

LCD screen – Liquid Crystal Display screen. The most common method of high resolution display of digital mammograms.

LROC analysis – Localised Receiver Operating Characteristic analysis. A method of measuring signal detection performance using data of both confidence level that the signal is present/absent, and the location of the signal, which only allows for one lesion per case. In breast screening the confidence level and signal location refers to the probability of malignancy and lesion location.

Malignant lesion – A lesion which is cancerous and has a tendency to metastasize

Mammogram – x-ray image of a breast

Mammography Reader – A medical professional qualified to read mammograms in the NHS Breast Screening Programme. Either a radiologist or Radiography Advanced Practitioner who specialises in reading mammograms.

Medio-lateral oblique mammogram – An x-ray view of the breast in a slanting direction from the woman's side towards the midline

Multi-viewer – A backlit device for displaying film mammograms upon which hundreds of mammograms can be hung at one time, and the display can be moved from one set of mammograms to another through means of an electronically controlled roller system.

Musculoskeletal disorder – Work related injury or disorder of the muscles, nerves, tendons, ligaments, joints, cartilage or spinal discs for example carpal tunnel syndrome.

Normal case – A set of mammograms from a woman who was not recalled from screening for further tests.

NASA TLX – National Aeronautics and Space Administration Task Load Index. A set of questions about a task designed to elicit workload in that task.

NASA RTLX - National Aeronautics and Space Administration Raw Task Load Index. Version of NASA TLX without weighting the importance of the subscales.

Previous mammograms – Mammograms taken in previous screening rounds in the breast screening programme. A woman aged 56 attending screening will have previous mammograms from 3 and 6 years previously if she attended the first two rounds of screening to which she was invited.

Prior mammograms – The most recent previous mammograms.

Radiography Advanced Practitioner – Breast screening radiographer who has undertaken advanced training and is therefore qualified to read mammograms in the NHS breast screening programme.

Radiologist – Doctor specialising in breast screening radiology, who is qualified to read mammograms in the NHS breast screening programme

Recalled case – A woman who has been recalled from the breast screening programme for further tests after a suspicious pattern was detected in her mammograms.

ROC analysis – Receiver Operating Characteristic Analysis. A method of measuring signal detection performance using ratings of confidence level that the signal is present/absent. In breast screening the signal refers to presence of a malignant lesion.

RULA – Rapid Upper Limb Assessment postural analysis tool. Tool for estimating postural risk factors contributing to the risk of developing musculoskeletal disorders. Uses data concerning joint angles, weights carried and repetition rate.

Screening bag – A bag containing the information concerning the breast screening history of one woman, including previous mammograms and any test results.

Workload – Portion of mental capacity expended on a task. If there is insufficient mental capacity for the task requirements then fatigue and or performance decrements may result.

1 Introduction

1.1 Breast Cancer Screening in the UK

Over 12,000 women die from breast cancer in the UK each year, this has fallen by 20% since the Breast Screening Programme was initiated (Forrest, 1986). Over 1.7 million women are screened (Health and Social Care Information Centre, 2009) and approximately 1,400 lives saved annually (Austoker *et al.*, 2006). There is a radiation risk associated with breast cancer screening, and for every 35 lives saved by screening approximately one fatal breast cancer is caused by the x-ray radiation exposure in the screening process, (Austoker *et al.*, 2006). The Breast Screening Programme was initiated in 1988 on the advice of the Forrest Report (1986), with women aged 50-64 invited for film screen mammography, taking one mediolateral oblique x-ray of each breast. This was increased to two views of each breast (mediolateral oblique and craniocaudal) in 2003, alongside the increase of the upper age limit for invitation to attend to 70 (Department of Health, 2000). The age range was extended again to invite women aged 57-73 in 2007, alongside a commitment to introduce digital mammography (Cancer Reform Strategy, 2007).

After screening the mammograms taken are reviewed by two qualified readers. In the UK this is either radiologists that specialise in breast screening or radiography advanced practitioners, (radiographers trained to read breast screening mammograms). If these two readers disagree then a third reader arbitrates and their decision is final, as shown in figure 1.1. Henceforth, both radiologists and radiography advanced practitioners will be referred to as mammography readers when a distinction between the two groups is not necessary.

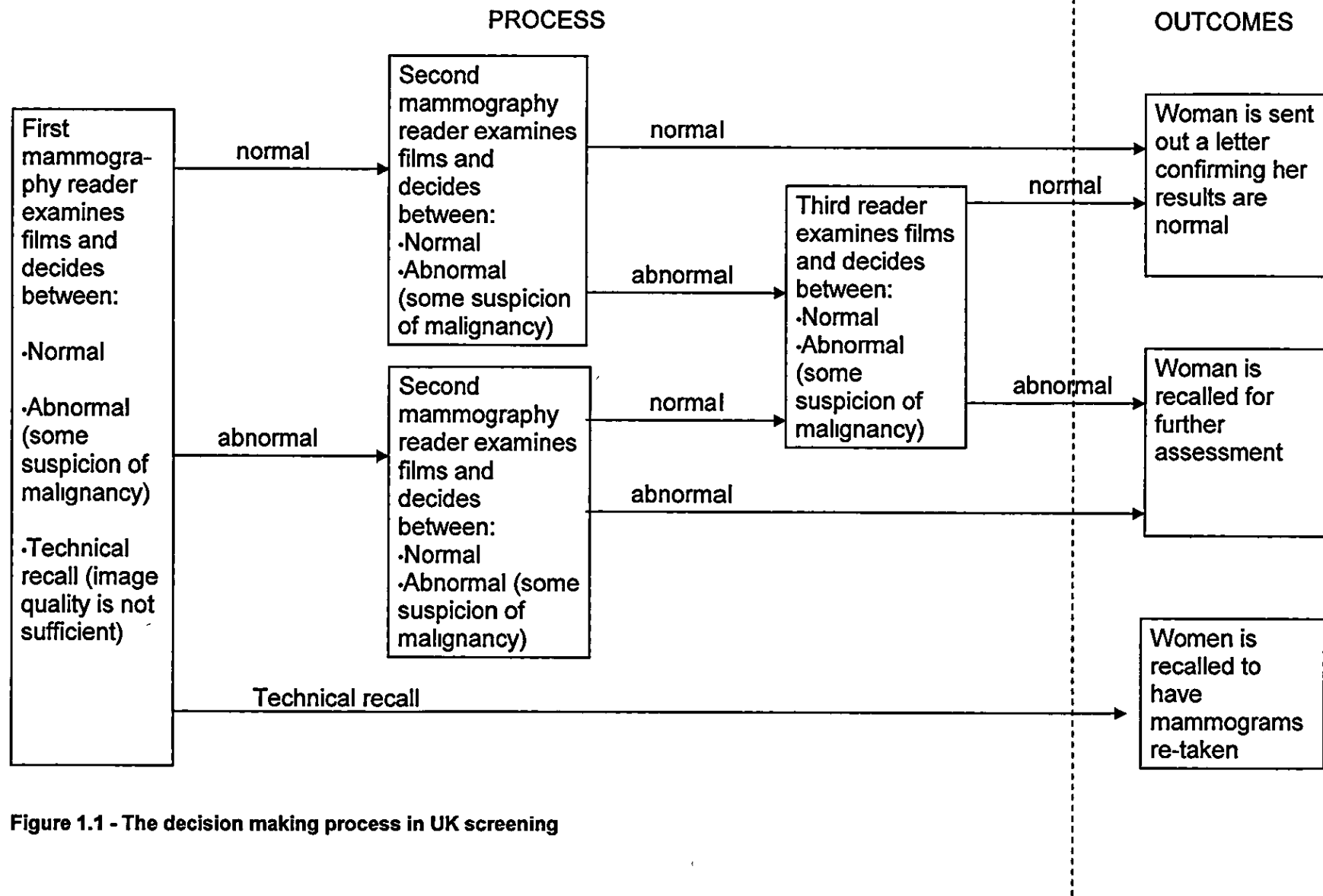


Figure 1.1 - The decision making process in UK screening

1.2 The Introduction of Digital Mammography

Several studies have investigated the performance of digital mammography in comparison to using screen film, the largest of which was the DMIST trial involving 49,528 cases in the USA (Pisano *et al.*, 2005). Each case was an asymptomatic woman who had both film and digital mammography. Cancer detection performance was not significantly different overall, but was superior using digital mammography for women under the age of 50 years ($p=0.002$), women with heterogeneously dense or extremely dense breasts on mammography ($p=0.003$), and premenopausal or perimenopausal women ($p=0.002$). An additional finding was that digital mammography required a lower radiation dose than did film screen. A similar study using 6736 cases, also in the US (Lewin *et al.*, 2002), found that recall rate was lower using digital mammography. In contrast, in a Norwegian study with 25,263 cases aged 45-69 (Skaane and Skjennald, 2004) randomly assigned to either digital or film screen mammography the reverse was found, with recall rates higher using digital mammography ($p<.05$), with an additional trend towards higher cancer detection rates using digital mammography ($p=.053$). However, this same research group had previously found no such differences in 3683 cases aged 50-69 which had both digital and film screen mammography (Skaane and Skjennald, 2003). These differing results could be due to differences in study design, differences between screening in the USA and Europe (the USA has a higher recall rate; Smith-Bindman *et al.*, 2003), or confounding variables such as outlined by Bick and Diekmann (2007, pg 1935) "differences in positioning

and reader performance far outweigh any difference in the acquisition technique”.

In film screen mammography x-rays (mammograms) of the breast are taken using photographic plate, which is developed and hung (displayed) on a backlit multi-viewer. In digital mammography the photographic plate is replaced by an electronic version, and the image stored digitally and displayed on high resolution computer screens. Examples of the workstations which display film and digital mammograms are shown in figure 1.2. Digital mammography provides a greater range of display options than film screen. When reading film mammograms a magnifying glass can be used, and the mammograms can be hung in any order on the multi-viewer. However digital mammography allows the introduction of image manipulation tools including electronic magnification, contrast adjustment, edge enhancement, and image rotation and resizing. Some of these tools are shown in figure 1.3.

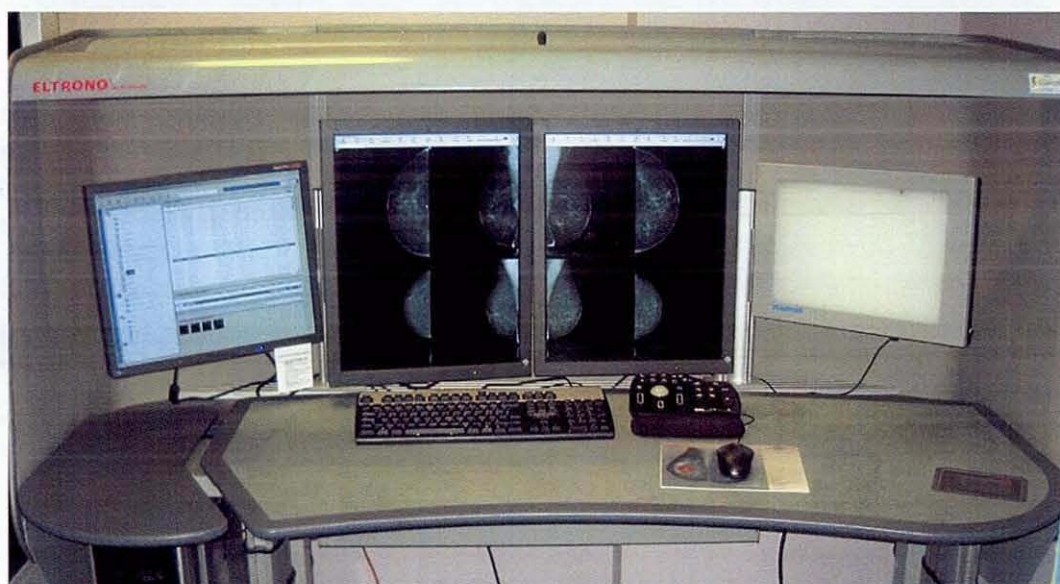
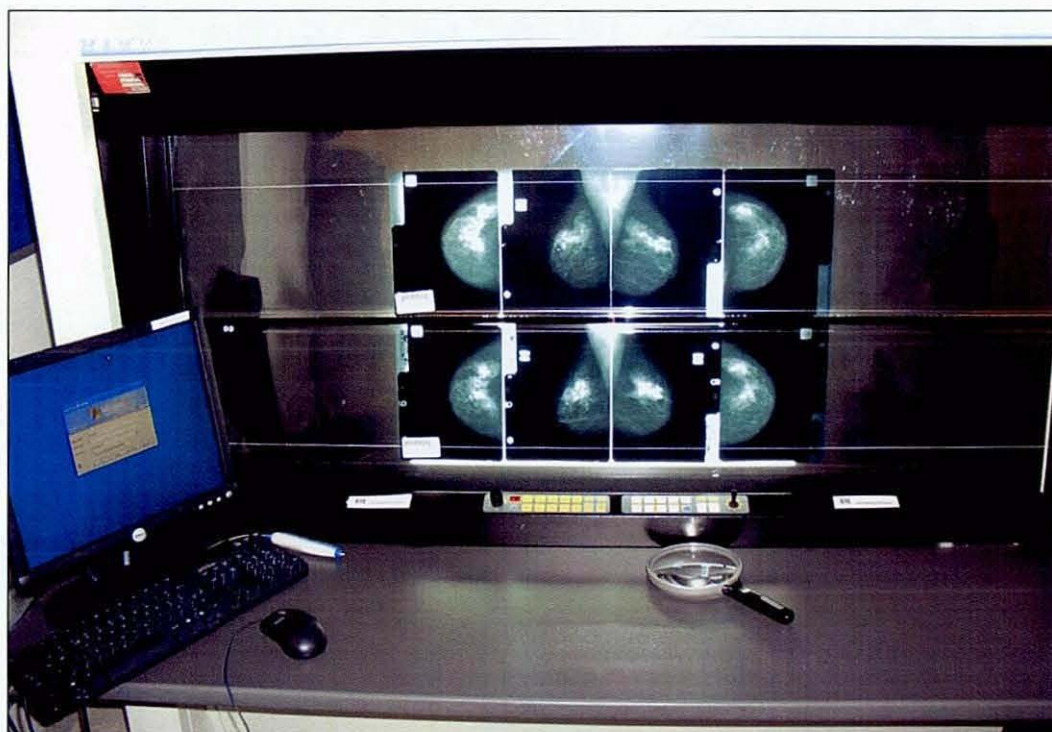


Figure 1.2 - Two workstations for reading mammograms in the Breast Screening Programme. The film workstation (above) has been used since the NHS Breast Screening Programme was set up in 1988, the digital workstation (below) will replace such film workstations over time. Both have current mammograms on the upper row and prior mammograms on the lower row, and a smaller screen to the left to show the details of the women screened and allow the reader to input results.

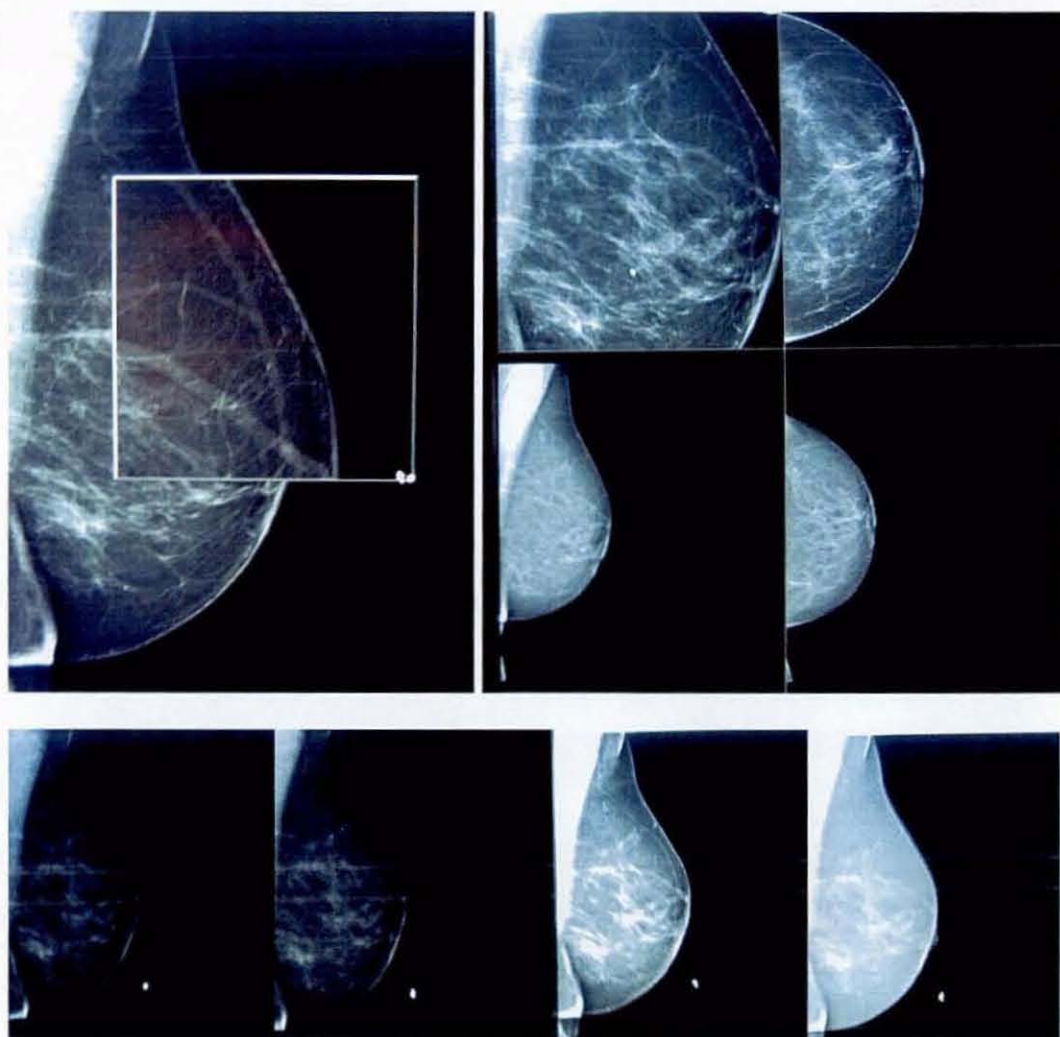


Figure 1.3 – Digital mammograms with image enhancement tools. Magnification with the electronic magnifying tool (above left), and with full image magnification (above right). Contrast adjustment for the same mammogram is shown on the lower row.

In the UK the NHS Breast Screening Programme (NHSBSP) is making the transition from film to digital mammography in 2009/10. The Cancer Reform Strategy (Department of Health, 2007, pg 47) states that the introduction of digital mammography "would allow the image to be manipulated so it improves the radiologist's ability to interpret breast tissue. Digital images could be exchanged electronically between radiologists at different hospitals to discuss difficult cases... [and] provide revenue savings in terms of reduced radiographer time and less chemicals or film handling and printing". Therefore this strategy commits every breast screening centre to have at least one full field digital mammography set by 2010. Simultaneously the age range for women invited for screening will be increased to 47-73 years by 2012, with an estimated increase in women screened per year of over 400,000 (24%). It is unclear when the transition to digital mammography will be completed. The strategy states that direct digital mammography will be introduced over the same time period as the implementation of the age extension, but it is unclear whether this means that all film screen equipment will be phased out by 2012. Therefore it is likely that many screening centres will be using both film and digital mammography equipment concurrently for several years.

In the transition to digital mammography, the current mammograms are displayed digitally on high resolution LCD screens, but the prior mammograms from the previous screening round remain in film format. These film prior mammograms can either be displayed in film format on a backlit multi-viewer or light box, or digitised and displayed onscreen alongside the current mammograms. There are several different designs and suppliers available for each of these options. The different display media for the prior mammograms are shown schematically in figure 1.4.

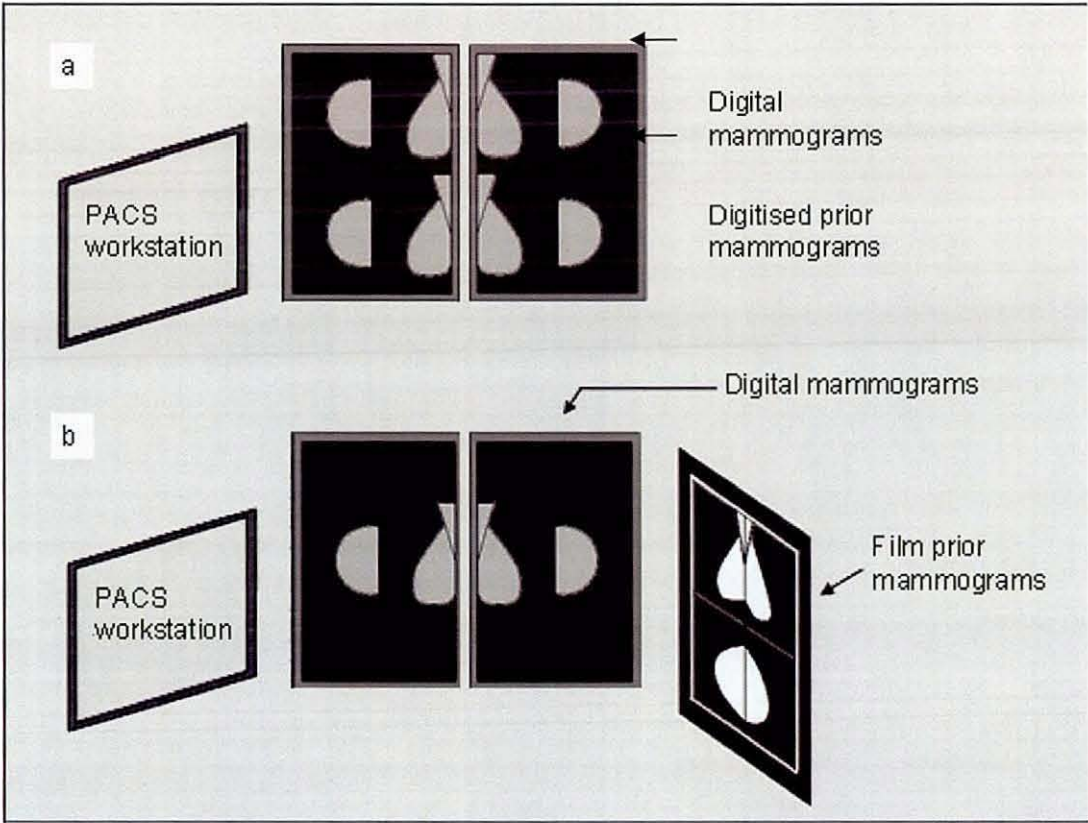


Figure 1.4 – Schematic view of (a) digitising the prior mammograms and displaying onscreen alongside the digital current mammograms, and (b) display in film format on an adjacent backlit multi-viewer.

At the time of commencement of the research in October 2006 there were three breast screening centres in the UK using digital mammography. In Coventry digital mammograms were acquired for screening using the Sectra MicroDose system, with film prior mammograms pre-hung on a multi-viewer adjacent to the Sectra digital workstation. In Nottingham the GE Digital mammography system was being used to acquire mammograms in the Breast Screening Programme, but there was no storage or display capacity so the digital images were being printed onto film and displayed on a multi-viewer. In Exeter digital mammography screening was being used, but by a private company due to the competence of two radiologists previously working there having been called into question. Many other breast screening centres at that time had digital mammography equipment for diagnostic but not screening use.

Since the beginning of the research many other screening centres in the UK have started using digital mammography for screening. Derby introduced the first fully integrated digital mammography system, which means that the computer systems for storage, display and reporting all work together. At this breast screening centre the prior mammograms are not displayed but are provided at the workstation in order for readers to hang them themselves when they think it necessary. At Nottingham breast screening centre they are using the Sectra MicroDose system to acquire mammograms, but still print them and display on a multi-viewer alongside the film prior mammograms. They plan to either hang the film prior mammograms for every case, or provide them at the workstation for readers to hang the images themselves when they have acquired a digital display capability. Great Yarmouth are the only screening centre who currently digitise all prior screening mammograms. They have one

screening van with the Hologic Selenia digital mammography system, and digitise all prior mammograms using the Hologic R2DX/LS digitiser. Film display of prior mammograms on a multi-viewer was not an option for this breast screening centre due to their room layout. At Manchester's Nightingale Centre the same problem with not being able to place a multi-viewer adjacent to the digital workstation was encountered, but they do not digitise the prior mammograms, they simply provide a light box for the reader to hang them when they consider it necessary. The Coventry, Solihull, and Warwickshire breast screening service is now fully digital, and hangs all prior mammograms in film format adjacent to the digital workstations. A trial of digitising all prior mammograms is proposed to start in 2009 at this breast screening centre.

When this research commenced there was no guidance in the UK about how to present the prior mammograms in the transition to digital mammography, or indeed about whether it is necessary to use them at all. In the American transition to digital mammography each breast screening centre made its own decision about whether and how to use the prior mammograms, and some decided not to use them at all, (E. Krupinski, personal communication). However, in the USA screening is every year whereas in the UK it is every three years, and specificity in the USA is lower and less of a concern to practitioners (Smith-Bindman et al., 2003). In the Dutch screening programme all prior mammograms from one screening round previously will be digitised, in light of evidence showing the importance of prior mammograms, and the difficulties for the reader in using digital current and film prior mammograms (N. Karssemeijer, personal communication, 2008).

Therefore there are four options under consideration by breast screening centres in the UK for display of the prior mammograms in the transition to digital mammography.

- Do not display the film prior mammograms at all
- Do not display the film prior mammograms for every case, but make them available so the reader can hang them on a light box as and when necessary
- Display the film prior mammograms for every case on a multi-viewer adjacent to the digital workstation
- Digitise the prior mammograms and display onscreen alongside the current mammograms at the digital workstation.

Adopting a particular approach, either nationally or locally has potential implications for departmental workflow, staff stress and workload, and most importantly cancer detection performance.

1.3 Aims and Objectives

The aims of this research are as follows:

1. To produce recommendations about how the analogue prior mammograms should be displayed in the transition to digital mammography.
2. To produce data to support these recommendations encompassing both reader comfort and reader performance.
3. To publish these recommendations through peer reviewed journals, practitioner conferences, and through reporting to the NHS Breast Screening Programme

The objectives are therefore:

1. To understand the literature on the use of prior mammograms in the transition to digital mammography
2. To measure the impact of the display medium of the prior mammograms on
 - a Physical comfort, and risk of musculoskeletal disorders in mammography readers
 - b. Mammography readers speed of reading and perceptions of workload
 - c. The amount that the mammography readers use the prior mammograms
 - d. Cancer detection performance

3. To determine whether the type of mammography reader (radiologist or radiography advanced practitioner) impacts on the metrics from objective 2
4. To test all findings by publishing in peer reviewed journals and presenting at both academic and practitioner conferences
5. To publish guidance which will assist UK breast screening centres to decide how to display prior mammograms in the transition to digital mammography.

1.4 Methodological Approach

The transition to digital mammography was already underway at the start of this research, with all breast screening centres targeted at having at least one digital mammography unit for screening by 2010. The decision of how to implement the transition to digital mammography will be made in each breast screening centre at that time, with what research evidence is available to them. No other researchers in the UK are known of who were simultaneously researching the same problem. Therefore, to ensure the results of this research were of practical use rather than simply an academic exercise the research must cover as many of the display options as possible, and be completed by 2009. To achieve this within the timescales it was necessary that the research considered outputs and results rather than the complex mechanisms by which those results were reached. The research was conducted as a series of field experiments which mirrored screening practice as closely as possible. As the research is designed to influence screening practice, a focus on producing

results which could be presented in formats which would influence breast screening practitioners was kept.

1.5 Thesis Outline

All four of the possible methods for displaying prior mammograms were investigated. However, in those experiments in which observations of behaviour in screening practice were made (chapters 2-4) only film and digitised format display for every case could be implemented. There is evidence from one previous study (Roelofs et al., 2007) that not displaying the prior mammograms for every case could be a detriment to performance, and therefore such an implementation would be unethical when participants are reading live screening cases. In chapters 2-4 an additional implementation was used, of film current and film prior mammograms displayed on a multi-viewer, which represents the workstation used all UK breast screening centres before implementing digital mammography, see table 1.1.

Table 1.1 – Summary of the workstations investigated in this thesis

Workstation	Display of Mammograms		Chapters in which Workstation is included				
	Current	Prior	2	3	4	5	6
Film	Film	Film	√	√	√		
Hybrid	Digital	Film	√	√	√	√	√
Digital	Digital	Digitised	√	√	√	√	√
No Priors	Digital	None				√	

Chapter 2 describes an investigation comparing workstation ergonomics at a digital workstation with film or digitised prior mammograms, and at a traditional film workstation. The Rapid Upper Limb Assessment (RULA, McAtamney and Corlett, 1993) postural analysis tool, and body part discomfort charts were used to ascertain any differences in risk levels of postures adopted, and discomfort at the different workstations.

Chapter 3 examines workload and speed of reading mammograms at a digital workstation with film or digitised prior mammograms, and at a traditional film workstation. The NASA Task Load index questionnaire (Hart and Staveland, 1988) was used to understand differences in perceptions of workload with different presentation media of prior mammograms. Analysis of videotape of live screening sessions was used to determine the time taken per case at the different workstations.

Chapter 4 focuses on the level of use of the prior mammograms at a digital workstation with film or digitised prior mammograms, and at a traditional film workstation with film prior mammograms. Video-tape of live screening sessions was analysed to determine whether the participant was looking at the current mammograms, the prior mammograms, or at something else. The number of times the participant looked at the prior mammograms per case, and the proportion of cases for which the prior mammograms were looked at on at least one occasion were compared across workstations.

Chapter 5 describes measurements of cancer detection performance at a digital workstation with film prior mammograms, with digitised prior

mammograms, and without prior mammograms. Difficult test cases were used to enable identification of performance differences. Jackknife Free Response Receiver Operating Characteristic (JAFROC, Chakraborty, 2006) analysis of performance was used, with an additional analysis of numbers of correct and incorrect recalls.

Chapter 6 compares behaviour (use of prior mammograms and speed of reading) with the test cases as described in chapter 5, and in screening practice as described in chapters 3 and 4. This is used to evaluate the applicability of the results of the performance experiment to screening practice.

1.6 Subject Matter Immersion

The methodological approach taken required modelling the screening situation closely, and therefore conducting research in a live screening environment. It was important to gain a thorough understanding of that screening environment so that the research could be implemented in a manner which did not impact on screening practice, and closely modelled it for experimental purposes. Therefore a series of activities involving subject matter immersion were undertaken.

The three breast screening centres in the UK which had digital mammography at the time of commencement of the study were all visited initially, and from this an understanding of the problems and opportunities of the introduction of digital mammography. The centre which was most advanced with the introduction of

digital mammography was University Hospital (Coventry) and this was selected as the study centre.

Informal interviews with members of staff at the study centre were conducted over a series of 10 visits, alongside job shadowing for 3 roles. The interviews were with two radiologists, two radiography advanced practitioners, one radiographer, one radiography assistant practitioner, the superintendent radiographer and two member of administrative staff. Job shadowing was with a radiographer taking mammograms in one of the screening vans, a radiologist reading screening mammograms and a member of the administrative team.

A task analysis was developed of the screening process at the study hospital using the amalgamation of these data. The task analysis was shown to the Programme Manager and a radiography advanced practitioner at the study hospital, and amendments were made based on feedback. The task analysis was used in three ways. Initially the development of the task analysis was used to immerse the investigator into the new field. Then when introducing digitisation of prior mammograms into the live screening situation it was used to negotiate how digitisation would be integrated into the departmental workflow. This enabled the experiment to go ahead without any adverse effects on the department achieving its targets, in particular to dispatch results within 2 weeks of the screening session. Finally the task analysis was used to identify events for the postural analysis in chapter 2. Appendix 1 details the task analysis.

1.7 Literature Review

A more detailed literature review will be given as the introduction to each chapter, with an overall introduction to the area described here. There is very little research concerned specifically with the transition to digital mammography, with an extensive literature review resulting in only one paper on the topic (Roelofs et al., 2007). This paper describes 12 radiologists reading 160 mammograms (50% malignant) both with and without prior mammograms. When reading without prior mammograms the participants were asked to identify for each case whether they considered the prior mammograms necessary. This created a third condition of 'with prior mammograms upon request'. The findings were that performance improved with prior mammograms, in comparison to either without prior mammograms, or with prior mammograms upon the readers' request. The implications of this study are that the options of not displaying the prior mammograms at all, or asking the mammography reader to hang them themselves as and when they consider it necessary in the transition to digital mammography may both be suboptimal.

There are several issues with the applicability of these findings to the Breast Screening Programme in the UK. Firstly all of the cases used were analogue in origin, digitised and stored for the study at 100 μ m, and displayed digitally. This means that image was stored as a set of square pixels each of which was 100 μ m in diameter, and therefore no matter how high the resolution of the monitor or the magnification used there is no information available to display

beyond this 100 μ m limit. In practice the current mammograms would be digitally acquired images at a resolution of 50 μ m, and the prior mammograms would be analogue film images that had been digitised. This is significant as it may be more difficult to make comparisons between digital and analogue images, than between two sets of analogue images which have been digitised, and this could influence performance. It could be argued that analogue prior mammograms are not useful in the transition to digital mammography, because analogue and digital images cannot be accurately compared, and the paper by Roelofs *et al.* (2007) could not answer such criticism.

There were only two reading sessions per participant, one with and one without prior mammograms. Each reading session included 160 cases, which is a very high number of cases to read all in one session, and is not the norm in the UK due to the potential for fatigue effects. Additionally, as the sessions with and without the prior mammograms were on different days at least one month apart confounding variables could have affected performance such as how much sleep the participant had had the night before, what duties were undertaken before reading the case set, and the time of day it was read.

The Localised Receiver Operating Characteristic (LROC) paradigm was used with lesion localised fraction at non-lesion localised fraction at less than 25% used as the performance metric. This means that for each case the participants' marked the location of any lesions they perceived and assigned a probability of malignancy. Only the highest rated lesion for each case was used in the analysis. The threshold for recall was determined from the normal cases as the point at which one incorrect lesion would be identified in 25% of cases,

i.e. recalling 25% of normal cases. This threshold was applied to the abnormal (malignant) cases, and the proportion which were rated above this threshold was used as the measure of performance. There are several issues with this. Firstly, in practice all lesion locations would be investigated further, even if there were more than one per case. If there were two suspicious lesions on the same woman in most cases a biopsy would be taken from both locations, rather than simply ignoring one. Secondly, choosing the recall threshold such that 25% of normal cases are recalled is arbitrary and not evidence based. The authors justify this by stating that "in breast cancer screening the rate of non-lesion locations generally is lower than 10%...a relatively large interval was chosen because our study sample was enriched with difficult normal and benign cases" (Roelofs *et al.*, 2007, pg 74). In the NHS Breast Screening Programme 4.6% of women over 45 are recalled for further tests, and 0.8% have cancer, therefore the threshold is such that 3.8% of normal cases are recalled (The Health and Social Care Information Centre, 2009). In the study by Roelofs *et al.* (2007) where 50% of the cases were cancerous the threshold for recall was chosen such that 25% of normal cases were recalled. This cannot be related to recall decisions in screening practice. Therefore, little can be concluded about the effect of the prior mammograms on number of correct and incorrect recalls in screening practice.

The participants were radiologists from all over Europe, however some of them were not familiar with soft copy reporting. To prepare them they read 150 digital mammography cases so that they could learn to use the magnification, contrast, and computer aided detection tools. The behaviour and performance of such participants with little experience reading soft copy mammograms may

differ from those who are experienced. The prior mammograms were presented behind the current mammograms, with participants able to toggle between the current and prior images. This presentation will be novel to participants who are not experienced in soft copy reading, as when using films it is necessary to display current and prior mammograms adjacent to one another. Again this inexperience may affect how the prior mammograms are used, and performance using the prior mammograms.

The option of displaying the prior mammograms on a multi-viewer adjacent to the digital workstation was dismissed in this study, on the grounds that "reading digital images in combination with film images is difficult to organize and may lead to a loss of efficiency" (Roelofs et al., 2007, pg 71). Whilst this is indeed true the same could be said about the digitisation of prior mammograms. In the UK presenting the prior mammograms in film format is a viable option worth investigating, and therefore more research is needed in this area.

Finally there are some practical differences between the study design and UK screening practice. In the UK mammograms are taken every three years, and most commonly only the prior mammograms from the most recent previous screening round are displayed. In this Dutch study mammograms were taken every 2 years, and two sets of previous mammograms were available in addition to the current mammograms.

1.8 Contribution to Knowledge

The research in this thesis makes an original contribution to knowledge in several areas. In the transition to digital mammography the research about the change in workstation ergonomics, workload, and the behaviour of participants is all novel, and forms the only information available in these areas for breast screening centres making decisions about how to undertake the transition to digital mammography. The tools of RULA postural analysis, and NASA TLX workload assessment are both well established, but have been applied here to a novel area, mammography workstation ergonomics.

There is some previous research about cancer detection performance in the transition to digital mammography (Roelofs *et al.*, 2007), the work in this thesis makes an original contribution in several ways. Firstly when using digital current mammograms the difference in performance between using film and digitised prior mammograms has not been investigated before. Secondly, whether analogue prior mammograms are still beneficial to performance when using digital current mammograms has not previously been investigated. And finally results were obtained both in terms of JAFROC figure of merit, and in terms of increases in recall rate. Providing results in terms of the effect on the recall rate in screening not only provides a contribution to knowledge, but one that practitioners will find relevant and easy to understand, and is therefore more likely to influence decisions.

The measurements of behaviour in screening practice in comparison to reading test cases is also novel. Many previous studies have investigated behaviour

reading mammograms using eye tracking equipment (for example Mello-Thoms, 2006, Krupinski and Nodine, 1994) However, no other studies could be found which measured behaviour both in screening practice and reading test cases and made comparisons between the two. This is an important area as it provides some indication of the applicability of the results of ROC type studies using test cases to real world screening.

2 Physical Comfort in the transition to Digital Mammography

2.1 Introduction

A large body of research is available concerning the dimensions and design of office workstations, a section of which is presented here. Some research attention has been paid to the design of radiology workstations, which differ from an office workstation both in some design aspects, and the safety critical nature of the work. Significant gaps in this research are described, particularly in both objective and subjective measurements of participants using the workstations.

Postural risk factors when using a standard office workstation have been established based on invasive measurements of joint pressure, and subjective measurements of discomfort and fatigue. Prolonged wrist flexion and extension, (Gelberman *et al* , 1981), wrist ulnar and radial deviation, (Werner *et al.*, 1997) and forearm pronation and supination (Werner *et al.*, 1997, Rempel *et al.*, 1998) increase the risk of carpal tunnel syndrome through increases in carpal tunnel pressure (Szabo and Chidgey, 1989, Gelberman *et al.*, 1981, Phalen, 1966). When the elbow is flexed beyond 90 degrees both intraneural (Gelberman *et al.*, 1998), and extraneural (MacNicol, 1982) pressures increase, and arm discomfort increases (Sauter *et al.*, 1991). Increases in upper arm flexion and abduction result in decreases in time to localised muscle fatigue (Chaffin, 1973). Increased neck flexion from 30⁰ to 45⁰ results in shorter time to

localised muscle fatigue (Chaffin, 1973), and increased time with the neck in greater than 20° flexion has been found to be correlated to greater neck discomfort ($p < .01$, Kilbom *et al.* 1986). Repeated neck extension has been linked with neck pain in fruit pickers (Sakakibara *et al.*, 1995). An extensive review of the literature concluded that there is “evidence that work-related awkward postures [twisting and bending] are associated with low-back disorders” (Putz-Anderson *et al.*, 1997, pg 373), with back disorder associated with trunk flexion, twist, and lateral bend in auto assembly workers (Punnett *et al.*, 1991) When seated, pressure in the nucleus pulposus of the intervertebral discs in the lumbar spine is lowest when adopting a slightly reclined posture resting against a backrest with a 50mm deep pad to support the lumbar region of the spine. An upright posture with no back support increases the intervertebral pressure, and this pressure is greatest when the back is in a slumped posture with no support from a backrest. (Andersson *et al.*, 1974). The available research data have been combined to form detailed recommendations for the dimensions and adjustability of office chairs, desks, monitors, keyboard and mouse (Kroemer and Grandjean, 2005). There is also evidence to suggest that moving between postures whilst at a workstation is beneficial, as regular movement of the back creates a diffusion gradient which enables nutrients to reach the centre of the intervertebral discs in the spine. (Kroemer and Grandjean (2005), pg 75). This is supported in legislation with the statement that “work organisation, job content, and furniture design should encourage user movement. This means that prolonged static sitting posture is minimized and that more or less continuous voluntary adjustments of posture can be made” (ISO 9241-5, 1998, pg 5)

Radiology-specific research about workstation design has primarily concerned lighting levels (for example Goo *et al.*, 2004), and monitor type and quality (for example Krupinski *et al.*, 2003). However, some design processes for the introduction of digital imaging have been documented. Ratib *et al.* (2000) used 3D modelling and iterative design with user input to redesign a radiology room. The final design had workstations in the middle of the room facing outwards, in an inverted cube shape. This layout prevents noise and light from straying from one workstation to another, however the design also necessitates positioning the light box for viewing films above the workstation. No consideration is given by the authors into the potential strain on the neck that this may cause, or the reaching involved to hang the prior mammograms. Nagy *et al.* (2003) conducted a paper-based user centred design. A hexagonal workstation layout was developed to give each workstation a 120° curve. This enables multiple monitors to be viewed easily, and light from one does not reflect on another. All previous mammograms were digitised to minimise stray light from film viewers. Adjustable chairs with lumbar support, wrist support, and stand up workstations were provided. There are some ergonomic recommendations for the radiology workstation, (Harisinghani *et al.*, 2004, Siddiqui *et al.*, 2006, Nagy *et al.*, 2003) in particular highlighting the issues of chair adjustability, neutral wrist position, and optimising monitor height. However, these are not based on new research and do not differ from those concerning an office workstation, (Kroemer and Grandjean, 2005).

If a worker has pre-existing musculoskeletal problems, then according to Cumulative Load Theory (Kumar, 2001) they are more likely to experience further musculoskeletal problems in the future. In the NHSBSP due to a

shortage of radiologists, radiographers are being trained to read mammograms in addition to taking them (The Department of Health, 2000). Therefore understanding the prevalence of musculoskeletal disorders in the population of breast screening radiologists and radiographers will provide information on the propensity of that population to further injury in particular body parts, and therefore inform workstation design. May *et al.* (1994) conducted a survey of 320 breast screening radiographers and found that 76% reported some pain, with over 20% experiencing pain in the lower back, 14% reporting pain in the neck and 13% in the upper back. There is anecdotal evidence of musculoskeletal disorders in four radiologists (Ruess *et al.*, 2003) showing evidence of both carpal and cubital tunnel syndrome.

Detailed ergonomic requirements for the layout of a standard office workstation are available based on objective joint pressure data. However the radiology workstation differs from this standard workstation because the monitor/screen area to be viewed is much larger, in some instances a magnifying glass is used, there is a greater need to view very small details, and the task is safety critical. Additionally the breast screening task differs from other radiology tasks in its repetitive nature. Investigation is required to ascertain whether these aspects have implications for mammography reader comfort and risk of musculoskeletal disorders, and how any risks can be minimised. The transition to digital mammography provides a unique problem of viewing prior mammograms, and whilst there has been some investigation into the implications of how the prior film mammograms are displayed in terms of performance (Roelofs *et al.*, 2007), there are no data available concerning how the display might affect workstation ergonomics.

2.2 Aims

1. To determine whether the change from film to digital mammography will impact radiology workstation comfort and risk of musculoskeletal disorders.
2. To determine the impact on comfort and risk of musculoskeletal disorders at the radiology workstation of digitising prior mammograms in preference to displaying them in film format during the transition to digital mammography
3. To determine whether there are any differences between radiologists and radiography advanced practitioners in relation to aims 1 and 2.

2.3 Choice of Methods

2.3.1 Design approach vs ergonomic assessment

The study was an assessment of an existing workstation rather than the development of a new one, so approaches such as iterative design, and fitting trials were not appropriate. The aim was to conduct an ergonomics assessment to establish the workstation comfort, and the risk of musculoskeletal disorders with prolonged use. The primary focus was not usability, and therefore approaches such as interview, focus groups, and heuristic evaluation were not used.

2.3.2 Workstation Evaluation without Participants

The people using the workstation are all experienced medical professionals, and therefore it is prudent to conduct any investigations that are possible without the use of participants. An analysis of a workstation can be conducted without participants in two ways: using anthropometric data; or using recommendations from studies on similar workstations. Anthropometric data give the normal distribution of a population's dimensions, such as popliteal height to inform chair height, or seated eye height to inform monitor placement. A standard approach is to accommodate a range of sizes from the 5th percentile female size to the 95th percentile male. This approach was not chosen because anthropometric data have already been combined with comfort rating data and joint pressure measurements to produce recommended dimensions for a

standard office workstation. There is little benefit in repeating this process for the radiology workstation, as there are many similarities between this and the standard office workstation. Therefore comparisons of the workstation dimensions to the recommended dimensions form the first stage in the research.

2.3.3 Participant Observation and Measurement Methods

Several participant observation methodologies are available: postural analysis; workspace envelopes; computerised position and velocity measurements; biomechanical analysis of stresses; and electromyography (EMG) measures of muscle activity.

Computerised systems are available to measure posture directly using a transmitter pad placed at various points on the body alongside an array of detectors around the participant. This equipment enables detection of both joint angles and movement velocity. This provides rich postural information, but was not appropriate for this study due to the limited space surrounding the workstation for detector placement, and the attachment of the detector pads would have been difficult to administer in a busy hospital department and may have interfered with the participants' natural movements. Muscle activity can be determined either through biomechanical analyses of joint angles and weights or directly using surface EMG equipment. However this approach is more appropriate to situations where there are large forces involved, and both methods would be difficult to apply in practice.

Measurement of workspace envelopes gives information about the efficiency of workstation design. The normal workspace is the area where objects can be moved with a sweep of the forearm, maintaining the upper arm in vertical position; the maximal workspace is that which can be reached with the arm extended but the torso remaining upright Farley (1955), and the extreme workspace defined as the area which can be reached by both extension of the arm, and tilting of the torso, (Das and Grady, 1983). There is evidence that working within the normal work envelope decreases worker physiological cost Sengupta and Das (2004), and increases performance for manual tasks (Ellis, 1951)

Postural analysis associates risk scores with different body joint angles, and provides methods of summing these to provide a comparative risk score. This is a quick and simple method of comparing different workstations and highlighting higher risk postures, and is used to highlight areas which require further investigation. Rapid Upper Limb Assessment (RULA, McAtamney and Corlett, 1993) is the postural analysis tool most appropriate for use with a seated task at a workstation. This is because the tool is designed for light office work, and has been tested in a VDU based task. Validity of the method was partially shown with an association found between postural scores for the neck and lower arm and discomfort. Reliability testing is reported as showing a "high consistency of scoring amongst subjects" (McAtamney and Corlett, 1993, pg 98). The postural analysis tools considered but not used were OWAS (Heinsalmi, 1986) because its scoring system is designed for heavy manual labour, and REBA (Hignett and McAtamney, 2000) which, whilst it has been validated for use in health care, is designed for standing tasks.

2.3.4 Subjective Comfort Reports

Comfort and discomfort are subjective concepts, and therefore subjective measurement is appropriate. However, it must be considered that reports of discomfort are linked not only to physical factors but also to psychosocial factors (Bongers *et al.*, 1993, Ferguson and Marras, 1997). Borg (1998) provides a rating of physical exertion scale, however the screening task is too sedentary for this to be appropriate. A Body Part Discomfort Survey (Corlett and Bishop, 1976) uses ratings of discomfort in different body areas, as defined by a body map. Ferguson and Marras (1997) describe a model of lower back pain where discomfort is the first symptom of a musculoskeletal disorder, followed by more severe pain, time off work and eventual disability. Therefore discomfort can be used as an early indicator of workstation design issues which can lead to musculoskeletal disorders.

The approaches used to analyse the workstation will be comparisons of workstation measurements to recommendations, and RULA postural analysis in conjunction with Body Part Discomfort scoring.

2.4 Method

Full NHS ethical approval was granted from the South East Research Ethics committee, reference number 07/MRE01/55. The participant information sheet and informed consent form are in appendix 2.

2.4.1 Workstations

The same three workstations are investigated in chapters 2-4, and two of these workstations (digital and hybrid) are investigated further in chapter 5 and 6. Therefore these will be introduced in depth in this chapter and referred to thereafter. The workstations which were investigated were: film; hybrid; and digital. The film workstation represents a typical workstation used in the NHS Breast Screening Programme currently. The hybrid and digital workstations represent two different methods of displaying the prior mammograms during the transition to digital mammography.

The film workstation consisted of a Mammolux XL (Planilux, Germany) backlit multi-viewer with both current film and prior film mammograms displayed together, see figure 2.1. The mammograms were acquired using a Mammomat 3000 Nova screening unit (Siemens, Germany), with Kodak MIN-R2000 mammography film, developed using a Kodak X-OMAT Multiloader 7000 (Carestream Health, Toronto, Canada). The chair was adjustable in seat height from 44cm to 54cm cm, the work surface was at a height of 75cm and not adjustable. Maximum span of areas to be viewed was 61cm in height and

145cm horizontally. A magnifying glass of weight 500 grams was provided. Screening decisions were entered into the computer using barcode and a barcode reader pen, and keyboard and mouse for recalled cases. Each case has a screening bag associated with it, containing previous mammograms and data for that woman.

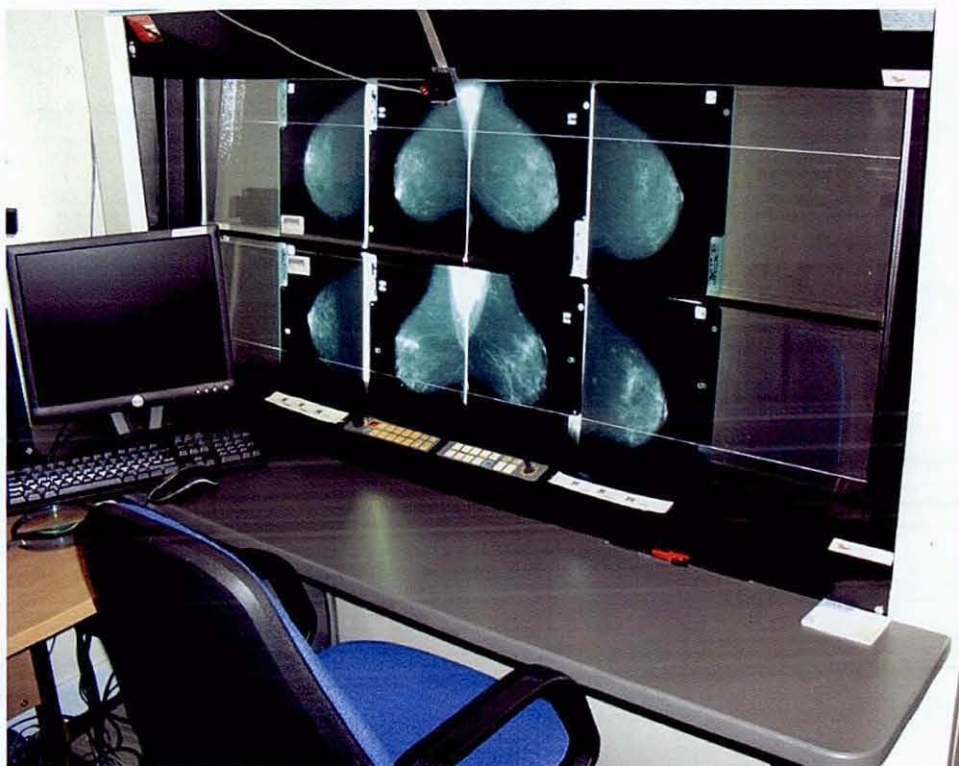


Figure 2.1 - The film workstation with film current mammograms on the top row and film prior mammograms on the lower row.

The hybrid workstation contained two Radiforce 54cm five megapixel LCD monitors (EIZO, Japan) to display the current digital mammograms, and an adjacent and perpendicular backlit multi-viewer (Mammolux XL, Planilux, Germany) to display the prior film mammograms, see figure 2.2. The LCD screens were 39cm wide and 47cm tall each, with viewing area 34 x 42.5cm. These were positioned vertically so that the lowest viewing surface of the screen was 8cm above the table. The chair was adjustable in seat height from

44 to 54 cm. Screening decisions were entered by signature on the paperwork contained in the screening bags.



Figure 2.2 – The hybrid workstation with current mammograms viewed on LCD screens to the left, and prior mammograms viewed on a multi-viewer to the right.

The digital workstation contained two Radiforce 54cm five megapixel LCD monitors (EIZO, Japan) displaying the current digital mammograms and the digitised prior mammograms, see figure 2.3. There were three hanging protocols (i.e. three layouts in which the mammograms were presented) set up on the workstation. The first hanging showed the digital current mammograms on the upper row and the film prior mammograms on the lower row, the second hanging showed medio-lateral oblique views of the current mammograms, and the third hanging showed cranio-caudal views of the current mammograms. The prior mammograms were digitised using an Array 2905 Laser Film Digitiser set to 75 μ m, 12 bit greyscale depth. The LCD screen, table, and chair

specifications and size were the same as at the hybrid workstation. Screening decisions were entered by signature on the paperwork contained in the screening bags.

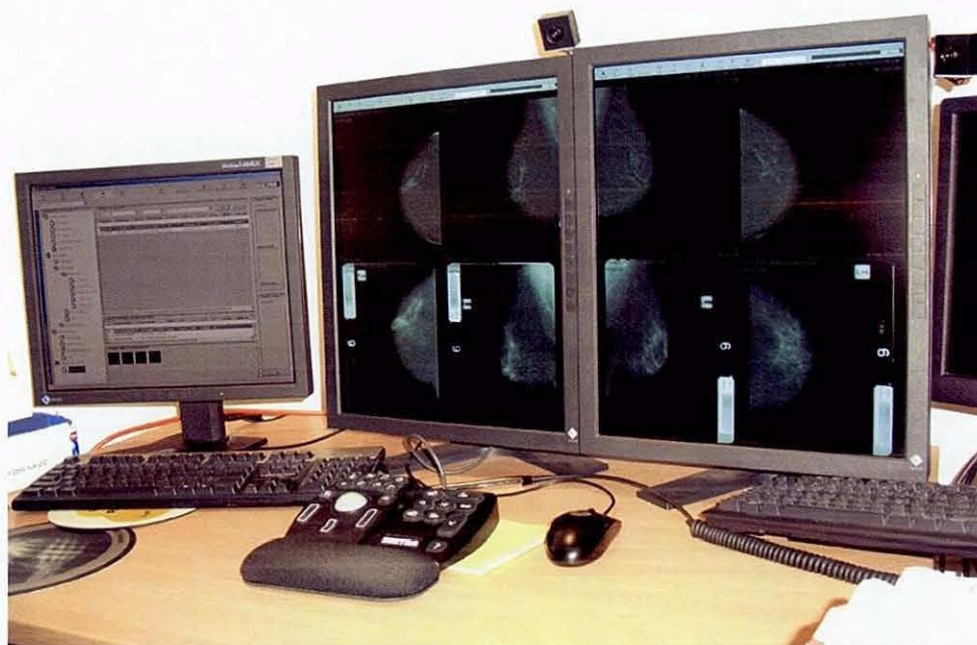


Figure 2.3 – The digital workstation with the first hanging displayed; current mammograms on the upper row and prior mammograms on the lower row

All three workstations were in fact part of the same workstation, but using different aspects. Room lighting was switched off during all experiments, and there were no windows, so the only light sources were the multi-viewer (when switched on), light from the LCD screens, and a small amount of light from another workstation in the same room.

2.4.2 Participants

At the time of commencement of the study only two centres in the UK regularly used digital technology for breast screening, one of which was at University Hospitals (Coventry). All eight mammography readers from that hospital were invited and agreed to take part, of which four were radiologists and four were radiography advanced practitioners. All were qualified film readers with experience ranging from 2.5 to 19 years, average eight years.

Participants had different levels of experience at each workstation. The film workstation had been used by the participants for as long as they had been film reading, the hybrid workstation had been used by the participants for a period of two years prior to the commencement of the study. The digital workstation was introduced for the purposes of the study, and therefore participants had no experience of it

Each participant took part in a total of six reading sessions each lasting 45 minutes, two sessions at each workstation. It was intended to counterbalance the order in which participants undertook the experiments at each of the three workstations, but due to delays in the digitisation process all measurements at the digital workstation were taken after those at the film and hybrid workstations.

2.4.3 Research methods

Workstation Dimensions Assessment

Recommendations were collated for workstation dimensions where a VDU is to be used. The dimensions of the three workstations investigated here were compared to these recommendations, to identify and resolve any ergonomic issues which were not generic, but rather related to the particular workstation dimensions implemented.

Discomfort Questionnaires

Discomfort questionnaires for low physical intensity work are typically conducted over the course of a whole day (Corlett, 2005) However, a 45 minute session was used for this study to model mammography readers' real world practices, and to enable differentiation between workstations. Radiologists are recommended to undertake direct clinical care, assessment, and follow up of suspicious and symptomatic cases including ultrasound and biopsy, in addition to reading 5000 screening cases per year (Liston *et al.*, 2005). In practice this mean that there is a limited time for screen reading and many interruptions. Therefore a 45 minute session provides an accurate model of real world circumstances.

A body part discomfort questionnaire was filled out both before and after each of the sessions. Discomfort was rated for twelve body parts on a scale as follows: 1=no discomfort; 2=very mild discomfort; 3=mild discomfort; 4=moderate discomfort; and 5=severe discomfort. The descriptions of the levels of discomfort were designed to detect small changes in discomfort,

because the reading sessions were so short. The data recording sheet can be found in appendix 3.

Postural Analysis

Rapid Upper Limb Assessment (RULA, McAtamney and Corlett, 1993) postural analysis tool was used because it is suitable for low intensity seated labour. Data were collected for postural analysis using four cameras, each perpendicular to one another, surrounding the participant, these four images were synchronised and displayed together. Event based, rather than time based analysis was used to enable direct comparisons between the workstations whilst minimising the data points required. This is made possible as the same activities are conducted at all of the workstations. To determine which events to analyse a task analysis was conducted from initial short unstructured interviews with all participants, and observing mammography readers at work. This task analysis was then reviewed for accuracy with three participants, it can be found in appendix 1. Information in the academic literature was sought to determine how many repeat measurements were necessary to produce each data point for analysis, i.e. how many measurements for each event for each participant at each workstation, but as event based RULA is not a common method no direct precedent was found. Three measurements of each session were taken, at the beginning, middle, and end, so that any effects of fatigue were recorded. Therefore, the events analysed were those closest in time to the following points, the earliest point in the timings, excluding the first case, as this may differ to the bulk of the cases analysed, 22.5 minutes through the 45 minute session, and the latest point excluding the last case. There are 17 events detailed, which would require

analysis of 1,224 postures, and 9,792 data points. In order to focus analysis attention on the higher risk postures, only those actions which either occur more than four times per minute for a participant, involve reach in the extreme reach envelope (i.e. reaching that requires bending of the torso), or involve weights of greater than 2kgs were considered for analysis, leaving nine events. For those events which occurred over a finite time period, for example looking at the current mammogram, the most extreme posture within that time period was analysed.

RULA is a subjective technique involving estimation of angles, and therefore it was necessary to check that the scoring for RULA assessment conducted as part of this study was in line with the scoring judgements made in the rest of the ergonomics community. Therefore of the 576 filming points which had been scored using RULA, a subset of 57 stills was taken from the film. These stills were given RULA scores by both the author, and an expert, Anna Jones. Anna has five years experience of applying postural analysis techniques in the field of medical ergonomics, with a focus on ambulance ergonomics. The most important measure of reliability is intra-observer reliability, because provided the scoring is consistent between the three conditions then valid comparisons can be made. To address this 57 stills of postures were scored by the author, and then one month later these same 57 postures were re-scored, and the results compared. The standard for inter-observer agreement is 75% as described by Heinsalmi (1986).

Statistical Analysis

Both body part discomfort and RULA scores are ordinal data and therefore non-parametric statistics were used, and where averages were taken the median value was used. The Wilcoxon test was applied to changes of discomfort over the 45 minute reading session at each workstation, and the Friedman test applied to the differences between workstations. The Friedman test was also applied to the differences in RULA score for each event between the three workstations, and where appropriate Friedman post hoc tests applied as described by Siegel and Castellan (1988, pg 180-181). All of these tests were repeated with the radiologists and the radiography advanced practitioners separately, and the results from these two groups compared.

2.5 Results

2.5.1 Workstation Dimensions

Both regulations and research data were combined to give recommendations for desk height, chair height, and maximum weight to lift. These were compared to the dimensions at the film, hybrid and digital workstations as shown in table 2.1.

One hundred screening bags were taken at random from the batches at the screening centre on 30/10/07 and weighed. The range of weights for each bag was from 44g for a woman with no prior mammograms to 901g for a woman who had attended several previous screening rounds. The mean weight was 168g with a standard deviation 127g. A batch is constructed of the screening bags for one screening van for one day, and so can contain up to sixty bags and therefore weighing over 10kgs.

Table 2.1 – Dimensions of the film, hybrid and digital workstations in comparison to recommendations.

Dimension	Source of Recommendation	Recommendation	Film workstation dimension	Hybrid Workstation Dimension	Digital workstation Dimension
Desk Height	BS EN ISO 9241-5:1999	720mm \pm 15mm	750mm	720mm	720mm
	Kroemer and Grandjean (1997): Based on height preferences for writing task	740mm			
Chair height	Pheasant and Haslegrave (2006): Based on a shod popliteal heights	380-535mm	440mm to 540mm	440mm to 540mm	440mm to 540mm
	Kroemer and Grandjean (1997): Based on preferred seating positions	270-300mm below desk height	210mm to 310mm below desk height	180mm to 280mm below desk height	180mm to 280mm below desk height
Maximum weight of batches of screening bags	Health and Safety Executive (2004)	If stored: at waist height 10kgs; below waist height 7kgs; and on the floor 3kgs.			
Desk Depth	Jachinski <i>et al.</i> (1998): based on preferred VDU viewing distances	60-100cm	41cm	54cm	54cm

Recommendations for screen height and angle (ISO 9241-5, 1998, pg 6) are that "the optimum position for the most important visual display is within $\pm 15^\circ$ in the vertical and horizontal direction from the line of sight" where the line of sight is inclined approximately 35° below the horizontal. These measurements are dependent upon the viewer's head height and line of sight angle and therefore require participant measurements. One still image was taken from the videotape for each participant sat upright at each workstation. The proportion of the current and prior mammograms that fell within 15° of the line of sight was measured for each image. At the digital workstation 15.5% of the current mammograms and 97.5% of the prior mammograms fell within this area. At the film workstation 25% of the current mammograms and 82.5% of the prior mammograms fell within the defined area. The mammograms were mounted higher at the film than the digital workstation, the top of the film display is 146cm from the ground whereas the top of the digital display is 124cm from the ground. However the desk in front of the digital display was greater in depth (54cm) than that in front of the film display (41cm) necessitating a greater viewing distance when seated upright.

2.5.2 Body Part Discomfort

All participants completed a survey designed by May et al. (1994) about existing musculoskeletal disorders prior to commencement of the study, as detailed in table 2.2. The survey design was that used by May *et al.* (1994).

Table 2.2 – Existing Musculoskeletal disorders in participants

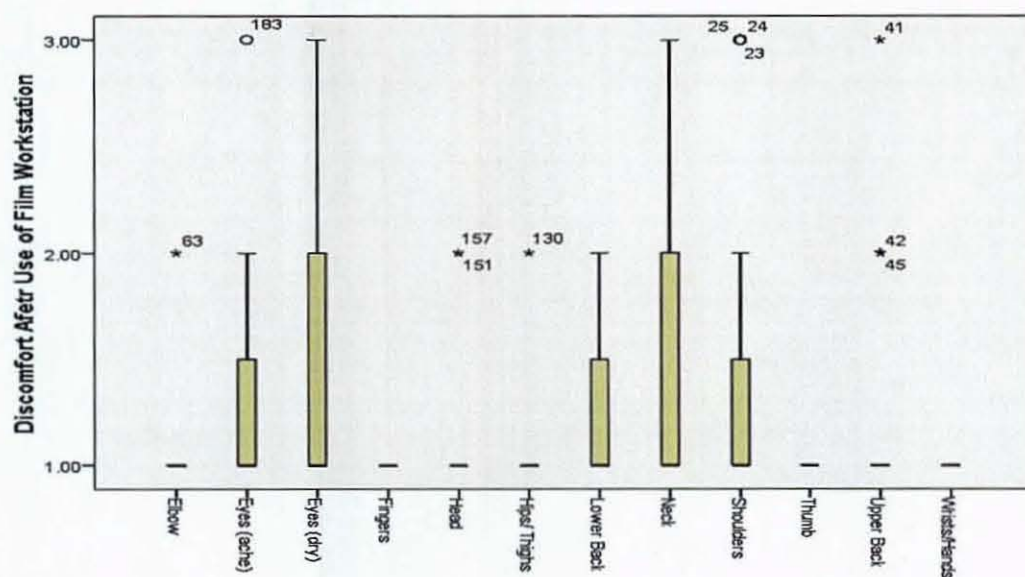
Participant	Discomfort areas	Necessitated change in work duties	Work Related
Radiographer 1	Lower back	Y	N
Radiographer 2	Lower back	Y	Y
Radiographer 3	Lower back, Neck, Thumb and finger	N	Y
Radiographer 4	Shoulder	N	Y
Radiologist 1	-		
Radiologist 2	Lower back	N	N
Radiologist 3	-		
Radiologist 4	-		

Levels of discomfort before the reading session began are shown in a boxplot in figure 2.4. This shows that both median and interquartile range of discomfort scores for all body parts was 1, which corresponds to no discomfort. The median discomfort score is the one that was selected most frequently by participants, and the interquartile range contains 50% of all responses (calculated using Tukey's hinges). All reports of discomfort appear as outliers. For the elbows, hips/thighs, fingers and thumbs there was no discomfort reported by any participants before any sessions at any of the workstations. After a 45 minute session at the film, hybrid, and digital workstations the interquartile range for discomfort scores includes some

2.5.



Figure 2.4 – Boxplot of discomfort scores for different body parts before the sessions began. A score of 1=no discomfort, 2=very mild discomfort, 3=mild discomfort, 4=moderate discomfort, 5=severe discomfort



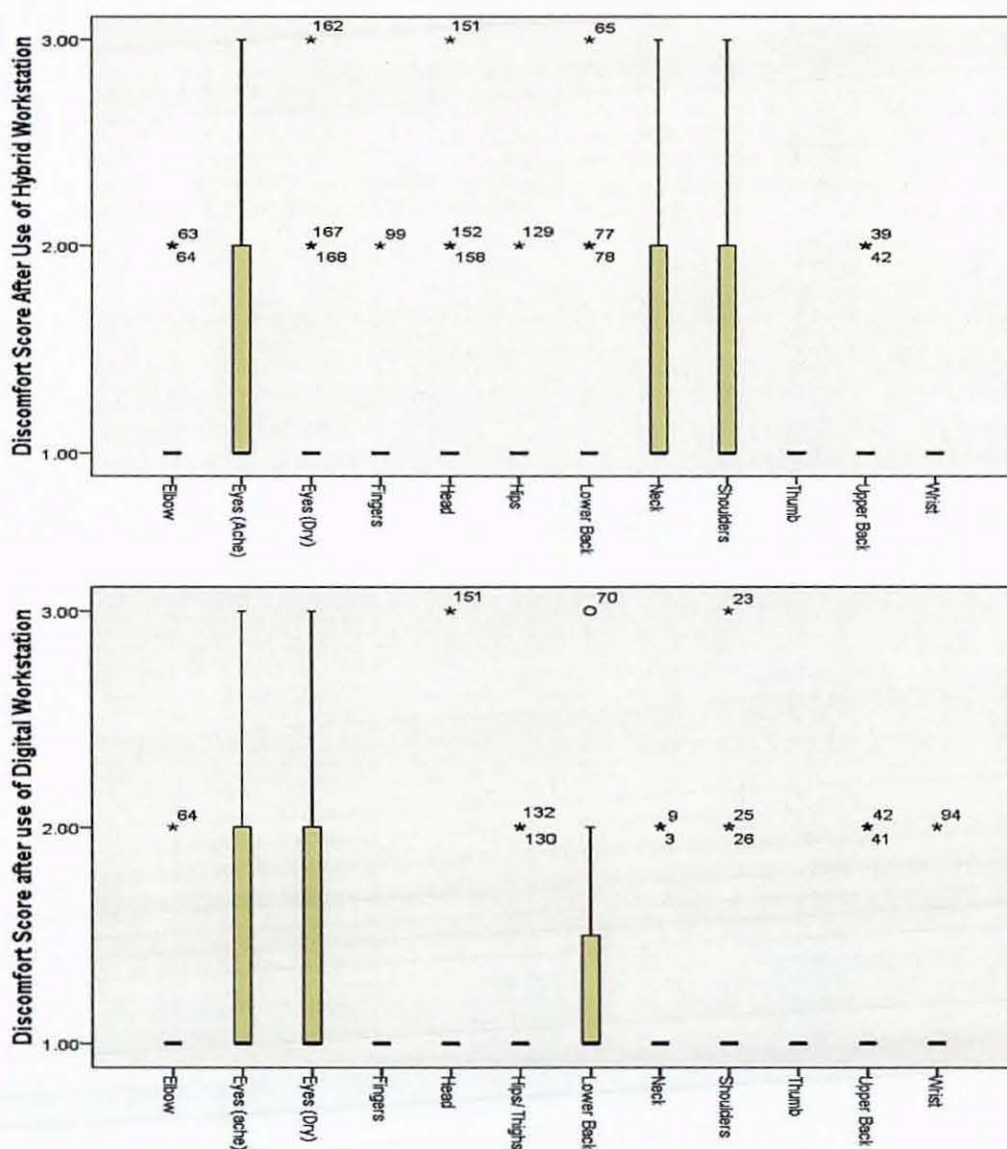


Figure 2.5 – Discomfort Scores after a 45 minute session at the film (top), hybrid (middle) and digital (bottom) workstation

Each participant completed two sessions at each workstation. The change in reported discomfort was analysed for the first and second session at each workstation separately, as the data are ordinal and so a mean could not be used. Results of the Wilcoxon signed rank tests for those combinations of body parts and workstations for which the interquartile range extends beyond a score of 1 are shown in table 2.3. Effect sizes were calculated using the Z score divided by the square root of the number of observations. None of the changes were statistically significant, however there were medium to large effect sizes across both sessions tested for increases in discomfort of the

eyes and lower back at the digital workstation, and the shoulder and eyes at the film workstation.

Table 2.3 - Change in discomfort scores after the 45 minute sessions for those combinations of workstation and body part for which the interquartile range extends beyond a score of 1.

Workstation	Body Part	Session	Change in Discomfort			
			Z score	Asymp. Sig. (2-tailed)	Effect size	Effect Size according to Cohen's Criteria (REF)
Digital	Lower Back	1	-1.000	.317	-0.35355	Medium
		2	-1.633	.102	-0.57735	Large
	Eye (Aching at back or middle)	1	-1.000	.317	-0.35355	Medium
		2	-1.342	.180	-0.47447	Medium to Large
	Eye (Dry)	1	-1.000	.317	-0.35355	Medium
		2	-1.000	.317	-0.35355	Medium
Hybrid	Eye (Aching at back or middle)	1	.000	1.000	0	
		2	-1.857	.063	-0.65655	Large
	Neck	1	-1.732	.083	-0.61235	Large
		2	.000	1.000	0	
	Shoulder	1	-1.000	.317	-0.35355	Medium
		2	.000	1.000	0	
Film	Shoulder	1	-1.414	.157	-0.49992	Large
		2	-1.000	.317	-0.35355	Medium
	Neck	1	-.378	.705	-0.13364	
		2	-1.342	.180	-0.47447	Medium to Large
	Eye (Aching at back or middle)	1	.000	1.000	0	
		2	-1.000	.317	-0.35355	Medium
	Eye (Dry)	1	-1.633	.102	-0.57735	Large
		2	-1.414	.157	-0.49992	Large

Friedman's ANOVA demonstrated no significant differences between the workstations for change in discomfort. This was true for all body parts, and for both sessions 1 and 2. Therefore the scores for all workstations were combined. The boxplot for discomfort scores after use of the workstations is shown in figure 2.6.

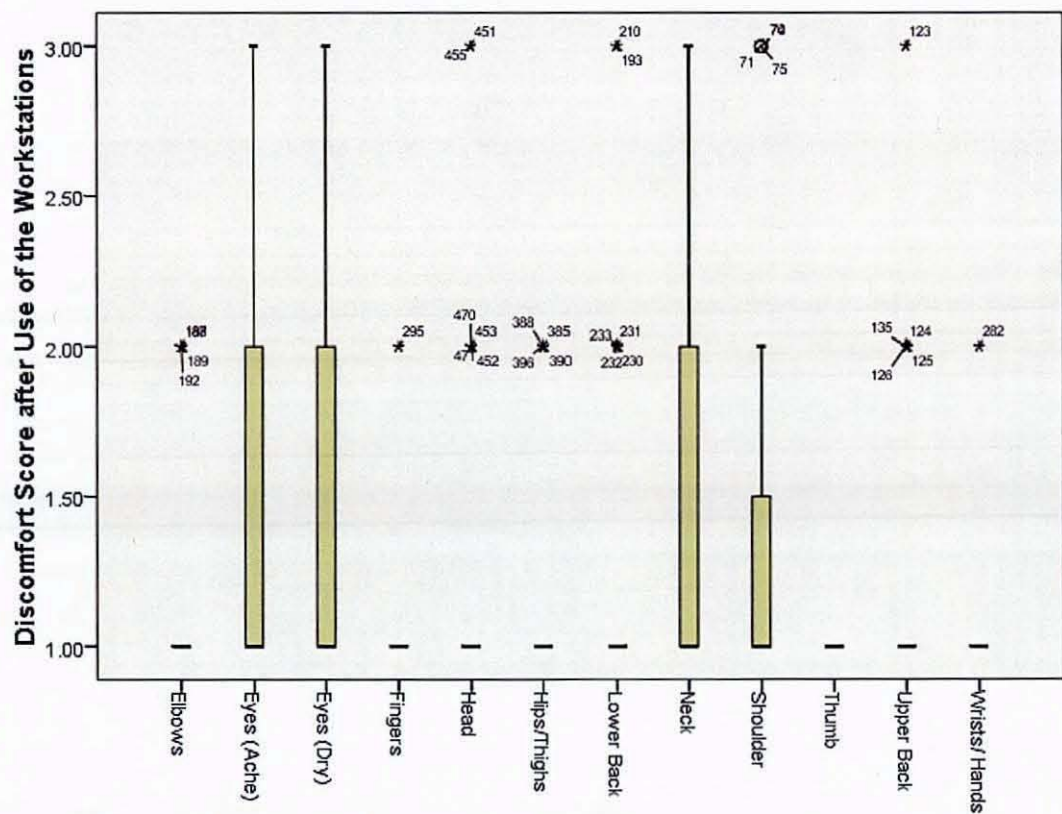


Figure 2.6 – Boxplot for discomfort scores after a 45 minute session at one of the workstations. Data for sessions 1 and 2 and for the film, hybrid, and digital workstations are combined here.

The data for radiologists and radiography advanced practitioners was divided and all analyses repeated. This revealed no significant effects for either group. A boxplot of the discomfort scores for radiologists and radiography advanced practitioners in each body part after a 45 minute session (at any workstation, in either session) is shown in figure 2.7, and the change in discomfort scores in figure 2.8.

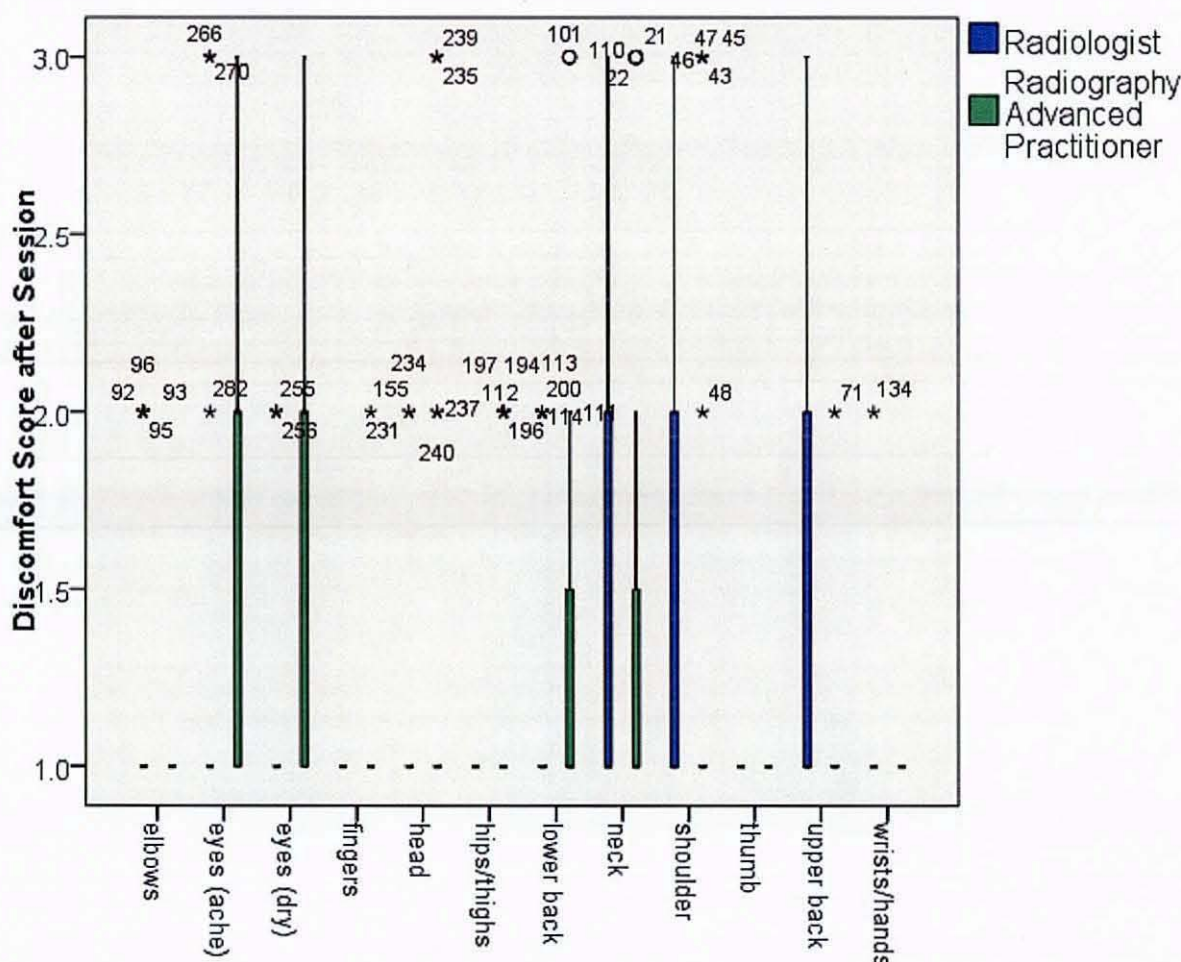


Figure 2.7 – Boxplot for discomfort scores after a 45 minute session at one of the workstations. Data for sessions 1 and 2 and for the film, hybrid, and digital workstations is combined here, but scores for radiologists and radiography advanced practitioners displayed separately.

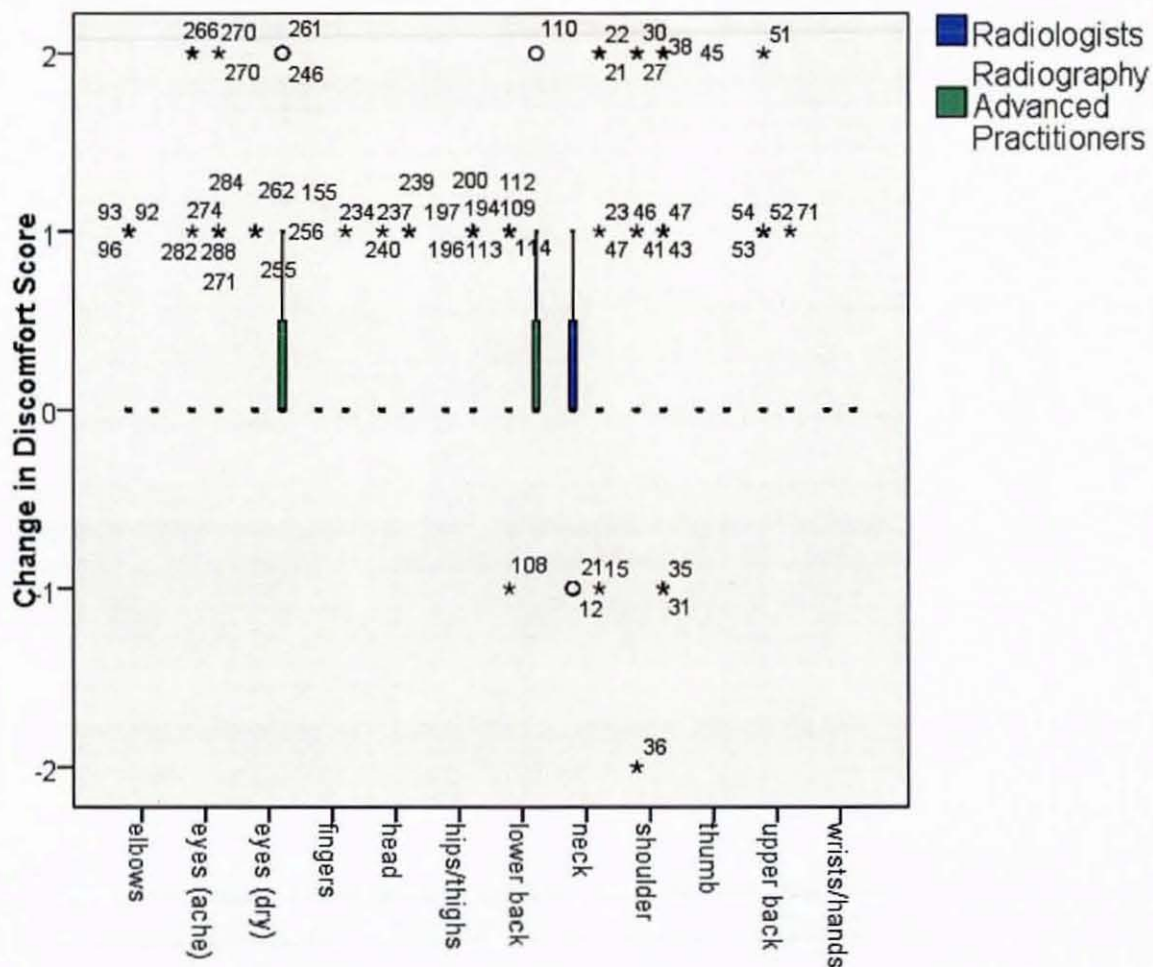


Figure 2.8 – Boxplot for the change in discomfort over a 45 minute session at one of the workstations. Data for sessions 1 and 2 and for the film, hybrid, and digital workstations is combined here, but scores for radiologists and radiography advanced practitioners displayed separately.

2.5.3 Postural Analysis

Intra-observer reliability testing showed 88% agreement on scores between the two sessions scored by the same person. Inter-observer reliability testing showed 78% agreement on scores, this is above the threshold acceptance level of 75% cited by Heinsalmi (1986) in reference to the OWAS method.

The RULA scores for the nine events at the three workstations ranged from a score 2 to 7. A score of 1 or 2 indicates that posture is acceptable if it is not maintained or repeated for long periods, 3 or 4 indicates that further investigation is needed and changes may be required, 5 or 6 indicates that investigation and changes are required soon. A score of 7 indicates that investigation and changes are required immediately (McAtamney and Corlett, 1993). Results of the postural analysis are shown in figure 2.9. There were no significant differences between the RULA risk scores of the radiologists and radiography advanced practitioners, and no trends towards any differences either, see figure 2.10.

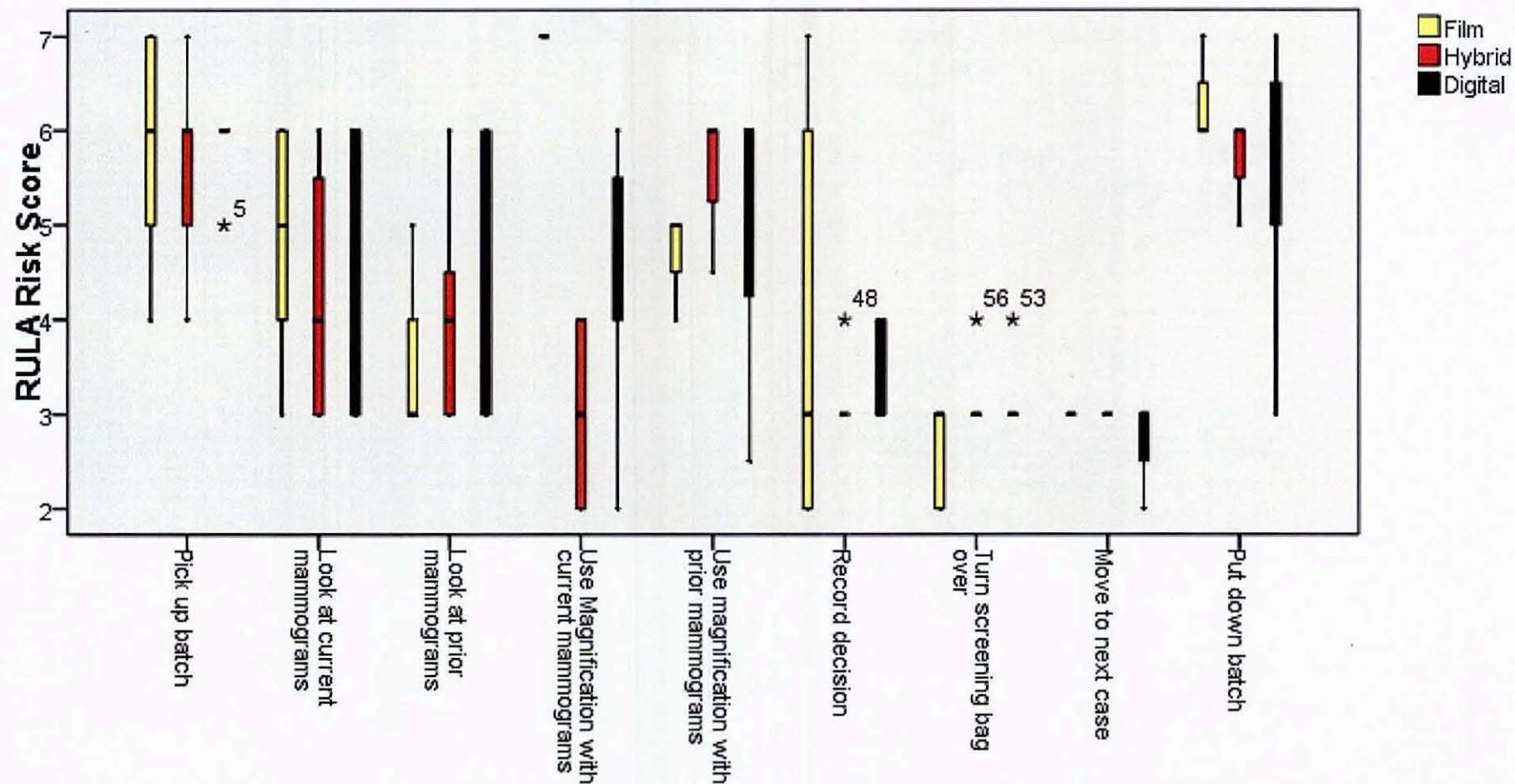


Figure 2.9 – Results of RULA postural analysis. A score of 1 or 2 indicates that posture is acceptable if it is not maintained or repeated for long periods, 3 or 4 indicates that further investigation is needed and changes may be required, 5 or 6 indicates that investigation and changes are required soon. A score of 7 indicates that investigation and changes are required immediately.

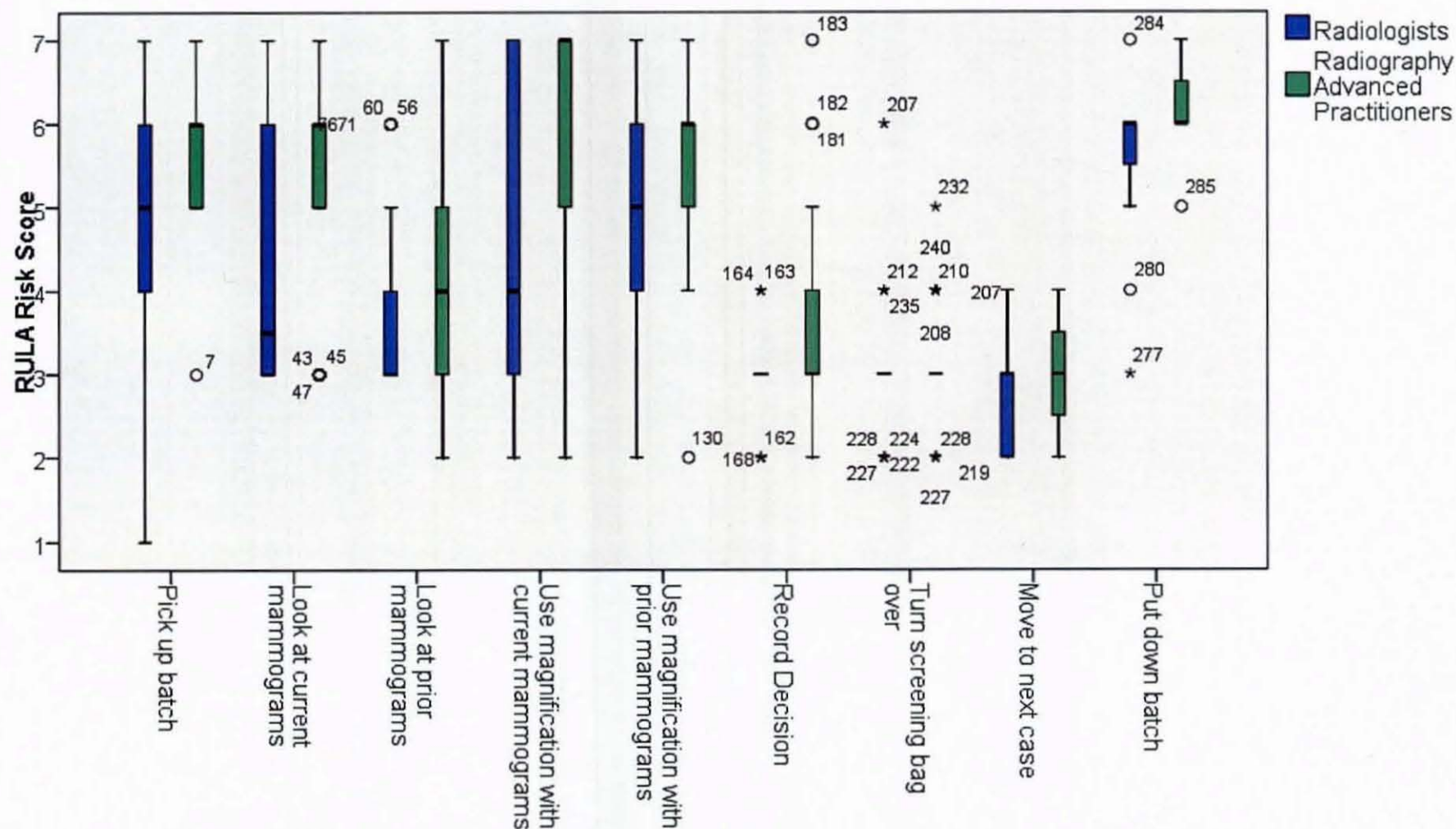


Figure 2.10 – Results of RULA postural analysis divided by participant type (radiologist or radiography advanced practitioner).

The actions of picking up and putting down the batches of film bags both had median risk scores of six, which indicates that investigation and changes are required soon. A batch can contain up to sixty bags, and such a batch will weigh over 10kgs. The bags are stored at a low level in this reading room; this may be because the room is very small and lacks storage space. Example postures for picking up and putting down the batches of screening bags are shown in figure 2.11.



Figure 2.11 – Two postures adopted when picking up and putting down the screening bags.

Looking at the current mammograms with a magnifying glass produced a median risk score of 7 at the film workstation, which is the highest risk score. Mauchly's test for sphericity was not significant for the comparisons between the workstations ($p=1.0$). Friedman's ANOVA showed that there was a difference in RULA scores between the different workstations ($\chi^2(2)=10.3$, $p=.006$). Post hoc tests showed that the RULA score was higher at the film than the digital workstation ($p<.05$), and a trend towards a higher RULA score at the film than the hybrid workstation (difference = 1.17, critical difference = 1.2). The vertical position of the current mammograms is higher at the film workstation than at both the hybrid and digital workstations, the viewing area is also wider, and a magnifying glass weighing 500g is used rather than a software magnification tool. This can result in higher scores due to flexion of the neck, side bending and twisting of the torso, and higher arm scores respectively. Two examples of the postures adopted are shown in figure 2.12.



Figure 2.12 – Two examples of a participant looking at the current mammograms at the film workstation

The interquartile range for recording a decision at the film workstation extends from RULA score 2 through to 6. Decisions are recorded at the film workstation using a barcode reader; some participants choose to read the barcode which identifies the woman from the label on the x-rays themselves rather than the screening bags. This increases the RULA score due to elevation of the arm, and sometimes is associated with increased scores for lower arm, wrist and torso. An example of such a posture is shown in figure 2.13. At the hybrid and digital workstations the decision is inputted through a signature on the screening bags.



Figure 2.13 – Recording the decision at the film workstation (above) and the digital workstation (below)

The RULA score for turning over the screening bags differed by workstation ($\chi^2(2)=4.7$, $p=.028$), but post hoc tests showed only a slight trend towards the hybrid and the digital workstations having higher RULA scores associated with them than at the film workstations (difference 0.75 and 0.75, critical difference 1.2). The desk space at the film workstation was greater than at the other two workstations, and so twisting and reaching was not required to put the bags on another work surface, see figure 2.14. However, the median RULA score was 3 at all workstations making it one of the lower risk activities analysed.

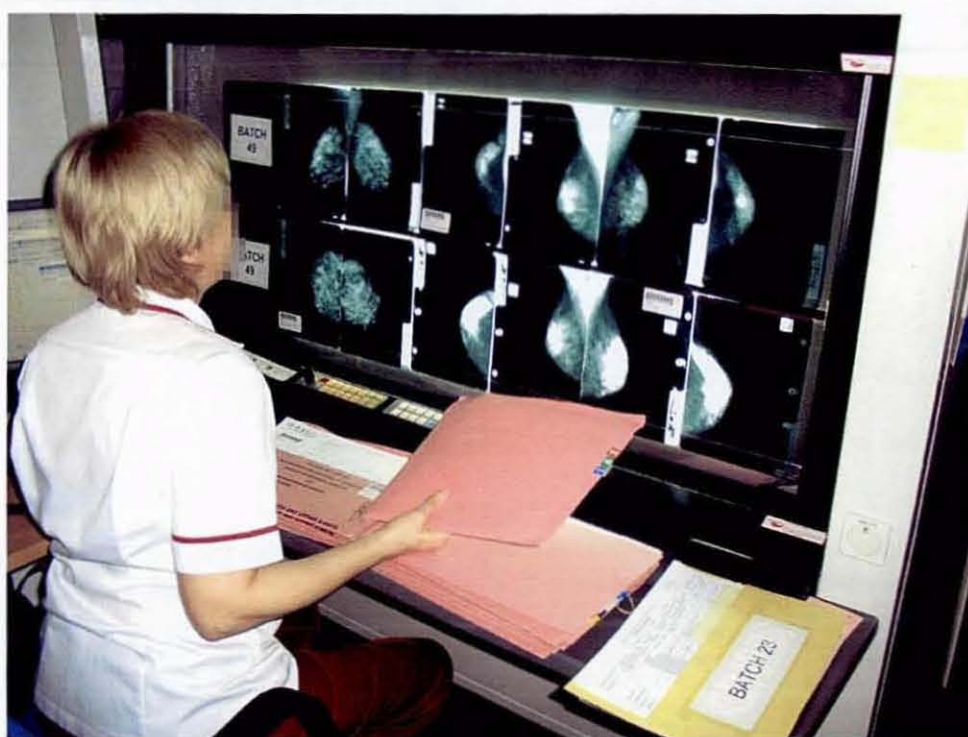


Figure 2.14 – A mammography reader turning over a screening bag to move to the next case at the film workstation (above) and the digital workstation (below).

The position of the prior mammograms is the key difference between the workstations. At the hybrid workstation the prior mammograms are both further away from the reader, and perpendicular to the main display. There were no significant differences in RULA risk score between the three workstations for looking at the prior mammograms ($\chi^2(2)=.1$, $p=.9$). Two postures adopted to look at the prior mammograms at the hybrid workstation are shown in figure 2.15. This shows that in some instances the participants leaned closer to the prior mammograms, causing twisting and side-bending of the torso but obtaining a better view due to the greater proximity to the screen, and in some instances only the head was turned, resulting in a degraded view versus that at the film workstation due to the greater eye to screen distance. In fact, the median RULA score for the torso for looking at the prior mammograms at the hybrid workstation was 1, indicating no twisting or side bending.



Figure 2.15 – Two postures adopted at the hybrid workstation when looking at the prior mammograms. Top image: the participant leans over both twisting and side bending the torso to get closer to the images. Bottom image: Participant simply turns his head resulting in less twisting and side bending of the torso, but a greater eye to image distance reducing detail perception.

There was no difference in RULA score between the different workstations for looking at the prior mammograms with a magnifying glass ($\chi^2(2)=.3$, $p=.7$). At the hybrid workstation in some cases the participant kept their seat position constant and twisted and leaned their torso, and in some cases they moved the whole chair, see figure 2.16. In the latter case whilst the posture adopted tended to have lower RULA scores particularly for the neck and torso, it was necessary to move the chair again before looking at the current mammograms, resulting in a time delay which may have affected ability to make comparisons between current and prior mammograms. At the digital workstation magnification was possible without lifting a magnifying glass, unlike at the other two workstations, and prior mammograms were presented on the same screen as the current mammograms unlike at the hybrid workstation, yet median RULA score was six, which is no lower than at either of the other two workstations. This may be due to a tendency for the participants to lean over the table to get closer to the screen, and therefore flexing the trunk and in some cases moving the neck into extension to accommodate the vertical screen orientation, see figure 2.17.



Figure 2.16 – Using the magnifying glass on the prior mammograms at the hybrid workstation, by leaning over whilst maintaining chair position (above) and by moving the chair (below)

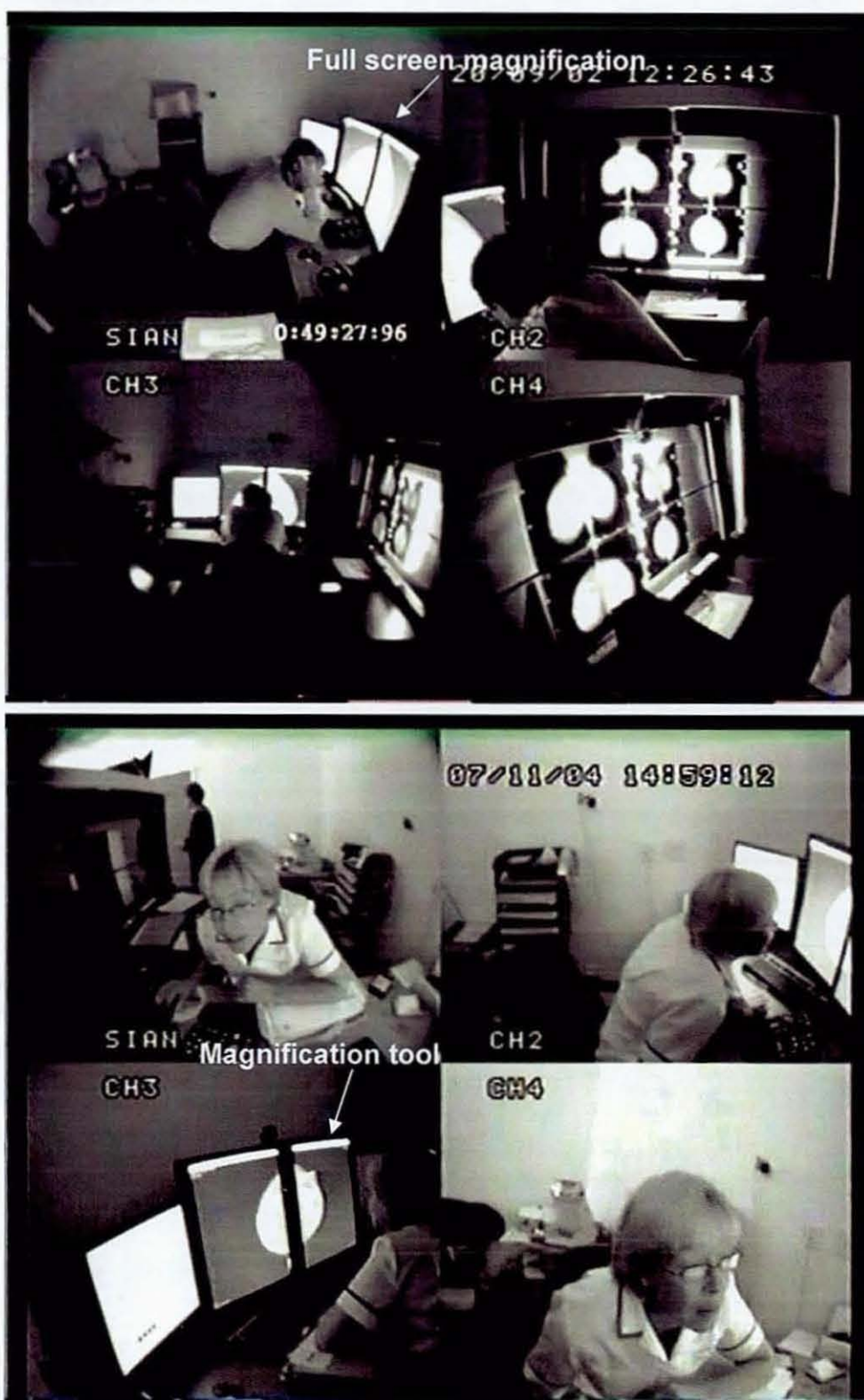


Figure 2.17 - Examples of using magnification at the digital workstation. Upper row shows use of full screen magnification (above) and use of the magnification tool (below) of hanging three, in both cases the participant is leaning forward to get closer to the screen.

2.6 Discussion

2.6.1 Workstation Dimensions and weights

At the digital workstation only 15.5% of current mammograms were within the optimal position for the visual display as defined by BS EN ISO 9241-5 (1999), and just 25% at the film workstation. This has highlighted two potential issues. Firstly that both displays are tall, and therefore viewing the top section may require neck extension. Secondly that whilst the film workstation extends 22cm higher than the digital workstation, the proportion of the current mammograms within the aforementioned optimal position for display was lower at the digital workstation. This would be in part due to the greater distance to screen at this workstation. The behavioural impact of this enforced increase in distance to screen is investigated further in the postural analysis. These data are based on a small number of measurements and therefore provide a tool to highlight potential issues rather than evidence of a problem.

A batch of 50 cases of average weight will weigh over 8kgs, therefore according to the Health and Safety Executive guidelines (2004) such a weight should be lifted from no lower than waist height. If the batch is lifted from the floor the maximum recommended weight for women is 3kg corresponding to 17 bags, and if lifted from a shelf below waist height 7kgs corresponding to 41 bags. Therefore either the batches should all be stored at waist height, or the number of cases per batch should be reduced to meet the HSE guidelines.

The optimal solution would be to introduce paperless reporting so no bags are necessary at all.

2.6.2 Body Part Discomfort

Whilst none of the increases in discomfort were statistically significant, there were some combinations of body parts and workstations for which the effect sizes were medium to large for both the first and second sessions (table 2.2). These were dry eyes at both the digital and film workstations, aching eyes at the digital workstation, shoulder discomfort at the film workstation, and lower back discomfort at the digital workstation. The study has highlighted that further research is needed concerning the effects of reading mammograms on the eyes, in particular considering visual performance and fatigue.

It is unusual to measure discomfort in a sedentary task after such a short time interval, Corlett (2005) recommends taking measurements over a whole day citing that recovery from static loading is slow and therefore discomfort is cumulative. However, in this case the aim was to determine if there are any differences in discomfort between the different workstations, and therefore it was necessary to measure changes in discomfort solely due to that workstation. The session length could not have been increased, both because it is a realistic representation of real world practices, and because it may have affected the participants cancer detection performance through fatigue and so would not be ethical. A more accurate measure of discomfort effects could be achieved by implementing each workstation for a whole week and measuring discomfort at the beginning and end of each day, however this was not

possible at the study hospital as there was only one set of digital equipment available out of a total of four viewers.

When comparing discomfort between radiologists and radiography advanced practitioners no significant differences were found. However, the interquartile range of discomfort scores after the sessions extended beyond no discomfort for the radiologists at the neck, shoulder and upper back, and for radiography advanced practitioners at the eyes, lower back and neck. The RULA risk scores show no difference between postures adopted by radiologists and radiography advanced practitioners and therefore any differences in discomfort are not likely to have their origins in differences in behaviour. When the changes in discomfort at the workstation are considered the interquartile range of discomfort extends beyond no change for radiologists in the neck area, and for radiography advanced practitioners the lower back and eye areas. May (1994) who reported that breast screening radiographers experience most discomfort in their lower back, followed by the neck and upper back, and cite that screening women may involve awkward postures as a potential reason for the discomfort experienced. Both radiologists and radiography advanced practitioners position women for mammography, biopsy and ultrasound as part of their regular work activities, and therefore any neck and back pain experienced by these participants may be more associated with this part of their work rather than reading mammograms.

The study could have been extended in several ways. The number of participants could have been increased. This would have been very difficult to

achieve as the study could not be taken to other centres, as there is not another centre in the UK which has a digital workstation for screening and is able to put a multi-viewer adjacent to it. Participants could have been transported to University Hospital (Coventry) from other hospitals, however this would have been very expensive and difficult to arrange, and they would have no experience using the workstations so may behave in a different manner. The number of sessions per participant could have been increased, and a median of the discomfort scores taken. This would have taken a lot of participants time, and may not have produced any significant results as the median change in discomfort may have been zero. The scale could have been changed to a five point scale with only the 'anchors' labelled, 0 as 'no discomfort' and 5 as 'extreme discomfort'. This would have allowed a mean to be taken of the sessions using the argument of Corlett and Bishop (1976) that discomfort is found to be linearly proportional to task time for a holding task, and therefore discomfort is a linear scale when the individual demarcations are not labelled. However this may have decreased the sensitivity of the test as the labels 'very mild discomfort' and 'mild discomfort' would have to be removed. Therefore, whilst the body part discomfort experiment could not reasonably be extended, it has highlighted the body parts upon which to focus attention when considering workstation design, namely the neck, shoulder, lower back, and eyes.

2.6.3 RULA – The transition to digital mammography

To understand the impact of the introduction of digital mammography on workstation ergonomics comparisons were made between the RULA scores at the film workstation and the hybrid/digital workstations. Two issues were highlighted: use of the magnifying glass on the current mammograms, and turning over the screening bags. Use of the magnifying glass on the current mammograms gave a higher RULA score at the film workstation than at the digital workstation, ($p < .05$). This may be because the height of the current mammograms requires flexion of the neck, and use of a magnifying glass requires weight bearing and flexion of the arm. This provides evidence that the workstation layout proposed by Ratib *et al.* (2000) would not be adequate for breast screening as positioning the prior mammograms above the workstation and viewing them with a magnifying glass would require extreme postures. The film workstation is being phased out in the NHSBSP so this result does not merit further investigation.

Turning over the screening bags resulted in higher RULA scores at the digital than at the film workstation. This may be due to insufficient desk space for the screening bags in addition to the keyboard, mouse and Sectra keypad on the work surface. There are plans in place to make screening paperless in the NHSBSP, and this would solve this problem.

Using the magnification tool at the digital workstation the median risk score was 5 when looking at the current mammograms and 6 when looking at the prior mammograms, indicating that investigation and changes are required

soon (McAtamney and Corlett, 1993). These high scores were due, at least in part, to participants moving their heads close to the screen whilst looking at the mammograms. In many cases, this resulted in flexion of the torso and extension of the neck. Some discomfort was reported in the neck after 29% of the sessions, and therefore postures involving extension of the neck merit further investigation. There are two approaches to deal with this issue: improve the workstation ergonomics so that participants are able to get close to the screen without adopting awkward postures; or provide additional magnification so it is not necessary to lean closer. Considering the first approach, the neck was in extension when looking at the prior mammogram for 7 out of 24 events at the digital workstation, and only two out of 24 events at the film workstation. The prior mammograms are situated higher up and at an angle of 6 degrees to the vertical at the film workstation, whereas they were positioned lower and vertically at the digital workstation. Therefore increasing screen tilt at the digital workstation could be investigated as a potential solution. Taking the second approach is preferable as leaning close to the screen is suboptimal both in terms of posture and strain on the eyes. The magnification tool is accessed via a menu screen requiring three mouse clicks, and therefore is more complex to use than a magnifying glass. The magnification tool may not be providing enough magnification, and the screens themselves may not be of optimal size. Whilst there is research to show that two screens are better than four for chest x-rays (Siegel and Reiner, 2002) there is no similar evidence in mammography, or evidence about optimal screen size. Accessibility of magnification tools may be an issue in the transition to digital mammography, particularly because using a

magnifying glass on a digital screen will not give the same detail as use of the magnification tool, as the limiting factor in spatial resolution is the pixel size of the LCD screen.

The postural analysis indicates that after the transition to digital mammography there will be some postural improvements when looking at a magnified view of the current mammograms, but there are issues with the accessibility of magnification tools, shortage of desk space if paperless reporting is not implemented, and screen height and angle should be optimised with consideration that readers are likely to lean in close to the screen.

2.6.4 RULA – Digital or Film Display of Prior Mammograms

The choice of whether to digitise the prior mammograms or display them on an adjacent multi-viewer is not likely to affect the incidence of musculoskeletal disorders in mammography readers as there were no significant differences in RULA scores for the nine events between the hybrid and digital workstations.

It would be reasonable to expect that the postures adopted looking at the prior mammograms at the hybrid workstation would have higher RULA scores associated with them than those at the other two workstations, because the prior mammograms are situated perpendicular to, and a distance from the rest of the images at the hybrid workstation. This may be due to two effects: poor postures adopted at both the film and digital workstations when looking at the

prior mammograms because of their low position; and participants adapting their behaviour to avoid uncomfortable postures at the hybrid workstation. The latter is a particular concern because changes in behaviour in this safety critical task could lead to changes in performance. The median RULA score for the trunk when looking at the prior mammogram at the hybrid workstation was 1, indicating no twisting or side bending. Therefore, for a large proportion of cases participants were simply turning their heads to look at the prior mammograms, and were not leaning closer to them or moving their chair to get closer. When a participant chooses to move only their head the distance to the film mammograms will be greater than that they are accustomed to, and they will not be able to see the same level of detail. If they twist and lean their torso they will have a better view but a less comfortable posture. If they move their whole chair each case will take much longer to report, and their concentration may be disturbed. This is a small data set and therefore requires further investigation. There is some evidence that posture and discomfort can affect performance with a trend towards a relationship between shoulder discomfort and performance on a VDU task ($p=.06$) reported by Straker *et al.* (1997), work height has been found to affect rate of manipulation performance ($p<.01$, Ellis, 1951) and intercorrelations in factor analysis between both discomfort, trunk angle and performance in a circuit board inspection task (Bhatnager *et al.*, (1985). However, other studies have not been able to repeat this effect, most notably in the inspection task of x-ray baggage screening (Drury *et al.*, 2008). Furthermore, it is unclear if any link between discomfort and reduced performance is simply a fatigue effect. Further research in this area is necessary.

2.6.5 Triangulation of Methods

The interquartile range of discomfort scores after the session extended into neck and shoulder discomfort at the film workstation, in comparison to no discomfort at the digital workstation. This corresponds with a higher RULA risk score at the film workstation for the task of using a magnifying glass on the current mammograms, which is a task repeated with greatest frequency in reading mammograms. This shoulder discomfort may be partly associated with the weight of the magnifying glass at the film workstation, in comparison to the button operated magnification tool at the hybrid workstation. The neck discomfort may be due to the larger viewing area at the film workstation which extends up to 146cm from the floor in comparison to the top of the viewing area of the digital workstation 124cm from the floor. Only 54% of the film display vertically was within 15° of the line of sight, and just 57% of the film display. Therefore, the task involves significant vertical neck movement to view both current and prior mammograms. The reason that the displays are so large is to increase display resolution so that subtle abnormalities such as microcalcifications can be seen, and because film mammograms are developed at standard sizes to optimise analogue image quality. With advances in display technology for digital mammography resolution will increase, and therefore microcalcifications could be visible in a smaller display, and greater use of magnification tools could also enable a smaller display to be used. However, the limits of the human visual system must be considered, alongside the acceptability of the electronic zoom tools. Further research in this area is required to understand how the technical possibilities

of digital mammography including different screen sizes, zoom tools, and other tools such as contrast adaption influence cancer detection performance.

2.6.6 Evaluation of Methods

The data presented here are from a field study which was designed to model realistic options for the introduction of digital mammography, rather than fundamental causes of postural and behavioural differences. The hybrid and digital workstations differ in both the location and the display medium of the prior mammograms. Therefore when considering the reasons for any differences in posture between the two workstations it cannot be known whether these were caused by the position or location of the prior mammograms. However, in practice there are few realistic display configurations, and so this extra information about cause of effects, although interesting, is not necessary for breast screening centres making display decisions. Digitised prior mammograms would always be displayed on the same viewers as the current digital mammograms, as high resolution LCD screens are very expensive and there is no evidence that display on separate screens would provide any clinical benefit. Film prior mammograms could not be displayed within the field of view of the LCD screens. This is because the extraneous light could be a detriment to performance, as film display has significantly higher luminance than digital display, and Wang and Gray (1998) demonstrated that multi-viewer masking of extraneous light improves diagnostic performance. Therefore film prior mammograms could realistically only be displayed perpendicular to (as investigated here) or above the digital

display, but the latter option is impractical as it necessitates excessive neck flexion, and reach.

This study was with 8 participants, and 24 measurements of each event at each workstation. The discomfort scores and postures adopted will be influenced both by the anthropometric dimensions of participants, and by any existing musculoskeletal disorders they have (table 2.2). All four radiography advanced practitioners reported pain, which is a reasonable approximation (with a small sample) to the population as surveyed by May *et al.* (1994) who found that 76% of breast screening radiographers reported pain, although there could be response bias in this with only 40% response rate. The study could be extended to include a greater number of recordings per participant, or more importantly a greater number of participants. However there are published precedents of within subjects RULA postural analysis with similar numbers of participants, including one participant performing gastric bypass surgery, (Lawson *et al.*, 2007), ten participants when investigating breast screening radiographers (May and Gale, 1998), 11 participants undertaking cytology screening, (Lomas, 1998), and 12 participants when investigating a VDU task (Mohammed *et al.*, 1999). The intention of this study was as an initial investigation to highlight any particular events which are of interest and may merit further attention. The issues which have been highlighted in the transition to digital mammography are the usability of the magnification tool, the display screen height and angle, and potential changes in behaviour looking at the prior mammograms at the hybrid workstation. This last area is

of greatest interest because it may have a bearing on cancer detection performance.

2.7 Conclusions

The first aim of this investigation was "to determine whether the change from film to digital mammography will impact radiology workstation comfort and risk of musculoskeletal disorders". Whilst there were no differences in the discomfort scores, there may be a reduction in the risk levels for musculoskeletal disorders when changing from film to digital mammography, as the action of looking at the current mammograms with a magnifying glass resulted in higher RULA risk scores at the film workstation (median 7) than at both the hybrid (median 3) and digital workstations (median 5) which may replace it. Lower RULA scores were recorded for turning over the screening bags at the film workstation, than at the hybrid and digital workstations which may replace it. However the median risk score was only three for all workstations, and the introduction of paperless reporting will result in screening bags no longer being necessary.

The second aim was "to determine the impact on radiology workstation comfort and risk of musculoskeletal disorders of digitising prior mammograms in preference to displaying them in film format during the transition to digital mammography". There was no evidence from the body part discomfort or postural analysis to suggest that there were any differences in comfort between the hybrid and digital workstations i.e. between digitising the prior mammograms, or displaying them in film format on an adjacent multi-viewer. However, the postures adopted at the hybrid workstation show that the readers are viewing the prior mammograms from a greater distance than

when they are digitised. This suggests that the readers may be adapting their behaviour to address the physical challenge of the large distance between the current and prior mammograms. This raises two questions: is behaviour in terms of level of use of prior mammograms also affected by this physical distance? And is there an impact on cancer detection performance from any changes in behaviour?

The third aim was to “to determine whether there are any differences between radiologists and radiography advanced practitioners in relation to aims 1 and 2”. There were no differences in RULA risk scores between radiologists and radiography advanced practitioners, indicating that the two groups are adopting similar postures to undertake the same tasks. There was some increase in discomfort for radiography advanced practitioners in the eyes and lower back, and for radiologists in the neck but it was not significant. This may indicate more sensitive areas related to other work activities, but equally could be attributed simply to random variation.

The results of this study have highlighted two key areas which merit further research. Firstly, the postural analysis at the hybrid workstation has highlighted a need to research whether reading behaviour is affected by workstation layout. Secondly, whether any behavioural changes or postural considerations affect performance in cancer detection.

3 Workload and Productivity in the Transition to Digital Mammography

3.1 Introduction

Mammography readers in the Breast Screening Programme will experience increases in case load alongside the introduction of digital technology, by the year 2012 (Department of Health, 2007). The available evidence of how this increase in case volume and change in display medium might affect mammography readers' speed of reading and experience of workload are discussed here. Additionally the impact of the display medium of the prior mammograms is discussed.

Case load per member of staff in breast screening is set to increase by 2012, due to a combination of the extension of the age range of women screened from 50-70 years to 47-73years, and an increased number of the 'baby boom' generation reaching screening age (Department of Health, 2007). The previous age extension was found to have "resulted in a 40% increase of the workload of the programme, which has only been possible because of new working practices" (Department of Health, 2007, pg 46-47). The new working practices referred to are the introduction of radiography advanced practitioners (radiographers trained to read mammograms), and assistant practitioners to assist with taking mammograms. The current age extension will be introduced alongside the introduction of digital mammography, which is expected to reduce the time required to take each set of mammograms

(Department of Health, 2007), and so for radiographers whilst the case load will increase the time required per case will decrease. However, for those reading the mammograms there will be an equivalent increase in case load, but no decrease in time taken per case according to research by Pisano *et al.*, (2002), where speed of reading was found to be the same for soft copy versus film display. There is a need to extend this research to determine whether digitising the prior mammograms or displaying them in film format on an adjacent multi-viewer will affect time taken per case.

An increase in case load does not necessarily result in an increase in subjective workload, or a reduction in performance. Workload is defined as "that portion of the operators limited capacity actually required to perform a particular task. The objective of workload measurement is to specify the amount of expended capacity... to avoid existing or potential overloads" (O'Donnell and Eggemeier, 1986, pg 42-2). Case load in the US is increasing (Bhargaven and Sunshine, 2002), with high case load (162 cases in a day) being successfully cited in a court case as 'reckless and wanton' (Berlin, 2000). However, in the UK each mammography reader must read at least 5000 cases per year, (Liston *et al.*, 2005), as an increase in volume of cases read has been associated with improved performance (Kan *et al.*, 2000, Esserman *et al.*, 2002). There is little research available which demonstrates the optimal balance of case volume: sufficient to improve expertise, but not too much to overwork the mammography readers. Subjective measures of workload can give some indication of how work volume is affecting the staff. Whilst there has been little research in radiology, subjective assessment of

workload has been conducted in hospital emergency departments using NASA-TLX, providing evidence of the variation of subjective workload with task and participant type (France *et al.*, 2005, Levin *et al.*, 2006). The current literature does not provide enough data to enable predictions of how mammography readers' experience of workload will change with increased case volume.

The transition to digital mammography will bring a change in display medium, alongside a change from using a magnifying glass to use of computerised magnification tools, and the additional availability of many other computerised tools such as contrast adjustment and image inversion. The effect of this change on perceptions of workload in breast screening has not been directly addressed, however Mayes *et al.* (2001) compared NASA TLX workload score for a paper and VDU based reading task, and found workload score was lower for the paper based task but this was not significant. Hancock (1996) found that subjective workload varied with input device in a computer based tracking task. This suggests that it is possible for workstation controls and the display medium to affect perceived workload when completing the same task, but there is no direct research to show whether perceived workload is affected by whether the radiology workstation is film or digital.

Whether the prior mammograms are digitised or displayed in film format on an adjacent multi-viewer will affect both their proximity to the current mammograms, their appearance, and light levels at the workstation. Subjective workload variation with workstation layout has been investigated:

Hancock and Scallen (1997) hypothesised that locating a control nearer to its functional equivalent would reduce workload, but found no such effect using the Subjective Workload Assessment Technique. This is analogous to the distance between the current and prior mammograms, however as the task and participant type are different it cannot be surmised that prior mammogram placement will not affect subjective workload. The appearance of film and digital mammograms is different as the digital images are pixellated. The cognitive effort required in making comparisons between digital and analogue mammograms may differ from that required comparing digital and digitised images, however no previous research in the subject could be found. The viewer for displaying film mammograms is brighter than the digital display, and therefore when film and digital images are compared there is also adaptation of the eyes necessary. Changes in pupil size can produce almost instantaneous light or dark adaptation to vary light levels of the order 30:1 (Overington, 1976), and therefore visual performance should not be affected unless one of the screens has a light levels of over 30x that of the other. Any effects of continuously changing light levels on workload are not well documented.

In the context of increasing case loads, it is important to understand whether the introduction of digital mammography will affect speed of reading and mammography readers' subjective workload. Additionally to understand if there are differences in mammography readers' speed and subjective workload between digitising prior mammograms and displaying them in film format on an adjacent multi-viewer in the transition to digital mammography.

The answer to these questions cannot be inferred from the currently available literature.

3.2 Aims

1. To measure any changes in subjective workload and mean time to read a case in the transition from film to digital mammography
2. To determine the impact on readers' subjective workload, and mean time to read a case at the radiology workstation of digitising prior mammograms in preference to displaying them in film format during the transition to digital mammography
3. To establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2.

3.3 Choice of Methods

Both objective and subjective measures are available to measure workload. These objective measures include eye blink rate, heart rate, primary task measures, secondary task completion, and EEG measurements. Blink rate is lower for tasks with higher visual demand and therefore can be related to visual workload, (Stern and Skelly, 1984; Veltman and Gaillard, 1996), however it is also related to humidity, and angle of gaze (Skotte *et al.*, 2007; Tsubota and Nakamori, 1995), and therefore there would be too many confounding variables in a mammography reading room to give an accurate

reading. Heart rate has been used as a measure of workload (Roscoe, 1992), however Hart *et al.* (1984) report that it is correlated with subjective reports of stress rather than workload, and it is affected by the ingestion of stimulants such as caffeine (Steinke *et al.*, 2009). Primary task measures are measures of the capability of the operator to complete the operation such as performance and speed. According to Wierwille and Eggemeier (1993) primary task measures should always be included in workload evaluation. However, O'Donnell and Eggemeier (1986, pg 42-4) describe a model where at levels of workload which do not exceed the information processing capacity of the operator performance remains constant with variations of workload, and so they conclude that "secondary task, subjective or physiological measures should be considered in preference to primary task measures in this [low workload] region". In the high workload region primary task measures provide a good measure of workload. A reproduction of the model they present is in figure 3.1. For the screening task as it is not known in which region of figure 3.1 the film readers would be operating, and therefore it is appropriate to consider a combination of primary task and other methods. Accurate measurements of performance cannot be obtained in live screening due to the low proportion of cancers, but speed of reading can easily be measured.

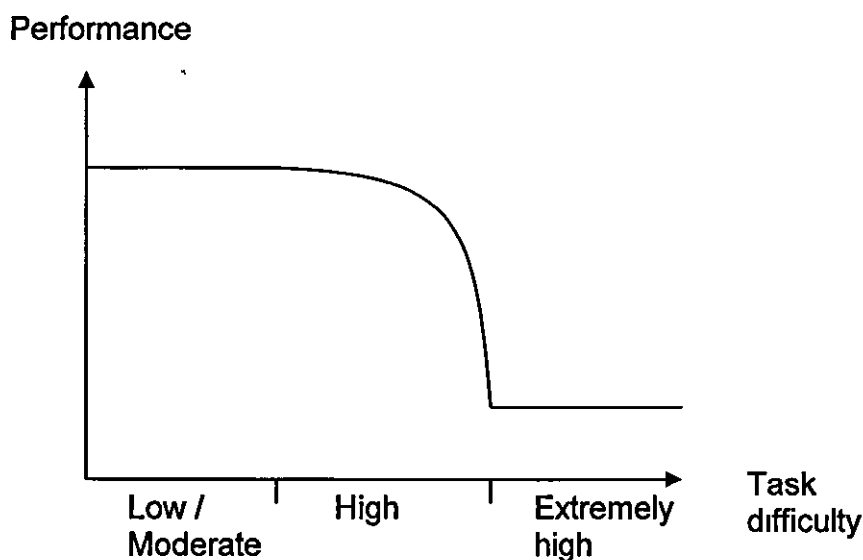


Figure 3.1 – Model of variation of workload as a result of task difficulty with primary task performance; adapted from O'Donnell and Eggemeier (1986)

Secondary task completion is a measure of spare processing capacity from the primary task. Participants are asked to complete a secondary task when they have available mental resources, with increases in productivity in the secondary task associated with decreases in workload (O'Donnell and Eggemeier, 1986). Whilst this may provide an account of the spare attentional resources, it would not be ethical to implement this in the Breast Screening Programme using live cases, as it may affect cancer detection performance. Similarly whilst there is evidence that EEG measurements are linked to mental demand (Hankins and Wilson, 1998), it is not a practically applicable technique for live screening.

Subjective rating scales which are multidimensional can provide information not only on the overall workload experienced, but can provide “some diagnostic information on the sources of workload represented by the

subscales" Wierwille and Eggemeier (1993, pg 267). According to Hill *et al.* (1992) the two multidimensional workload rating scales which have been most validated are the NASA Task Load Index (NASA TLX, Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (SWAT, Reid and Nygren, 1981). NASA TLX uses scoring from 0 to 100 of six subscales of workload: mental demand; physical demand; temporal demand; performance; effort; and frustration level. SWAT uses three levels of scoring on three elements of workload. time load; mental effort load; and psychological stress load. SWAT has fewer increments on the scale, and the zero point on the scale does not represent zero workload, and therefore Nygren (1991) suggests that NASA TLX is a more sensitive tool, particularly for low workload tasks. Hill *et al.* (1992) measured sensitivity of both SWAT and NASA TLX measures, using factor analysis to establish a single factor for workload, and using the correlation of each scale with the operator workload factor (the factor validity) as the measure of sensitivity. NASA TLX was found to be more sensitive than SWAT across five tasks. NASA TLX has been found to exhibit lower between subjects variability than SWAT (Vidulich and Tsang, 1986). Nygren (1991, pg 30) argues that the "lower between subjects variability property ...does indeed make (NASA TLX) better suited for many applications but not all. The applications in which it is well suited include those in which individual differences are expected to be minimal or of no concern, or in which more global predictions for a specific population of judges (e.g. highly trained helicopter pilots) are to be obtained." This property makes NASA TLX more suited to the task of comparing workload between workstations for participants who are all highly trained film readers.

However, there are several issues with the NASA TLX weighting which is applied to add relative importance to the six subscales. The process involves pairing the subscales into all 15 possible combinations, the participant chooses the most important of each pair, and the weighting for each subscale is the proportion of the 15 comparisons for which it was chosen. This results in a maximum weighting of 33% and minimum weighting of 5%. The obvious issues with this are that the most important subscale may be considered more significant than 33% of overall workload, and a participant who scores consistently will always rate one subscale zero and it will be excluded from the analysis, even though it was scored the least important rather than unimportant (Nygren, 1991, pg 32). The authors conclude that the NASA TLX "dimensional weighting procedure is ineffective, and should generally be ignored". Byers *et al.* (1989) compared weighted NASA-TLX scores with an unweighted equivalent, NASA Raw Task Load Index (NASA RTLX) and found very high correlation ($R_s=0.96-0.98$), and therefore conclude that they are essentially equivalent. Furthermore, Hendy *et al.* (1993, pg 596) investigated other methods of using the NASA TLX paired comparison data to give weightings, and concluded that "a simple unweighted additive model provides an adequate model for combining the individual factor ratings into an estimate of overall workload. It is not expected that procedures such as the TLX (Weighting) PCA (Principal Component Analysis) Thurstonian and dual scaling methods would reliably yield better results". Therefore because global predictions for a population of expert film readers are required, and individual differences are not of great interest, an unweighted NASA TLX or NASA

RTLX workload questionnaire will be used, alongside the primary task measurement of reading speed.

The validity and reliability of measures of workload are important for the interpretation of any results. Nygren (1991) argues that predictive and concurrent validity are of paramount importance in applied workload research, i.e. the extent to which the NASA TLX score predicts actual workload, and correlates with a measure that has already been validated. Several studies (Hill *et al.*, 1992, Warm and Hancock, 1991) have shown that NASA TLX workload score increases with increasing task difficulty and time demands, Nygren (1991) cites that this shows some evidence of construct validity and not predictive validity, i.e. that the scale correlates with a psychological model of workload, but has not been shown to predict actual workload. It should also be noted that there has been some dissociation found between subjective workload measures and performance in the presence of factors such as dual tasks or very high levels of task difficulty (Yeh and Wickens, 1988). A correlation between SWAT and NASA TLX scores for workload has been found (Vidulich and Tsang, 1986) but Nygren (1991) argues that this demonstrates criterion validity rather than concurrent validity as these other measures of workload have not been validated. Predictive validity requires a correlation between the workload score and actual workload, and therefore the measurement of predictive validity requires the measurement of actual workload, which is the use of information processing resources in the brain, and therefore Hendy *et al.*, (1993) say that it is difficult to validate a measure of workload without an external representation of information processing

resources. There is evidence that NASA TLX has face validity, i.e. that participants perceive it to be a better measure of workload than SWAT or univariate scales (Hill *et al.*, 1992), which demonstrates that the NASA TLX overall workload score is a good measure of participants perceptions of their own workload, but does not show a direct link to information processing capacity. Therefore whilst NASA TLX has been validated as much as any other subjective measure of workload, in the interpretation of results it must be considered that there is no evidence of predictive validity and so it is not proven to measure the information processing demands in the brain.

3.4 Method

The film, hybrid, and digital workstations were investigated as detailed in chapter 2. Four radiologists and four radiography advanced practitioners took part in the study, with range of experience 3 to 18 years, mean 8 years. Each participant undertook two 45 minute sessions at each of the film, hybrid and digital workstations. During these sessions they read current screening cases.

3.4.1 Workload Method

Immediately after every session each participant completed the NASA RTLX workload questionnaire, which is equivalent to the NASA TLX workload questionnaire but with no weighting applied to the subscales. The subscales of mental demand, physical demand, temporal demand, performance, effort,

and frustration were scored from 0 to 100% along a 10cm line. A description of each subscale as defined by Hart and Staveland (1988) was provided above the scale to be marked. The data recording sheet is in appendix 4. Each participant undertook two sessions at each workstation, and a mean of the scores between the two sessions was taken for each subscale. Perceptions of high performance are associated with lower perceptions of workload, whereas a higher score on any of the other subscales is associated with higher workload. Therefore, overall workload was calculated by taking a mean of the scores for mental demand, physical demand, temporal demand, effort, frustration, and 100 minus performance.

3.4.2 Speed of Reading Method

For each participant the time taken to report every screening case was recorded over the two sessions at each of the three workstations. This was achieved by analysis of the video of the experiment, which was imprinted with a time stamp correct to the nearest second. The start time and end time for each case was defined as when the participant put the screening bag down for the previous case. The start time for the first case was defined as when the participant first looked at the mammograms. If the participant stopped looking at the case, for example to answer a question from a colleague, then the timer was stopped for the duration of the interruption. The mean time taken to read a screening case for each participant at each workstation was calculated. This calculation was repeated with recalled cases removed

because these cases took longer, and it was not possible to control the number of recalled cases for each participant at each workstation.

3.4.3 Statistical Analysis

Both workload scores and time taken per case can be treated as ratio data and the use of parametric statistics was appropriate where the data quality criterion were met. The primary comparison of interest was between the hybrid and digital workstations, because these are the two possible future implementations of digital mammography. Therefore, a priori Student's *t* tests were used to test the difference between workload scores and mean time taken per case at hybrid and digital workstations. The Kolgorov-Smirnov and Shapiro-Wilk statistics were used to check that the differences between the scores at the hybrid and digital workstations followed a normal distribution for both workload score and time taken per case. The power of these statistics for measuring deviations from normality is also low, because the number of participants in the study is low, and therefore the Q-Q plots and boxplots were examined to identify any skewness/kurtosis or outliers in the data set respectively.

Mixed design analysis of variance was conducted to establish any differences in workload score or time taken per case in the transition to digital mammography through comparisons with the traditional film workstation. Workstation type was the within subjects independent variable, and participant type (radiologist or radiography advanced practitioner) was the

between subjects independent variable. Data quality was tested using Mauchly's test for sphericity

To assess which aspects of workload were contributing to trends in the overall workload score the correlation between each of the subscales and the overall workload was assessed.

3.5 Results

3.5.1 Workload Results

A priori comparison found NASA-RTLX workload scores were higher at the hybrid than at the digital workstation ($t(7)=2.83$, $p=.03$, $r=.73$). These data passed the tests for use of parametric statistics, as the differences between the scores obtained for each subject were normally distributed, see appendix 5.

Analysis of variance for the workload scores for the three workstations showed a significant main effect of workstation type, ($F(2,12)=5.26$, $p=.02$), but pairwise post hoc tests were not significant. The mean workload scores are shown in figure 3.2. The main effect of participant type was not significant ($F(1,6)=.47$, $p=.5$). There was a trend towards an interaction between participant type and workstation type ($F(2,12)=3.05$, $p=.09$), which indicates that the variation in workload score across workstations may differ by participant type, see figure 3.3. Mauchly's test for sphericity was not significant ($\chi^2(2)=1.626$, $p=.4$).

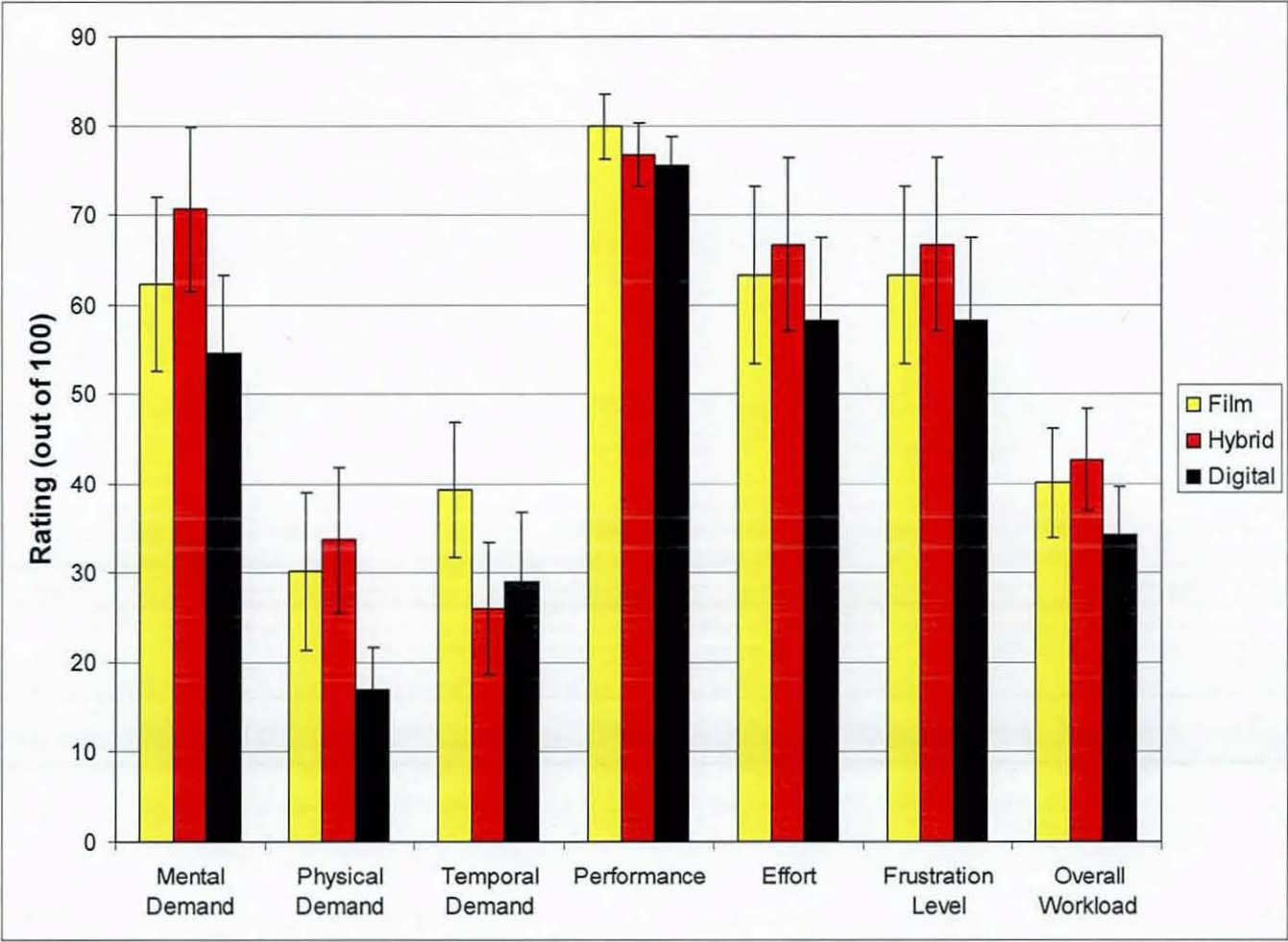


Figure 3.2 – Mean NASA RTLX workload scores for each of the workstations. Error bars represent ± 1 standard error.

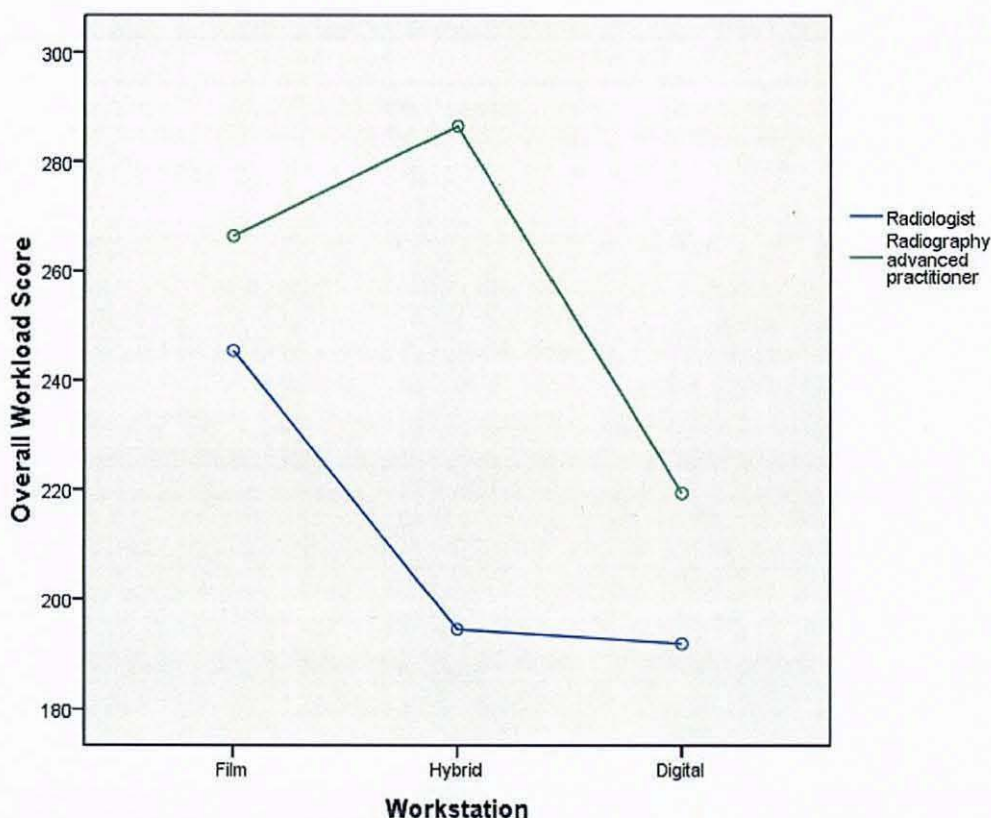


Figure 3.3 - Plot of the interaction between workstation type and participant type for workload score. There is a trend towards an interaction ($p=.09$).

To assess which aspects of workload are contributing to the difference in workload scores between the hybrid and digital workstations the correlation between each of the subscales and the overall workload was assessed. The distributions of both overall workload, and all of the subscales were not normally distributed, as shown in figure 3.4, and detailed in appendix 5. Participant 7 was removed from the analysis, as this participant was the source of all of the outliers on the boxplots. With participant 7 removed all variables met the normality of distribution criteria, with the exception of performance, for which the Shapiro-Wilk test ($p=.6$) determined the condition was not violated but the Kolmogorov-Smirnov test ($p=.045$) determined the condition of normality was violated. There was a significant positive

correlation between overall workload and mental demand ($r=.71$, $p<.0005$), physical demand ($r=.58$, $p=.005$), temporal demand ($r=.60$, $p=.004$), effort ($r=.46$, $p=.04$), and frustration ($r=.53$, $p=.01$). There was no correlation between overall workload and performance ($r=.22$, $p=.3$), as shown in table 3.1.

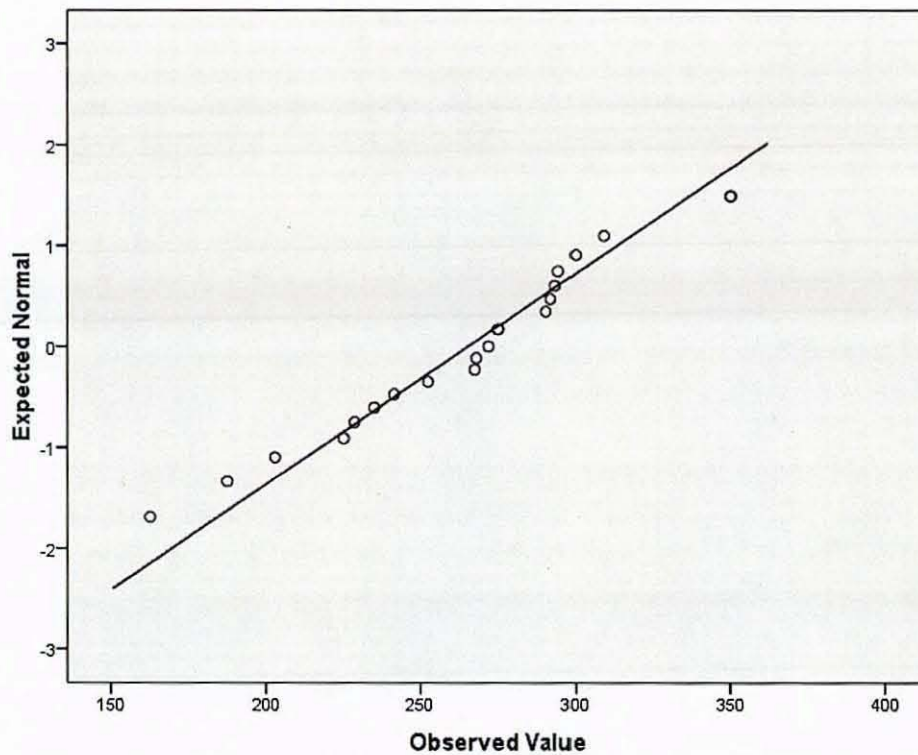
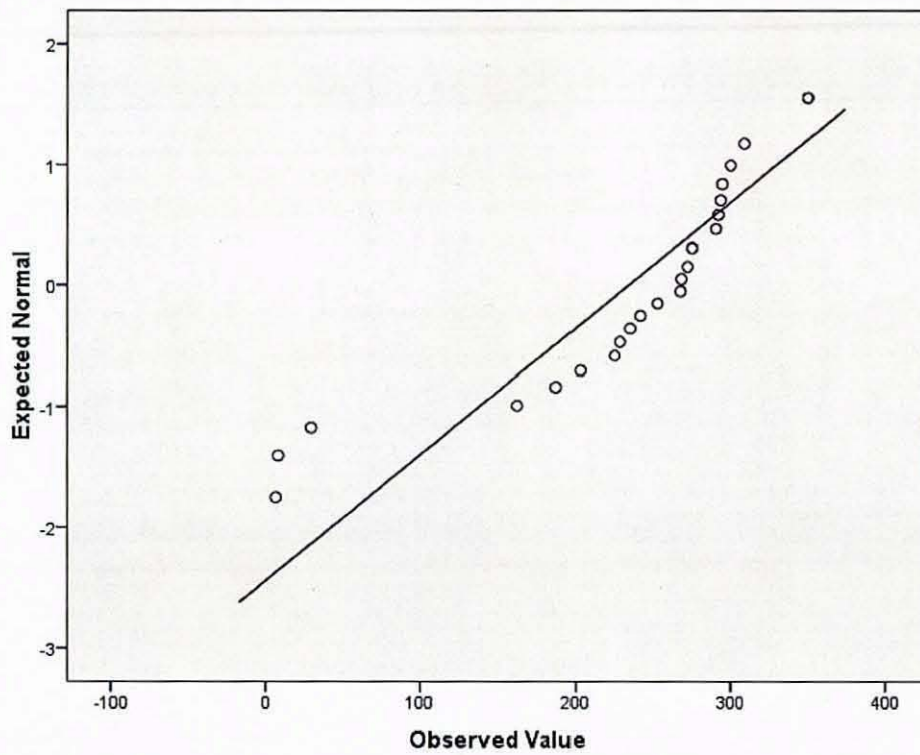


Figure 3.4 – Q-Q plot of workload scores with participant 7 included (above) and excluded (below). Inclusion of participant 7 results in outliers and skewness.

Table 3.1 - Correlations between subscales of workload and overall workload, with participant 7 removed from the analysis, *denotes correlation is significant at the 0.05 level (2-tailed), ** denotes correlation is significant at the 0.01 level (2-tailed).

Correlations		Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	Overall Workload
Mental Demand	Pearson Correlation	1.000	.504*	.010	-.444*	.726**	.136	.712**
	Sig. (2-tailed)		.020	.967	.044	.000	.557	.000
Physical Demand	Pearson Correlation		1.000	.157	-.447*	.331	-.222	.584**
	Sig. (2-tailed)			.497	.042	.143	.333	.005
Temporal Demand	Pearson Correlation			1.000	.043	-.095	.403	.602**
	Sig. (2-tailed)				.854	.682	.070	.004
Performance	Pearson Correlation				1.000	-.744**	.389	-.215
	Sig. (2-tailed)					.000	.081	.349
Effort	Pearson Correlation					1.000	-.048	.463*
	Sig. (2-tailed)						.835	.035
Frustration	Pearson Correlation						1.000	.528*
	Sig. (2-tailed)							.014

3.5.2 Speed of Reading

Use of parametric statistics was found to be appropriate, as the differences between the scores obtained for each subject were normally distributed, see appendix 5. There was a trend towards a lower time taken per case at the full digital workstation (35 seconds per case) than at the hybrid workstation (44 seconds per case) but this difference was not significant ($t(7)=2.3$, $p=.053$). The analysis was repeated with recalled cases excluded, and the time taken was lower at the full digital workstation (32 seconds per case) than the hybrid

workstation (39 seconds per case) by 18% ($t(7)=2.5$, $p=.04$), as shown in figure 3.5.

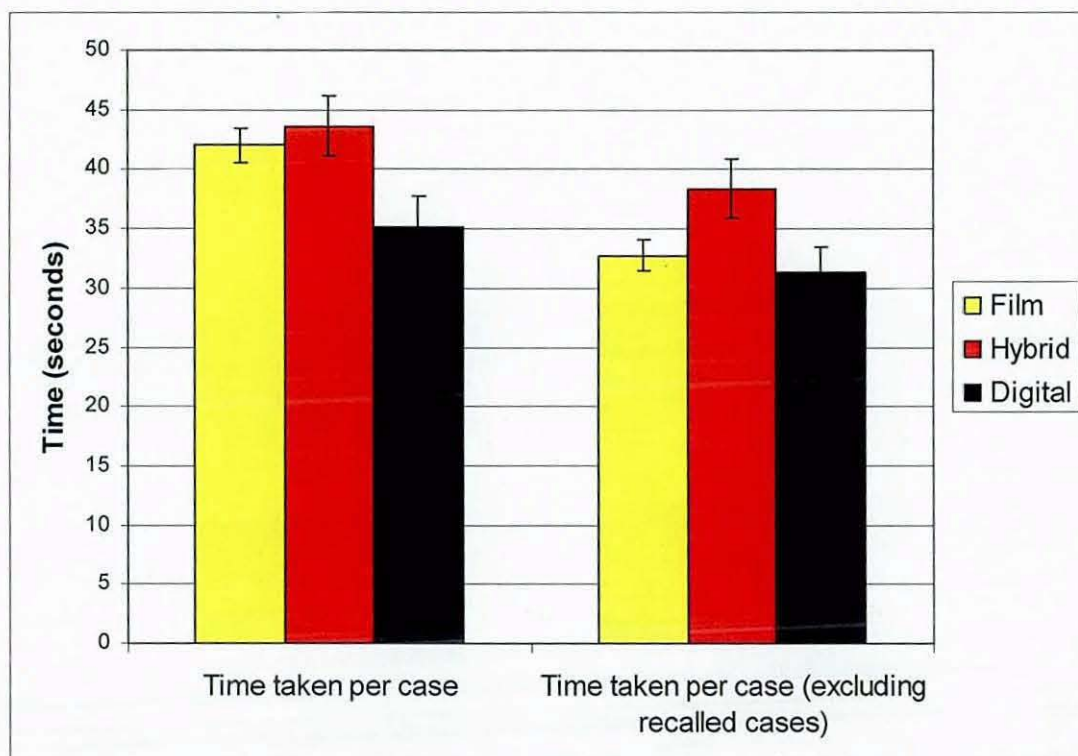


Figure 3.5 - Mean time taken per case at each workstation. Error bars represent \pm one standard error from the mean.

The analysis of variance showed a significant main effect of workstation type, ($F(2,12)=4.64$, $p=.03$), but pairwise post hoc tests were not significant. There was a significant main effect of participant type ($F(1,6)=10.9$, $p=.02$), with radiography advanced practitioners taking more time per case than radiologists. There was no interaction between participant type and workstation type ($F(2,12)=.047$, $p>.9$), which indicates that the difference in speed of reading at different workstations did not differ by participant type, see figure 3.6. Mauchly's test for sphericity was not significant ($\chi^2(2)=2.72$, $p=.3$).

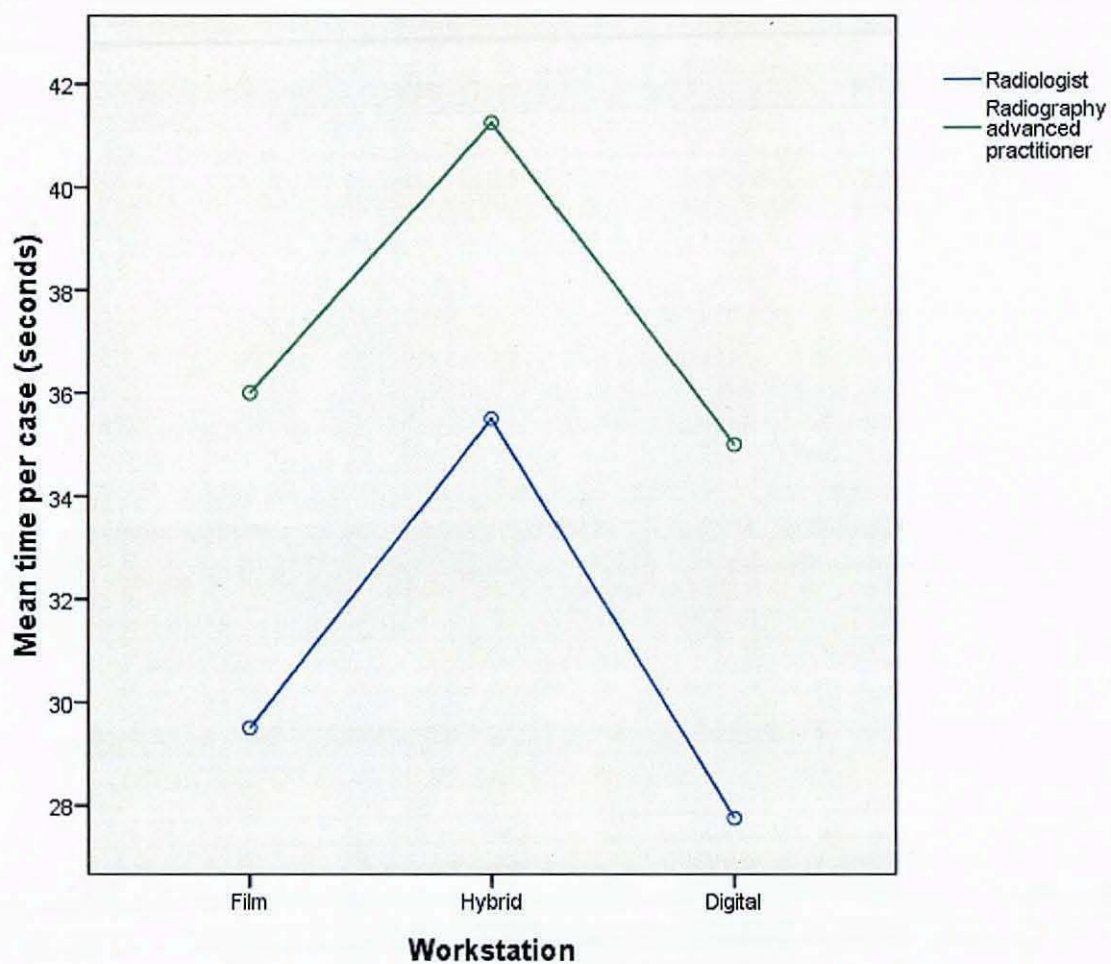


Figure 3.6 – Plot of the interaction between workstation type and participant type. Both workstation type and participant type are significant main effects but there is no interaction between them.

3.6 Discussion

3.6.1 Speed of reading

No evidence was found to show a change in the time taken per case when film is replaced by digital mammography. This is in agreement with the findings of Pisano *et al.* (2002). In fact, the mean time per case at the hybrid workstation was higher than that at the film workstation, but this result was not significant. If the introduction of digital mammography is unlikely to provide a reduction in time taken to read each case then the requirement for hours of mammography reading time in the Breast Screening Programme will increase with the increase in case load with the age extension. If there are no additional staffing resources then each mammography reader will have greater demands placed upon them. The effect this will have on individuals stress levels, work load and performance are unclear.

Mean time taken to read a normal case was 18% shorter when the prior mammograms were displayed in digitised rather than film format. In the context of an increase in case load due to the age extension to 47-73 years this may be an important result. The cause of the difference in speed of reading is important as it may have implications for performance. It could be simply due to the extra time taken to physically move between the digital current mammograms and the film prior mammograms. However the postural analysis data detailed in chapter 2 suggest that in the majority of cases the participant simply turns their head rather than moving their whole body, which

would not account for a seven second time difference. It may be due to the adjustment between perceiving digital current mammograms and film prior mammograms which are quite different in appearance, or the time taken to adapt the eyes between the different light levels of the film alternator and the digital LCD screen. However, another explanation is that the participants are simply using the prior mammograms fewer times per case when they are digitised, due to either their small display size or digitisation quality. If this is the case then it may degrade performance and therefore the time savings are not an advantage. Further investigation of the level of use of the prior mammograms when in film and digitised format is necessary.

The mean time taken to read each normal case was shorter for radiologists than radiography advanced practitioners by 6 seconds per case. There was no interaction between participant type and workstation type for speed of reading. Therefore, there is a reduction in mean time to read a normal case when the prior mammograms are displayed in digitised rather than film format for both types of participants.

3.6.2 Workload

There was no evidence found to suggest that subjective workload will increase in the transition to digital mammography if case load were to remain the same. No evidence was found to suggest that the change in equipment and availability of extra functionality at the digital workstation in comparison to the film workstation produces any change in perceptions of workload. This

was in participants with two years experience using digital mammography for screening, and therefore these results may not apply for the period immediately after the introduction of the new technology. The effect of the increase in case load with the age extension on mammography readers' perceptions of their own workload was not investigated.

Workload scores were lower when prior mammograms were displayed in digitised rather than film format. The reason for this difference has implications for its interpretation. It may be due to an increase in effort necessary at the hybrid workstation to make comparisons between the current and prior mammograms due to the physical distance between them, and the adjustment between the different light levels and the analogue and digital nature of the two displays. If this is the case, and the quality of information displayed in the film and digitised prior mammograms is the same, then to achieve equivalent performance the mammography reader would have to use more resources at the hybrid than the digital workstation. This can be modelled by performance resource curves, as shown in fig 3.7. The upper left area of the chart is the most desirable to operate in because maximum performance is obtained with minimum resources.

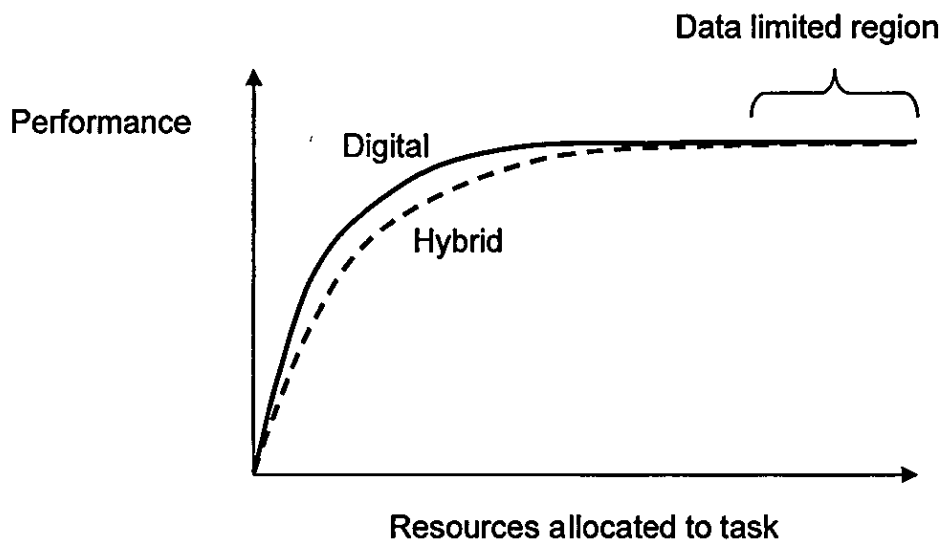


Figure 3.7 – Hypothetical relationship between performance and resource allocation reading mammograms with prior mammograms in digitised and film formats. Relationship is based on the theory that the higher perceived workload at the hybrid workstation is due at least in part to the greater mental resources required to make comparisons to the film prior mammograms.

There is evidence to suggest that performance is improved with the use of prior mammograms (Burnside *et al.*, 2002, Thurfjell *et al.*, 2000, Sumkin *et al.*, 2003, Varela *et al.*, 2005, Roelofs *et al.*, 2007). Wickens (1991) proposes that adoption of different strategies in response to the same task can produce different performance response curves. Two theoretical performance resource curves are proposed to model mammography reading with and without prior mammograms, see figure 3.8. With prior mammograms, greater performance is possible and it takes a greater amount of resources to reach the data limited stage as more data is available. If the prior mammograms are difficult to access then further resources are required for an equivalent increase in performance than when the prior mammograms are easy to access. This model predicts that prior mammograms will be used more at the digital than at the hybrid workstation because they are more accessible, and therefore the

performance advantage of using them can be accessed with little effort (as the performance resource curve will be nearer to the desirable upper left region of the chart). It also predicts that when committing high levels of resources performance will be better using prior mammograms, but when committing only low levels of resources this advantage may no longer be present, particularly if the prior mammograms are difficult to access.

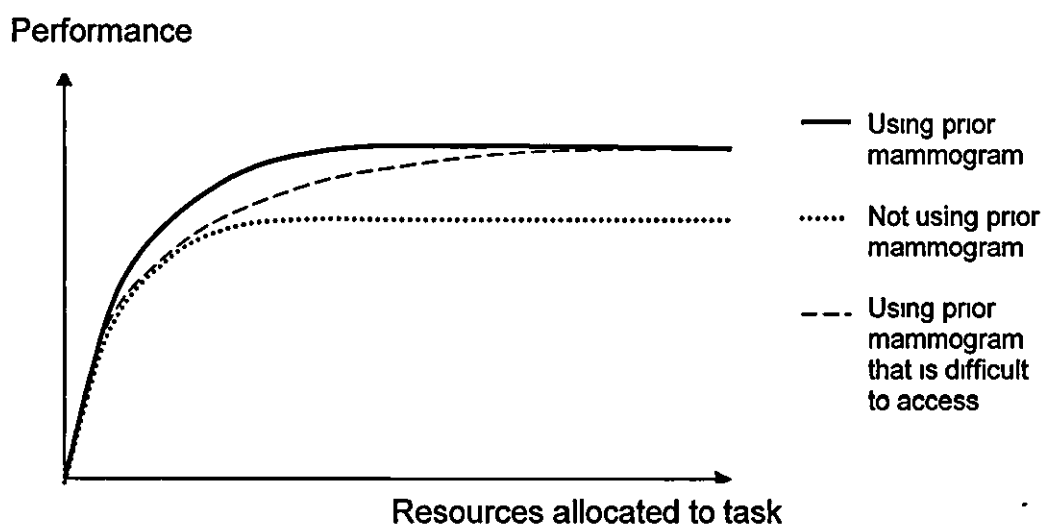


Figure 3.8 – Hypothetical relationship between effort and performance in reading mammograms with or without the prior mammograms. At very low levels of effort performance may be the same without prior mammograms, but at higher levels of effort without the prior mammograms performance becomes data limited.

These hypotheses are dependent on the idea that looking at the current and prior mammograms use the same single resource, which appears reasonable as according to Wickens (1992) multiple resource theory these are both perceptual visual spatial tasks. However, in practice there is an additional task required when using the hybrid workstation, as the film multi-viewer must be moved to the next case at the same time as the digital display. There is only one display to move when using either the film or the digital workstation.

Keeping the extra viewer in the correct place could be considered a concurrent task to the film reading task, and therefore a dual task is carried out at the hybrid workstation when only a single task is required at the film and digital workstation. Yeh and Wickens (1988) report a dissociation between performance and workload when comparing difficult single tasks to easier dual tasks. The measured performance was higher for simpler dual tasks, but the workload was also reported to be higher. Therefore the higher workload scores at the hybrid workstation may simply be due to the simple extra task of keeping the second viewer in the correct place. However, it is also possible that with experience automaticity would develop for this second task and it would no longer require resources to be expended. There are also other potential dissociations between workload and performance including automaticity, motivation, (Vidulich and Wickens, 1985) and participant overload (Yeh and Wickens, 1988). These factors may be also be present in the breast screening situation. The NASA RTLX performance subscale was not correlated with overall workload score suggesting that participants' perceptions of their own performance were not related to either their perceptions of overall workload or the presentation medium of the prior mammograms. However, perceptions of performance are not necessarily indicative of actual performance. Measurements of levels of use of the prior mammograms and performance would establish whether the performance resource model is relevant, and whether dissociation is occurring between workload and performance.

Whilst participant type (radiologist or radiography advanced practitioner) is not a significant main effect for workload score, there may be an interaction between participant type and workstation type. Figure 3.3 shows that for radiography advanced practitioners the mean workload score at the hybrid workstation was higher than at the film workstation, for radiologists the reverse relationship was found. Radiologists have very different training to radiography advanced practitioners, which involves reading both analogue and digital x-rays of a range of body parts. This may reduce their experience of workload when both analogue and digital mammograms are displayed at the hybrid workstation than their radiography advanced practitioner colleagues who do not have such a breadth of experience. However, this is just a trend rather than a significant effect and therefore no conclusions can be drawn from it. It does however highlight how the different experiences of staff may lead to different perceptions of the new digital equipment.

3.6.3 - Conclusions

The first aim of this experiment was to determine if there will be any changes in subjective workload and mean time to read a case in the transition from film to digital mammography. There was no significant change in the participants' subjective workload scores and mean time to read a case between the film workstation which represents the current setup, and either the hybrid or digital workstations which represent two possible digital mammography implementations. There is no evidence to suggest that the transition from film to digital mammography will result in a change in time taken per case or

mammography readers' workload per case. However, the effects of the increase in case volume which will be implemented alongside the introduction of digital mammography have not been tested here.

The second aim was to measure the impact of displaying digitised prior mammograms on the digital workstation or film prior mammograms on an adjacent multi-viewer upon workload, and mean time to read a case. Displaying the prior mammograms in film format was found to increase both mean time taken per normal case, and NASA RTLX workload scores in comparison to digital display.

The third aim was to establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2. Radiography advanced practitioners average time to read each case was longer than their radiology counterparts but there was no interaction with workstation type, and therefore speed of reading was slower at the hybrid workstation than the digital workstation across all types of participants. There was a trend towards an interaction between workstation type and workload score, but it was not significant. Therefore it is unclear whether perceptions of workload at the different workstations differ between radiologists and radiography advanced practitioners.

4 Behavioural Use of Prior Mammograms

4.1 Introduction

The use of prior mammograms in breast screening has been found to improve performance (Roelofs *et al.*, 2007, Varela *et al.*, 2005, Burnside *et al.*, 2002, Thurfjell *et al.*, 2000, Sumkin *et al.*, 2003), this will be considered in more detail in chapter 5. Breast screening mammography readers may underestimate the value of prior mammograms to their own performance according to one study (Roelofs *et al.*, 2007). Twelve radiologists read 160 cases twice, with and without prior mammograms, and it was noted whether the participants thought the prior mammograms necessary. This created effectively 3 conditions: reading with prior mammograms for every case; reading without prior mammograms for every case; and reading with prior mammograms upon request (which was calculated using a combination of the data for the first two conditions plus whether the prior mammograms were considered necessary). Prior mammograms were considered necessary for 24% of normal cases, 33% of benign cases and 28% of malignant cases. Localised Receiver Operating Characteristic (LROC) analysis showed that performance with prior mammograms for all cases was superior to performance with prior mammograms only when requested. This indicates that the radiologists underestimated the proportion of cases for which they needed the prior mammograms. In light of this finding it would be valuable to understand whether in clinical practice difficulty in accessing prior mammograms affects the proportion of cases for which they are used, and

this may impact on performance. In particular whether displaying prior mammograms in digitised format at the workstation or in film format adjacent to the workstation affects mammography readers behaviour in accessing them.

The findings from the postural analysis detailed in chapter 2 indicate that participants may be simply turning their heads to view the film prior mammograms rather than leaning over to get a better view, which suggests that the physical distance to the film prior mammograms may be affecting reading behaviour. The finding that NASA RTLX workload ratings were higher with film rather than digitised prior mammograms may also influence viewing behaviour in accessing the prior mammograms. However, the manner in which it would be influenced is dependent on the reason for the difference in perceptions of workload, which is unknown.

Whilst there is a plethora of research about performance and eye movement behaviour in reading mammograms, very little could be found concerning behaviour or eye movements in the use of prior mammograms alongside current mammograms. The search engines used in reviewing this literature included Web of science, Medline, Articlefirst, Zetoc, and the SPIE digital archive. Search terms used were "prior mammograms" and "previous mammograms". Where many results were found the first 50 were viewed, and then the search was refined further to remove those with the word "CAD" in the title to remove those focused on computer aided detection algorithms.

There are a number of studies investigating the effect of the presence of prior mammograms on performance, but no eye tracking studies were found.

4.2 Aims

1. To measure any changes in level of use of prior mammograms in the transition to digital mammography
2. To determine the impact on level of use of prior mammograms of digitising them in preference to displaying them in film format during the transition to digital mammography
3. To establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2.

4.3 Choice of Methods

To achieve the aims detailed above, measurements of behaviour in real world breast cancer screening are required, and therefore ecological validity is of paramount importance. Eye tracking equipment can measure fixation duration, saccadic eye movement patterns, and gaze trails between current and prior mammograms, and therefore can provide rich behavioural data. However, remote eye tracking is not appropriate when the participant moves between two workstations which is the case at the hybrid workstation. Head

mounted eye tracking was trialled, but the head mounted apparatus, and time to calibrate the equipment was judged to modify the comfort and behaviour of participants to an unacceptable degree. The effects of wearing head mounted eye tracking equipment on behaviour and cancer detection performance are unknown, and therefore head mounted eye tracking equipment was considered unethical for use in live screening.

Prior to the development of Purkinje reflection based eye tracking equipment there were a wide range of techniques used to measure eye movements. Those techniques which were least intrusive are direct visual observation and photographic recording. An early review of eye tracking techniques describes several studies which employ direct observation of the eye to detect direction of gaze and saccadic length, and determine that "such experiments have the merit of simplicity and have been extensively employed, but small movements cannot be observed without some optical magnification" (Lord and Wright, 1950, pg 10). Early photographic recording techniques involved taking successive photographs with the participants head held still and the eye marked with a dot of Chinese white or even mercury, which was found to give gaze direction accurate to 5 degrees of arc (Barlow, 1952). This is obviously a technique that is inappropriate with the advent of modern technology, but does demonstrate that video recorded data can be used to a high degree of accuracy. In fact Yarbus (1967) in a review of eye tracking methods states that using videotape of the eyeball the direction of gaze can be calculated correct up to 1° of visual angle. This level of accuracy is far greater than

necessary for distinguishing whether the reader is looking at current or prior mammograms.

With modern eye tracking equipment unsuitable, and any interference with the participant unethical as behaviour in live screening is to be analysed, eye movements were measured simply by video taping the participants eyes and manually calculating what the participant was looking at. This would maximise ecological validity, through the introduction of just a small unobtrusive video camera. This approach was considered the most appropriate if sufficient data quality could be achieved.

4.4 Method

4.4.1 Pilot Studies

A series of pilot studies were conducted. The first pilot study involved two participants, one male and one female. The aim was to ascertain if it is possible to detect whether gaze is directed at current or prior mammograms using only remote video cameras, and if so how many video cameras were necessary, whether distance to screen information is necessary to determine gaze location, and which lens type to use on the video cameras. Each participant sat in front of a multi-viewer displaying eight numbered mammograms, see figure 4.1, and was asked to look from current to prior mammograms as directed. Four different camera placements and two types of lens were trialled by looking at the video feed to see if eye position could be determined.

The second pilot aimed to determine whether the results of the first could be replicated for different participants, and when the activity was reading mammograms, and whether individual calibration was necessary before each session. Five participants took part, all of whom were radiological novices, and had not taken part in the first study. Each participant was shown one example of a spiculated mass, and one example of microcalcifications and instructed that these were indicative of cancer if they had appeared since the previous mammograms. They were then asked to first look at each of the eight positions on the multi-viewer in whichever order they chose and test whether the experimenter could tell from the video feed whether they were looking at the top or bottom row. Then they read a series of ten cases at a mammography multi-viewer and look for signs of cancer. The experimenter viewed the video output and determined whether it was possible to detect whether the participant was looking at the current or prior mammogram at all times.



Figure 4.1 – Experimental set up for camera placement trials. Numbers 1 to 4 indicate the four camera placements trialled. The inset shows the camera used.

4.4.2 Main Study

For the main study the film, hybrid, and digital workstations were investigated as detailed in chapter 2. Four radiologists and four radiography advanced practitioners took part in the study, with range of experience 3 to 18 years, mean 8 years. Each participant was video-taped undertaking two 45 minute routine screen reading sessions at each of the three workstations. This involved recording whether each case was 'normal' i.e. returned to screening, or 'abnormal' in which case a recall form was filled out to call the woman back for further tests. The cases were not pre-defined but were part of the readers' routine work within the clinical context. Prevalent screens (women at their first

screening appointment and therefore without prior mammograms) were excluded from the analysis. The number of cases examined per session varied according to the participants' reading speed, the mean was 41. All cases had both cranio-caudal and medio-lateral oblique views for the current mammograms, and the majority of cases (over 95%) had the same views for the previous mammograms, with less than 5% having only medio-lateral oblique views for the previous mammograms. Video recording was completed using four unobtrusive miniature cameras (one videotaping each of the face and the display at the digital and multi-viewer workstations), attached to synchronisation equipment located under the workstation, see figure 4.2.



Figure 4.2 – The hybrid workstation with miniature video cameras highlighted by the blue arrows, and video recording display below the workstation.

The number of times the participants looked at the previous mammograms was recorded for every case read by analysing the participants' gaze

positions on the video-recorded data. Every reading session involved video-taping the participants whilst reporting current screening cases, rather than reading known test cases. This enabled more accurate measurements of real life behaviour, as measurements were taken during normal work activities. However, in these circumstances the number of cases recalled could be a confounding variable, as it is likely that cases for recall would be dealt with differently, and the old films referred to more frequently when completing recall paper work. Therefore video analysis of the number of visual comparisons to previous mammograms was stopped as soon as the participant started to fill in a recall form. The analysis was completed twice, including and excluding recalled cases.

4.4.3 Statistical analysis

The number of times the participants looked at the prior mammograms per case, and the proportion of cases for which the prior mammogram was looked at were the variables analysed. Measurements for both variables were compared for the hybrid or digital workstation using an 'a priori' Student's paired t test. The purpose of this was to provide information about whether the prior mammograms should be digitised. These data were analysed to check that the distribution of differences between the hybrid and digital workstations were normally distributed using the Kolgomorov-Smirnov and Shapiro-Wilk tests alongside boxplots and Q-Q plots. A mixed design analysis of variance was conducted with both workstation type (film, digital, or hybrid) and participant type (radiologist or radiography advanced practitioner) as

independent variables. These were followed by pairwise post hoc Student's t tests with a Sidak correction for multiple comparisons where appropriate. For the ANOVA the assumption of sphericity was tested using Mauchly's test.

4.5 Results

The pilot studies demonstrated that with a hanging protocol of current mammograms on the upper and prior on the lower row it was possible to detect whether participants were looking at current or prior mammograms with one video camera at location 1 on figure 4.1. Measuring distance to screen and individual calibration were found to be unnecessary. A wide angle lens was necessary due to the large display size. For the digital workstation a second synchronised camera was required as there were three hanging protocols to view, in two of which the prior mammograms were not hung. When looking at the current mammograms and then subsequently at the workstation controls the saccade passes over the prior mammograms. In these circumstances the prior mammogram is not considered looked at unless a fixation upon the prior mammograms is clearly visible from the video-tape. A fixation was defined as such when the experimenter could see the motion of the eyes stop on the video tape. When eye tracking equipment is used then eye position is recorded electronically using Purkinje reflections from the participants eyes, and therefore fixation duration can be defined accurately by the experimenter. Fixation duration (including both minimum pause duration of the eye and stimulus processing time) can be defined as anything from 60 to 500msec, but more typically around 200msec (Salthouse and Ellis, 1980).

However in this study using such equipment may have changed the manner in which the task was performed to an unacceptable degree, and biased the very measure which was being reported. Therefore the perception threshold of the experimenter to detect pauses in motion was considered an appropriate metric for defining fixation duration, with accuracy in defining fixation duration sacrificed in pursuit of minimising systematic error associated with the experimenter influencing participant behaviour. In the second pilot study the experimenter was 100% correct in measuring from the video whether the participant was looking at the current or prior mammograms for all eight positions for all five participants.

At the hybrid workstation for each fixation on the prior mammograms it was not known whether the intention was to examine the mammograms, or examine the case number. This problem was unique to the hybrid workstation as the prior mammograms were displayed separately to the current mammograms, and therefore identification was sometimes required. As a result of this the number of comparisons to the prior mammograms at the hybrid workstation may be overestimated.

The Kolgomorov-Smirnov and Shapiro-Wilk tests alongside boxplots and Q-Q plots showed that the assumption of normality was not violated for any of the metrics used in the Student's *t* tests. Mauchy's test showed that the assumption of sphericity was not violated for any of the metrics used in the ANOVA.

The proportion of cases for which the previous mammograms were consulted (i.e. visually fixated at least once) was found to be higher at the full digital workstation (82%) than at the hybrid workstation (63%, $t(7)=2.5$, $p=.04$), i.e. higher when the previous mammograms are digitised rather than displayed in film format on a multi-viewer, as shown in figure 4.3. The average number of times participants looked at the previous mammograms per case was greater at the full digital than the hybrid workstation ($t(7)=2.73$, $p=.03$). This is due to a combination of consulting the previous mammograms for a higher proportion of cases at the full digital workstation, and when these are consulted, looking at them a greater number of times per case ($t(7)=2.98$, $p=.02$).

For the proportion of cases for which the prior mammogram was used the main effect of workstation type was significant ($F(2,6)=8.6$, $p=.02$), but post hoc tests showed that although mean number of comparisons is higher at the film than the hybrid workstation this was not significant ($p=.2$), see figure 4.3. The main effect of participant type was not significant ($F(1,3)=3.0$, $p=.2$), but the interaction between workstation type and participant type was significant ($F(2,6)=12.0$, $p=.008$), see figure 4.4.

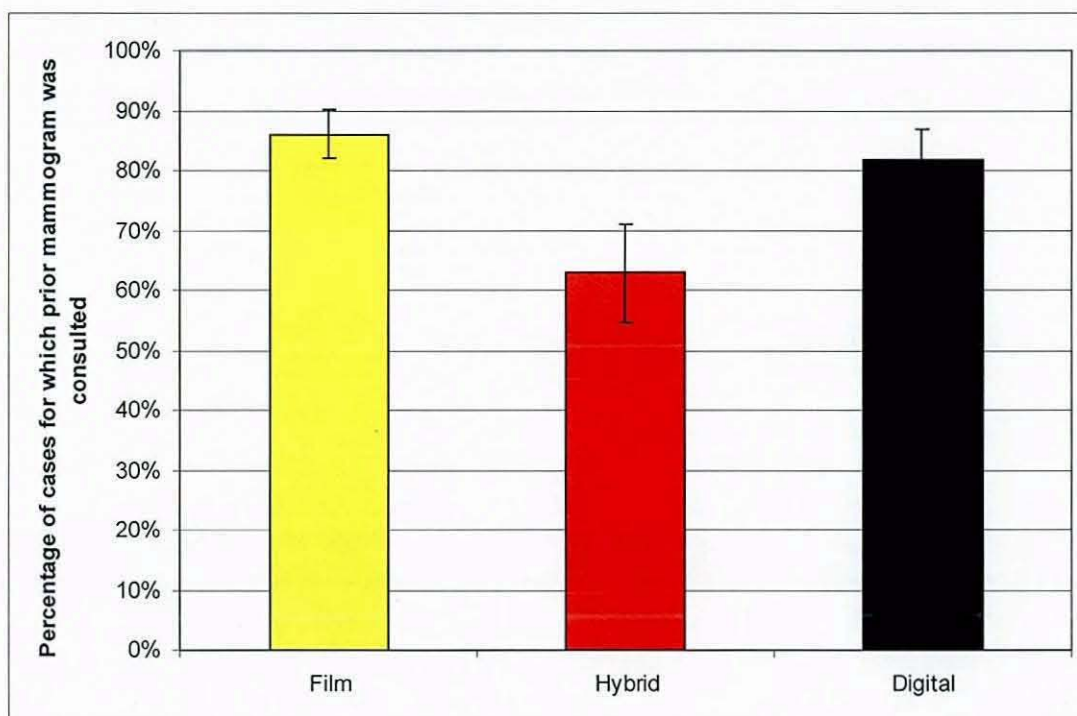


Figure 4.3 – The proportion of cases for which the prior mammograms were used (looked at once or more) at the film, hybrid and digital workstations. Error bars represent \pm one standard error.

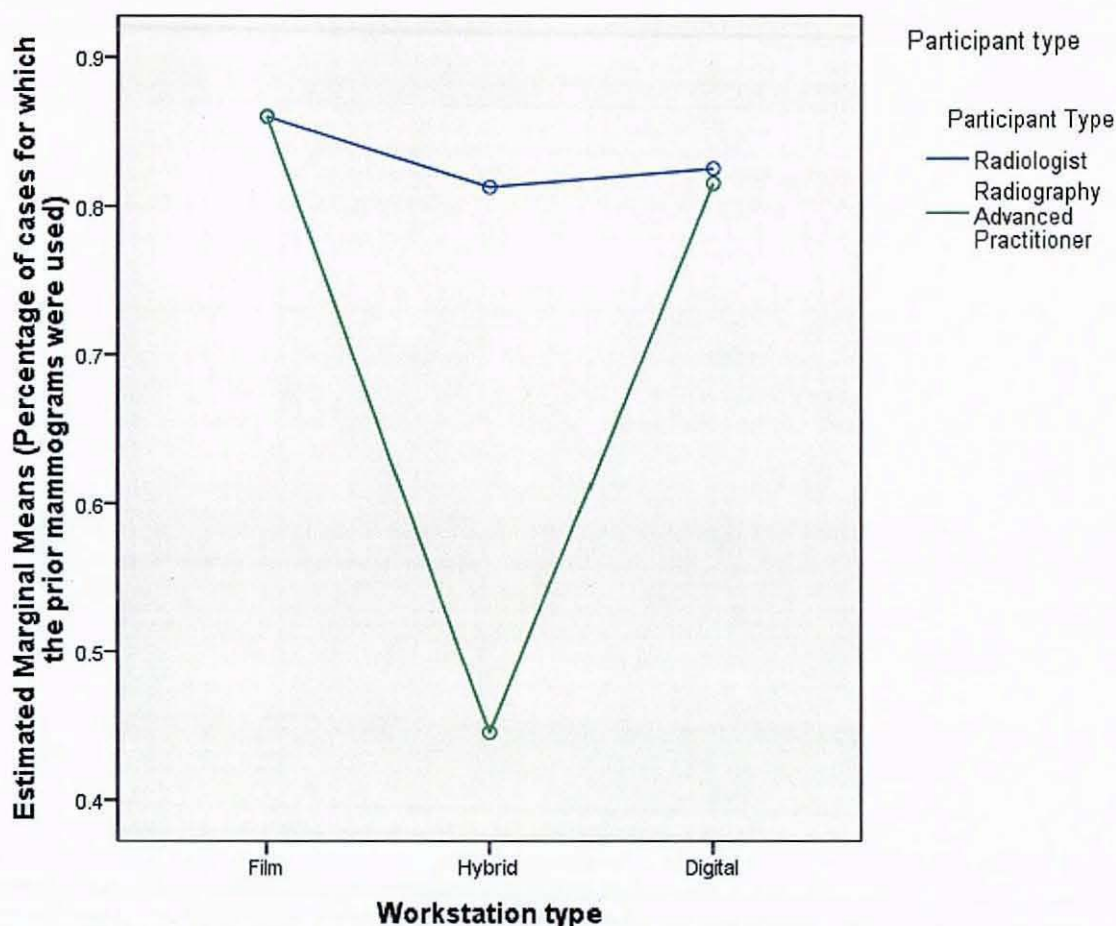


Figure 4.4 - Plot of the interaction between workstation type and participant type for proportion of cases for which the prior mammograms are used. Type 1 is radiologists and type 2 is radiography advanced practitioners. There is a significant interaction ($p=.008$).

The mean number of comparisons to the previous mammograms per case differed between the workstations ($F(2,6)=8.7$, $p=.02$), see figure 4.5. There was a trend towards the previous mammograms being looked at a greater number of times per case when displayed on the film workstation rather than on the hybrid workstation but this was not significant ($p=.1$). When only considering cases for which the previous mammograms were consulted then similar results were obtained. Analysis of variance found differences between the workstations ($F(2,6)=5.2$, $p=.049$), and the number of times participants

looked at the previous mammograms per case was higher at the film than the hybrid workstation but this was not significant.

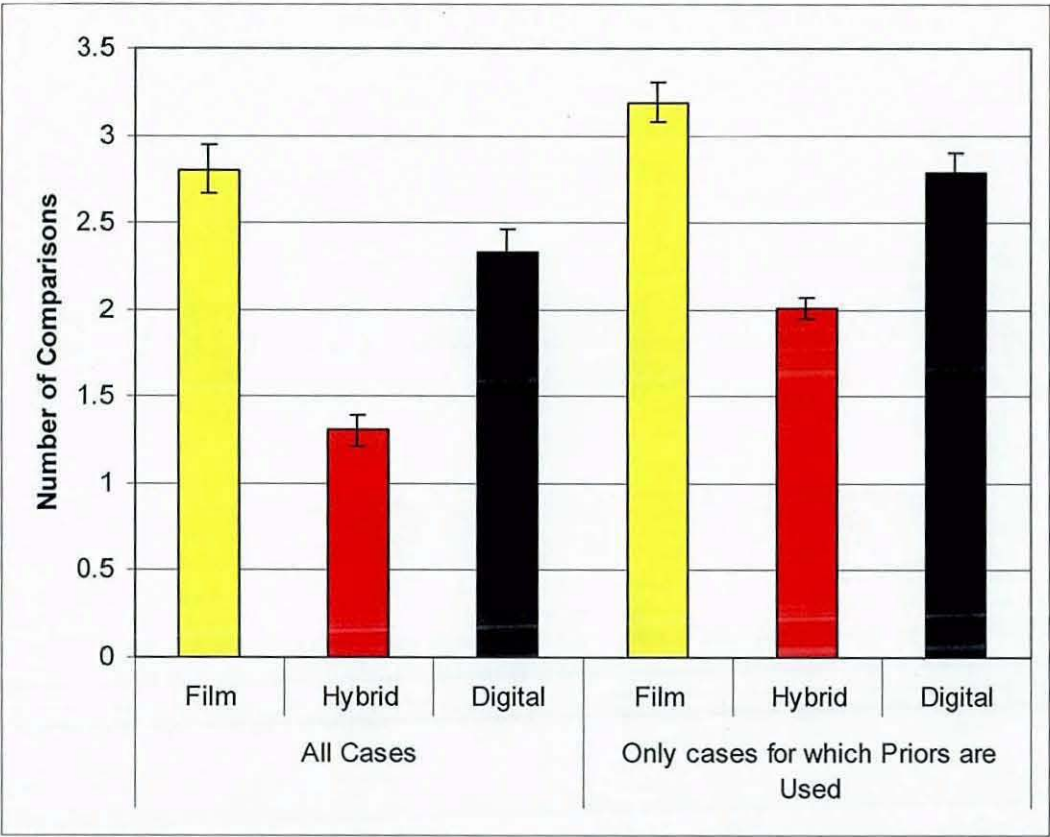


Figure 4.5 – Mean number of times per case that the participants looked at the prior mammograms. Error bars represent \pm one standard error.

For the number of comparisons to the prior mammograms the main effect of participant type was not significant ($F(1,3)=.12, p=.8$). The mean number of comparisons was higher for radiologists than radiography advanced practitioners at the hybrid workstation, but there was not a significant interaction between workstation type and participant type ($F(2,6)=2.1, p=.2$), see figure 4.6.

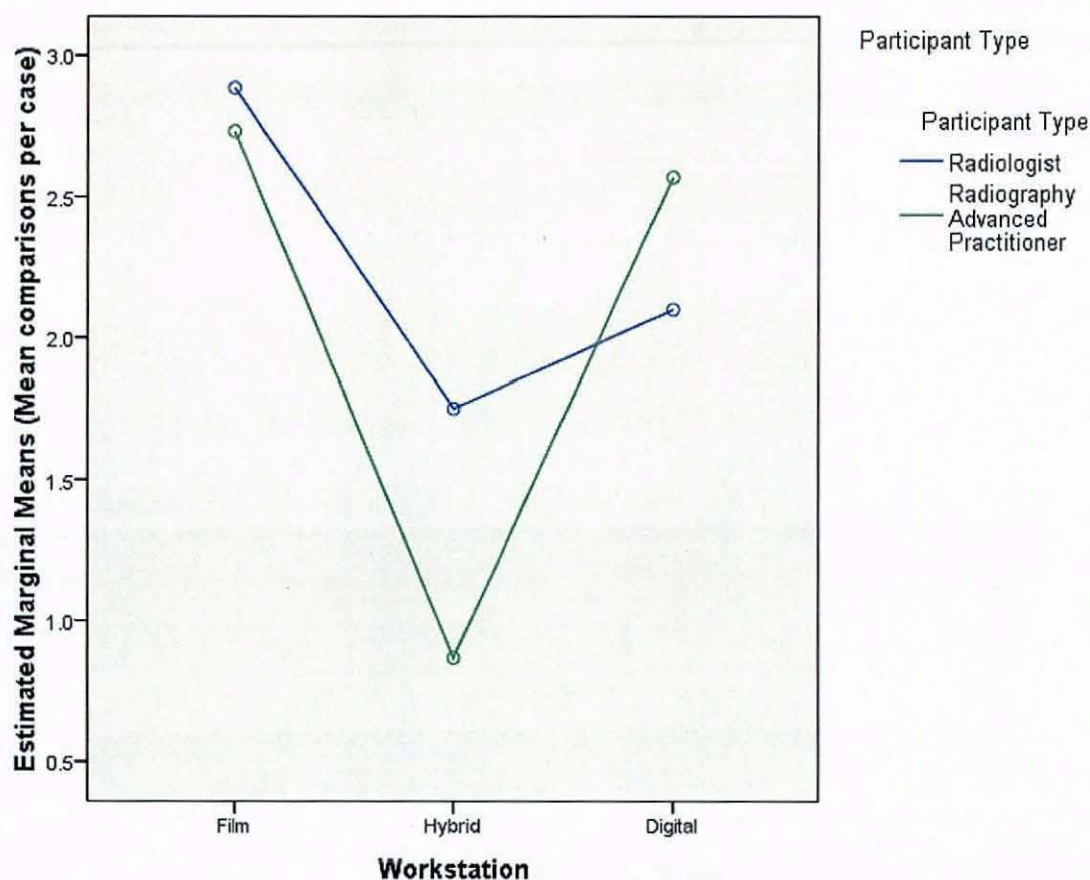


Figure 4.6 - Plot of the interaction between workstation type and participant type for mean number of comparisons per case to the prior mammograms. Type 1 is radiologists and type 2 is radiography advanced practitioners. There is not a significant interaction ($p=.2$).

As there was no main effect or interactions for participant type when analysing the number of comparisons to the prior mammograms the analysis was repeated with participant type removed as an independent variable. Using a one way repeated measures ANOVA, the mean number of comparisons to the previous mammograms per case differed between the workstations ($F(2,12)=8.7$, $p=.004$), and the previous mammograms were looked at a greater number of times per case when displayed on the film workstation rather than on the hybrid workstation ($p=.03$). When only considering cases for which the previous mammograms were consulted then similar results were obtained. Analysis of variance found differences between the workstations

($F(2,14)=10.2$, $p=.002$), and the number of times participants looked at the previous mammograms per case was higher at the film than the hybrid workstation ($p=.01$), see figure 4.5.

The proportion of cases which were recalled differed by workstation reflecting the random variation in case presentation inherent in live screening. At the film workstation 4% of cases were recalled, 2% were recalled at the hybrid workstation, and 1% at the full digital workstation. The above analyses were repeated with recalled cases excluded; the results were not significantly different to those reported.

4.6 Discussion

When the current mammograms are digital, the readers are using the prior mammograms for a greater proportion of cases when they are digitised (82%) versus displayed on a multi-viewer (63%). This suggests that for 19% of cases the readers are deciding, either consciously or subconsciously, to consult the prior mammograms if they are digitised but not if they remain in film format. This may have implications for performance as Roelofs (2007) showed that readers underestimate the proportion of cases for which using the prior mammogram would have improved their performance. Secondly, when only cases where the prior mammogram was consulted at least once are considered, the reader looks at the digitised priors a greater number of times per case than if the priors were displayed in film format. This suggests a change in reading behaviour and approach, depending on the display medium

of the prior mammograms. For example, the reader may be looking fewer times per case at the prior mammograms in film format, but absorbing more information each time. Equally it is possible that readers are simply taking less information from the film mammograms.

The interaction between participant type and workstation type indicates that the reduction in the proportion of cases for which the prior mammograms were used at the hybrid workstation is a much larger effect for radiography advanced practitioners than it is for radiologists. In fact three of the four radiologists demonstrated this effect, but to a lesser degree than the advanced practitioners, and one demonstrated the reverse effect. Radiography advanced practitioners experience a completely different course of training to radiologists, which does not include experience in other radiology departments. The behavioural differences reported here may be associated with the extra experience radiologists have of viewing film prior mammograms in a variety of situations, and reading a range of both film and digital radiographs, resulting in radiologists adapting better to making comparisons between film and digital images.

The performance resource curve for the hybrid workstation postulated in chapter three overlaps with that of not using the prior mammogram at lower levels of effort. This predicts that at low levels of effort not using the prior mammogram produces (or appears to produce) at least equivalent performance, but at higher levels of effort performance not using the prior mammogram becomes data limited, and the prior mammogram is required for

performance to improve. All mammography readers are aware of the benefit of prior mammograms to performance, and therefore the behaviour of using prior mammograms for a greater proportion of cases when displayed in digitised rather than film format may be as a result of attempting to balance performance outputs with workload and effort levels in a busy breast screening department. If the performance resource model stands, then the mammography readers who took part in this study are operating at reasonably high levels of workload, and are not operating in the data limited region. Therefore an increase in workload such as the introduction of the hybrid workstation or an increase in case load could lead to degraded performance for these participants.

Number of comparisons to the prior mammograms was higher at the film than at the hybrid workstation, but there was no statistically significant difference between the film and digital workstations. This suggests that if the hybrid workstation is implemented at this breast screening centre then the readers would adapt their reading behaviour to look at the prior mammograms fewer times. This adaptation of behaviour would not occur if the digital workstation was implemented. The cognitive processes behind such a change in behaviour, or the effects on cancer detection performance are unknown.

The process of data collection may well have affected the results of this study according to the Hawthorne effect. This was minimised where possible by using small cameras, positioning the cameras as unobtrusively as possible whilst maintaining sufficient video quality for analysis, and leaving the

cameras in situ but not recording between experimental reading sessions so that the cameras became part of the environment. However, the participants were all aware that they were being video-taped, and therefore are likely to have made more effort to look at the prior mammograms than they would have otherwise. In reality use of the prior mammograms may be lower in all three conditions, but particularly at the hybrid workstation where use of the prior mammogram requires greater effort from the mammography reader.

There are other potential issues with the hybrid workstation which were not investigated here. It is known that prior mammograms improve performance overall, and in particular reduce errors in decision making during image interpretation. In this study it has been shown that the prior mammograms are looked at fewer times and for fewer cases if they are displayed in film format rather than digitised. This could increase decision making errors not only through reduced use of the prior mammogram, but also because comparisons are more difficult due to the distance between the images to be compared, and the differences in levels of illumination of the two images requiring adjustment of the eyes. Furthermore it is not known whether the presence of prior mammograms influences search strategy through providing information to the initial global processing stage of image interpretation. The global processing stage is the first microseconds of viewing an image, in which mammography readers have been shown to be able to detect the majority of lesions (Nodine and Kundel, 1987). If prior mammograms are of use at the global processing stage of image interpretation then performance may be

degraded at the hybrid workstation as the prior mammograms are not within the field of view when looking at the current mammograms

4.7 Conclusions

The first aim of this study was "to measure any changes in level of use of prior mammograms in the transition to digital mammography". It was found that when the prior mammograms were displayed in film format rather than digitised then the average number of times the participants looked at the prior mammograms was reduced by over a third.

The second aim was "to determine the impact on level of use of prior mammograms of digitising them in preference to displaying them in film format during the transition to digital mammography". Displaying the prior mammograms in digitised rather than film format was found to increase the proportion of cases for which the prior mammogram was looked at by 19%, and when the prior mammogram was used it was found to increase the average number of times per case that the prior mammogram was looked at from two to nearly three.

The third aim was "to establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2". An interaction between workstation type and participant type was found for the proportion of cases for which the prior mammogram was used. Therefore this effect of using the prior mammograms for a greater

proportion of cases when digitised is more profound in radiography advanced practitioners than radiologists, and for one of the radiologist participants the reverse effect was found.

5 Chapter 5 – The use of Prior Mammograms and Performance

5.1 Introduction

In the transition to digital mammography digitisation of prior mammograms has been found to be preferable to film display in terms of mammography readers' perceptions of workload, speed of reading, and level of use of prior mammograms. However, there is no available evidence to date of which produces superior cancer detection performance. This chapter will investigate cancer detection performance using film, digitised, and no prior mammograms, and present the results using both receiver operating characteristic analyses and recall rates.

Prior mammograms are known to improve cancer detection performance through an increase in specificity (Roelofs, *et al.*, 2007, Varela *et al.*, 2005, Sumkin *et al.*, 2003, Burnside *et al.*, 2002, Thurfjell *et al.*, 2000). Several prospective studies using ROC based methods have shown an increase in cancer detection performance when prior mammograms are used (Roelofs *et al.*, 2007, Varela *et al.*, 2005, Sumkin *et al.*, 2003, Thurfjell *et al.*, 2000), however the ROC figures of merit cannot be directly translated into changes in the number of women recalled. This may make these studies less influential in decisions about how to display the prior mammograms. A retrospective study in the USA (Burnside *et al.*, 2002) reviewed 38,456 screening cases. All cases had prior mammograms available but they were not used for 6743

cases. The recall rate was lower when the prior mammograms were used (3.8% versus 4.9%, $p=.0001$), with no significant change in cancer detection rate. This work has not been undertaken in the northern European population screening setting.

The effect of the presentation medium of the prior mammograms on cancer detection performance using digital mammography has not been studied. Equivalent performance using digital mammograms with soft copy and printed film display has been demonstrated (Pisano *et al.*, 2002), but no such study reports performance using digitally acquired and displayed current mammograms with film versus digitised prior mammograms. In the UK Breast Screening Programme prior mammograms were found to be used for a greater proportion of cases when displayed in digitised (82%) rather than film (63%, $p=.04$) format, (Taylor-Phillips *et al.*, 2009). This implies that digitising prior mammograms may improve specificity in cancer detection, because Roelofs *et al.* (Roelofs *et al.*, 2007) demonstrated using a localised receiver operating characteristic study that performance was superior when radiologists were presented with the prior mammograms for every case, rather than just the cases for which they deemed the prior mammograms necessary.

5.2 Aims

1. To measure any changes in cancer detection performance with or without prior mammograms, and present any changes in terms which will influence clinicians' practice such as recall rate
2. To determine the impact on cancer detection performance of digitising prior mammograms in preference to displaying them in film format during the transition to digital mammography.
3. To establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2.

5.3 Choice of Methods

The cancer detection task is essentially one of determining whether there is a 'signal' present or not, with the signal being evidence of cancer. Therefore signal detection theory is the most appropriate framework for measurement of cancer detection performance, with receiver operating characteristic (ROC) analysis the tool. In this paradigm the participant rates the probability of malignancy for each case along a linear scale. Each possible threshold for recall is applied and the number of false positive, false negative, true positive and true negative decisions is calculated, giving a measure for sensitivity and specificity. A true positive case is a correctly recalled cancer, a false positive case is a normal case incorrectly recalled (type I error), a true negative case is a normal case correctly not recalled, and a false negative case is a

cancerous case incorrectly not recalled (type II error). When sensitivity is plotted against 1-specificity for each possible threshold an ROC curve is produced, as shown in figure 5.1. The larger the area under an ROC curve the better the performance, as it corresponds to the fraction of correct choices in a two alternative fixed choice experiment. ROCFIT software can be used to analyse single reader single case ROC data (Dorfman and Alf, 1969)

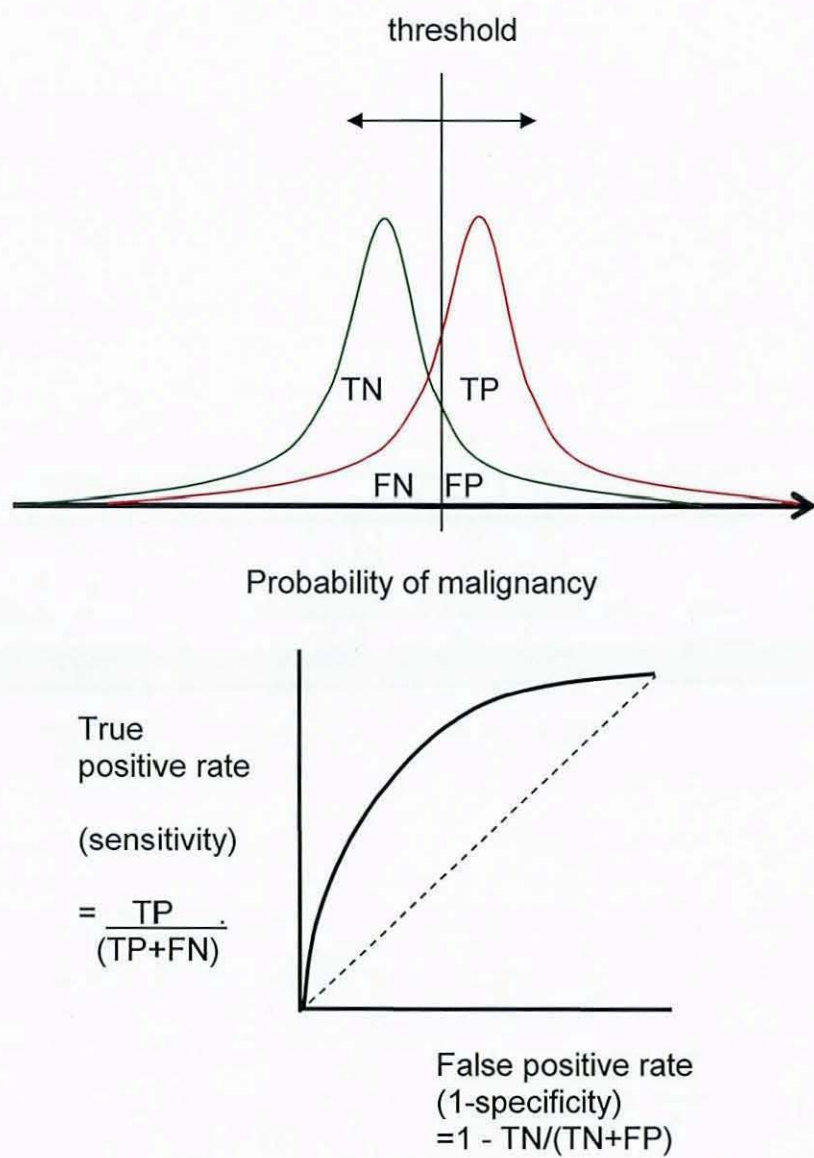


Figure 5.1 – The formation of ROC curves. At each possible threshold value the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are calculated and plotted in terms of sensitivity and 1-specificity on the ROC curve. The larger the area under the ROC curve the better the performance. The dotted diagonal line represents chance performance.

There are two disadvantages of ROC analysis when applied to breast cancer screening. Firstly it can only be used to measure one participant at a time, which means results cannot be generalised to a population of readers. Secondly the participant does not have to correctly identify the location of the cancerous lesion, they only have to report whether the case is cancerous. Therefore if the participant incorrectly identifies normal tissue as abnormal, and fails to identify some abnormal tissue elsewhere on the same image the answer is considered correct. When this occurs in breast screening the biopsy is taken from the area of normal tissue and so the cancer is missed

To enable generalisation to a population of readers Multiple Reader Multiple Case (MRMC) ROC analysis was developed (Dorfman *et al.*, 1992). This allows performance metrics to be calculated for a combination of several readers using Jackknife methods to create pseudovalues. However there is still no requirement for the participant to give correct localisation information.

Several methods are available which require lesion localisation in addition to a higher than threshold malignancy rating for true positive results. These include the region of interest (ROI) approach (Obuchowski *et al.*, 2000), Localised Receiver operating Characteristic (LROC, Swenson, 1996), and Free-Response Receiver Operating Characteristic (FROC, Egan *et al.*, 1961) These were all considered for use in this experiment, with emphasis on achieving accurate modelling of breast screening practice. Taylor-Phillips *et al.* (2009) describe how the level of use of the prior mammograms in screening is affected by their display medium, and therefore modelling

screening conditions as accurately as possible is a pre-requisite of the experimental design.

The Region of Interest (ROI) approach introduces some aspects of localisation into the modelling of the mammographic screening task by dividing the mammogram into five regions of interest: upper outer, upper inner; lower outer; lower inner; and retro-areolar. The mammography reader rates a probability of malignancy for each region instead of each case (Obuchowski *et al.*, 2000). ROC analysis is then performed on classification performance for every region of interest instead of for every patient. The data are clustered, with a cluster for each case to overcome the problem of independence of cases. ROI methodology has a firm mathematical foundation, and the smaller areas of interest do provide additional location information, but it is not a sufficient approximation to the screening task for the purposes of this experiment. Firstly there is no marking the location of the abnormality, and therefore it is essentially a classification task with smaller areas to classify. Secondly the division of the mammogram into five areas will necessarily change the mammography reader's search strategy, and therefore heavily influence the data produced.

Localised Receiver Operating Characteristic (LROC, Swensson *et al.*, 1996) methods are in practice very similar to ROC, using case based analysis with an additional requirement for correct localisation. In LROC experiments the participant is asked to both classify each case according to probability of malignancy, and to mark the location of the most suspicious lesion on the

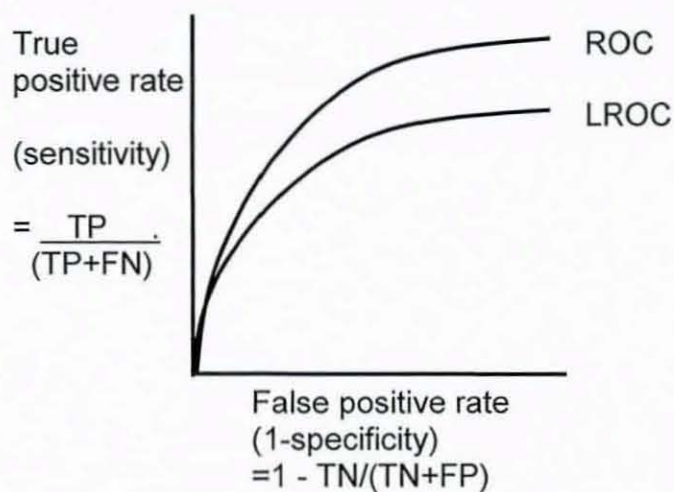
image. Therefore to achieve a true positive in the LROC paradigm requires both correct classification and localisation. The method of maximum likelihood analysis is used in a similar manner to ROC analysis, and therefore the area under the LROC curve will always be equal to or smaller than the area under the ROC curve. One of the assumptions of the method detailed by Swensson (1996, pg 1713) is that "the perceived appearance of the most suspicious actual target does not affect that of the most suspicious nontarget location (and vice versa)" i.e. that the perception of a non-lesion location in an image is not affected by the presence or absence of a lesion in that same image. This may not always hold true due to the phenomenon of 'satisfaction of search' (Berbaum *et al.*, 1990). LROC provides a better approximation to screening practice than ROC methods, however only one lesion can be identified per case. In breast screening some cases have several malignant lesions, all of which need to be correctly identified.

The free response receiver operating characteristic (FROC) paradigm most closely mirrors breast screening practice. In a FROC experiment the mammography reader only responds when they have located an abnormality, they mark the location and the probability of malignancy and then search for the next abnormality. There is the freedom to mark several abnormalities on the same mammogram, or no abnormalities at all, as is the practice in breast screening. The methodology was first introduced by Egan *et al.*, (1961) but failed to gain wide acceptance due to difficulties in analysing the collected data. FROC curves can be plotted in a similar way to ROC curves with λ , the mean number of false positive responses per image on the ordinate, and ν ,

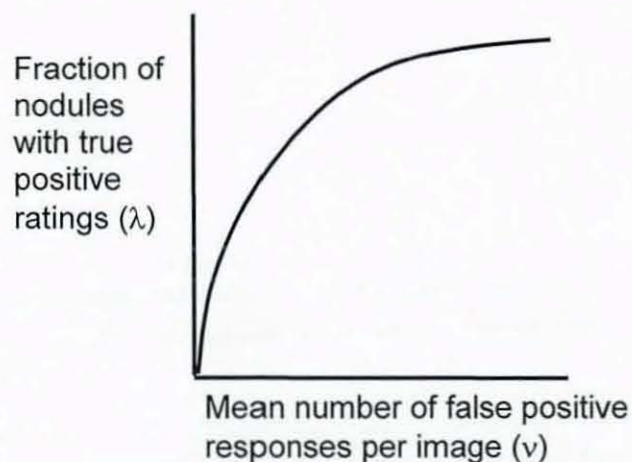
the fraction of nodules with true positive ratings on the abscissa, see figure 5.2.

FROC analysis using FROCFIT software (Chakraborty, 1989) uses the method of maximum likelihood analysis. There are two assumptions which this data analysis relies on which may be violated, (Chakraborty and Winter, 1990). Firstly independence of cases, i.e. that the true positive and false positive responses are independent random events. This is violated when there is more than one lesion per patient. Secondly the Poisson assumption, that for a given threshold the probability of generating a given number of false positive responses per image is given by the Poisson distribution. The Poisson assumption is necessary because FROC data is lesion based rather than case based, and whilst there are a finite number of possible false positive cases, there are an infinite number of possible false positive lesions, and so a conversion to case based data is required to give a measure of specificity. Alternative Free Response Receiver Operating Characteristic (AFROC) (Chakraborty and Winter 1990) analysis was developed to remove the need for the Poisson assumption, by counting false positive images rather than false positive lesions. However the assumption of independence of cases still applies, and this is not met when there is more than one lesion per case. The AFROC curve plots the fraction of nodules with true positive ratings on the ordinate, and probability that a case has at least one false positive finding on the abscissa, as shown in figure 5.2.

ROC
and
LROC



FROC



AFROC

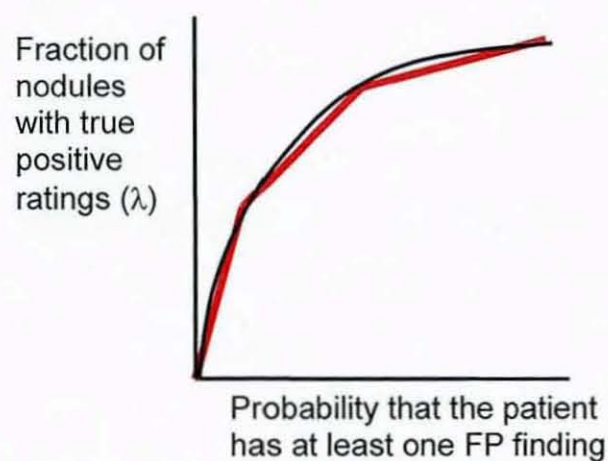


Figure 5.2 – Comparison of ROC and LROC, FROC and AFROC curves. The red line denotes the trapezoidal approximation to the AFROC curve made by JAFROC analysis.

Jackknife free response receiver operating characteristic (JAFROC) provides a FROC analysis method that meets the requirement of independence of cases, and does not rely on the Poisson assumption. JAFROC uses similar principles to MRMC ROC analysis. The pseudovalues for JAFROC are figures of merit (obtained using non-parametric Wilcoxon-Mann-Whitney U statistic, rather than area under curve values).

$$\theta = \frac{1}{N_N N_L} \sum_{i=1}^{N_N} \sum_{j=1}^{N_L} \psi(X_i, Y_{jk})$$

$$\begin{aligned} \psi(X_i, Y_{jk}) &= 1 \quad \text{if } Y > X \\ &= 0.5 \quad \text{if } Y = X \\ &= 0 \quad \text{if } Y < X \end{aligned}$$

Where θ is the figure of merit, N_N is the number of normal images, N_L is the number of lesions, X_i is the rating of the highest rated noise site on normal image i , Y_j is the rating assigned to the j th lesion. The principle of jackknifing is used by removing one case at a time and calculating the figure of merit, then removing a different case until a matrix of pseudo-values is created for each experimental condition. Analysis of variance is then conducted on these matrices to determine whether there is any effect of treatment. The jackknifing takes a whole case out at a time rather than individual lesions and therefore JAFROC can meet the independence of cases assumption for cases with multiple lesions in a way that FROC and AFROC cannot. However, JAFROC is a non-parametric method, in contrast to ROC and LROC which are

parametric methods. As a result of this the figure of merit is the non-parametric trapezoidal approximation to the area under the AFROC curve, rather than a smooth curve, see figure 5.2. The trapezoidal nature of the approximation may provide inconsistencies, as for the same curve if the location of the trapezoidal points are different then the area under the curve will differ, see figure 5.3. Therefore in a JAFROC experiment it is important to ensure a smooth spread of trapezoidal points across the ROC curve, which can be achieved by participants using a wide spread of ratings categories.

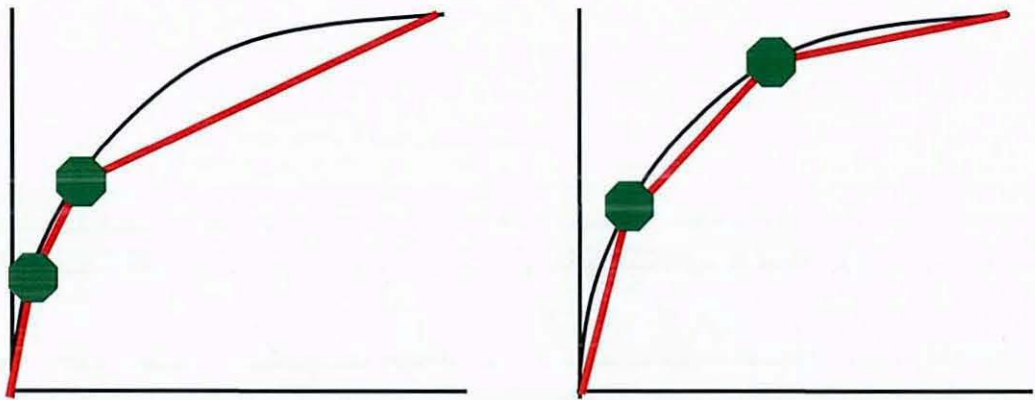


Figure 5.3 – Two identical AFROC curves with different trapezoidal approximations. The less accurate approximation on the left would result from participants not using the lower malignancy ratings.

The disadvantage of lesion rather than case based analysis is that if one case contains more than one lesion that case will be given greater weight than others with only one lesion. There is a version of JAFROC which contains case weighting to ameliorate this effect, however this version did not pass null hypothesis testing (i.e. not having a rejection rate the same as the significance level of the test) (Chakraborty, 2006). Therefore the most recent

version of JAFROC has no weighting factor and so the problem of one case having undue weight remains.

JAFROC can be used to analyse data from an experimental setup which mirrors breast screening practice to a greater extent than either ROC or LROC analyses will permit. Its limitations are twofold. Firstly, it is a non-parametric method which provides a trapezoidal approximation to the AFROC curve, and therefore a wide range of reported probability of malignancy ratings are required for accurate results. Secondly, there is no weighting system for cases with multiple lesions so these cases may have undue influence on overall results. Therefore, as accurate modelling of screening behaviour is one of the primary considerations in the design of this experiment, JAFROC analysis will be used with consideration of the two limitations listed above.

5.4 Method

Full NHS ethical approval was granted from the South East Research ethics Committee reference number 08/H1102/35. The participant information sheet and informed consent form are in appendix 6.

5.4.1 Ambient Lighting conditions

Ambient lighting levels have been linked to performance reading x-rays (McEntee *et al.*, 2006). Therefore this must be kept constant across

conditions. The European guidelines state that ambient light should be less than 10 lux for primary display devices, however this is qualified by the statement that "the maximum ambient light actually depends on the reflection characteristics and minimum luminance of the monitor, but for reasons of simplicity this is ignored" (Health and Consumer Protection Directorate, 2006, pg 134). Therefore, the appropriate level of ambient lighting was determined using recommendations from the American Association of Physicists in Medicine Task Group 18 (Badano *et al.*, 2005), as these guidelines do consider reflection characteristics. The coefficient of specular reflection was determined using the ratio of luminance reflected from the screen (at an angle of 15° to the perpendicular) to direct luminance, using an average of ten measurements. The coefficient of diffuse reflection was determined as the ratio of luminance to illuminance using the setup shown in figure 5.4. Again an average of ten measurements were taken. The luminance meter used was the Minolta LS 1110 (Kodak, USA). Maximum and minimum screen luminance were measured using ten measurements from each of the TG18-LN12-01 and TG18-LN12-18 test patterns respectively. Recommended maximum ambient lighting level for specular reflection was calculated by using the coefficient of specular reflection alongside the maximum and minimum luminance values in the table provided by task group 18 (Badano *et al.*, 2005, pg 76). Recommended maximum ambient lighting for diffuse reflection was calculated as the product of 0.25 and the minimum luminance, divided by the coefficient of diffuse reflection. These recommended maximum ambient lighting levels were compared to actual ambient lighting levels with and without dimmed lights, and with and without the film multi-viewer switched on.

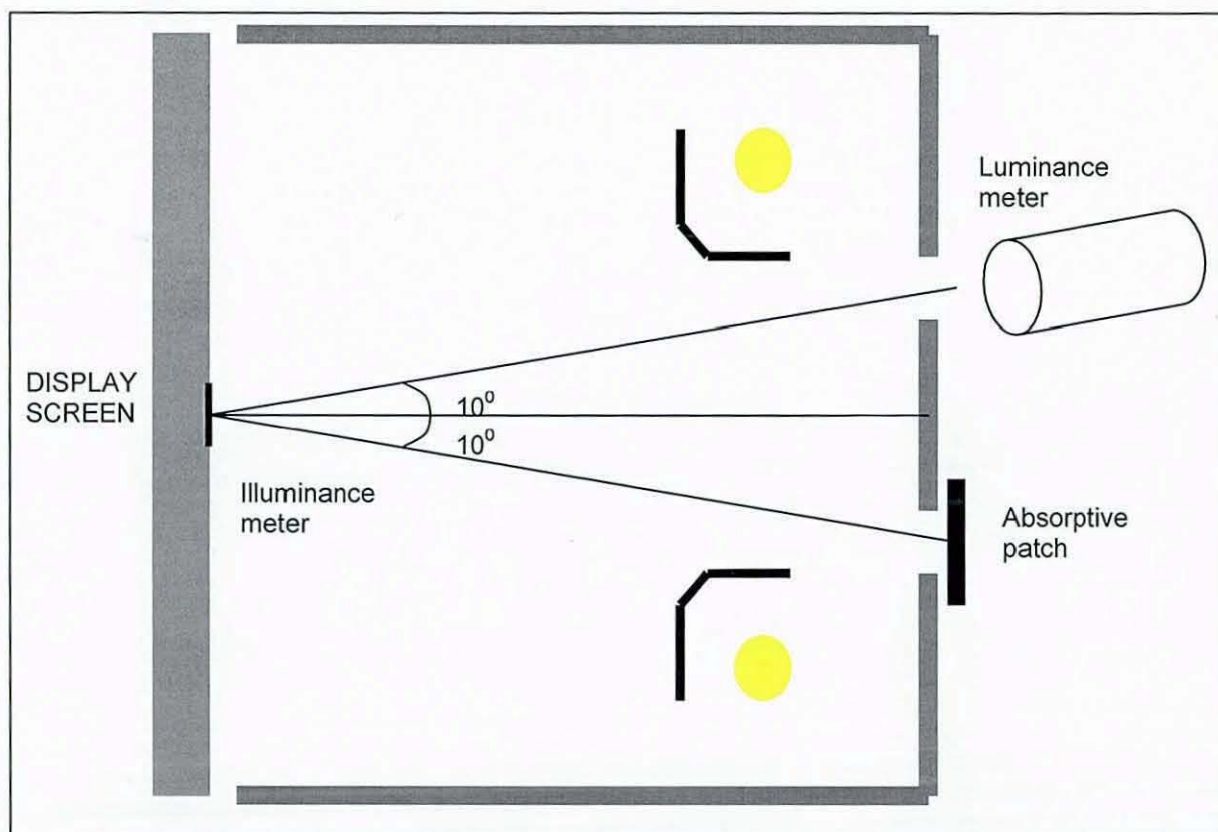


Figure 5.4 – Experimental apparatus for measurement of coefficient of diffuse reflection.

5.4.2 Case Selection

A set of 160 cases was assembled from the UK Breast Screening Programme, consisting of 66 malignant and 94 benign/normal cases. Of the benign/normal cases 58 were recalled for further tests in the UK Breast Screening Programme (36 biopsy and 22 mammography/ultrasound) and 36 were not recalled (30 of which had been discharged after arbitration by a third reader), see table 5.1 for a further breakdown of case type. All incident round cancers detected at digital screening between March 2005 and June 2007 as part of the Warwickshire, Solihull, and Coventry Breast Screening Programme were considered for inclusion in the study (79 in total) benign/normal cases

were selected at random from a database of difficult benign/normal cases from the same time period. This database included all cases which went to third reader arbitration, or were recalled and subsequently found to be normal or benign. Classification of cases as normal or abnormal was carried out by an expert radiologist with 20 years experience in breast screening. Normal cases were defined as such by screening results, the results of any follow up tests (mammography, ultrasound, and biopsy) for those cases which were recalled after screening, and two years after screening free from the development of interval cancers (34% of which had a subsequent negative screening round). All abnormal cases were proven by biopsy. The same expert radiologist marked the outline of any lesions on a paper print out of the mammograms, and advised whether each case was appropriate for inclusion in the experiment. Some 19 cases were not appropriate for inclusion due to either: being mammographically occult; only having single view prior mammograms; technical problems; or being normal cases subsequently presenting with an interval cancer. This left 160 cases: 64 which contained one malignant lesion; two which contained two malignant lesions; and 94 which were normal or benign.

Table 5.1 – Breakdown of the 160 cases used in the experiment by case type, difficulty rating as assigned by an expert radiologist, and suspicious pattern type. Calc. is an abbreviation of calcifications.

Case Type (Outcome of assessment in the NHS Breast Screening Programme)	Number of Cases													
	Total	Difficulty Rating					Suspicious Pattern Type							
		1	2	3	4	5	Well Defined Mass	Ill Defined Mass	Spiculate Mass	Architectural Distortion	Asymmetry	Suspicious Calc.	Diffuse Suspicious Calc.	Benign Calc.
Normal (not recalled by either reader)	6	0	3	2	1	0	1	0	0	0	1	0	0	0
'Difficult Normal' (went to arbitration but not recalled)	30	0	2	13	15	0	7	3	5	6	7	0	0	3
'Very Difficult Normal' (Recalled but not biopsied)	22	0	0	4	16	2	0	4	8	6	6	0	0	0
Benign (Recalled and biopsy was negative)	36	2	4	17	13	0	5	6	2	3	4	11	0	6
Malignant (recalled and biopsy was positive)	66	12	23	19	11	1	6	13	27	1	4	13	2	2
Total	160	14	32	55	56	3	19	26	42	16	22	24	2	11

5.4.3 Participants and methods

Eight participants from one breast screening centre in the UK took part in the study, four radiologists and four radiography advanced practitioners. All were qualified to interpret mammograms in the UK NHS Breast Screening Programme, with average experience reading mammograms of 7 years. The same 160 cases were read three times on a digital workstation with film; digitised; and without prior mammograms. At least one month elapsed between participants re-reading the same cases. Each participant read every session on the same day of the week and at the same time of day to reduce confounding variables. Each session involved reading no more than 54 cases to reduce the effects of fatigue. For each case the participant was asked to mark the location of any lesions with a cross on 6cm x 5.5cm paper print outs of the mammograms, rate the probability of malignancy of each lesion from 0 to 100% on a linear scale, and report whether they would recall the case for further tests if they encountered it in the Breast Screening Programme. An example of the data recording sheets is in appendix 8. There were no restraints on how many lesions the participants could mark. Jackknife Free Response Receiver Operating Characteristic (Chakraborty 2006, Chakraborty, 2008a) was used for the analysis of performance. Lesion location was considered correct if the centre of the cross was within 2mm of the lesion outline as defined by the expert radiologist. To encourage a spread of malignancy ratings to ensure a good JAFROC trapezoidal approximation to the curve participants were instructed that "it is very important that you err on

the side of marking too many rather than too few lesions. If there are any indications of possible malignancy then please mark the lesion". The full participant instructions are in appendix 7.

Counterbalancing was applied between the film and digitised prior mammograms conditions. The cases were matched into pairs by similarity in terms of case type, lesion type, and case difficulty. These pairs of cases were split into three parts A, B, and C by random selection stratified by case type, so that each reading session lasted no longer than 1 hour to reduce the effects of fatigue. These were split again by random selection into parts A1 and A2, B1 and B2, and C1 and C2. The effects for which counterbalancing was considered were: fatigue as each session progressed, participant type (radiologist or radiography advanced practitioner), and case order. The priority was to enable a fair comparison between performance using digital or film prior mammograms. Therefore in each session half of the cases read were with digitised and half were with film priors, so if a participant had a particularly good or bad session this would impact both conditions. To ameliorate the effects of fatigue within a session if a participant read with digitised prior mammograms first for the first three sessions, they read with film prior mammograms first for the second three sessions. To counterbalance for participant type and ability, the two most experienced radiologists were paired together, so if one read with the digitised prior mammograms first in the first sessions, the other read with the film prior mammograms first in the first three sessions. Counterbalancing for case order was given the lowest priority, as it was considered less likely to be a confounding variable than participant

type and experience, and fatigue. The digital mammography workstation did not have the functionality to present the cases in each set in a random order each time, and therefore the order for each set was randomly selected once, and the cases presented in that order. The most experienced radiologist read the cases with part 1 first, and was matched with the third most senior radiologist who read part 2 first, see table 5.2. The sessions without any prior mammograms were completed after those with the prior mammograms. However, number of cases per session and case sets were maintained. To measure whether fatigue within each session affected the results performance was compared using number of true positive and false positive cases between the first 27 cases and the second 27 cases of each session. Additionally to test whether either participants tired of the experiment, or started learning the cases performance was compared when the cases were read for the first and second times.

Table 5.2 - Counterbalancing applied between the conditions of film or digitised prior mammograms. Priority was given for counterbalancing by participant type and experience, and whether cases were read with film or digitised prior mammograms first within each session (to ameliorate the effects of fatigue)

Participant	Type	Order of digitised or film priors within session (fatigue)	Part of case set	Session					
				1	2	3	4	5	6
1	Radiologist	Digitised then film	1 then 2	D(A1) H(A2)	D(B1) H(B2)	D(C1) H(C2)	H(A1) D(A2)	H(B1) D(B2)	H(C1) D(C2)
2	Radiologist	Film then digitised	1 then 2	H(A1) D(A2)	H(B1) D(B2)	H(C1) D(C2)	D(A1) H(A2)	D(B1) H(B2)	D(C1) H(C2)
3	Radiologist	Digitised then film	2 then 1	D(A2) H(A1)	D(B2) H(B1)	D(C2) H(C1)	H(A2) D(A1)	H(B2) D(B1)	H(C2) D(C1)
4	Radiologist	Film then digitised	2 then 1	H(A2) D(A1)	H(B2) D(B1)	H(C2) D(C1)	D(A2) H(A1)	D(B2) H(B1)	D(C2) H(C1)
5	Radiography Advanced Practitioner	Digitised then film	1 then 2	D(A1) H(A2)	D(B1) H(B2)	D(C1) H(C2)	H(A1) D(A2)	H(B1) D(B2)	H(C1) D(C2)
6	Radiography Advanced Practitioner	Film then digitised	1 then 2	H(A1) D(A2)	H(B1) D(B2)	H(C1) D(C2)	D(A1) H(A2)	D(B1) H(B2)	D(C1) H(C2)
7	Radiography Advanced Practitioner	Digitised then film	2 then 1	D(A2) H(A1)	D(B2) H(B1)	D(C2) H(C1)	H(A2) D(A1)	H(B2) D(B1)	H(C2) D(C1)
8	Radiography Advanced Practitioner	Film then digitised	2 then 1	H(A2) D(A1)	H(B2) D(B1)	H(C2) D(C1)	D(A2) H(A1)	D(B2) H(B1)	D(C2) H(C1)

To calculate whether participant type affected the JAFROC performance results a mixed model analysis of variance of the JAFROC figure of merit was conducted with both workstation type (film, digital, or hybrid) and participant type (radiologist or radiography advanced practitioner) as independent variables.

For each case in each condition the participant determined whether they would recall the case for further tests in the UK Breast Screening Programme. This was converted into number of false positive cases, and false negative cases. A within subjects analysis of variance for the number of false positive cases, number of false positive lesions per case, and the number of false negative cases was conducted, and where appropriate post hoc paired comparisons with a Sidak correction. The assumption of sphericity was tested using Mauchly's test.

The normal cases used in this study can be categorized into three groups: cases which were referred to arbitration but were not recalled; cases which were recalled and had a benign biopsy; and cases which were recalled but did not have a biopsy. An analysis of variance was carried out to determine whether this case classification affected the probability of a false positive recall in the study. This is of interest because the proportions of these three types of case in the study did not exactly mirror those in the UK Breast Screening Programme, as the random sampling was not stratified by case type. The number of false positive cases found in the study was then recalculated using a weighting to convert from the experimental proportions to

the proportions encountered in the UK Breast Screening Programme. Results are presented for both weighted and unweighted false positive rates.

The results for false positive rates and recall rates are from single readers in the experiment, but in the NHS Breast Screening Programme double reading with arbitration is used. Therefore an additional analysis was carried out to convert the results from single reader to double reader with arbitration. For each case in each modality all possible combinations of pairs of readers were made, and the outcome from each pairing classified as 'return to screen' if both readers indicated they would not recall the case, 'recall' if both readers indicated they would recall the case, and 'arbitration' if one reader would recall the case and the other reader would not. Therefore for each case in each modality instead of eight recall decisions there were 36. For each case which was determined for arbitration each possible selection of a third reader was made (six possible readers), and their decision of whether to recall was final. This models the third reader arbitration used in the UK Breast Screening Programme. An additional analysis was conducted with the initial recall, return to screening, or arbitration decision made without the prior mammograms, and the arbitration decision made with the film prior mammograms.

An option that is being implemented in one breast screening centre already in the UK is to provide the prior mammograms in screening bags at the digital workstation, for readers to hang when they wish. Another version of this strategy is to allow the readers to fetch the prior mammograms from storage

in the filing room when they wish. For each case when reading without the prior mammograms participants were asked to state whether "If you were reading this case as part of the NHSBSP would you hang the prior mammograms if they were available at the desk? Fetch and hand the prior mammograms if they were filed in an adjacent room?" Participants were asked to circle yes or no to both of these questions for every case.

All participants were familiar with the equipment, but unfamiliar with reporting on a percentage confidence scale. Before starting the experiment each participant was given a set of three practice cases to report, and an opportunity to ask questions about any aspect of the study.

5.4.4 Cost Analysis

A cost analysis was conducted for this breast screening centre to compare the projected cost of digitising prior mammograms, film display of prior mammograms, and no display of prior mammograms. Calculations for cost of equipment, staff time, and any increases in recall rate were made for each implementation, per 10,000 women screened at this centre. Only 82% of women screened in the UK have previous mammograms, the remainder are prevalent screens (The Health and Social Care Information Centre, 2009). This was factored into the appropriate calculations.

Digitisation equipment was costed to last for the three year transition to digital mammography in a UK screening centre. The purchase and maintenance

contract prices for July 2009 were used. The cost of purchase of film multi-viewers was not included as these would already be in place. The maintenance cost for multi-viewers per 10,000 women screened is taken from Legood and Gray (2004) and updated using the retail price index to 2009 figures. To cost for increases in recall rates the monetary value of extra mammography, ultrasound and biopsy equipment was calculated. As the equipment in the study hospital is used for both assessment of screening recalls, and symptomatic and specialist screening, the proportion of time the equipment was used for assessment was calculated, and only that percentage of the total equipment costs was used. The cost of consumables for biopsy were included, using the mean number of FNA, core, and Vacora needles used per 10,000 women for assessment, multiplied by the cost per needle.

The cost of staff time was calculated using a combination of measurements of time taken, and reports from other research papers. One novice member of staff was trained in digitising prior mammograms, and hanging film prior mammograms for a period of one week. The time this member of staff took to digitise 30 cases, and to hang 30 cases of film prior mammograms was measured. Time taken to retrieve and replace records was taken from Legood and Gray (2004). Hamermesh (1990) reports that 9% of working time is spent in breaks and other time on the job not spent working, and therefore each time estimate was increased by 9% to account for this. For digitising, retrieving and replacing records, and hanging film prior mammograms the median pay for a NHS band 1 or 2 administrative staff member was used.

Staffing for extra recalls was calculated using the staff complement at the study hospital per assessment clinic, multiplied by the number of extra assessment clinics necessary. The number of extra assessment clinics was calculated by multiplying the current number of clinics by the percentage increase in recalls. Staff per clinic includes one consultant radiologist, one advanced practitioner, three radiographers, one breast care nurse, a receptionist and a typist. Pay was calculated using 2009 pay scales (The NHS Staff Council, 2009) and working hours and salary oncosts from Curtis and Netten (2006).

5.4.5 Equipment

The digital mammograms were obtained from the MicroDose Mammography system (Sectra, Sweden) displayed using Sectra mammography PACS on twin five megapixel LCD screens (EIZO, Japan). The previous mammograms were acquired using a Mammomat 3000 Nova (Siemens, Germany), with Kodak MIN-R2000 mammography film, developed using a Kodak X-OMAT Multiloader 7000 (Carestream Health, Toronto, Canada). The film prior mammograms were digitised using an Array 2905 Laser Film Digitiser, (Array Corporation, New Hampshire, USA), set to 75µm, standard resolution, bit depth 12. Mammographic film display was on a Mammolux XL multi-viewer (Planilux, Germany), which was positioned adjacent and perpendicular to the digital workstation. Reading conditions were identical for both experimental conditions with the room darkened. Participants had access to the multi-viewer for all experimental conditions, which they could dim or turn off as necessary.

5.5 Results

5.5.1 Ambient lighting levels

The maximum acceptable ambient lighting was 31 lux when considering diffuse reflection and 14 lux when considering specular reflection. The actual ambient lighting levels were 8.5 lux ($\sigma=.3$) with the dimmed room lighting, 20 lux ($\sigma=11$) with the multi-viewer with one set (of 4) prior mammograms illuminated, 41 lux ($\sigma=9$) with the multi-viewer two sets of prior mammograms illuminated, and 32 lux ($\sigma=12$) with one set of prior mammograms illuminated and the dimmed lights on. The standard deviation of measurements is high when the multi-viewer is turned on as the characteristics of the mammograms illuminated affects the light levels from the multi-viewer. Light from the multi-viewer exceeded the recommended limits for specular reflection when only displaying the prior mammograms from one woman, however as the light source is perpendicular to the LCD screen specular reflection from this angle may not be an issue. As a result of this analysis, and in consultation with the participants the room lighting was switched off for every condition.

5.5.2 Performance

JAFROC1 is the most recent method for analyzing free-response data which has advantages when the number of normal cases is relatively small. However, there were almost four times the number of false positive marks on normal cases than on abnormal cases: 0.669 per normal case vs. 0.171 per abnormal case, averaged over all readers and modalities. Since JAFROC1 analysis was validated (Chakraborty, 2008b) assuming equal probability of false positive marks on normal and abnormal cases we were advised (Chakraborty, 2009, personal communication) to use JAFROC analysis (Chakraborty 2008a) instead. The JAFROC figure of merit ignores false positive marks on abnormal cases and is therefore insensitive to this asymmetry.

The reason for the lower false positive rate on abnormal images is not clear, but has been reported by others (Gur and Rockette, 2008) and is consistent with a satisfaction of search effect where having found a lesion the participants are less likely to report more lesions. Tuddenham (1962) originally introduced this concept, and hypothesised that it occurred because having found one lesion the radiologists were 'satisfied' and therefore did not search further. Berbaum et al. (1990) first demonstrated this phenomenon experimentally by measuring performance in interpreting chest x-rays both with and without synthesised pulmonary nodules added to the images. The addition of the nodules reduced performance in detecting a range of other more subtle abnormalities (native abnormalities) on the same image.

Berbaum et al. (1991) extended this study to measure search times before and after finding the first abnormality. Again they found a 'substantial decrement in detection of lesions when more than one abnormality was present' ($p=.01$), but search did not cease after finding the first abnormality, and the time taken to detect an abnormality was not affected by the presence of another abnormality. This suggests it is more likely to be a failure of recognition or decision making rather than the error of search suggested by Tuddenham (1962). Samuel et al. (1995) conducted a similar study and whilst in contrast they found the detection of native abnormalities to be unaffected by the introduction of nodules, the detection of nodules was decreased by the presence of native abnormalities (37%, $n=10$ more missed nodules, $p< .005$). This study also tracked the eye movements of participants and measured the length of fixations on the missed lesions. The missed lesions due to 'satisfaction of search' were classified as recognition errors (fixated but for less than 1000msec).

The most common application of the theory of satisfaction of search is in understanding why subtle lesions are missed more frequently when there are obvious lesions in the same image. However in the study presented in this thesis the phenomenon was the likely cause of fewer false positive marks on abnormal than normal images. As a result of this the latest version of JAFROC was not appropriate for use, and indeed JAFROC software is in the process of being upgraded to account for the satisfaction of search effect as a direct result of the results presented here (Dev Chakraborty, personal communication).

Cancer detection performance as measured by JAFROC figure of merit differed between the conditions ($F(2,65.2)=5.6, p=.006$). Performance was superior using both digitised prior mammograms (95%CI = .01-.06) and film prior mammograms (95%CI = .009-.05) than using no prior mammograms. There was no difference in performance between using film or digitised prior mammograms. This is illustrated by a free response receiver operating characteristic curve as shown in figure 5.5. JAFROC analysis was repeated with cases 144 and 158 removed as they both contain two lesions, and therefore could have disproportionate influence on the results. This made no significant difference to the results ($F(2,116)=6.2, p=.003$).

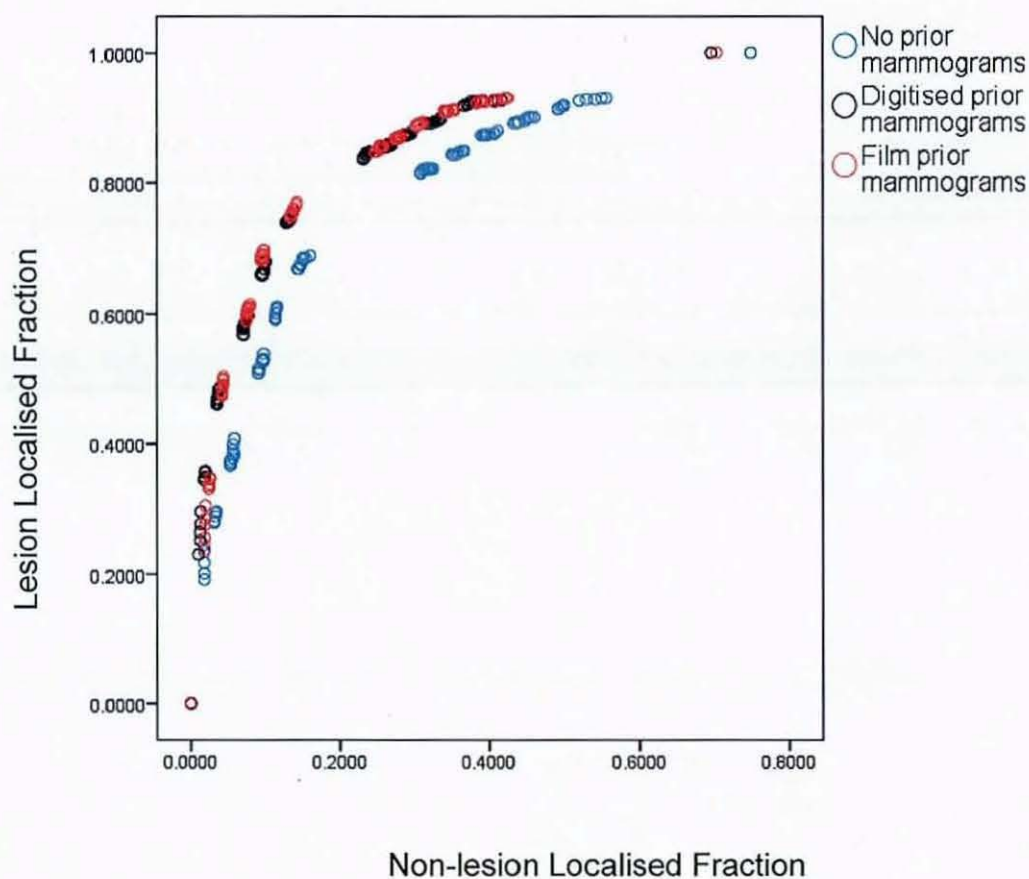


Figure 5.5 – Free Response Receiver Operating Characteristic (FROC) curves for the conditions: no prior mammograms; digitised prior mammograms; and film prior mammograms. Lesion localised fraction is the proportion of lesions correctly localised at a threshold, and non-lesion localised fraction is the number of non-lesions localised per image at that threshold.

The analysis of variance of the JAFROC figures of merit showed no main effect of participant type, but a trend ($F(2,6)=3.7$, $p=.09$) towards an interaction between participant type (radiologist or radiography advanced practitioner) and prior mammogram display type (film, digitised, or no display), this trend is illustrated in figure 5.6

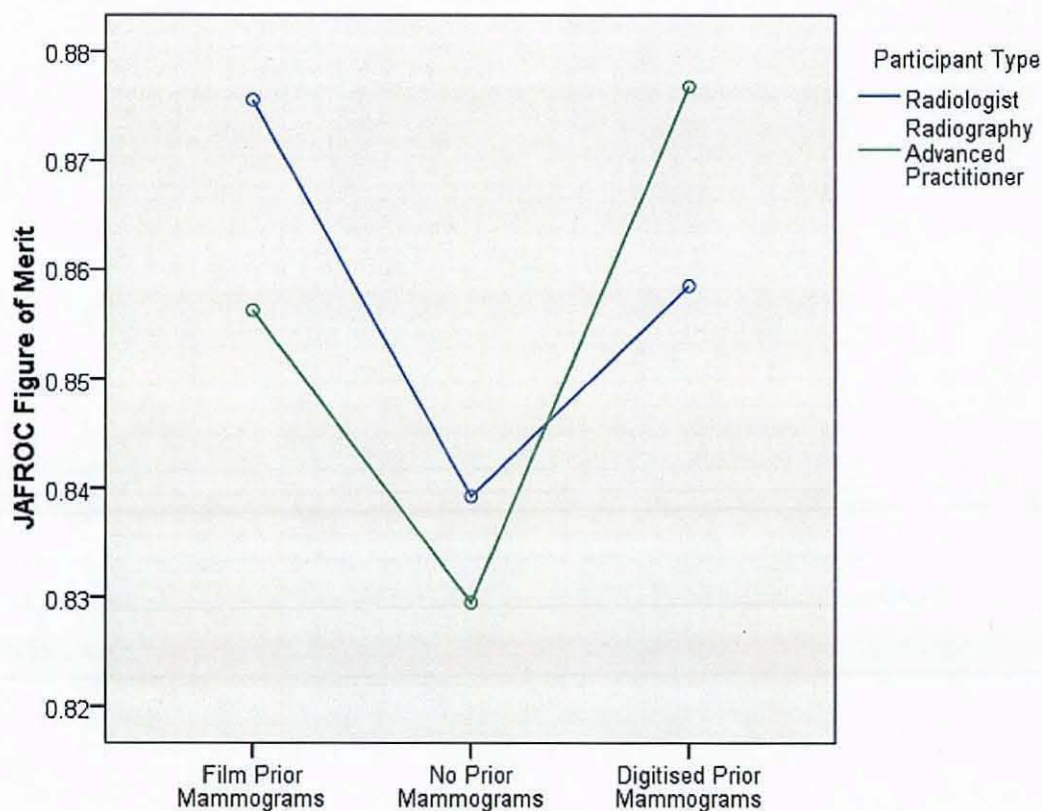


Figure 5.6 - Plot of the interaction between presentation of the prior mammograms and participant type for JAFROC Figure of Merit. There is a trend towards an interaction ($p=.09$).

There were no effects of fatigue found. The number of correct recalls was 470 in the first half of the session and 482 in the second half, the number of incorrect recalls was 977 in the first half of the session and 975 in the second half of the session. Neither of these differences were statistically significant. Performance did not differ when the participants read the cases for the first

time or the second time, so there were no indications of either learning from previous experience with the same cases or increasing boredom with the experiment being a detriment to performance.

Mauchly's test showed that the assumption of sphericity was not violated for any of the analyses of variance. The number of false positive cases (i.e. normal cases recalled) differed by case type ($F(2,14)=17.0, p<.0005$), with post hoc tests showing that those cases which had been recalled in the Breast Screening Programme were more likely to be recalled in the experiment ($p=.01-.001$). However, whether the recalled cases had a biopsy or alternative tests in the UK Breast Screening Programme did not affect the probability of them being recalled in the experiment. The proportion of 'difficult' normal cases in the study in comparison to the screening programme is detailed in table 5.3. A weighting system as described above was applied to the findings to calculate the false positive rates based on the proportions from the true screening situation and therefore to enable application of the findings to recall rates in the Breast Screening Programme. The weighting did not make a significant difference to any of the results

Table 5.3 – The proportion of three types of normal cases present in the Breast Screening Programme in comparison to the study.

	Outcome of screening		
	Recalled cases which had a biopsy	Recalled cases which did not have a biopsy	Cases which were arbitrated and not recalled
Proportion of each type of normal case in the UK Breast Screening Programme	14%	50%	36%
Proportion of each type of normal case in the study	42%	24%	34%

The number of false positive cases (normal cases recalled) differed between the three conditions ($F(2,14)=7.3$, $p=.007$), see table 5.4. Post hoc tests showed 28% more false positive cases when using no priors than using digitised priors ($p<.05$), this difference decreased to 26% when the weighting for case type was applied ($p=.04$). There was also a trend towards more false positive cases when using no priors than using film priors (24% increase, $p=.09$), and the same trend when the weighting was applied (26% increase, $p=.07$) but these results were not significant.

The number of false positive lesions (i.e. lesions marked as potentially malignant which were normal/benign) differed between the three conditions ($F(2,14)=8.7$, $p=.004$) and post hoc tests showed a 36% increase when not using prior mammograms in comparison to using digitised prior mammograms ($p=.04$), which decreased to 30% when the case weighting was applied ($p=.02$). There was also trend towards an increase in false positive lesions when not using prior mammograms in comparison to using film prior mammograms (30%, $p=.07$, weighted for case type 29%, $p=.1$), but these were not significant. The number of false negative lesions, and false negative cases (missed cancers) did not differ between using film, digitised or no prior mammograms.

Table 5.4 – Four measures of performance using film prior mammograms, digitised prior mammograms and no prior mammograms.

	Film priors	Digitised Priors	No priors	Significance level
JAFROC figure of merit	0.86	0.87	0.83	.006
Number of false positive lesions per case	0.38	0.36	0.49	.004
Proportion of normal cases recalled (false positive cases)	46%	44%	57%	.01
Proportion of normal cases recalled weighted for case type	40%	40%	50%	
Proportion of abnormal lesions not recalled	8%	9%	9%	

Converting the results for the normal cases from single reader to double reader with arbitration did not change them significantly, see table 5.5. The model of double reading with arbitration projected a 30% increase in the number of false positive cases when not using prior mammograms in comparison to using film priors, and a 37% increase in comparison to using digitised priors. The model of using the film prior mammograms for the arbitration process but not for screening projected an increase in the number of false positive recalls of 20% in comparison to using film prior mammograms throughout, and 27% in comparison to using digitised prior mammograms throughout.

Table 5.5 – Results of the model converting single reader results to double reader with arbitration

		Film priors	Digitised Priors	No priors
Proportion of normal cases sent to arbitration		39%	37%	39%
Proportion of normal cases recalled	Using single reader	46%	44%	57%
	Using double reader	27%	26%	37%
	From arbitration	47%	45%	54%
	Using double reader with arbitration (using film priors for arbitration)	45%	43%	59%
Proportion of abnormal lesions not recalled	Single reader	8%	9%	9%
	Double reader with arbitration	4%	4%	4%

When reading the cases without prior mammograms participants said they would hang the prior mammograms for 84% of cases if they were stored at desk, and 51% of cases if they were stored in the other room. For the six normal screening cases participants said that they would hang the prior mammograms for 66% of cases if they were stored at the desk and 39% of cases if they were stored in the other room.

5.5.3 Cost Analysis

At the study hospital in 2007/08 the recall rate was 3.9% with a cancer detection rate of 0.7%. If the results of the study were correlated with the screening situation, and ceasing to use prior mammograms were to result 26% increase in the number of normal cases recalled, the recall rate would increase from 3.9% to 4.6%, corresponding to an additional 63 women recalled per 10,000 screened, with no change to the cancer detection rate.

The mammography, ultrasound and biopsy equipment at the study hospital was used for assessment of women recalled from the screening programme for 40% of the time. The cost of the extra assessments would be £6,338 in staff, £2312 in equipment and maintenance, and £501 in consumables for biopsy.

The centre screens 35,000 women per year, for which one digitiser is sufficient. The cost of digitisation equipment including 3 years maintenance would be £3932 per 10,000 women. Roller viewer maintenance costs would be £2571 per 10,000 women.

Digitisation of mammograms took a mean time of 1 minute 37 seconds to sort, insert and repack the films, and attach the correct DICOM data to the files, whilst concurrently the digitiser took 2 minutes 15 seconds per case. In practice the employee is likely to slow their pace of work to match the digitiser speed. Hanging and taking down the film prior mammograms took the same member of staff mean 27 seconds per case. Legood and Gray (2004) report that retrieving and replacing records takes 2 minutes per case. The costs are detailed in table 5.6.

Table 5.6 – Projected costs per 10,000 women at a UK breast screening centre for implementing three different approaches to the display of prior film mammograms.

		Digitised prior mammograms	Film prior mammograms	No Prior Mammograms
Equipment				
Purchase		£3,538	N/A	£1,591
Maintenance (inc physics)		£394	£2,571	£721
Consumables				£501
Labour				
Displaying prior mammograms	prior	£3,560	£712	
Retrieving and replacing files		£3,164	£3,164	
Assessment clinic				£6,338
Total		£10,656	£6,447	£9,151

5.6 Discussion

Removing prior mammograms was found to increase the number of benign or normal recalls by 26% in comparison to using digitised prior mammograms, corresponding to an increase in recall rate from 3.9% to 4.6%. This is comparable to the findings of a retrospective study in the US which found that without prior mammograms the recall rate increased from 3.8% to 4.9%. Such a large shift would necessitate extra working hours to accommodate the extra women recalled. Any increase in numbers of benign or normal recalls is undesirable as false positive recalls have been found to increase short term distress, increase worry about cancer in the year after screening (Aro *et al.*, 2000), increase the psychological cost of attending screening again, and almost double the chances of non-attendance at the next screening round (Brett and Austoker, 2001). To avoid an increase in recall rate from non-display of prior mammograms the threshold for recall could be adjusted,

however this would inevitably lead to an increase in the number of missed cancers.

Participant type (radiologist or radiography advanced practitioner) was not a main effect for JAFROC performance score. This is in line with other research findings that performance of the two groups is equivalent when stratified by years of experience, (Scott and Gale, 2006). However there was a trend towards an interaction between participant type and method of display of the prior mammograms, with radiologists performing better with film prior mammograms and radiography advanced practitioners performing better with digitised prior mammograms. In breast screening practice the radiography advanced practitioners use the prior mammograms less in film format than the radiologists (Taylor-Phillips *et al.*, 2009) which may provide an explanation for this effect. Analysis of level of use of prior mammograms in the experiment is necessary to test this hypothesis.

The counterbalancing used in this experiment was only between using film and digitised prior mammograms. Therefore it is important to establish whether the performance decrement in the third condition was due to not having access to the prior mammograms, or because this condition was investigated last, and participants experienced boredom or fatigue with the experiment by then. There was no significant change in performance between when the cases were read for the first and the second times, and therefore it is likely that it was the removal of the prior mammograms causing the decrement to performance when the cases were read for the third time.

In this study the abnormal cases were so proven by biopsy, this is the gold standard of truth. Some of the normal cases were determined so by the results from the screening programme, double reader with arbitration, plus the opinion of the expert radiologist. For these cases the standard of truth was the majority 3 of 4 readers, plus 2 years without interval cancers developing. The rest of the normal cases were recalled in the screening programme, but additional tests determined that they were normal. The truth here is determined by the results of the additional tests including biopsies, the opinion of the expert radiologist, and two years without the development of interval cancers. This falls short of the gold standard for normal cases, which would be a subsequent screening round with no abnormal findings. This gold standard was not achievable here because at the time of the study there was only one breast screening centre in the UK with an archive of digital images, and these spanned just two and a half years, and therefore to achieve a subsequent negative screening round would have necessitated waiting another 3 years, by which time the results of the experiment would be of little or no use. Revesz et al. (1983, pg 461) found that using different standards of truth for chest radiographs could result in different conclusions being drawn from an experiment, and therefore advise that "strategies that define the presence or absence of disease only by the diagnostic tests under evaluation are inadequate", however Miller et al. (2004) refute this and cite a similar study in which standard of truth did not affect the study results. In the study presented in this chapter 13% of the cases do not have evidence of their status (normal/abnormal) above and beyond that provided by the mammograms themselves and are therefore subject to this criticism, i.e. those

cases that went to arbitration in the Breast Screening Programme and were not recalled for further tests, and have not yet subsequently re-attended screening. These cases could not be removed from the analysis, as their inclusion means the study case set represents a cross section of the type of cases encountered in the Breast Screening Programme.

It has been proposed in some breast screening centres that although not using prior mammograms will increase the recall rate, the effect will be reduced by the process of double reading with arbitration, as this is known to reduce recall rates. The modelling employed in this study projected the reverse of this. Conversion to double reading with arbitration amplified the effect of the prior mammograms, and resulted in a greater difference in the number of unnecessary recalls between presenting prior mammograms for every case or not. This may be because using prior mammograms increased the readers specificity, and this effect was applied twice, once in screening and again in arbitration. In practice the knowledge that the mammography reader is reading an arbitration case may encourage them to read in a different manner, whereas in the model presented here the arbitration reader is unaware that their decision is being used for arbitration. However, it does demonstrate that the mechanism of double reading does not render the prior mammograms obsolete, it may even amplify their impact. Another suggestion for a low cost solution to the problem of presenting the prior mammograms is to provide the film prior mammograms for the arbitration cases only. The modelling in this study projects that this approach would have increased the number of false positive cases at the study centre by 20% in comparison to

using film priors for every case, and 27% in comparison to using digitised priors for every case, and therefore is also a suboptimal solution.

Making the prior mammograms available for the reader to hang them themselves when they consider it necessary was found by Roelofs *et al.* (2007) to be a detriment to performance in comparison to displaying the prior mammograms for every case. Nonetheless a breast screening centre in the UK has adopted this strategy, and others may follow and therefore it merited further investigation. When reading without prior mammograms participants stated that they would hang the prior mammograms for 84% of cases if they were stored at the desk and 51% of cases if they were filed in an adjacent room. These appear to be high proportions, but the case set used was difficult so prior mammograms would be used for a greater proportion of these difficult cases. Of the normal cases in the study participants said they would hang the prior mammograms for 66% of cases if stored at the desk, and 39% of cases if filed in another room. If this is a good representation of intentions for behaviour in breast screening practice then readers would either have to compromise on the number of prior mammograms that they use, or take significantly longer to read each case and therefore be unable to read their case load. In the Roelofs study prior mammograms were only considered necessary for around 30% of cases, so an important issue is why in this study participants stated they would hang them for 84% of cases. Both studies used a difficult case set with around 50% malignant cases. In the Roelofs study participants were actually asked to distinguish whether the prior mammograms were 'needed' or 'helpful', so they may in practice have wished

to hang some of those cases for which the prior mammograms were considered 'helpful'. This demonstrates the difficulty in asking participants to verbalise their intentions and behaviours rather than measuring them directly. The behaviour study detailed in chapter 4 shows that in film screen mammography prior mammograms are used for 86% of screening cases, therefore this is the best benchmark available for the proportion of cases the readers wish to use the prior mammograms for. This cannot be achieved in practice if mammography readers are required to hang them themselves, particularly with the increase in workload with the latest screening age extension, as they simply will not have the time.

The cost calculations indicate that digitising the prior mammograms is £1505 more expensive per 10,000 women than not displaying them at all. The majority of the costs for the digitisation solution are associated with administration time to complete the digitisation process and the retrieval and return of films, whereas without the prior mammograms the burden shifts to the mammography readers who conduct and interpret the mammograms, ultrasound examinations, and biopsies on the extra women recalled. Hanging the film prior mammograms on a multi-viewer adjacent to the digital workstation in advance of the reading session was associated with the lowest costs. Whilst this is the optimal solution in terms of cost effectiveness, digitisation is preferable for the mammography reader, as it has been found to take less time per case, reduce perceptions of workload, and encourage the mammography reader to use the prior mammograms for a greater proportion of cases (Taylor-Phillips *et al.*, 2009). The cost of staff time was calculated in

a top down manner for the additional recalls associated with not using prior mammograms, i.e. using knowledge of the actual staffing levels used in assessment clinics, but in a bottom up manner for the cost of digitisation i.e. by calculating the time taken to digitise the prior mammograms per woman and multiplying by the number of women. This may be unfair because in practice there are many other work activities which are accounted for in the top down but not the bottom up approach, such as breaks, talking to colleagues, and some administration tasks. The calculations for staff time spent digitising could not be made in a top down manner as it has not been implemented in a large UK breast screening centre, and the cost of staffing the extra recalls would have been extremely complex to calculate in a bottom up manner due to the number of different procedures involved in recalls, and the numbers of staff involved in each procedure. A correction was applied to the bottom up numbers to account for time spent in other activities, but a more thorough cost calculation is required through implementation of a trial of digitising all prior mammograms in a UK breast screening centre, and therefore calculating the costs of digitisation in a top down manner.

This study used a case set biased to be much more difficult than a typical screening session. However, when extrapolating the data to real world screening the results are still applicable as it is from these difficult cases that the false positive recalls arise, rather than from the simpler normal cases. If more of the simple normal cases were included in the case set it can be argued that this would not increase the number of false positive lesions, and so all of the FROC curves would simply shift to the left, and the net result

would be the same. Additionally, the weighting implemented to mimic the proportions of arbitration cases not recalled, recalled cases which had a biopsy, and recalled cases which did not have a biopsy did not significantly alter the results. Therefore the type of difficult normal case did not present a confounding variable in this study. However, the difficulty of cases and knowledge that performance was being measured may have influenced behaviour, in particular increasing vigilance and level of use of the prior mammograms. In light of evidence that in breast screening prior mammograms are used for 19% more cases when digitised than displayed in film format (Taylor-Phillips *et al.*, 2009), there may be a difference between performance using film or digitised prior mammograms in real world screening, even though one was not found under experimental conditions. The equivalence in performance between using film or digitised prior mammograms indicates that digitisation at 75µm, bit depth 12 may be sufficient for screening mammography. There is a need to investigate mammography readers' behaviour reading mammograms in 'live' screening in comparison to an experimental scenario with difficult test cases to further understand the applicability of ROC studies to real life.

5.7 Conclusions

The first aim of this study was to measure any changes in cancer detection performance with or without prior mammograms, and present any changes in terms which will influence clinicians practice such as recall rate. The use of digitised prior mammograms was found to improve performance in terms of

both JAFROC figure of merit, and number of normal cases recalled. Not using prior mammograms may increase the recall rate at the study hospital from 3.9% to 4.6% with no corresponding increase in cancer detection rate. Additionally, it has been projected that double reading is more likely to amplify than reduce this effect.

The second aim of this study was to determine the impact on cancer detection performance of digitising prior mammograms in preference to displaying them in film format during the transition to digital mammography. There was no difference in performance found between using film or digitised prior mammograms. A difference may still exist due to the greater level of use of prior mammograms in digitised rather than film format as described in chapter 4. Analysis of the level of use of prior mammograms in this experiment in comparison to real life is necessary to resolve this.

The third aim was to establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2. There was a trend towards an interaction between participant type and presentation medium of the prior mammograms, with radiography advanced practitioners performing better with digitised prior mammograms, and radiologists performing better with film prior mammograms.

6 Comparison of Behaviour in Experimental Setting and Screening Practice

6.1 Introduction

The studies detailed so far show that in breast screening practice at the study hospital the prior mammograms are used for 19% more cases when they are displayed in digitised rather than film format. However, when performance is measured using test cases there is no significant difference between using film or digitised prior mammograms. Therefore it is necessary to understand the patterns of use of prior mammograms when measuring performance using test cases, and whether these mirror that of screening practice. This will enable correct interpretation of the performance measurements, and determine whether the results of the performance study can be directly applied to the Breast Screening Programme.

6.2 Aims

1. To compare the level of use of prior mammograms and the time taken per case between using digitised or film prior mammograms in the experimental setting (when participants are reading difficult test cases and having their performance measured).
2. To compare the level of use of prior mammograms and the time taken per case between screening practice and the experimental setting.

3. To establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2.
4. To determine whether the behaviour observed in the experiment sufficiently modelled the actual behaviour observed in screening practice. If this is so then it is possible to enable direct generalisation of the performance results to the screening programme at least in the study centre considered here.

6.3 Method

Three metrics of behaviour were measured during the performance experiment detailed in chapter 5: the percentage of cases for which the prior mammograms were used; the mean number of times the prior mammograms were looked at per case; and the time taken per case. The method of recording these metrics involved analysis of video-tape of the participants reading the mammograms, taken from four different angles in the same manner as detailed in chapters 3 and 4.

The analysis was in two parts. Firstly the comparison of these three behavioural metrics between the hybrid and digital workstations during the performance study. Secondly, the comparison of these behavioural metrics between the performance study and screening practice.

For the first part of the analysis a mixed model analysis of variance was conducted for the three dependent variables: the percentage of cases for which the prior mammograms were used, the mean number of times the prior mammograms were looked at per case; and the time taken per case. The two independent variables were workstation type (digital or hybrid) and participant type (radiologist or radiography advanced practitioner). There is a requirement for within subjects analysis of variance that the data should follow a normal distribution at each level of the independent variable. This was tested using the Kolmogorov-Smirnov and Shapiro-Wilk statistics. As these statistics are not very sensitive to deviations from normality when the number of participants is small, box plots and Q-Q plots were also used to check for deviations from normality.

For the second part of the analysis comparisons were made between behaviour in the performance study and in screening practice. Seven of the participants in the performance experiment detailed in chapter 5 were the same as those whose behavioural use of prior mammograms and time taken per case was measured as described in chapters 3 and 4. Therefore within subjects comparisons were made between behaviour and time taken in a real screening setting and in the experiment for these seven participants. This can give a measure of whether the participants are behaving in a similar manner in the experiment to screening practice, and therefore a measure of whether the experiment can give an accurate representation of performance in practice. More specifically a mixed model analysis of variance was conducted with three independent variables, presentation medium of the prior

mammograms (digitised or film), setting (experiment or screening practice), and participant type (radiologist or radiography advanced practitioner). The three dependent variables tested were proportion of cases for which the prior mammograms were used, mean number of times the prior mammograms were looked at per case, and mean time taken per case. The normality of the data at each level of the independent variable was tested using the Kolmogorov-Smirnov and Shapiro-Wilk statistics, alongside box plots and Q-Q plots.

6.4 Results

6.4.1 Behaviour in the Experimental Setting

For the proportion of cases for which the prior mammograms were used the condition of normality was not violated as the Kolmogorov-Smirnov and Shapiro-Wilk tests were not significant and the boxplots showed no outliers. For the numbers of comparisons to the film prior mammograms the Kolmogorov-Smirnov test ($p=.01$) determined the condition of normality was violated, see table 6.1, and the boxplots indicated that participant 8 was an outlier, see figure 6.1. Therefore participant 8 was removed from the analysis of the numbers of comparisons to the prior mammograms. With the remaining seven participants the Kolmogorov-Smirnov test did not show a deviation from normality, however the boxplots indicated that participant 1 may be an outlier. Participant 1 was not removed from the analysis because whereas participant

8 was not within 2 standard deviations of the mean (participant 8 = 7.8, \bar{x} =3.6 σ =1.9), participant 1 was within 2 standard deviations of the mean (participant 1 = 4.5 \bar{x} =2.9 σ =0.9), and therefore was considered acceptable for inclusion.

Table 6.1 – Tests for normality for the number of comparisons to the prior mammograms at the hybrid and digital workstations. * denotes Lilliefors Significance Correction *. denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Digitised prior mammograms	.198	8	.200*	.923	8	.451
Film prior mammograms	.277	8	.070	.769	8	.013
Digitised prior mammograms (participant 8 removed)	.214	7	.200*	.964	7	.856
Film prior mammograms (participant 8 removed)	.276	7	.114	.895	7	.302

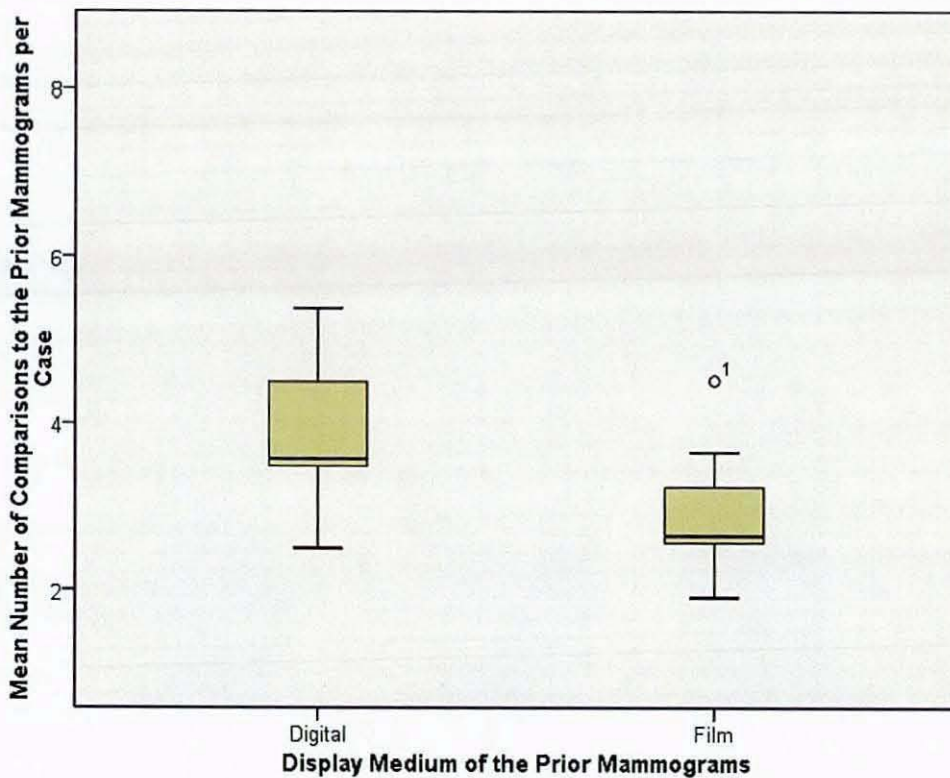
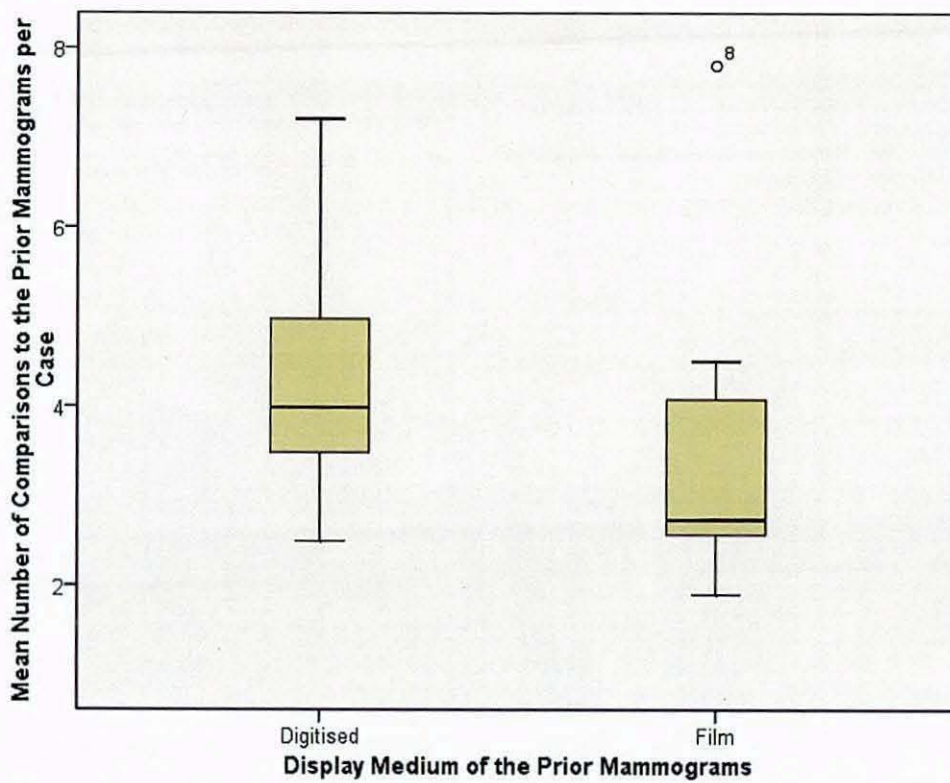


Figure 6.1 – Boxplots of the mean number of comparisons to the film and digitised prior mammograms including participant 8 (above) and excluding participant 8 (below).

For the time taken per case using both digitised and film prior mammograms the Kolmogorov-Smirnov ($p=.01-.0005$) and Shapiro-Wilk tests ($p=.002-.0005$) determined the condition of normality was violated, see table 6.2, and the boxplots indicated that participant 8 was an outlier, see figure 6.2. Therefore participant 8 was removed from the analysis of the time taken per case. With the remaining seven participants none of the tests showed a deviation from normality.

Table 6.2 – Tests for normality for the time taken per case at the hybrid and digital workstations. ^a denotes Lilliefors Significance Correction *. denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Digitised prior mammograms	.437	8	.000	.587	8	.000
Film prior mammograms	.323	8	.014	.689	8	.002
Digitised prior mammograms (participant 8 removed)	.239	7	.200*	.880	7	.225
Film prior mammograms (participant 8 removed)	.172	7	.200*	.975	7	.931

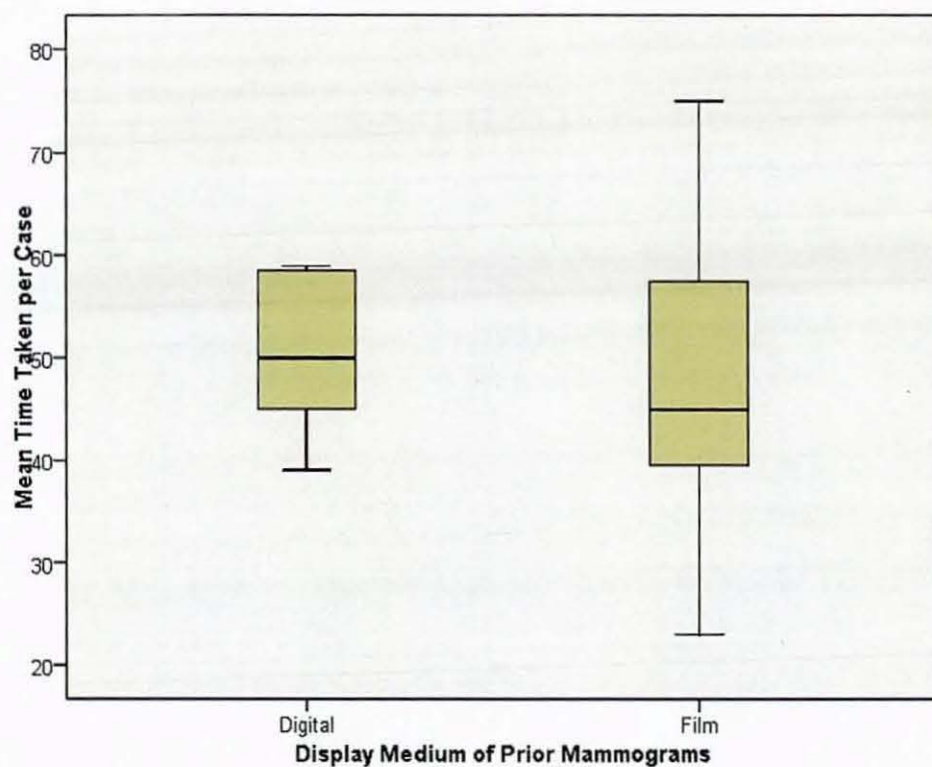
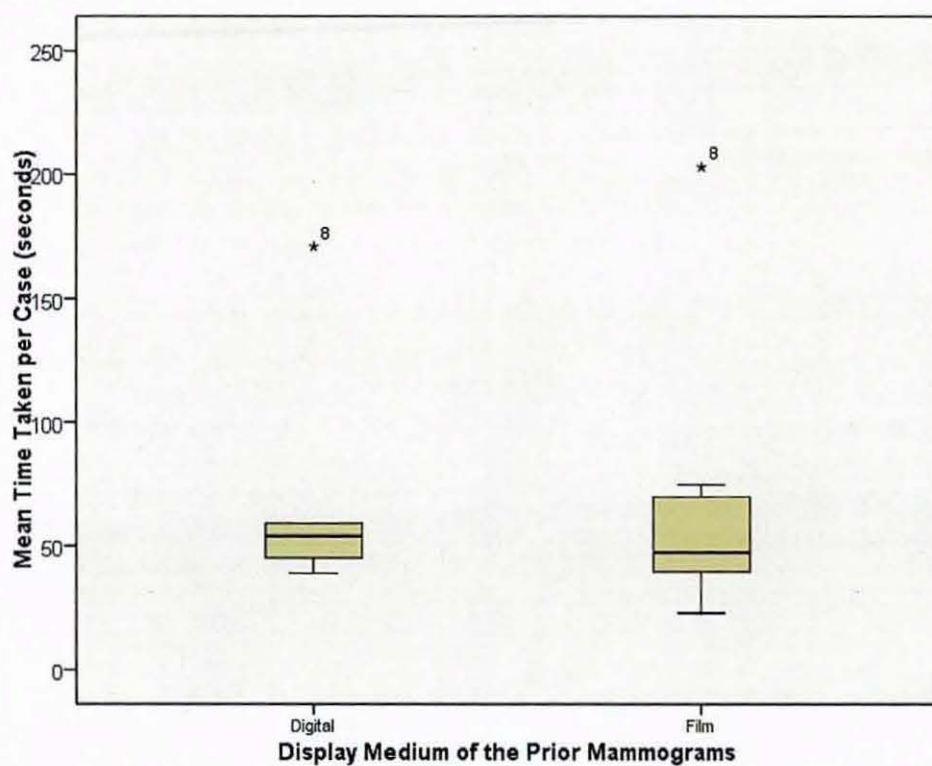


Figure 6.2 – Boxplots of the mean time taken per case using film and digitised prior mammograms, including participant 8 (above) and excluding participant 8 (below).

As a result of these tests the analysis of the proportion of cases for which the prior mammograms were used included all eight participants, but the analysis for mean number of comparisons and time taken per case excluded participant 8 as an outlier. There was no main effect of participant type (radiologist or radiography advanced practitioner) for any of the metrics.

The percentage of cases for which the prior mammograms were used did not differ with the presentation medium of the prior mammograms ($F(1,6)=1.6$, $p=.2$), with prior mammograms used for 96% of cases when in digitised format and 93% of cases when in film format. However, there was an interaction between presentation medium of the prior mammograms and participant type ($F(1,6)=11.6$, $p=.01$), with radiologists using the film prior mammograms for a greater proportion of cases, and radiography advanced practitioners using the digitised prior mammograms for a greater proportion of cases, see figure 6.3.

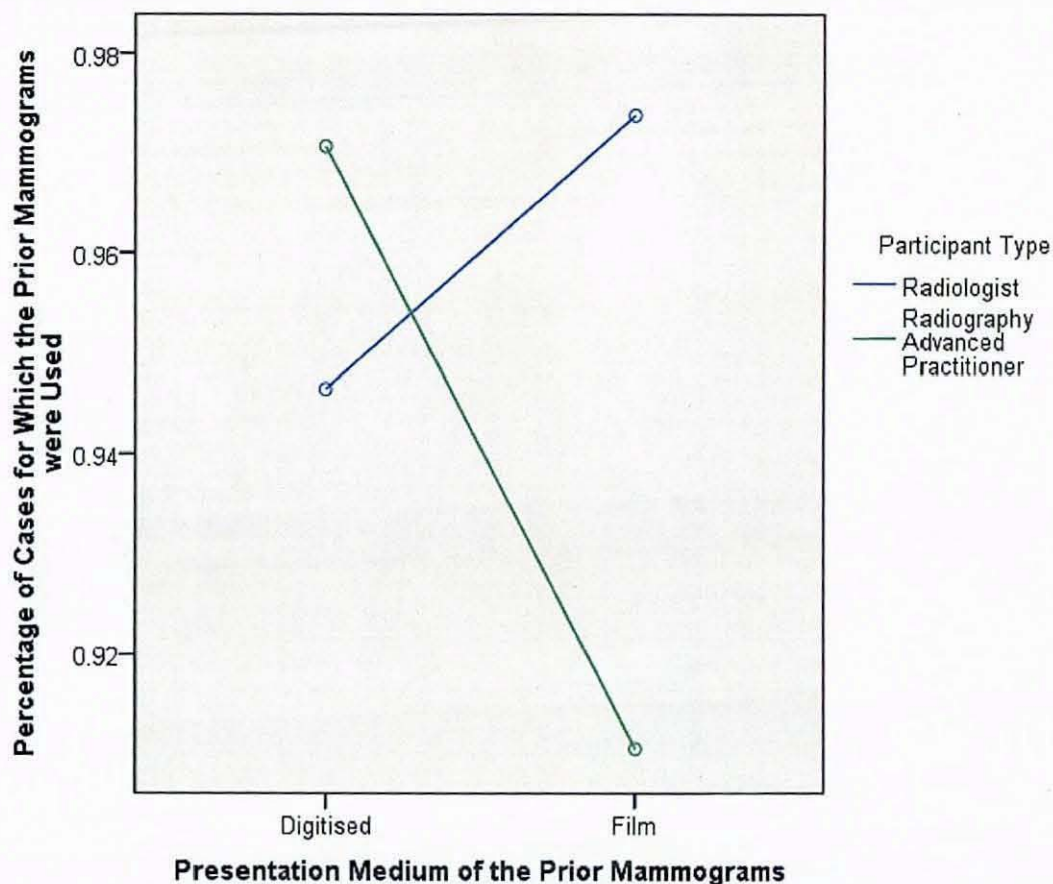


Figure 6.3 – The interaction between participant type and presentation medium of the prior mammograms for the percentage of cases for which the prior mammogram was used ($F(1,6)=11.6$, $p=.01$).

The mean number of comparisons to the prior mammograms per case differed by presentation medium of the prior mammogram ($F(1,5)=6.6$, $p<.05$), with greater number of comparisons made using digitised prior mammograms (mean 3.9 per case) than film prior mammograms (mean 2.9 per case). There was no interaction between participant type and display medium of the prior mammograms.

The time taken per case was not affected by presentation medium of the prior mammograms. Radiologists took less time per case using film prior mammograms, whereas radiography advanced practitioners were faster using digitised prior mammograms, see figure 6.4, but the interaction between participant type and presentation medium of the prior mammograms for time taken per case was not significant.

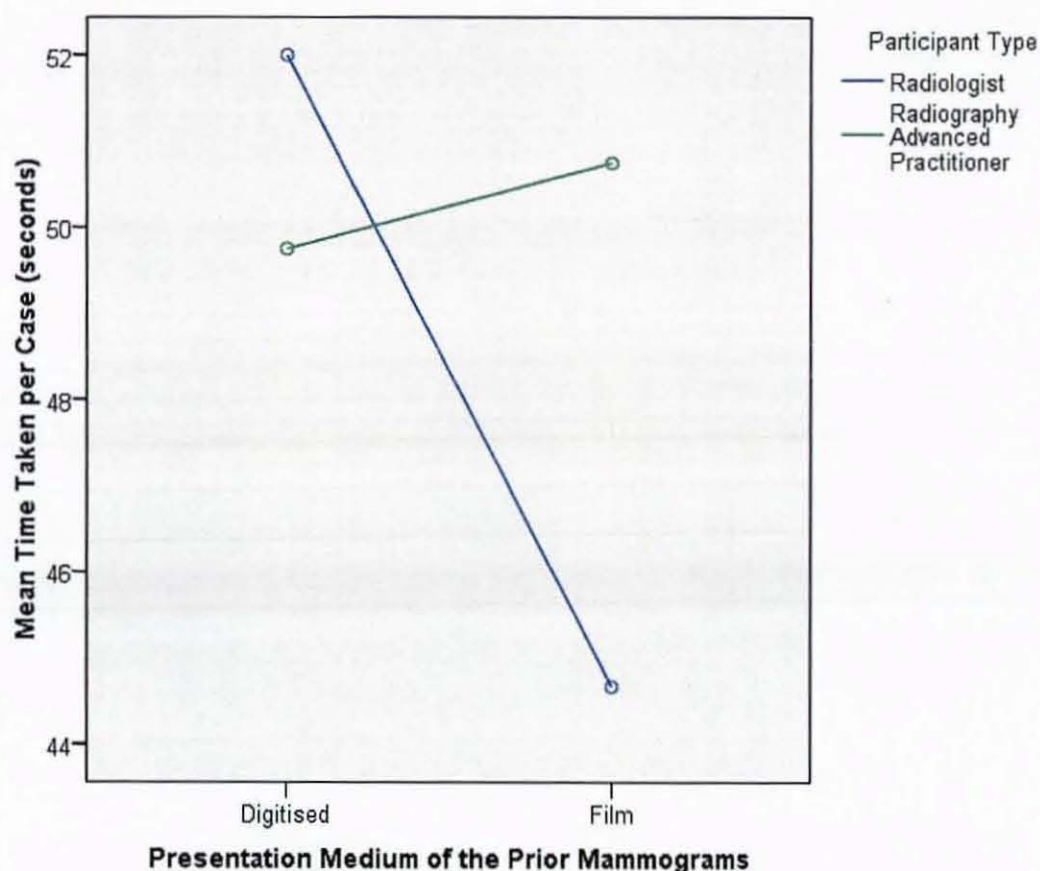


Figure 6.4 – Mean time taken per case by participant type and presentation medium of the prior mammograms. There is no significant interaction.

6.4.2 Comparison of Behaviour in the Experimental and Live Screening Settings

There were no deviations from normality detected by either the Kolmogorov-Smirnov and Shapiro-Wilk statistics, or the box plots and Q-Q plots. For the proportion of cases for which the prior mammograms were used there were main effects of both setting (experiment or screening practice, $F(1,5)=23.0$, $p=.005$) and presentation medium of the prior mammograms (film or digital, $F(1,5)=17.6$, $p=.009$). The prior mammograms were used for 95% of cases in the experiment in comparison to 71% of cases in screening practice, and for 88% of cases with digitised priors in comparison to 78% of cases with film priors, see table 6.3 for a further breakdown.

Table 6.3 – Mean proportion of cases for which the prior mammograms were used, number of comparisons and time taken per case, for both screening and experimental setting with film and digitised prior mammograms. Data is provided for all cases, and for just the normal screening cases (a normal screening case is one which was not recalled in breast screening practice by either reader). Data is for the seven participants who took part in both the observations of screening practice and the experiment. Significance tests are within subjects t tests between using digitised and film prior mammograms, a blank field represents no significant difference.

		Screening Practice			Experiment		
		Digitised Prior Mammograms	Film Prior Mammograms	Sig	Digitised Prior Mammograms	Film Prior Mammograms	Sig
For all cases (160 cases in the experiment, including recalled cases in screening practice)	Proportion of cases for which the prior mammograms were used	81%	59%	.04	96%	93%	
	Mean number of comparisons per case	2.4	1.3	<.05	3.9	2.9	.03
	Time taken per case (seconds)	35	45	.02	51	48	
For just normal screening cases (6 cases in the experiment, excluding recalled cases in screening practice)	Proportion of cases for which the prior mammograms were used	80%	58%	.04	90%	93%	
	Mean number of comparisons per case	2.3	1.2	.04	2.7	2.6	
	Time taken per case	31	40	.007	37	42	

For the proportion of cases for which the prior mammograms were used there were several interactions. There was an interaction between participant type and presentation medium of the prior mammograms ($F(1,5)=18.7, p=.008$). Radiography Advanced Practitioners used the prior mammograms for a smaller proportion of cases when in film format, and a greater proportion of cases when in digitised format in comparison to the radiologists, see figure 6.5.

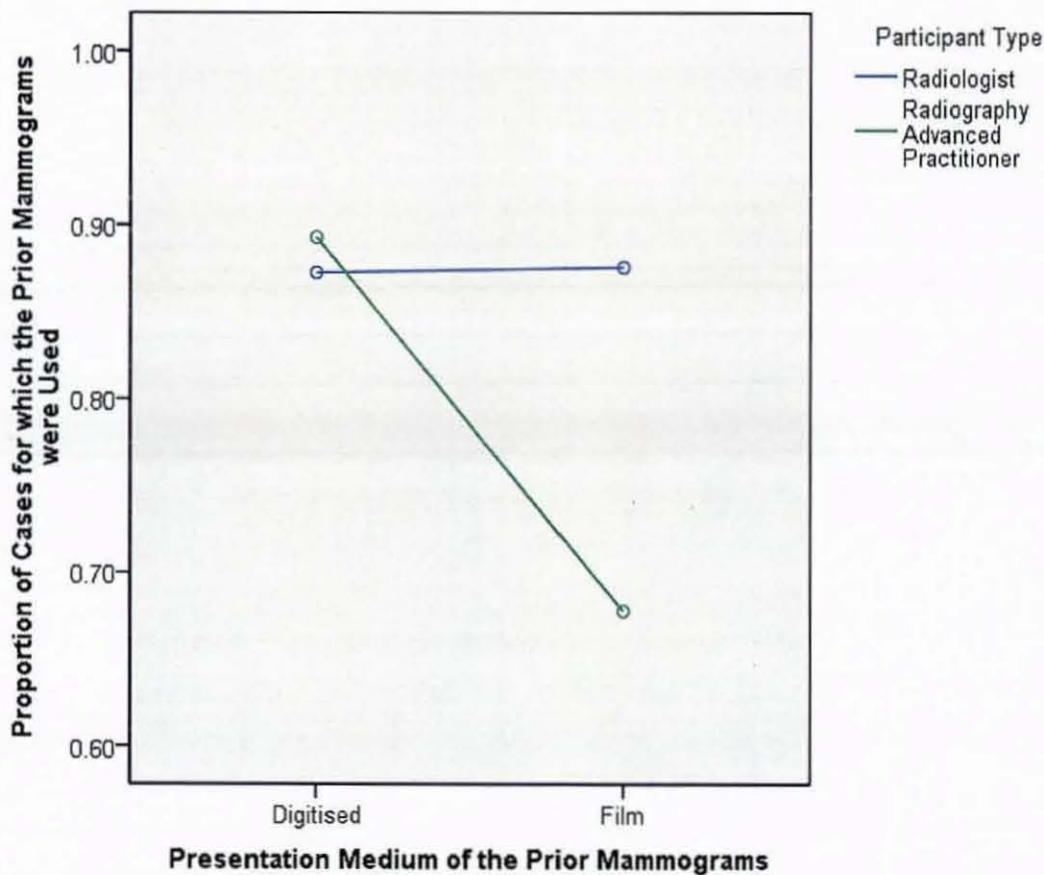


Figure 6.5 – The interaction between participant type and presentation medium of the prior mammograms ($F(1,5)=18.7, p=.008$) for proportion of cases for which the prior mammograms were used.

There was an interaction between setting (experimental or screening practice) and presentation medium of the prior mammograms (film or digital) for the proportion of cases for which the prior mammograms were used ($F(1,5)=13.8$, $p=.01$), see figure 6.6. For the same seven participants, in screening practice presenting the prior mammograms in digitised format resulted in them being used for 22% more cases ($p=.04$), whereas this difference was not mirrored in the experimental setting, with the digitised prior mammograms being used for 3% more cases than the film prior mammograms (not a significant difference).

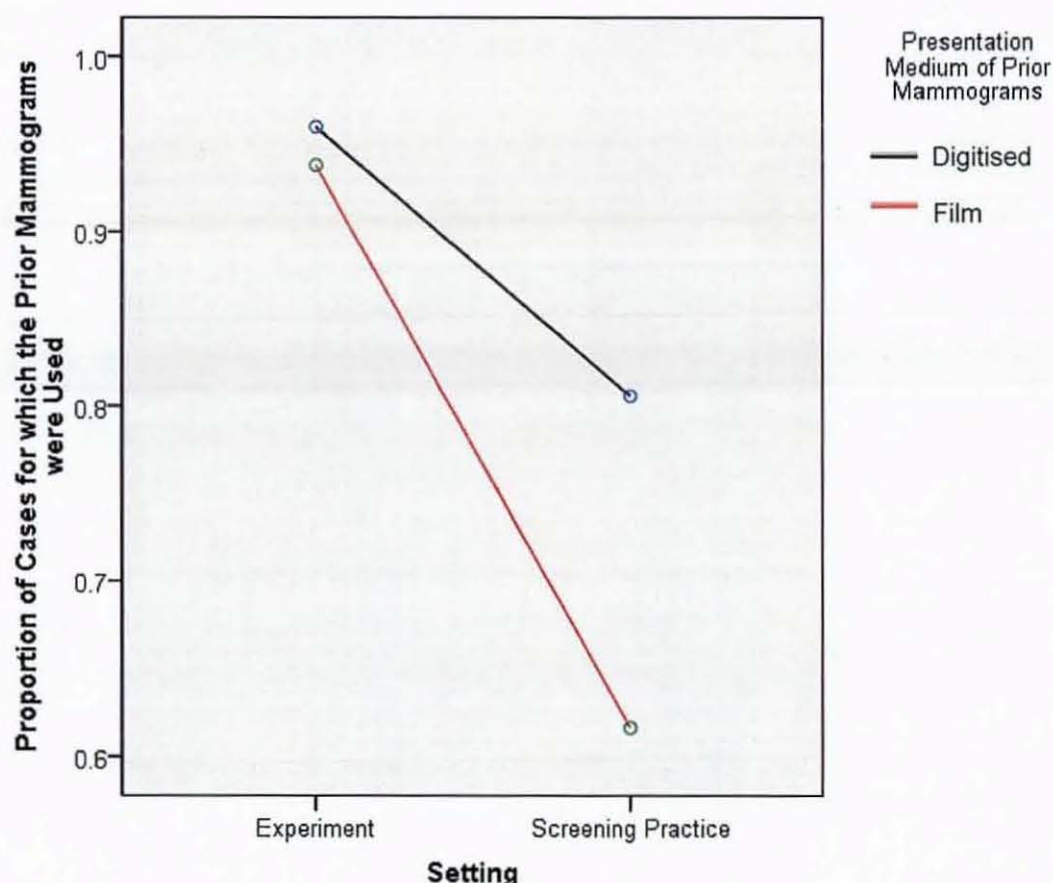


Figure 6.6 – The interaction between the setting (experimental or screening practice) and the presentation medium of the prior mammograms for proportion of cases for which the prior mammograms were used ($F(1,5)=13.8$, $p=.01$)

There was no interaction between setting and participant type for proportion of cases for which the prior mammograms were used, so any change in behaviour between the experimental setting and screening practice were not directly influenced by whether the participant was a radiologist or radiography advanced practitioner. However, there was a three way interaction between setting, presentation medium of the prior mammograms and participant type ($F(1,5)=9.8, p=.03$), see figure 6.7. The interaction between setting and presentation medium of the prior mammograms is large for radiography advanced practitioners and very small for radiologists. i.e. for radiologists the difference between the proportion of cases for which the prior mammograms were used in the experiment and in screening practice is not affected by the presentation medium of the prior mammograms, whereas for radiography advanced practitioners it is dependent on the presentation medium of the prior mammograms. The Radiography Advanced Practitioners used the prior mammograms for 36% more cases when digitised than displayed in film format in screening practice, but this difference was reduced to 8% in the experiment.

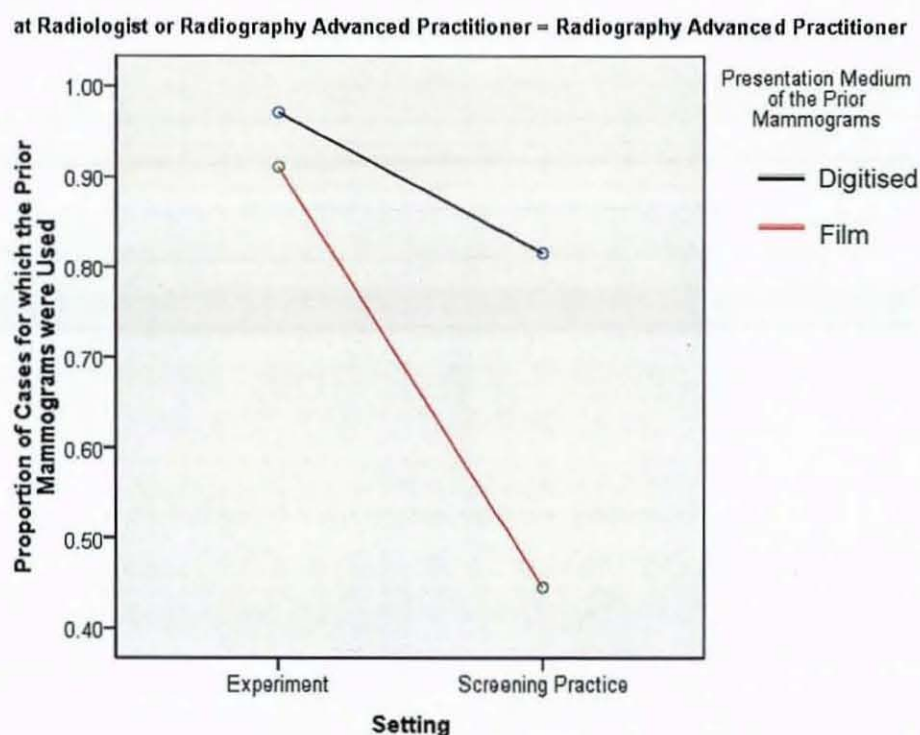
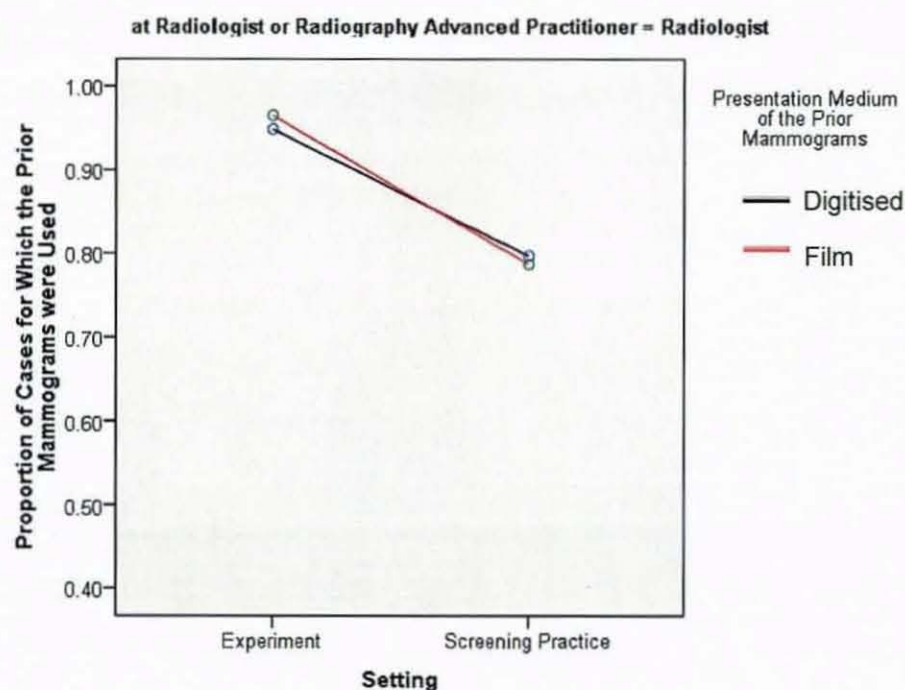


Figure 6.7 – The interaction between setting, presentation medium of the prior mammograms and participant type ($F(1,5)=9.8, p=.03$) for proportion of cases for which the prior mammograms were used. The relationship between setting and presentation medium of the prior mammograms is shown for both radiologists and radiography advanced practitioners.

For the number of comparisons to the prior mammograms there were main effects of both setting ($F(1,5)=30.1$, $p=.003$) and presentation medium of the prior mammograms ($F(1,5)=12.5$, $p=.02$). The mean number of comparisons to prior mammograms per case was in 3.5 the experiment in comparison 1.9 in screening practice, and 3.1 with digitised priors in comparison to 2.2 with film priors. There were no interactions, see figure 6.8.

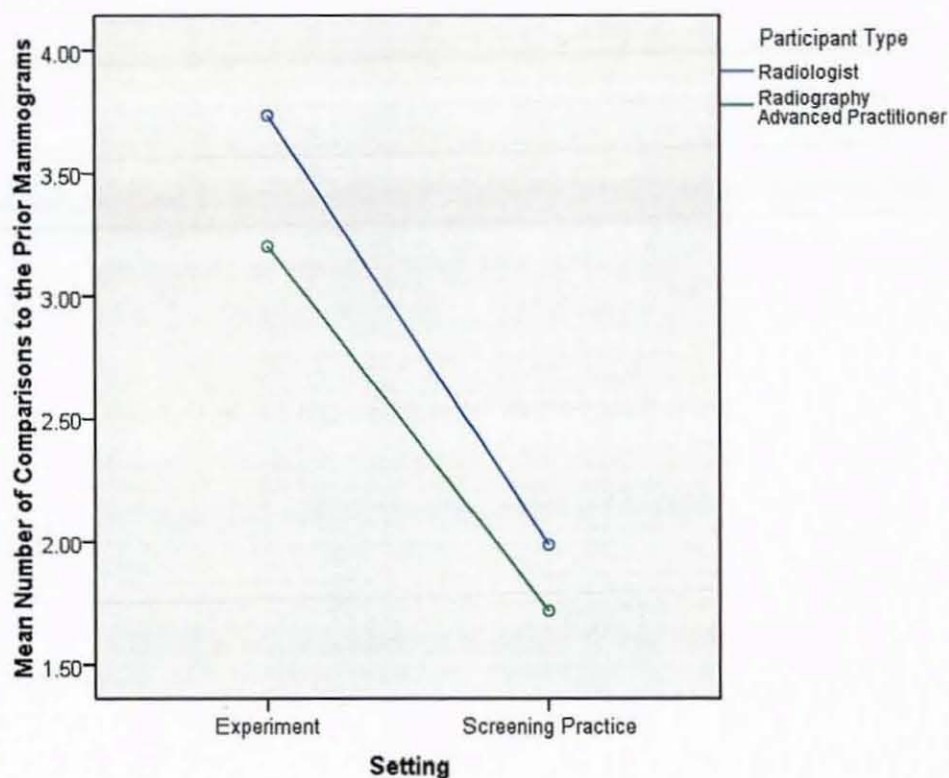
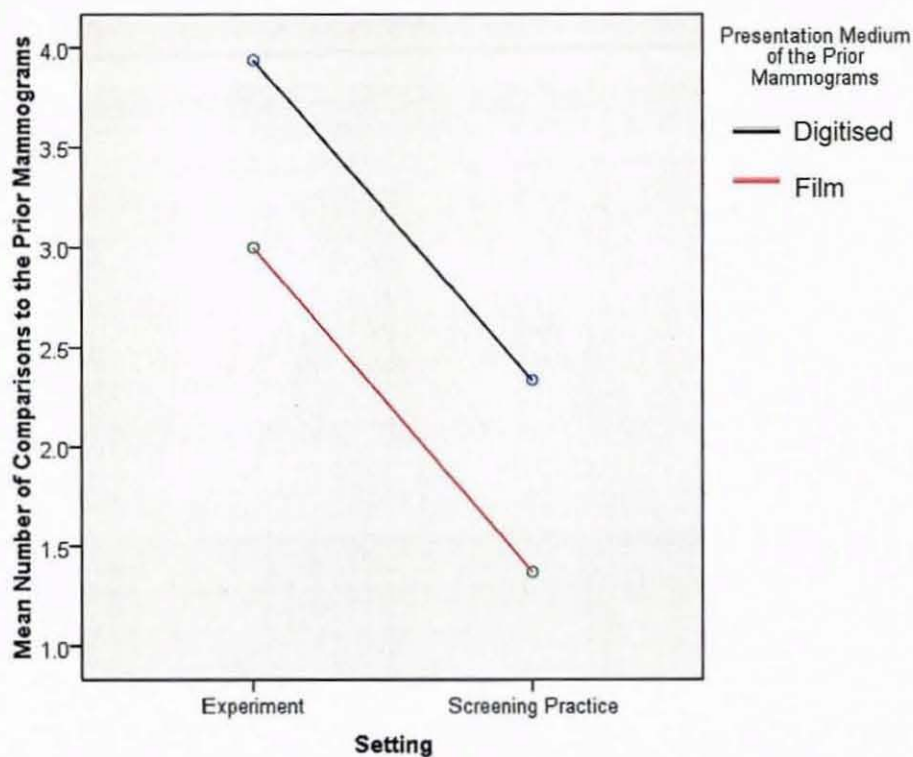


Figure 6.8 – Plots of the variation of the mean number of comparisons to the prior mammograms with setting (experimental or screening practice), stratified by participant type and presentation medium of the prior mammograms. There were no significant interactions.

The mean time taken per case was 49 seconds in the experiment in comparison to 40 seconds in screening practice but this difference was not significant ($F(1,5)=3.9, p=.1$). There was a trend towards an interaction between setting (experimental or screening practice) and presentation medium of the prior mammograms ($F(1,5)=4.6, p=.08$). The time saving in screening practice using digitised prior mammograms in comparison to film prior mammograms was not apparent in the experiment, see figure 6.9.

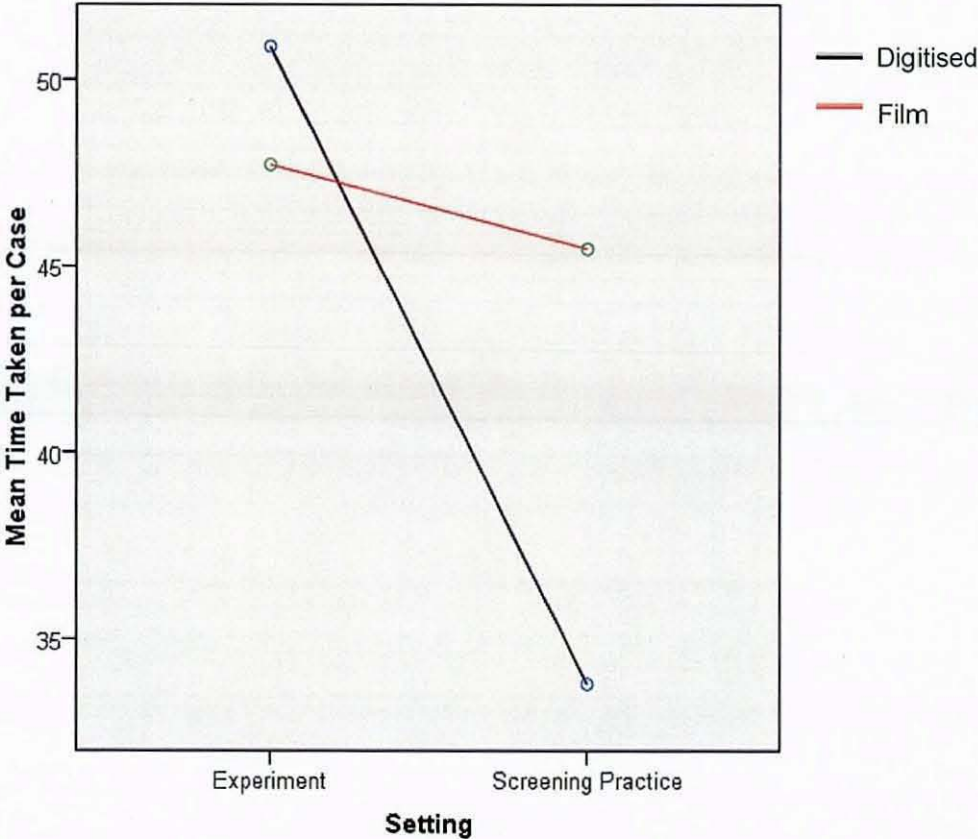


Figure 6.9 – The relationship between time taken per case in the experiment and in screening practice for both film and digitised prior mammograms. The interaction was not significant ($F(1,5)=4.6, p=.08$).

6.5 Discussion

The most important behavioural metric in this study when considering potential effects on performance is the proportion of cases for which the prior mammogram was used. This is because if a participant does not look at the prior mammograms we know they cannot have gleaned any information from them, and use of prior mammograms is known to improve specificity. When comparing proportion of cases for which the prior mammograms were used in the experiment to screening practice there was an interaction between the setting and the display medium of the prior mammograms. Whilst in screening practice the proportion of cases for which the prior mammogram was used was higher using digitised rather than film priors this difference was not apparent in the experiment. This result may limit the applicability of the experiment, as the experimental measurements of performance using film prior mammograms may be higher than would be attained in screening practice due to this behavioural difference. Whether it does limit the applicability of the experiment depends on the reason for the change, as discussed below.

There could be several explanations for the prior mammograms being used for a greater proportion of cases when using digitised rather than film priors in screening practice, but not in the experiment. These include: firstly in screening practice the case difficulty may differ between the conditions causing a confounding variable; secondly the greater case difficulty in the experiment could increase the need for the prior mammograms; thirdly

patterns of viewing of the current case could be influenced by the previous case read; and fourthly the knowledge that performance was being measured in the experiment could increase vigilance.

Considering the first explanation, in the experiment the same cases were read with film as with digitised prior mammograms, and therefore the case difficulty could not provide a confounding variable. In screening practice the cases were not the same when using film and digitised prior mammograms, and therefore it is possible that the behavioural differences in the screening setting were due to case type rather than presentation medium of the prior mammograms. The time taken per case was shorter using film than digitised prior mammograms in screening practice but not in the experiment, which appears to provide evidence for the notion that the measurements taken in screening practice using digitised priors were of more difficult cases than those using film priors. However, this is unlikely to provide a confounding variable as the number of cases per condition was large (mean 82), the case selection was random by set from the screening programme, and an additional analysis with just the simpler normal screening cases which were not recalled for further tests found that the effect was still present. Considering these factors it seems unlikely that in the observations of screening practice the cases for which the prior mammograms were film were more difficult than those for which the prior mammograms were digitised.

The second explanation cites that the case set used in the experiment was composed of such difficult cases that the prior mammograms were used for

over 90% of cases regardless of the presentation medium. Therefore, whilst for over 90% of these difficult cases the prior mammograms would be consulted in screening practice, these difficult cases are rare, they make up less than 5% of all cases at the study centre, (Duncan, personal communication) and therefore would not significantly affect the measurements of screening practice. Case difficulty could also explain why in screening practice reading with digitised prior mammograms was faster, whereas in the experiment there was no significant difference. This explanation can be tested. In the experiment 6 of the 160 cases were normal cases which were randomly selected from the screening programme, and were not recalled by either reader in the screening programme. If it is simply case difficulty driving the discrepancy between experiment and screening practice then the level of use of the prior mammograms for these 6 cases would be the same as the equivalent non-recalled cases in screening practice for the seven participants who took part in both studies. In the experiment the prior mammograms were looked at for over 90% of these 6 normal cases when in both film and digitised formats, in comparison to just 80% of digitised priors and 58% of film priors in screening practice, see table 6.3. Therefore case difficulty alone cannot explain the discrepancy between behaviour in screening practice and in the experiment, either for use of prior mammograms or time taken per case.

The third explanation that readers subconsciously revert to patterns of viewing, whereby the manner in which the current case is read is influenced by the manner in which the case before it was read. Therefore, in screening practice many cases could be quickly returned to screening, and therefore the

reader could get into a pattern of reading quickly and without the prior mammograms for several cases in a row. In the experiment the majority of cases are very difficult and therefore the pattern would be to use the prior mammograms for the majority of cases.

The fourth explanation is that the knowledge that their performance was being measured motivated the participants to greater vigilance, and therefore to greater use of the prior mammograms. Whilst the participants were video-taped in both the experiment and in screening practice, and therefore the Hawthorne effect would be present to some extent in all measurements, performance was measured only in the experimental condition, and all participants were aware that use of prior mammograms improves performance. High vigilance in the experimental setting would therefore explain the use of prior mammograms for over 90% of cases despite the display medium. In screening practice if vigilance was lower then prior mammograms would only be accessed when the participant crossed a threshold of effort to perceived benefit ratio, which would explain why they were used for a greater number of cases when digitised as they require less effort to access in this format.

If the first or second explanations describe fully the discrepancy between experiment and screening practice, then the performance measurements taken in the ROC experiment can be directly applied to screening practice. However, if the third or fourth explanations influence the discrepancy between the experiment and screening practice then the performance measurements

for using film priors in the experiment may be overestimates when applied to screening practice. Therefore, as the first and second explanations are unlikely, performance using film priors in the experiment may be an overestimate if applied to screening practice. The uncertainty is because Roelofs *et al.* (2007) found that increasing the use of prior mammograms from the 30% that the radiologists felt were necessary, to 100% improved performance. However, using digitised priors rather than film priors in screening practice was found to increase the proportion of cases for which they were used from 63% to 82%. It is possible that the particular subset of cases for which the digitised prior mammograms would be used and the film prior mammograms would not be used may not improve performance. However, on the balance of probability it is estimated that the prior mammograms for these cases would improve performance, because Roelofs *et al.* (2007) found that radiologists underestimate the usefulness of prior mammograms to their own performance, and there is no reason to believe that this applies to only a subset of cases.

This problem of applicability of results from experiments using test cases to screening in real world applies not only to the experiment outlined in this thesis, but to all ROC analysis studies using enriched case sets. To overcome this either a greater knowledge of the relationship between behaviour in real world screening and ROC studies is required, and its impact on performance, or verification of results with randomised controlled trials in screening practice. The latter is a very expensive option. There is some research in this field. Gur *et al.* (2003) measured performance in detecting abnormalities in chest x-rays

with different levels of prevalence of disease in the test set, and found no effect of prevalence. However in a later paper (Gur *et al.*, 2007) the results were re-analysed and it was found that although overall ROC score was unaffected by increases in disease prevalence in the test set, the average confidence score did decrease. It is not known why confidence in performance is poorer with increased prevalence, or indeed why behaviour in looking at the prior mammograms differed between a test situation with increased prevalence and a real screening situation, and therefore this area deserves more attention.

For the proportion of cases for which the prior mammograms were used there was an interaction between participant type and display medium of the prior mammograms. Radiography advanced practitioners used the prior mammograms for a greater proportion of cases when in digitised format, and radiologists used the prior mammograms for a greater proportion of cases when in film format. In screening practice there is a similar significant interaction with radiography advanced practitioners using the prior mammograms for a greater proportion of cases when digitised. When the proportion of cases for which the prior mammogram was used was compared between the experiment and screening practice there was a significant interaction between the setting (experiment or screening practice), the display medium of the prior mammograms (digital or film), and participant type (radiologist or radiography advanced practitioner). This demonstrates that whilst both groups of participants showed an increase in the proportion of cases for which the film prior mammograms were used in comparison to the

digitised priors when moving from screening practice to the experiment, the change was greatest for the radiography advanced practitioners. This may be because the radiography advanced practitioners had a lower starting point, i.e. in screening practice radiography advanced practitioners only used the film prior mammograms for 45% of cases, whereas radiologists used them for 81% of cases, in the experiment radiography advanced practitioners used the prior mammograms for 91% of cases and radiologists used them for 97% of cases. The difference in the proportion of cases for which the film and digitised prior mammograms were used in screening practice was not present in the experiment, and this difference was greater for radiography advanced practitioners so the discrepancy between screening practice and experiment is greatest for this group.

Interestingly, there were no interactions for number of comparisons per case to the prior mammograms. In the experiment there were 2.5 more comparisons per case than in screening practice, but this increase was not dependent on participant type or presentation medium of the prior mammograms. This suggests that the fact that the prior mammograms are looked at for a greater number of times per case when in digitised rather than film format does not result in a difference in cancer detection performance, as performance using film and digitised prior mammograms was equivalent. This supports the theory proposed in chapter 4, that although the film prior mammograms are looked at fewer times in comparison to the digitised priors, more information is taken in each time. This also provides a potential explanation for the greater perceived workload using film rather than digitised

prior mammograms, as taking in more information each time would make greater demands on the working memory. Equally it may be simply that looking at the prior mammogram the first one or two times is enough to elicit all of the useful data and the extra looks do not add any information which contributes to cancer detection performance. It is not clear why there are interactions for the proportion of cases for which the prior mammograms were used, but not for the mean number of times the participant looks at the prior mammograms per case. It may simply be because there is a limit to the proportion of cases for which the prior mammograms were used (i.e. 100%), but there is no limit to the mean number of comparisons to the prior mammograms.

6.6 Conclusions

The first aim was "to compare the level of use of prior mammograms and the time taken per case between using digitised or film prior mammograms in the experimental setting (when participants are reading difficult test cases and having their performance measured)." In the experimental setting no significant differences between using film or digitised priors were found for time taken per case, or proportion of cases for which the prior mammograms were used. There were a great number of comparisons per case to the prior mammograms when displayed in digitised rather than film format.

The second aim was "to compare the level of use of prior mammograms and the time taken per case between screening practice and the experimental

setting". The proportion of cases for which the prior mammograms were used was higher in the experiment than in screening practice, and there was an interaction with the presentation medium of the prior mammograms (film or digitised). In screening practice the prior mammograms were used for a greater proportion of cases when displayed in digitised (81%) rather than film (59%) format, whereas in the experiment there was no significant difference between digitised (96%) and film (93%) display. The number of comparisons to the prior mammograms was higher in the experiment than in screening practice, but there was no interaction with presentation medium of the prior mammograms. In screening practice the number of comparisons per case to the prior mammograms was 1.3 using film priors and 2.4 using digitised priors, in the experiment this increased to 2.9 using film priors and 3.9 using digitised priors. The mean time taken per case was greater in the experiment (49seconds) than in screening practice (40 seconds) but this was not significant.

The third aim was "to establish whether participant type (radiologist or radiography advanced practitioner) is a factor in any of the changes identified by aims 1 and 2". In the experiment there was an interaction between participant type and presentation medium of the prior mammograms for the proportion of cases for which the prior mammogram was used. The radiologists used the prior mammograms for a greater proportion of cases when displayed in film format, and the radiography advanced practitioners used the prior mammograms for a greater proportion of cases when displayed in digitised format. When comparing the experimental setting to screening

practice there was a three way interaction between setting, presentation medium of the prior mammograms, and participant type. i.e. the difference in proportion of cases for which the prior mammogram were used between the film and digitised prior mammograms was smaller in the experiment than in screening practice, and this difference was greater for radiography advanced practitioners than for radiologists. Therefore, although there were no performance differences between radiologists and radiography advanced practitioners, their behaviour in terms of proportion of cases for which the prior mammograms was used differed quite substantially, and this difference was dependent on both the setting and the presentation medium of the prior mammograms.

The fourth aim was "to determine whether the behaviour observed in the experiment sufficiently modelled the actual behaviour observed in screening practice. If this is so then it is possible to enable direct generalisation of the performance results to the screening programme at least in the study centre considered here". The performance results are applicable when the participants are in a state of high vigilance as prior mammograms were used for over 90% of the cases in the experiment. There was no significant difference between the proportion of cases for which the digitised or film prior mammograms were used in the experiment, whereas in screening practice the digitised prior mammograms were used for 19% more cases. Therefore, whilst equivalence in performance between using film or digitised prior mammograms was obtained with the participants in a state of high vigilance,

in screening practice when vigilance may decrease the performance using film prior mammograms may also decrease.

7 Discussion

The UK Breast Screening Programme is starting to make the transition from film to digital mammography. Each breast screening centre in the UK will be making choices about what to do with the film mammograms from the previous screening round, and this research provides data to inform the choice. The three options investigated were; displaying the prior mammograms in digitised format, displaying the prior mammograms in film format or not displaying the prior mammograms at all.

The findings from this body of research are that prior mammograms should be presented for every case in the transition to digital mammography, as not displaying them in a test situation resulted in a 26% increase in unnecessary recalls of normal women. Where possible the prior mammograms should be presented in digitised rather than film format as this was found to lower readers' perceptions of workload, increase the speed of reading and may improve cancer detection performance.

These findings have been published extensively in conference proceedings and a journal paper, presented orally at three practitioner conferences and one practitioner training course, and presented in the form of a poster at one practitioner conference. The supporting data were tested by peer review in the European Radiology submission process, and by presenting at two SPIE Medical Imaging conferences. All publications were also checked by the co-authors Prof. Alastair Gale and Dr Matthew Wallis. Any erroneous conclusions

which strayed beyond the evidence should have been picked up through this process, and therefore it is likely that there are sufficient data to support the conclusions.

7.1 Progression of the Research Direction

The first experiment detailed in chapter 2 used RULA postural analysis and body part discomfort charts to assess discomfort and the prevalence of awkward postures which could lead to musculoskeletal disorders. No significant differences in workstation comfort and ergonomics were found between using film and digitised prior mammograms. However, there was some evidence that the reason that there were no comfort differences was that when viewing the prior mammograms in film format participants were adapting their behaviour to avoid awkward postures, i.e. by viewing the film prior mammograms from a greater distance than they would normally view film mammograms. As a result of this greater distance they may have been able to perceive less detail in the images. Hence reading behaviour merited further study.

The second study detailed in chapter 3 measured perceived workload using the NASA RTLX tool and time taken per case. These were both higher using film than digitised prior mammograms. This shows that there is an advantage to digitising prior mammograms unless the time savings and lower workload are due to the readers using the prior mammograms to a lesser extent when

digitised. This further highlighted the need to investigate behavioural use of prior mammograms.

Behavioural use of prior mammograms was investigated by analysing eye movements from video recorded data, as detailed in chapter 4. The findings were that participants used (i.e. looked at at least once) the prior mammograms for 18% more cases when in digitised rather than film format. Therefore the presentation medium of the prior mammograms was influencing behaviour, which may in turn influence cancer detection performance.

Cancer detection performance was tested directly in a JAFROC performance experiment as detailed in chapter 5. Performance was not found to differ between using film or digitised prior mammograms, but further investigation was found to be necessary to ascertain whether behaviour in the experiment mirrored that of screening practice. The second finding of the performance experiment was that when prior mammograms were not available it resulted in a reduction in cancer detection performance in comparison to having them available in film or digitised formats. This reduction of performance was evident both in a reduction in the JAFROC figure of merit, and a 26% increase in recalls of normal women with no change in the number of cancers detected.

Behavioural use of prior mammograms in screening practice when compared with the performance experiment was analysed in chapter 6. Behaviour in the experiment did not mirror that of real life screening practice, as the prior mammograms were used for 95% of cases in the experiment and just 71% of

cases in screening practice. In screening practice the prior mammograms were used for a greater proportion of cases when digitised rather than displayed in film format, and this was not replicated in the experiment. Therefore although equivalence was found in the experiment, performance may still be superior using digitised rather than film prior mammograms in screening practice due to their use for a greater proportion of cases.

7.2 Aims and Objectives

The overall aims of this research were to produce and publish recommendations with supporting data about how the analogue prior mammograms should be displayed in the transition to digital mammography. The recommendations are to display the prior mammograms for every case in the transition to digital mammography, and where possible in digitised format. This follows from the research detailed in chapter 5 which demonstrates that the use of analogue prior mammograms (displayed in either film or digitised format) in digital mammography does improve performance, and chapters 3-4 show how using digitised prior mammograms resulted in reduced perceptions of workload, shorter mean time to read each case, and use of prior mammograms for a greater proportion of cases. These findings have been extensively published as detailed in appendix 9.

These overall aims were broken down into five objectives which were achieved as follows. The first objective was "to understand the literature on

the use of prior mammograms in the transition to digital mammography". The literature review was used to identify the gaps in knowledge and therefore guided the research design. There was one very pertinent study by Roelofs *et al.* (2007) which found that cancer detection performance was superior with prior mammograms, in comparison to either without prior mammograms, or with prior mammograms only upon the readers request (Roelofs *et al.*, 2007). This suggests that the option of not displaying prior mammograms, or that of asking the mammography reader to hang them for the cases where they feel they are necessary, may both be sub-optimal in terms of cancer detection performance. However, the cases used in the Roelofs study were all digitised. Therefore, the potential loss of information when making comparisons between digitally acquired current mammograms and analogue prior mammograms (displayed either in film or digitised format) was not accounted for. Additionally, there were no measurements of performance using film in comparison to digitised prior mammograms. This paper guided the research presented here towards focussing on the difference between digitised and film presentation of prior mammograms, and towards more accurate modelling of the situation which would be encountered in breast screening centres when measuring performance.

The radiology literature details case studies of mammography workstation design but there are no measurements of participant behaviour or opinion, and the ergonomics literature details application of a range of measurement techniques in other fields. Therefore, the research presented here used these ergonomics techniques to investigate comfort at the mammography

workstations. An extensive review of methods for the measurement of workload and cancer detection performance was undertaken to ensure that the most appropriate methods were chosen.

Objective 2a was "to measure the impact of the display medium of the prior mammograms on physical comfort, and risk of musculoskeletal disorders in mammography readers". No differences were found in body part discomfort scores or RULA postural analysis risk scores between using film and digitised prior mammograms at the digital workstation. However, there was some evidence that the participants were simply turning their heads rather than moving closer to attain a better view of the film prior mammograms, and therefore, because of the distances involved, would be able to see less detail in the images than when they were viewed at a purely film workstation. This behavioural adaptation may be due to the workstation ergonomics.

Objective 2b was "to measure the impact of the display medium of the prior mammograms on mammography readers speed of reading and perceptions of workload". Through analysis of videotape of participants at a digital mammography workstation, speed of reading of normal mammograms was found to be 18% faster using digitised rather than film prior mammograms. NASA RTLX workload scores showed that perceived workload was higher using film rather than digitised prior mammograms.

Objective 2c was "to measure the impact of the display medium of the prior mammograms on the amount that the mammography readers use the prior

mammograms". Analysis of videotape of participants reading screening mammograms at a digital mammography workstation showed two effects of presentation medium of the prior mammograms. The number of comparisons to prior mammograms per case was higher using digitised rather than film prior mammograms. More significantly, the proportion of cases for which the prior mammograms were used was higher using digitised (82%) rather than film (63%) prior mammograms.

Objective 2d was "to measure the impact of the display medium of the prior mammograms on cancer detection performance". JAFROC analysis of performance showed no difference in performance when reading difficult digital cases with either film or digitised prior mammograms. In the experiment the proportion of cases for which the prior mammograms were used did not differ between using film (93%) and digitised (96%) prior mammograms, and therefore behaviour in the experiment did not mirror that of screening practice. Whilst performance in the experiment showed equivalence, in screening practice performance using digitised prior mammograms may be slightly better than when using film prior mammograms.

The third objective was "to determine whether the type of mammography reader (radiologist or radiography advanced practitioner) impacts on the metrics from objective 2". Whilst the performance of radiologists and radiography advanced practitioners did not differ, there was a trend towards an interaction between participant type and presentation medium of the prior mammograms ($p=.09$), with radiography advanced practitioners performing

better with digitised prior mammograms, and radiologists performing better with film prior mammograms. There was also a three way interaction between participant type, presentation medium of the prior mammograms and setting (screening practice or experiment) for a proportion of cases for which the prior mammograms were used ($p=.03$). Participants used the prior mammograms for a greater proportion of cases in screening practice when digitised rather than in film format, but this difference was not present in the experiment, and this difference between experiment and practice was greater for the radiography advanced practitioners than for the radiologists.

The fourth objective was to “test all findings by publishing in peer reviewed journals and presenting at both academic and practitioner conferences”. The research work has been presented at SPIE Medical Imaging, the Ergonomics Society Annual Conference, the UK Radiology Congress, Symposium Mammographicum, the Royal College of Radiology Breast Group, in the journal European Radiology, and at a training course about the introduction of digital mammography. Feedback from practitioners through conference attendance influenced both the data collection and analysis techniques for the performance experiment. The condition of not using prior mammograms was added through feedback from practitioners voicing the opinion that although there was a lot of research showing that prior mammograms improved cancer detection performance, they still intended to undertake the transition to digital mammography without them. Their reasoning for this decision was that it was not proven that analogue prior mammograms improved performance when the current mammograms were digitally acquired due to the differences

in appearance of the images. Another piece of feedback from practitioners was that my performance results were in single reader format, and the NHS Breast Screening Programme uses two readers with arbitration. The practitioners wondered whether the benefit of prior mammograms would still be present to such an extent with double reading. This inspired the model to convert the results from single reader to double reader with arbitration by iteration through all possible reader combinations.

The fifth and final objective was to publish guidance which will assist UK breast screening centres to decide how to display prior mammograms in the transition to digital mammography. A journal paper has been published in European Radiology and another has been submitted for peer review in the same journal. An additional five conference papers have been published, see appendix 9.

7.3 Choice of Methods

In carrying out any investigation decisions have to be made about which method(s) to use. Here the investigation into workstation ergonomics used postural analysis and body part discomfort scores. There were no differences between the RULA risk scores or body part discomfort scores when using film or digitised prior mammograms. This could either be because there were genuinely no differences, or because the research methods were insufficiently powerful, either statistically or in the research design. There were no significant increases in body part discomfort over the 45 minutes sessions

with the exception of the eyes, indicating that the session time was too short. Unfortunately the session time could not be lengthened as the measurements were of screening practice using live cases and therefore lengthening the sessions may have induced fatigue and would therefore be unethical. Body part discomfort scores taken over the course of a day or week would not have been useful as during that time period the participants would have read both digital and film cases, with both digitised and film prior mammograms, and participated in a range of other activities. The RULA postural analysis also found no difference in postural risk score using either film or digitised prior mammograms, and this tool is more sensitive than body part discomfort scores over such a short time period. This research could have been extended to investigate presenting the film prior mammograms in different positions such as above and behind the digital workstation, as both of these implementations are available commercially, and different display positions of the LCD screens which are adjustable both in tilt and vertically, or expanded to utilise a wider range of ergonomics methods. However, with little indication of ergonomics issues from the initial study, and interesting information about behaviour using film versus digitised prior mammograms, the research direction was diverted towards behavioural and performance studies.

Analysis of eye movements was conducted manually using videotape of live screening rather than using eye tracking equipment. This was significantly more labour intensive, and provided lower depth of information than using head mounted eye tracking equipment. Eye tracking equipment could provide more accurate information about number of fixations, and more detailed

information about saccades and fixation duration, but would have necessarily been more intrusive and therefore may have influenced the behaviour it was measuring. Head mounted eye tracking could not be implemented in screening practice for ethical reasons, and so could only be used on test cases which would be an approximation to real screening behaviour. Remote eye tracking would have been very unreliable with such a large workstation. The differences in behaviour found between experiment and screening practice indicate that the manual eye tracking using videotape was the more appropriate implementation here. The analysis of behaviour could have been extended to investigating different hanging protocols at the digital workstation, or use of the contrast and magnification workstation tools, but measurements of performance were prioritised.

The calculations of cost compare a bottom up approach with a top down approach which may be unfair. The calculations for the staffing costs of extra recalls are based on actual staffing levels in assessment clinics, and therefore are calculated in a top down manner. The cost of digitisation is calculated in a bottom up manner, with measurements of the time taken per case multiplied by the number of cases to be digitised. This may be unfair because in practice there are many other work activities which are accounted for in the top down but not the bottom up approach, such as breaks, talking to colleagues, and some administration tasks. A correction factor was applied to try to account for these differences. However to obtain a more accurate cost comparison digitisation of prior mammograms could be implemented in a breast screening centre, and therefore top down estimates of costs for digitisation obtained.

7.4 Limitations of the Study and Further Research

There were three main limitations to the research: that it was all completed at one breast screening centre; that the behaviour in the performance experiment did not mirror that of screening practice; and that there were flaws in the counterbalancing for the performance experiment.

The research was all conducted at one breast screening centre.

If the performance study had been carried out in several breast screening centres then the findings would have been generalisable to the population of readers. Coventry was the only breast screening centre in the UK at that time with an archive of digital mammography cases, Nottingham had digital mammography for the same time period but no storage capacity. Therefore the cases used were all acquired using the Sectra Microdose system, and a compatible system was required to display them. By the start of the performance experiment the breast screening centre in Manchester had a Sectra digital mammography workstation, and agreed to take part in the research. However they could not create a workstation with a multi-viewer for the film prior mammograms adjacent to the digital workstation due to the room sizes and floor plan. Therefore, whilst the research would have been improved by extension to other breast screening centres this was not possible. The participants, whilst not selected at random from the population of UK breast screening readers do represent a cross section of this population with a wide range of experience (3 to 18 years) and include both radiologists and radiography advanced practitioners. The performance of the study centre in terms of recall rate and cancer detection rate is typical for a UK screening

centre. Therefore there is no reason why results from the participants should be atypical of results from other UK breast screening centres. Therefore the research should be of use to other UK breast screening centres, and provides the only available evidence comparing use of film and digitised prior mammograms in digital mammography.

The behaviour in the performance experiment did not mirror that of screening practice.

It is reasonable to assume that this is a problem inherent in all ROC studies, as they are all weighted with a larger number of abnormal cases, and participants are always aware that their performance is being measured. To avoid this instead of an ROC study with test cases a clinical trial of digital mammography in screening practice could have been carried out. This would have had its own disadvantages, namely that confounding variables would be difficult to control, and it would have taken longer to carry out as it would take three months at the breast screening centre to encounter 60 cancerous cases, and therefore all three conditions would take at least nine months. There was an additional complication that at that time only one of the three screening vans at the study hospital was digital, the other three used film screen technology so potentially it could have taken up to 27 months if conducted at just one breast screening centre. If conducted across several breast screening centres then each of these would require a digitiser and a member of staff to do the digitisation, and therefore additional funding would have to be applied for as well as additional ethics and hospital trust agreements, which again would have taken increased time. Therefore, in the

circumstances a trial of digital mammography in screening practice was not a practical option in the circumstances, as the results would not have been delivered before 2010.

Counterbalancing

In the performance experiment the counterbalancing was thorough for the comparison between film and digital prior mammograms, however the sessions without the prior mammograms were all after the sessions with the prior mammograms. This was because in the original experimental design the condition of no prior mammograms was not planned to be included, because it was assumed that breast screening centres would not be considering implementation of the transition to digital mammography without prior mammograms. This assumption was based on the weight of evidence available showing that prior mammograms improve performance. The third condition of reading without prior mammograms was added after consultation with practitioners at conferences, and through contacts made when initiating the research. In hindsight, not including the condition of no prior mammograms in the planning stage, and therefore not including it in the counterbalancing was an error. However, there was no difference in performance between sessions 1 to 3 and sessions 4 to 6, and therefore the degradation in performance in sessions 7 to 9 is likely to be due to the absence of the prior mammograms rather than the counterbalancing order.

None of these three limitations could be remedied with further research which could provide results soon enough to influence the majority of UK screening

centres in the transition to digital mammography. However, there are three areas of further research of interest which have emerged from this research.

Firstly, the relationship between behaviour and performance in ROC studies in comparison to screening practice could be investigated further, so that this information could be used when applying the results of further ROC studies to screening practice. ROC studies are a significantly quicker and cheaper method of measuring performance than clinical trials, and therefore are likely to continue to form a large role in performance research. Therefore, a greater understanding of how ROC results relate to screening practice would be beneficial to the research community.

Secondly, the implementation of digital mammography in a UK breast screening centre to demonstrate how it would work in practice, and to measure departmental workflow changes and costs. One reason for this is there have been several breast screening centres interested in the research presented here, but are unsure how digital mammography would work in practice, and struggling to cope with the complexity of the implementation of digital mammography. A demonstration of digitising prior mammograms in practice would simplify the process for other breast screening centres.

Thirdly, there is new technology available which takes the digitised images of analogue mammograms, and makes them similar in appearance to digitally acquired mammograms. The intention of this is to ease comparisons between the digitally acquired current mammograms, and the digitised prior

mammograms. This technology should be tested to ensure that it improves the usability of prior mammograms, and is not a detriment to performance.

This research provides information to NHS practitioners so that they can make an informed choice about what to do with the analogue prior mammograms in the transition to digital mammography. The aim was to provide a large breadth of information in short timescales to inform this decision, and this aim has been achieved. Information covering workstation ergonomics, workload, speed of reading, use of prior mammograms and cancer detection performance has been reported. A clear recommendation that prior mammograms should be used for every case, and that digitisation is preferable to film display has been formulated, and the research and its conclusions disseminated widely.

8 **References**

- ANDERSSON, B.J., ORTENGREN, R., NACHEMSON, A. and ELFSTROM, G., 1974. Lumbar disc pressure and myoelectric back muscle activity during sitting. I. Studies on an experimental chair. *Scandinavian journal of rehabilitation medicine*, **6**(3), 104-114.
- ARO, A.R., PILVIKKI ABSETZ, S., VAN ELDEREN, T.M., VAN DER PLOEG, E. and VAN DER KAMP, L.J.T., 2000. False-positive findings in mammography screening induces short-term distress—breast cancer-specific concern prevails longer. *European journal of cancer*, **36**(9), 1089-1097.
- AUSTOKER, J., BERAL, V., BERRINGTON, A , BLANKS, R.G., DAY, N.E., DAY, T.J., ELLIS, I.O., FAULKNER, K., MØLLER, H., MOSS, S M , PATNICK, J., QUINN, M., WALLIS, M.G. and WILSON, A R.M., 2006. *Screening for Breast Cancer in England: Past and Future NHSBSP Publication No 61*. 61. Sheffield: NHS Cancer Screening Programmes.
- BADANO, A., CHAKRABORTY, D., COMPTON, K., CORNELIUS, C., CORRIGAN, K., FLYNN, M.J., HEMMINGER, B., HANGIANDREOU, N., JOHNSON, J., MOXLEY-STEVENS, D.M., PAVLICEK, W., ROEHRIG, H., RUTZ, L., SAMEI, E., SHEPARD, J., UZENOFF, R.A., WANG, J., WILLIS, C.E.,

2005, *Assessment of display performance for medical imaging systems*, [report], Maryland: American Association of Physicists in Medicine Task Group 18. Available at: http://deckard.mc.duke.edu/~samei/tg18_files/tg18.pdf [accessed on 19th September 2009]

BARLOW, H.B., 1952. Eye movements during fixation. *The Journal of physiology*, **116**(3), 290.

BERBAUM, K.S., FRANKEN, E.A., JR, DORFMAN, D.D., ROOHOLAMINI, S.A., KATHOL, M.H., BARLOON, T.J., et al., 1990, Satisfaction of search in diagnostic radiology, *Investigative Radiology*, **26**, 640-648.

BERBAUM, K.S., FRANKEN, E.A., JR, DORFMAN, D.D., ROOHOLAMINI, S.A., KATHOL, M.H., BARLOON, T.J., BEHLKE, F.M., SATO, Y., LU, C.H. and EL-KHOURY, G.Y., 1990. Satisfaction of search in diagnostic radiology. *Investigative radiology*, **25**(2), 133-140.

BERLIN, L., 2000. Liability of interpreting too many radiographs. *American Journal of Roentgenology*, **175**(1), 17-22.

BHARGAVAN, M. and SUNSHINE, J.H., 2002. Workload of radiologists in the United States in 1998-1999 and trends since 1995-1996. *American Journal of Roentgenology*, **179**(5), 1123-1128.

- BHATNAGER, V., DRURY, C.G. and SCHIRO, S.G., 1985. Posture, postural discomfort, and performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **27**(2), 189-199.
- BICK, U. and DIEKMANN, F., 2007. Digital mammography: what do we and what don't we know? *European radiology*, **17**(8), 1931-1942.
- BONGERS, P.M., DE WINTER, C.R., KOMPIER, M.A. and HILDEBRANDT, V.H., 1993. Psychosocial factors at work and musculoskeletal disease. *Scandinavian journal of work, environment & health*, **19**(5), 297-312.
- BORG, G., 1998. *Borg's perceived exertion and pain scales*, Illinois :Human Kinetics.
- BRETT, J. and AUSTOKER, J., 2001. Women who are recalled for further investigation for breast screening: psychological consequences 3 years after recall and factors affecting re-attendance. *Journal of Public Health*, **23**(4), 292-300.
- BURNSIDE, E.S., SICKLES, E.A., SOHLICH, R.E. and DEE, K.E., 2002. Differential value of comparison with previous examinations in diagnostic versus screening mammography. *American journal of roentgenology*, **179**(5), 1173-1177.
- BYERS, J., BITTNER, A. and HILL, S., 1989. Traditional and raw task load index (TLX) correlations: Are paired comparisons

necessary? *Advances in industrial ergonomics and safety*, **1**, 481–488.

CHAFFIN, D B , 1973. Localized muscle fatigue-definition and measurement. *Journal of occupational medicine.: official publication of the Industrial Medical Association*, **15**(4), 346-354.

CHAKRABORTY, D P., 1989. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Medical physics*, **16**(4), 561-568.

CHAKRABORTY, D.P. and WINTER, L.H., 1990. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, **174**(3 Pt 1), 873-881.

CHAKRABORTY, D P , 2006. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Physics in Medicine and Biology*, **51**(14), 3449-3462.

CHAKRABORTY, D.P., 2008a. Jackknife Free-Response Receiver Operating Characteristic Analysis Software, [computer software], Available at: www.devchakraborty.com [accessed on 1st July 2009].

CHAKRABORTY, D P., 2008b. Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Academic Radiology*, **15**(12), 1554-1566.

- CORLETT, E.N., 2005. Static muscle loading and the evaluation of posture.
In: WILSON, J.R., and CORLETT, E.N., 2005. *Evaluation of Human Work*, Florida: CRC Press. 3rd Edn. Ch.16.
- CORLETT, E.N. and BISHOP, R.P., 1976. A technique for assessing postural discomfort. *Ergonomics*, **19**(2), 175-182.
- CURTIS, L. and NETTEN, A., 2006. *Unit costs of health and social care*. Kent: University of Kent.
- DAS, B. and GRADY, R.M., 1983. The normal working area in the horizontal plane. A comparative analysis between Farley's and Squires' concepts. *Ergonomics*, **26**(5), 449-459.
- DEPARTMENT OF HEALTH, 2007. *Cancer Reform Strategy*. London: Department of Health.
- DEPARTMENT OF HEALTH, 2000. *The NHS Cancer plan: a plan for investment, a plan for reform*. London: Department of Health.
- DORFMAN, D.D. and ALF, E., 1969. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of mathematical psychology*, **6**(3), 487-496.
- DORFMAN, D.D., BERBAUM, K.S. and METZ, C.E., 1992. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investigative radiology*, **27**(9), 723-731.

- DRURY, C.G. and FRANCHER, M., 1985. Evaluation of a forward-sloping chair. *Applied Ergonomics*, **16**(1), 41-47.
- EGAN, J.P., GREENBERG, G.Z. and SCHULMAN, A I , 1961. Operating Characteristics, Signal Detectability, and the Method of Free Response. *The Journal of the Acoustical Society of America*, **33**, pp. 993.
- ELLIS, D.S., 1951 Speed of manipulative performance as a function of work-surface height. *The Journal of applied psychology*, **35**(4), 289-296.
- ESSERMAN, L., COWLEY, H., EBERLE, C., KIRKPATRICK, A., CHANG, S , BERBAUM, K. and GALE, A., 2002. Improving the Accuracy of Mammography: Volume and Outcome Relationships. *Journal of the National Cancer Institute*, **94**(5), 369-375
- FARLEY, R R., 1955. Some principles of methods and motion study as used in development work. *General Motors Engineering Journal*, **2**(6), 20-25.
- FERGUSON, S.A. and MARRAS, W.S., 1997. A literature review of low back disorder surveillance measures and risk factors. *Clinical biomechanics*, **12**(4), 211-226.
- FORREST, P., 1986. *Breast cancer screening: Report to the Health Ministers of England, Wales, Scotland and Northern Ireland*. London: HMSO.

- FRANCE, D.J., LEVIN, S., HEMPHILL, R., CHEN, K., RICKARD, D.,
MAKOWSKI, R., JONES, I. and ARONSKY, D., 2005.
Emergency physicians' behaviors and workload in the
presence of an electronic whiteboard. *International journal of
medical informatics*, **74**(10), 827-837.
- GELBERMAN, R.H., HERGENROEDER, P.T., HARGENS, A.R.,
LUNDBORG, G.N. and AKESON, W.H., 1981. The carpal
tunnel syndrome. A study of carpal canal pressures. *The
Journal of Bone and Joint Surgery*, **63**(3), 380-383.
- GELBERMAN, R H., YAMAGUCHI, K., HOLLSTIEN, S B , WINN, S.S.,
HEIDENREICH, F.P., BINDRA, R.R., HSIEH, P. and SILVA,
M.J., 1998. Changes in Interstitial Pressure and Cross-
Sectional Area of the Cubital Tunnel and of the Ulnar Nerve
with Flexion of the Elbow. An Experimental Study in Human
Cadavera. *The Journal of Bone and Joint Surgery*, **80**(4),
492-501.
- GOO, J.M., CHOI, J., IM, J., LEE, H J., CHUNG, M.J., HAN., D., PARK, S.H.,
KIM, J H., NAM, S., 2004, Effect of monitor luminance and
ambient light on observer performance in soft-copy reading
of digital chest radiographs, *Radiology*, **232**, 762-766.
- GUR, D., ROCKETTE, H.E., ARMFIELD, D R., BLACHAR, A., BOGAN, J.K.,
BRANCATELLI, G., 2003, Prevalence effect in a laboratory
environment, *Radiology*, **228**, 10-14.

- GUR, D., BANDOS, A.I., FUHRMAN, C.R., KLYM, A.H., KING, J.L., and
ROCKETTE, H.E., 2007, The prevalence effect in a
laboratory environment: Changing the confidence ratings,
Academic Radiology, **14**(1), 49-53
- GUR, D. and ROCKETTE, H.E., 2008. Performance Assessments of
Diagnostic Systems Under the FROC Paradigm:
Experimental, Analytical, and Results Interpretation Issues.
Academic Radiology, **15**(10), 1312-1315.
- HAMERMESH, D.S., 1990. Shirking or productive schmoozing: wages and
the allocation of time at work. *Industrial and Labor Relations
Review*, , 121-133.
- HANCOCK, P.A., 1996. Effects of control order, augmented feedback, input
device and practice on tracking performance and perceived
workload. *Ergonomics*, **39**(9), 1146-1162.
- HANCOCK, P.A. and SCALLEN, S.F., 1997. The performance and workload
effects of task re-location during automation. *Displays*, **17**(2),
61-68.
- HANKINS, T.C. and WILSON, G.F., 1998. A comparison of heart rate, eye
activity, EEG and subjective measures of pilot mental
workload during flight. *Aviation, Space, and Environmental
Medicine*, **69**(4), 360-367.

- HARISINGHANI, M.G., BLAKE, M.A., SAKSENA, M , HAHN, P.F., GERVAIS, D., ZALIS, M., DA SILVA DIAS FERNANDES, L. and MUELLER, P.R., 2004. Importance and effects of altered workplace ergonomics in modern radiology suites. *Radiographics*, **24**(2), 615-627.
- HART, S.G., HAUSER, J.R. and LESTER, P.T., 1984. Inflight evaluation of four measures of pilot workload. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, **28** (11), 945-949.
- HART, S.G. and STAVELAND, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: P.A. HANCOCK and N. MESHKATI, *Human mental workload*. Oxford, England: North-Holland, 139-183.
- HEALTH AND CONSUMER PROTECTION DIRECTORATE, 2006. European guidelines for quality assurance in breast cancer screening and diagnosis, Fourth edition, Belgium: Office for Official Publications of the European Communities.
- HEALTH AND SAFETY EXECUTIVE, 2004. *Getting to Grips with Manual Handling: A Short Guide*. INDG143(rev2). Sudbury: Health and Safety Executive.
- HEINSALMI, P., 1986. Method to measure working posture loads at working sites (OWAS), *The Ergonomics of Working Postures: Models, Methods and Cases: the Proceedings of the First*

*International Occupational Ergonomics Symposium, 15-17
April 1985, Zadar, Yugoslavia: CRC Press, pp. 100.*

HENDY, K.C., HAMILTON, K.M. and LANDRY, L N., 1993. Measuring
subjective workload: when is one scale better than many?
Human factors, **35**(4), 579-601.

HIGNETT, S. and MCATAMNEY, L , 2000. Rapid entire body assessment
(REBA). *Applied Ergonomics*, **31**(2), 201-205.

HILL, S.G., IAVECCHIA, H.P., BITTNER JR, A.C., BYERS, J.C., ZAKLAD,
A.L. and CHRIST, R.E., 1992. Comparison of four subjective
workload rating scales. *Human factors*, **34**(4), 429-439.

INTERNATIONAL ORGANISATION FOR STANDARDISATION, 1998.
*ISO 9241-5: Ergonomic requirements for office work with
visual display terminals (VDTs) – Part 5: Workstation layout
and postural requirements*. Geneva: International
Organisation for Standardisation.

JACHINSKI, W, HEUER, H., and KYLIAN, H., 1998, Preferred position of
visual displays relative to the eyes: A field study of visual
strain and individual differences, *Ergonomics*, **41**(7), 1034-
1049.

KAN, L., OLIVOTTO, I A., WARREN BURHENNE, L.J., SICKLES, E A. and
COLDMAN, A.J., 2000. Standardized Abnormal
Interpretation and Cancer Detection Ratios to Assess

Reading Volume and Reader Performance in a Breast
Screening Program 1. *Radiology*, **215**(2), 563-567.

KILBOM, Å , PERSSON, J. and JONSSON, B.G., 1986. Disorders of the
cervicobrachial region among female workers in the
electronics industry. *International Journal of Industrial
Ergonomics*, **1**(1), 37-47.

KROEMER, K.H.E. and GRANDJEAN, E., 2005. *Fitting the Task to the
Human: A Textbook of Occupational Ergonomics*. 5th edn.
Great Britain:Taylor & Francis.

KRUPINSKY, E A. and NODINE, C.F., 1994. Gaze duration predicts the
location of missed lesions in mammography, *Digital
Mammography: Proceedings of the 2nd International
Workshop on Digital Mammography*, 10-12 July 1994, York,
England:Elsevier Science Ltd, pp. 399.

KRUPINSKI, E., JOHNSON, J., ROEHRIG. H., LUBIN, J., 2003, Using a
human visual system model to optimize soft-copy
mammography display: Influence of display phosphor,
Academic Radiology, **10**, 161-166

KUMAR, S., 2001. Theories of musculoskeletal injury causation. *Ergonomics*,
44(1), 17-47.

LAWSON, E H., CURET, M.J., SANCHEZ, B.R., SCHUSTER, R., BERGUER,
R., 2007, Postural ergonomics during robotic and

laparoscopic gastric bypass surgery: a pilot project, *Journal of Robotic Surgery*, 1, 61-67

LEGOOD, R. and GRAY, A., 2004. *A Cost Comparison of Full Field Digital Mammography (FFDM) with Film-Screen Mammography in Breast Cancer Screening*. NHSBSP Equipment Report 0403. Sheffield: NHS Cancer Screening Programmes.

LEVIN, S., FRANCE, D.J., HEMPHILL, R., JONES, I., CHEN, K Y., RICKARD, D., MAKOWSKI, R. and ARONSKY, D., 2006. Tracking workload in the emergency department. *Human factors*, 48(3), pp. 526.

LEWIN, J.M., D'ORSI, C.J., HENDRICK, R.E., MOSS, L.J., ISAACS, P.K , KARELLAS, A. and CUTTER, G.R., 2002. Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. *American Journal of Roentgenology*, 179(3), 671-677.

LISTON, J , WILSON, R., COOKE, J., DUNCAN, K , GIVEN-WILSON, R., KUTT, E., MICHELL, M , PATNICK, J., THOMPSON, W., WALLIS, M. and WILSON, M., 2005. *Quality Assurance Guidelines for Breast Cancer Screening Radiology*. 59. Sheffield: NHS Cancer Screening Programmes.

LOMAS, S.M., 1998, An investigation into the postures adopted by the upper limbs in cervical screening, In: *Global ergonomics: proceedings of the Ergonomics Conference*, Cape Town,

Scott, P. A. ,Bridger, R.S., Charteris, J (Eds), Oxford:Elsevier Science Ltd, 353-356.

LORD, M P. and WRIGHT, W D., 1950. The investigation of eye movements. *Reports on Progress in Physics*, **13**, 1-23.

MACNICOL, M.F., 1982. Extraneural pressures affecting the ulnar nerve at the elbow. *Journal of Hand Surgery (European Volume)*, **14**(1), 5-11.

MAY, J.L., GALE, A G., HASLEGRAVE, J., CASTLEDINE, J and WILSON, A R.M., 1994. Musculoskeletal problems in breast screening radiographers. *Contemporary Ergonomics*, , 247-252.

MAY, J.L., GALE, A G., 1998, Mammography - Painful for whom?, In: Global ergonomics: proceedings of the Ergonomics Conference, Cape Town, Scott, P. A. ,Bridger, R.S., Charteris, J.(Eds), Oxford:Elsevier Science Ltd, 291-294

MAYES, D.K., SIMS, V.K. and KOONCE, J.M., 2001. Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, **28**(6), 367-378.

MCATAMNEY, L. and CORLETT, E.N., 1993. RULA: a survey method for the investigation of work-related upper limb disorders. *Applied Ergonomics*, **24**(2), 91-99.

McENTEE, M., BRENNAN, P., EVANOFF, M., PHILLIPS, P., O'CONNOR, W.T., MANNING, D., 2006, Optimum ambient lighting conditions for the viewing of softcopy radiological images, *in Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, 2006*, Jiang, Y., and Eckstein, M.P., (eds), Proceedings of SPIE, 6146, (SPIE, Bellingham, WA, 2006), 6146-31.

MELLO-THOMS, C., 2006. How does the perception of a lesion influence visual search strategy in mammogram reading? *Academic Radiology*, **13**(3), 275-288.

MILLER, D.P., O'SHAUGHNESSY, K.F., WOOD, S.A. and CASTELLINO, R.A., 2004. Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions, *in Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, Chakraborty, D.P., and Eckstein, M.P. (eds), Proceedings of SPIE, 5372, (SPIE, Bellingham, WA, 2006), 173-184.

MOHAMMED, O., SHELL, R., SWANSON, N., 1999, Investigating the postural stress associated with sedentary work, in: *Advances in occupational ergonomics and safety*, Lee G.C.H. (Ed), Amsterdam: IOS press, 401-406

- NAGY, P , SIEGEL, E., HANSON, T., KREINER, L., JOHNSON, K. and REINER, B., 2003. PACS reading room design. *Seminars in roentgenology*, **38**(3), 244-255.
- NODINE, C.F. and KUNDEL, H.L., 1987. Using eye movements to study visual search and to improve tumor detection. *Radiographics*, **7**(6), 1241-1250.
- NYGREN, T.E., 1991. Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. *Human factors*, **33**(1), 17-33.
- OBUCHOWSKI, N A., LIEBER, M.L. and POWELL, K.A., 2000. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Academic Radiology*, **7**(7), 516-525.
- O'DONNELL, R.D. and EGGEMEIER, F.T., 1986. Workload assessment methodology. In: K. BOFF, L. KAUFMAN and J. THOMAS, eds, *Handbook of Perception and Human Performance*. pp. 42.
- OVERINGTON, I., 1976. *Vision and acquisition*. London: Pentech Press.
- PHALEN, G.S., 1966. The carpal-tunnel syndrome. Seventeen years' experience in diagnosis and treatment of six hundred fifty-

four hands. *The Journal of Bone and Joint Surgery*, **48**(2), 211-228.

PHEASANT, S. and HASLEGRAVE, C.M., 2006. *Bodyspace: anthropometry, ergonomics, and the design of work*. Florida: CRC Press.

PISANO, E D., GATSONIS, C., HENDRICK, E., YAFFE, M., BAUM, J.K., ACHARYYA, S , CONANT, E.F., FAJARDO, L.L., BASSETT, L. and D'ORSI, C., 2005. Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, **353**(17), 1773-1783.

PISANO, E.D., COLE, E.B., KISTNER, E.O., MULLER, K.E., HEMMINGER, B M., BROWN, M.L., JOHNSTON, R E., KUZMIAK, C.M., BRAEUNING, M.P., FREIMANIS, R I., SOO, M.S., BAKER, J A. and WALSH, R., 2002. Interpretation of digital mammograms: comparison of speed and accuracy of soft-copy versus printed-film display. *Radiology*, **223**(2), 483-488.

PUNNETT, L., FINE, L.J., KEYSERLING, W.M., HERRIN, G D. and CHAFFIN, D.B., 1991. Back disorders and nonneutral trunk postures of automobile assembly workers. *Scandinavian journal of work, environment & health*, **17**(5), 337-346.

PUTZ-ANDERSON, V., BERNARD, B.P., BURT, S.E., COLE, L.L., FAIRFIELD-ESTILL, C., FINE, L.J., GRANT, K.A., GJESSING, C., JENKINS, L., HURRELL JR., J.J., NELSON, N., PFIRMAN, D., ROBERTS, R., STETSON, D., HARING-

SWEENEY, M. and TANAKA, S., 1997. *Musculoskeletal disorders and workplace factors: a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back*. Cincinnati: National Institute for Occupational Safety and Health (NIOSH).

RATIB, O., VALENTINO, D.J., MCCOY, M.J., BALBONA, J A., AMATO, C.L. and BOOTS, K., 2000. Computer-aided design and modeling of workstations and radiology reading rooms for the new millennium. *Radiographics*, 20(6), 1807-1816.

REID, G.B. and NYGREN, T.E., 1988. The Subjective Workload Assessment Technique. In: Hancock, P.A. and Meshkati, N., *Human Mental workload*, Amsterdam:Elsevier Science Publishers B.V. ch 8.

REMPEL, D., BACH, J.M., GORDON, L. and SO, Y., 1998. Effects of forearm pronation/supination on carpal tunnel pressure. *Journal of Hand Surgery*, 23(1), 38-42.

REVESZ, G., KUNDEL, H L. and BONITATIBUS, M., 1983. The effect of verification on the assessment of imaging techniques. *Investigative radiology*, 18(2), 194-8.

ROELOFS, A.A., KARSSEMEIJER, N., WEDEKIND, N., BECK, C., VAN WOUDENBERG, S., SNOEREN, P.R., HENDRIKS, J.H., ROSSELLI DEL TURCO, M., BJURSTAM, N.,

JUNKERMANN, H., BEIJERINCK, D., SERADOUR, B. and
EVERTSZ, C.J., 2007. Importance of comparison of current
and prior mammograms in breast cancer screening.
Radiology, **242**(1), 70-77.

ROSCOE, A.H., 1992. Assessing pilot workload. Why measure heart rate,
HRV and respiration? *Biological psychology*, **34**(2-3), 259-
287.

RUESS, L., O'CONNOR, S C., CHO, K.H., HUSSAIN, F.H., HOWARD,
W J.,3RD, SLAUGHTER, R.C. and HEDGE, A., 2003. Carpal
tunnel syndrome and cubital tunnel syndrome: work-related
musculoskeletal disorders in four symptomatic radiologists.
American journal of roentgenology, **181**(1), 37-42.

SAKAKIBARA, H., MIYAO, M., KONDO, T A. and YAMADA, S.Y., 1995.
Overhead work and shoulder-neck pain in orchard farmers
harvesting pears and apples. *Ergonomics*, **38**(4), 700-706.

SALTHOUSE, T. A , ELLIS, C.L., 1980, Determinants of eye-fixation duration,
American Journal of Psychology, **93**(2), 207-234.

SAMUEL, S, KUNDEL, H.L., NODINE, C.F., TOTO, L.C., 1995, Mechanism of
satisfaction of search: Eye position recordings in the reading
of chest radiographs, *Radiology*, **194**, 895-902

- SAUTER, S.L., SCHLEIFER, L.M. and KNUTSON, S.J., 1991. Work posture, workstation design, and musculoskeletal discomfort in a VDT data entry task. *Human factors*, **33**(2), 151-167.
- SCOTT, H.J. and GALE, A.G., 2006. Breast screening: PERFORMS identifies key mammographic training needs. *British Journal of Radiology*, **79**(Special Issue 2), 127-133.
- SENGUPTA, A.K. and DAS, B., 2004. Determination of worker physiological cost in workspace reach envelopes. *Ergonomics*, **47**(3), 330-342.
- SIDDIQUI, K.M., CHIA, S., KNIGHT, N. and SIEGEL, E.L., 2006. Design and Ergonomic Considerations for the Filmless Environment. *Journal of the American College of Radiology*, **3**(6), 456-467.
- SIEGEL, E. and REINER, B., 2002. Radiology reading room design: the next generation. *Applied Radiology*, **31**(4), 11-16.
- SIEGEL, S. and CASTELLAN, N.J., 1988. *Nonparametric statistics for the behavioural sciences*. 2nd edn. Singapore: McGraw Hill Inc.
- SKAANE, P. and SKJENNALD, A., 2004. Screen-film mammography versus digital mammography with soft-copy reading: randomized trial in a population based study - the Oslo II study. *Radiology*, **232**, 197-204.
- SKAANE, P., YOUNG, K. and SKJENNALD, A., 2003. Population-based mammography screening: comparison of screen-film and

full-field digital mammography with soft-copy reading-Oslo I study. *Radiology*, **229**(3), 877-884.

SKOTTE, J.H., NOJGAARD, J.K., JORGENSEN, L.V., CHRISTENSEN, K.B. and SJOGAARD, G., 2007. Eye blink frequency during different computer tasks quantified by electrooculography. *European journal of applied physiology*, **99**(2), 113-119.

SMITH-BINDMAN, R., CHU, P.W., MIGLIORETTI, D.L., SICKLES, E.A , BLANKS, R , BALLARD-BARBASH, R., BOBO, J.K., LEE, N.C., WALLIS, M.G. and PATNICK, J., 2003. Comparison of screening mammography in the United States and the United Kingdom. *The Journal of the American Medical Association*, **290**(16), 2129-2137.

SMITH-BINDMAN, R., BALLARD-BARBASH, R., MIGLIORETTI, D.L., PATNICK, J. and KERLIKOWSKE, K., 2005. Comparing the performance of mammography screening in the USA and the UK. *Journal of medical screening*, **12**(1), 50-54.

STEINKE, L., LANFEAR, D.E., DHANAPAL, V. and KALUS, J.S., 2009. Effect of "Energy Drink" Consumption on Hemodynamic and Electrocardiographic Parameters in Healthy Young Adults. *The Annals of Pharmacotherapy*, **43**(4), 596.

STERN, J.A. and SKELLY, J.J., 1984. The eyeblink and workload considerations. *Proceedings of the human factors society*,

- STRAKER, L. and MEKHORA, K., 2000. An evaluation of visual display unit placement by electromyography, posture, discomfort and preference. *International Journal of Industrial Ergonomics*, **26**(3), 389-398.
- STRAKER, L.M., POLLOCK, C.M. and MANGHARAM, J.E., 1997. The effect of shoulder posture on performance, discomfort and muscle fatigue whilst working on a visual display unit. *International Journal of Industrial Ergonomics*, **20**(1), 1-10.
- SUMKIN, J.H., HOLBERT, B.L., HERRMANN, J.S., HAKIM, C.A., GANOTT, M.A., POLLER, W.R., SHAH, R., HARDESTY, L.A. and GUR, D., 2003. Optimal reference mammography: a comparison of mammograms obtained 1 and 2 years before the present examination. *American journal of roentgenology*, **180**(2), 343-346.
- SWENSSON, R G., 1996. Unified measurement of observer performance in detecting and localizing target objects on images. *Medical physics*, **23**(10), 1709-1725.

SZABO, R.M. and CHIDGEY, L.K., 1989. Stress carpal tunnel pressures in patients with carpal tunnel syndrome and normal patients. *The Journal of hand surgery*, 14(4), 624-627.

TAYLOR-PHILLIPS, S., WALLIS, M.G., DUNCAN, A. and GALE, A.G., 2009. Should previous mammograms be digitised in the transition to digital mammography? *European Radiology*, 19(8), 1890-1896.

THE HEALTH AND SOCIAL CARE INFORMATION CENTRE, 2009. *Breast Screening Programme, England 2007-08*. Leeds, UK: The NHS Information Centre.

THE NHS STAFF COUNCIL, 2009. *Pay Circular (AforC) 1/2009*. NHS terms and conditions of service handbook Amendment number 13. S.I.:NHS Employers.

THURFJELL, M.G., VITAK, B., AZAVEDO, E., SVANE, G. and THURFJELL, E., 2000. Effect on Sensitivity and Specificity of Mammography Screening with or Without Comparison of Old Mammograms. *Acta radiologica*, 41(1), 52-56.

TSUBOTA, K. and NAKAMORI, K., 1995. Effects of ocular surface area and blink rate on tear dynamics. *Archives of Ophthalmology*, 113(2), 155-158.

- TUDDENHAM, W.J., 1962, Visual search, image organisation, and reader error in roentgen diagnosis: studies of the psychophysiology of roentgen image perception, *Radiology*, 78, 694-704
- VARELA, C., KARSSEMEIJER, N., HENDRIKS, J.H.C L. and HOLLAND, R., 2005. Use of prior mammograms in the classification of benign and malignant masses. *European Journal of Radiology*, 56(2), 248-255.
- VELTMAN, J.A. and GAILLARD, A.W.K., 1996. Physiological indices of workload in a simulated flight task. *Biological psychology*, 42(3), 323-342.
- VIDULICH, M., 1986. Techniques of subjective workload assessment: a comparison of s. w. a. t. [subjective workload assessment technique] and the NASA-Bipolar methods. *Ergonomics*, 29, 1385-1398.
- VIDULICH, M.A. and TSANG, P.S , 1986. Techniques of subjective workload assessment: a comparison of SWAT and the NASA-Bipolar methods. *Ergonomics*, 29(11), 1385-1398.
- VIDULICH, M A. and WICKENS, C.D., 1985. Causes of dissociation between subjective workload measures and performance- Caveats for the use of subjective assessments, 3rd Symposium on Aviation Psychology, Ohio:Wright State University, 22-25th April 1985, 223-230.

- WARM, J.S., DEMBER, W.M., HANCOCK, P.A , 1996, Vigilance and Workload in Automated Systems, In: Parasuraman, R., and Mouloua, M , *Automation and human performance: theory and applications*, Philadelphia: Lawrence Erlbaum Associates.
- WERNER, R., ARMSTRONG, T.J., BIR, C. and AYLARD, M.K., 1997. Intracarpal canal pressures: the role of finger, hand, wrist and forearm position. *Clinical Biomechanics*, **12**(1), 44-51.
- WICKENS, C.D., 1992. *Engineering psychology and human performance*. 2nd edn. New York: Harper Collins.
- WICKENS, C.D., 1991. Processing resources and attention. In: D.L. DAMOS, ed, *Multiple Task Performance*. Basingstoke: Taylor & Francis, pp. 3-34.
- WIERWILLE, W.W. and EGGEMEIER, F T., 1993. Recommendations for mental workload measurement in a test and evaluation environment. *Human factors*, **35**(2), 263-281.
- YARBUS, A.L., 1967, *Eye Movements and Vision*, New York: Plenum Press.
- YEH, Y.Y. and WICKENS, C.D., 1988. Dissociation of performance and subjective measures of workload. *Human Factors*, **30**(1), 111-120.

Appendix 1 – Task Analysis of Screening at the Study Hospital

1.
Breast
Screening

Plan 1

Do 1,2,3,4, then 5,6 for film screen mammography,7,8

1.
Organise
Screening
clinic

2.
Prepare S X
bags for
women
attending
screening

3.
Radiographers
take
mammograms

4
Image
Processing

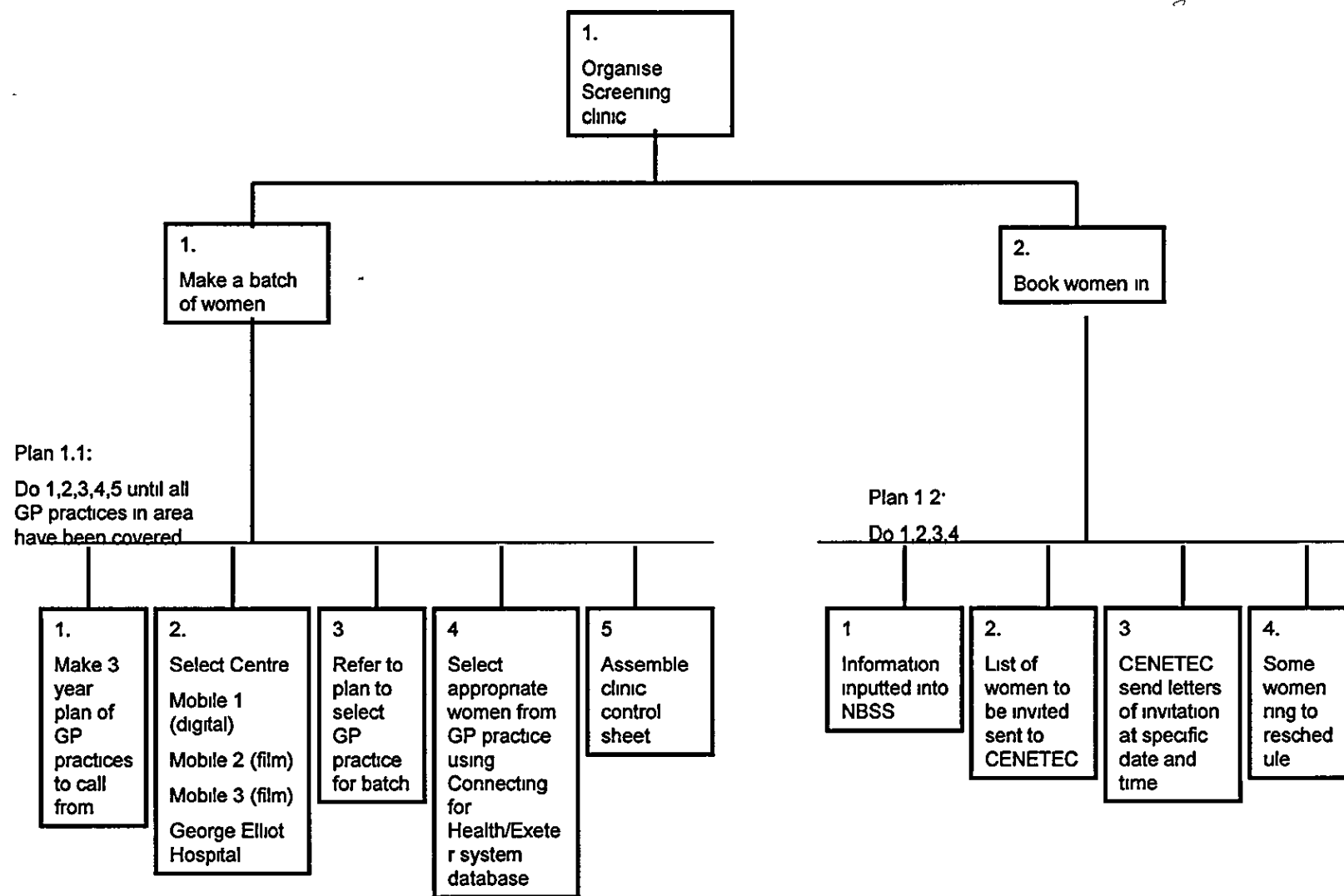
5.
Prepare batch
for hanging

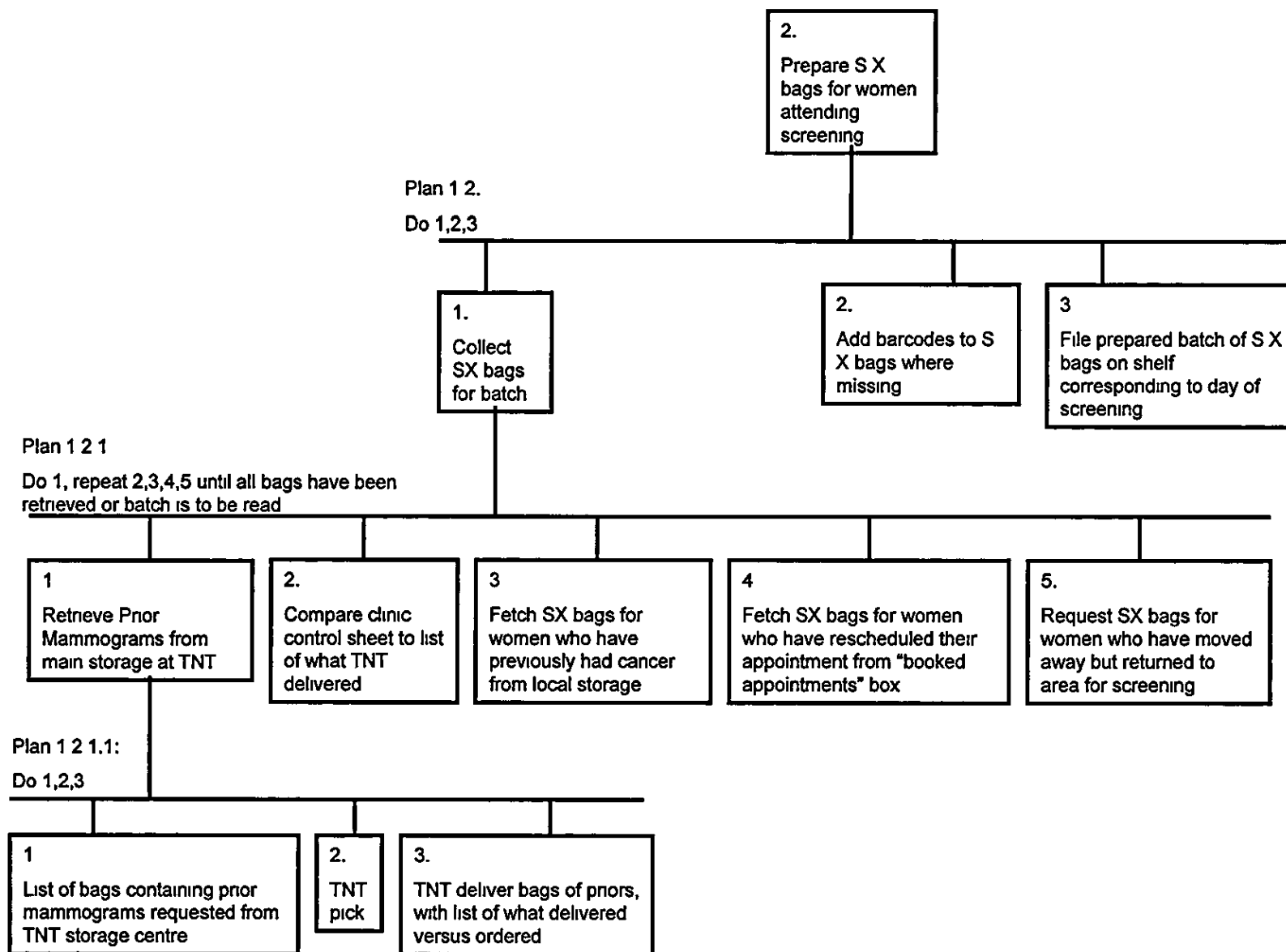
6
Hang Batch

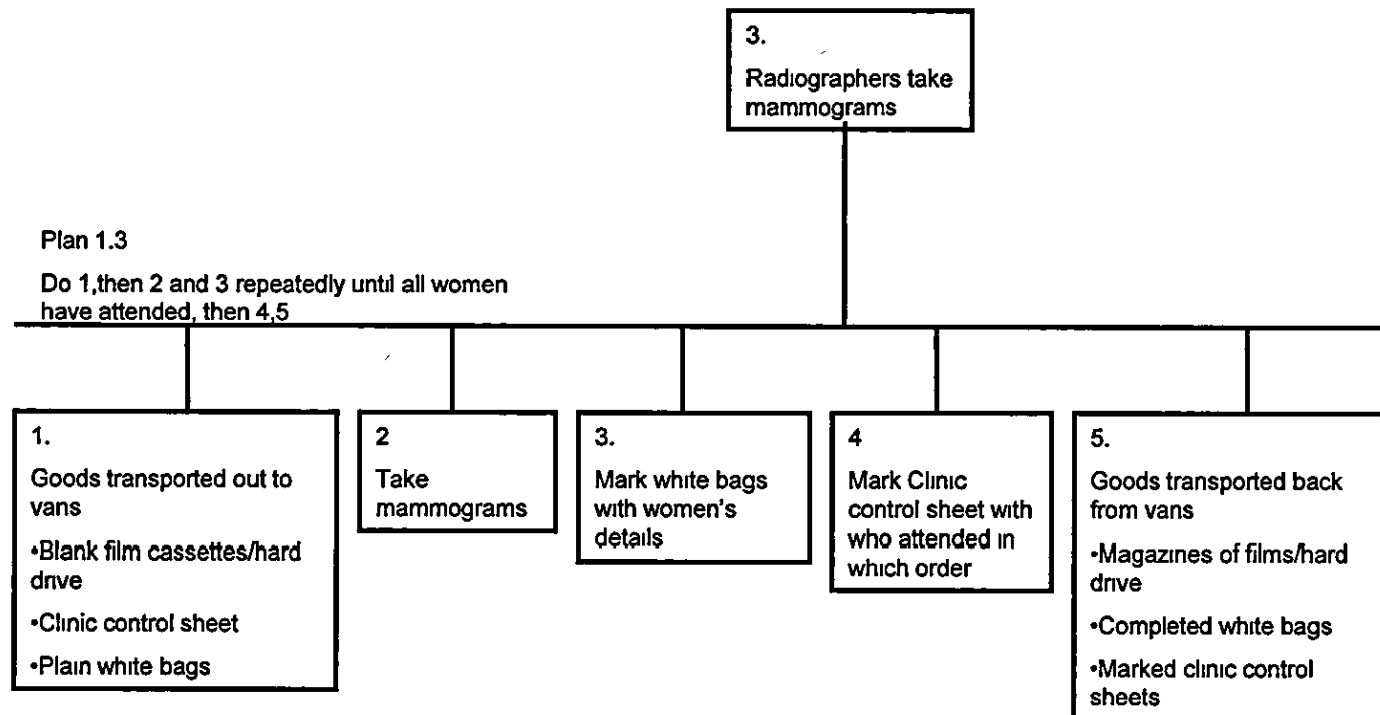
7
Read a batch

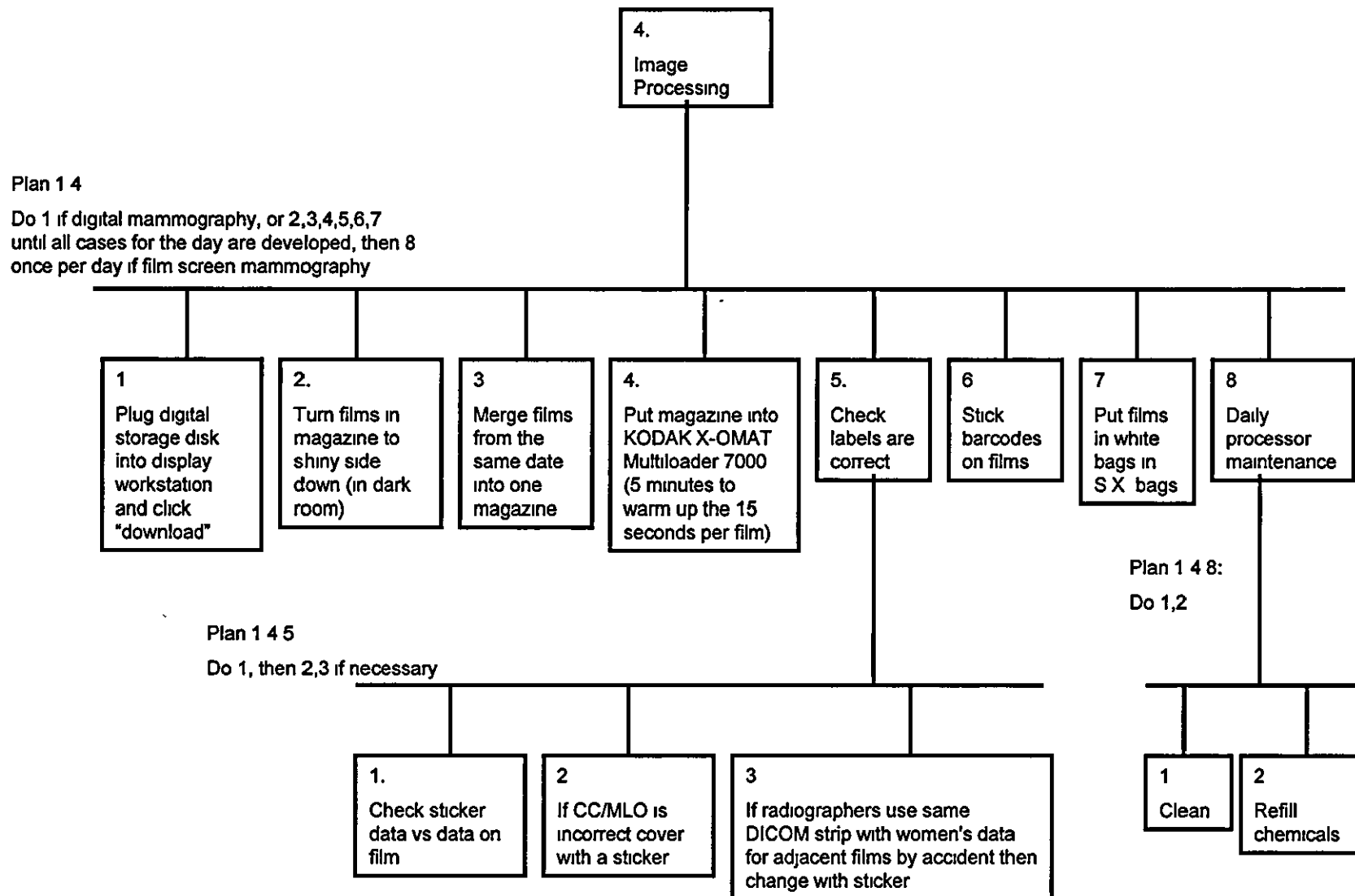
8
Send women
results
(normal/recall for
further tests)

Maximum 2 weeks



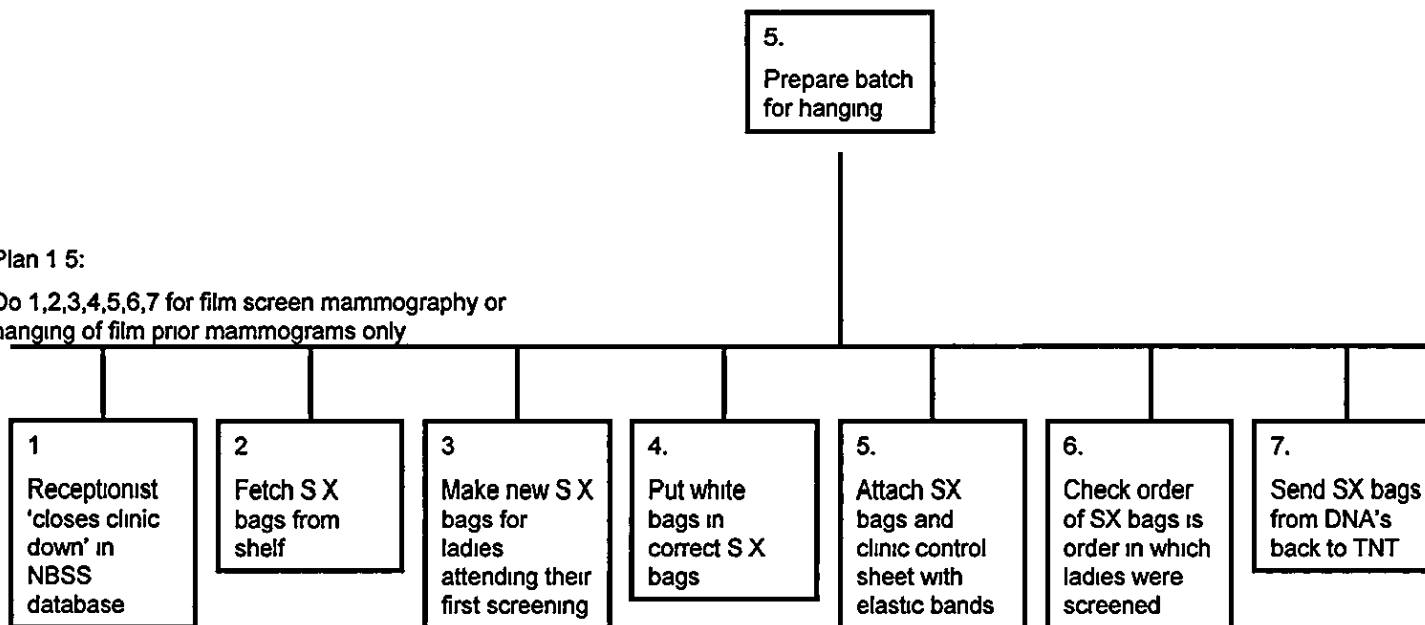






Plan 1 5:

Do 1,2,3,4,5,6,7 for film screen mammography or hanging of film prior mammograms only





Plan 1 6

Do 1,2,3,then 4,5 until all cases in batch are hung, then 6 for film screen mammography or hanging of film prior mammograms only

1.
Put folder with
randomly assigned
batch number at
front

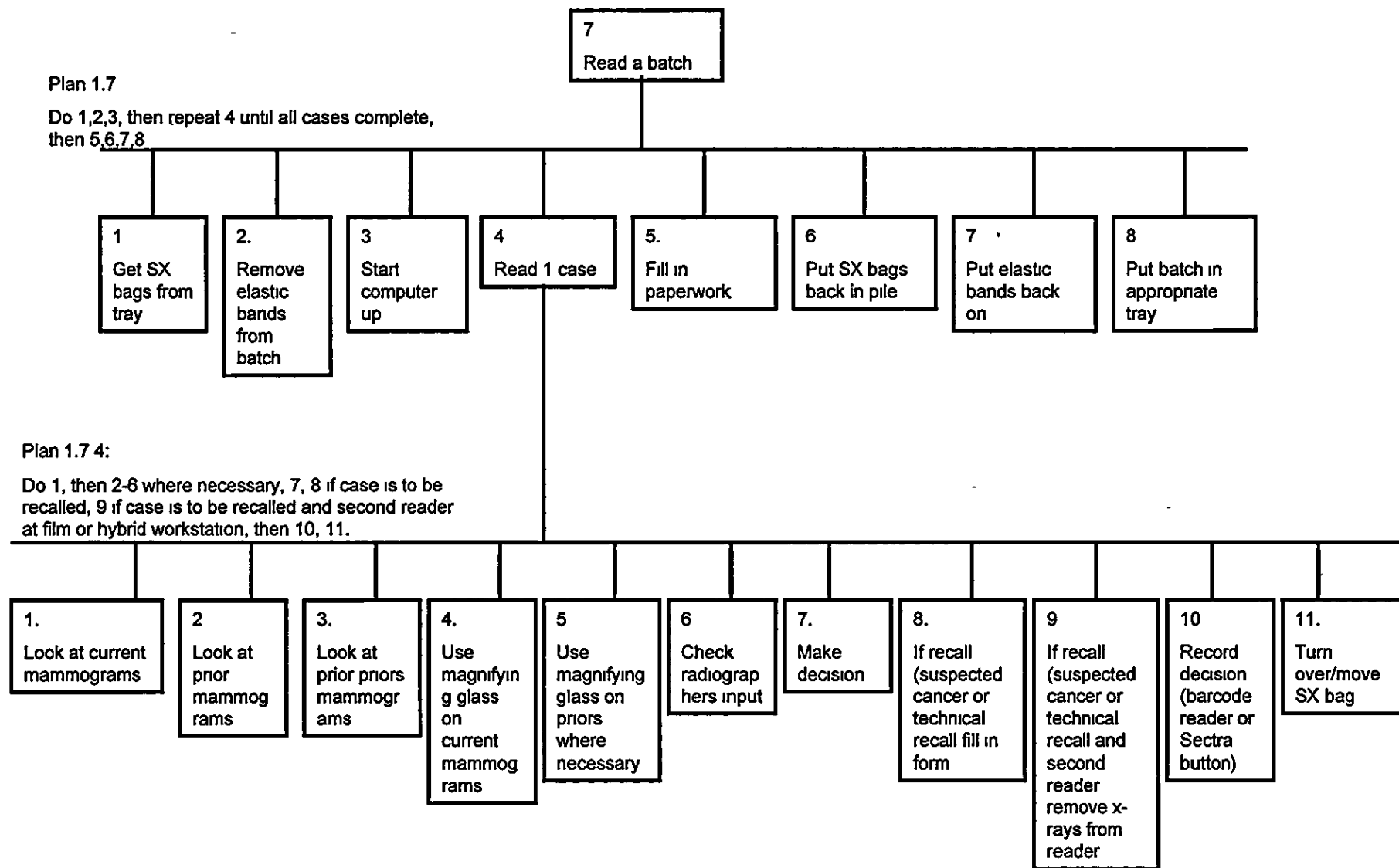
2.
Fill in reporting
flowchart sheet,
and reporting
flowchart wall chart

3.
Put reporting
flowchart sheet,
and clinic control
sheet, in batch
folder

4
Hang films on
viewer
Top row current
images RCC,
RMLO, LMLO,
LCC Bottom row
prior images RCC,
RMLO, LMLO, LCC

5.
Move viewer to
next position

6.
Return viewer
position to start of
batches loaded



Appendix 2 - Participant Information Sheet and Informed Consent Form

Maintaining Optimal Health and Performance of Radiologists in the Transition from Film to Digital Mammography in the NHS Breast Screening Programme

PARTICIPANT INFORMATION SHEET

The aim of this study is to understand how radiologists and level four advanced practitioners will be affected by the change from reading film to digital mammography images.

Taking part in the study will involve the following

- You will be video recorded whilst carrying out your normal work reading mammographic images for a period of 2 hours at each of 3 workstations. The filming will be by fixed video cameras located in unobtrusive places in the radiology reading room. The video will be used solely to analyse the postures adopted by you at each workstation. Video data will be treated as confidential, it will be securely stored for 6 years after the end of the investigation at which point it will be destroyed. Anonymised stills/clips from the video will be used in publications only with your prior written consent.
- You will be asked to answer a series of questions before and after one of your regular 1 hour reading sessions on 2 separate occasions at each of 3 different workstations. These questions will relate to comfort and workload. This will take around 5 minutes on each occasion.
- You will be asked to complete a one hour session at each of 3 workstations. This will involve reading a set of known cases for the first half hour whilst wearing light-weight head mounted eye tracking equipment, and explaining to the investigator the methods used to read the test cases for the second half hour. This is to understand methods and approaches used at the different workstations.

The total time commitment for this study is less than 5 hours spread over several months

All data will be anonymised immediately after data collection

You have the right to withdraw from this study at any stage for any reason, and you will not be required to explain your reasons for withdrawing.

Researcher contact details

Sian Taylor-Phillips

Tel: 07725000262

Email: s.phillips2@lboro.ac.uk

Project Supervisor: Prof. Alastair Gale

Tel: 01509635703

Email: a.g.gale@lboro.ac.uk

Maintaining Optimal Health and Performance of Radiologists in the Transition from Film to Digital Mammography in the NHS Breast Screening Programme

INFORMED CONSENT FORM

(to be completed after Participant Information Sheet has been read)

The purpose and details of this study have been explained to me. I understand that this study is designed to further scientific knowledge and that all procedures have been approved by the Loughborough University Ethical Advisory Committee, and the Caldecott Guardian at Coventry Hospital.

- I have read and understood the information sheet and this consent form.
- I have had an opportunity to ask questions about my participation.
- I understand that I am under no obligation to take part in the study.
- I understand that I have the right to withdraw from this study at any stage for any reason, and that I will not be required to explain my reasons for withdrawing.
- I understand that all the information I provide will be treated in strict confidence.
- I understand that all data including video and eye tracking data will be anonymised.
- I agree to participate in this study.

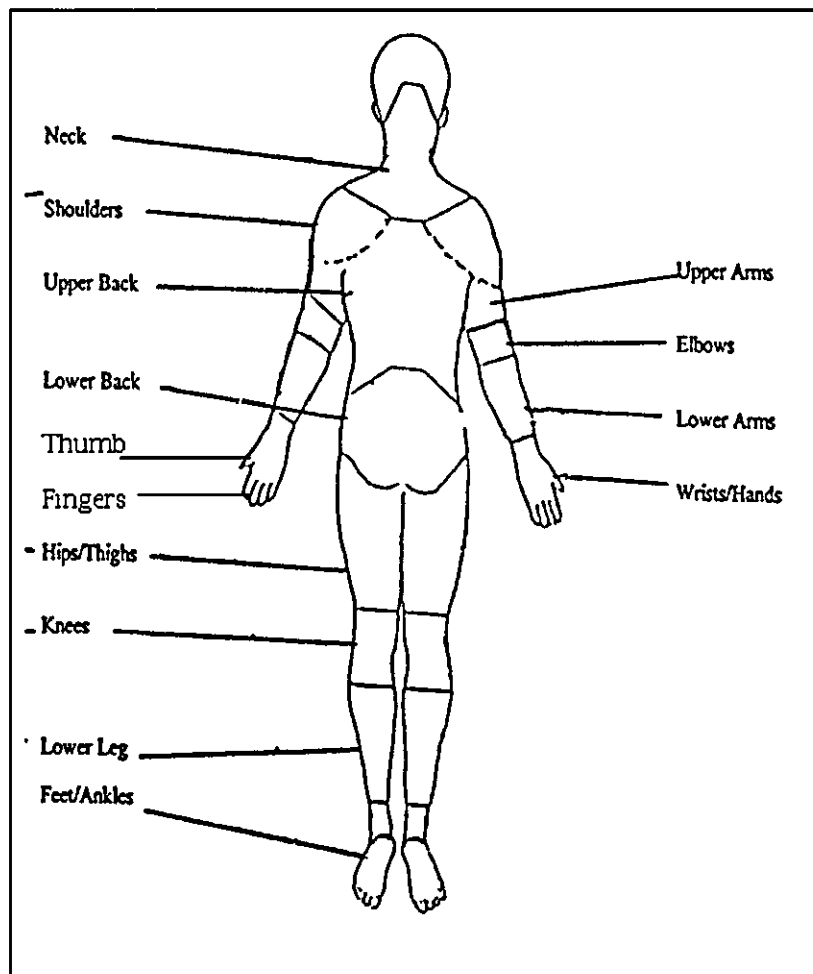
Your name _____

Your signature _____

Signature of investigator _____

Date _____

Body Part Discomfort Chart



Name _____

Date _____ Time _____

Please state the level of discomfort you currently feel in each of the following body parts from 1 to 5 (circle as appropriate)

1=no discomfort,
2=very mild discomfort,
3=mild discomfort,
4=moderate discomfort,
5=severe discomfort

Neck	1	2	3	4	5
Shoulders	1	2	3	4	5
Upper Back	1	2	3	4	5
Elbows	1	2	3	4	5
Low Back	1	2	3	4	5
Wnsts/Hands	1	2	3	4	5
Fingers	1	2	3	4	5
Thumb	1	2	3	4	5
Hips/Thighs	1	2	3	4	5
Head	1	2	3	4	5
Eyes (dry, burning, or sore at front surface)	1	2	3	4	5
Eyes (aching at back or middle)	1	2	3	4	5

Appendix 4 - NASA TLX Workload Questionnaire

Please place a mark on the scale to represent the magnitude of each of the following factors in the task you just performed.

Mental Demand: How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exact or forgiving?



Physical Demand: How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?



Temporal Demand: How much time pressure did you feel due to the rate or pace at which tasks or task elements occurred? Was the pace slow and leisurely or rapid and frenetic?



Performance: How successful do you think that you were in accomplishing the goals of the task? How satisfied were you with your performance in accomplishing these goals?



Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?



Frustration Level: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?



Appendix 5 – Normality Tests for Workload and Time Taken per Case

NASA RTLX workload

A priori comparison of workload at the digital and hybrid workstations. For a within subjects t test the difference between the two scores obtained for each subject at the hybrid and digital workstations should be normally distributed. The Kolmogorov-Smirnov test ($p=.2$) and the Shapiro-Wilk test ($p=.6$) both showed no deviation from a normal distribution, see table A4.1.

Table A5. 1 – Tests for normality for the comparisons between workload scores at the hybrid and digital workstations. ^a denotes Lilliefors Significance Correction *. denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Difference between workload scores at the hybrid and digital workstations	.210	8	.200*	.939	8	.597

There are only a small number of participants so the probability of a significant test result for deviations from normality is low, and therefore the Q-Q plot was also examined, as shown in figure A4.1. None of the values differ radically from the expected values for a normal distribution, and there is not a distinct pattern of skewness or kurtosis and therefore use of parametric statistics is appropriate.

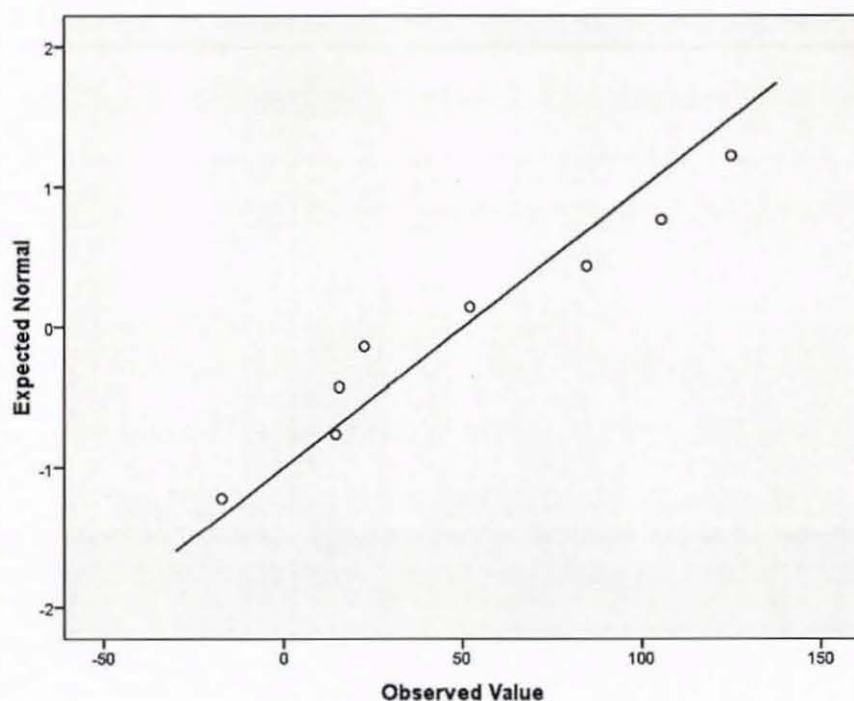


Figure A5. 1 – A Q-Q plot of the difference between the two scores for each participant at the hybrid and digital workstations.

NASA RTLX Workload Correlations

To analyse significance from Pearson's correlations between the subscales of workload and overall workload requires these data to be normally distributed. This was measured using both Kolgorov-Smirnov and Shapiro-Wilk tests for normality, along with examination of both the Q-Q plots and boxplots.

Table A5. 2 – Tests for normality for the correlations between subscales of workload and overall workload. ^a denotes Lilliefors Significance Correction *. denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Overall workload	.213	24	.006	.814	24	.000
Mental demand	.178	24	.048	.905	24	.028
Physical demand	.115	24	.200*	.936	24	.134
Temporal demand	.114	24	.200*	.936	24	.133
Performance	.223	24	.003	.921	24	.063
Effort	.250	24	.000	.771	24	.000
Frustration	.138	24	.200*	.933	24	.114

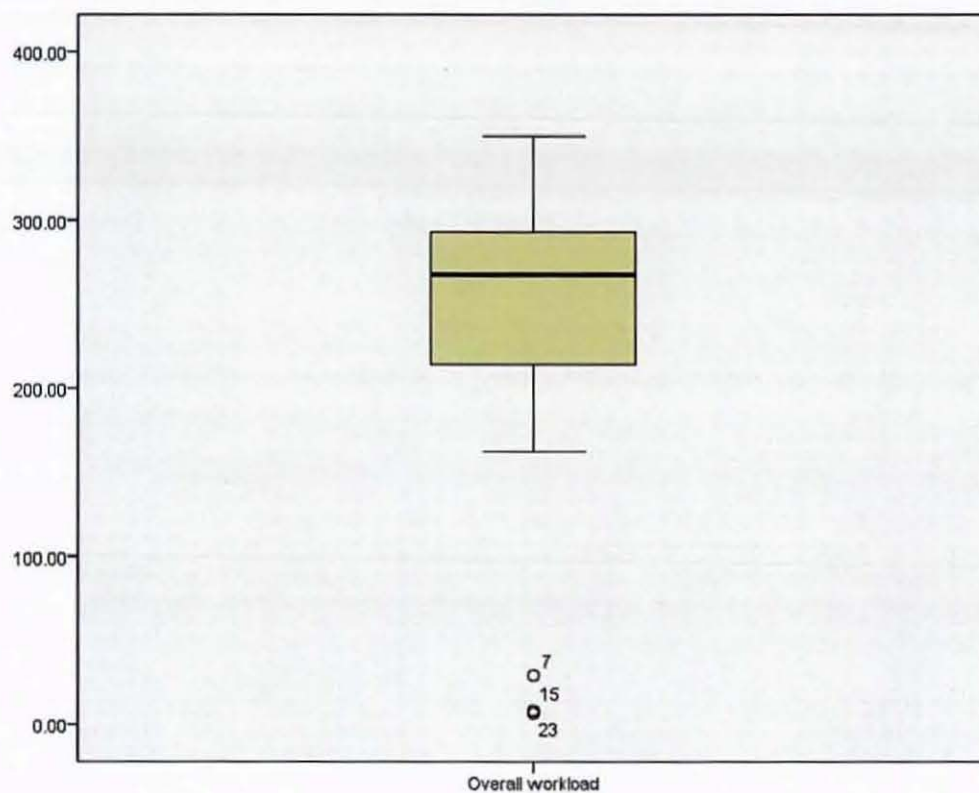
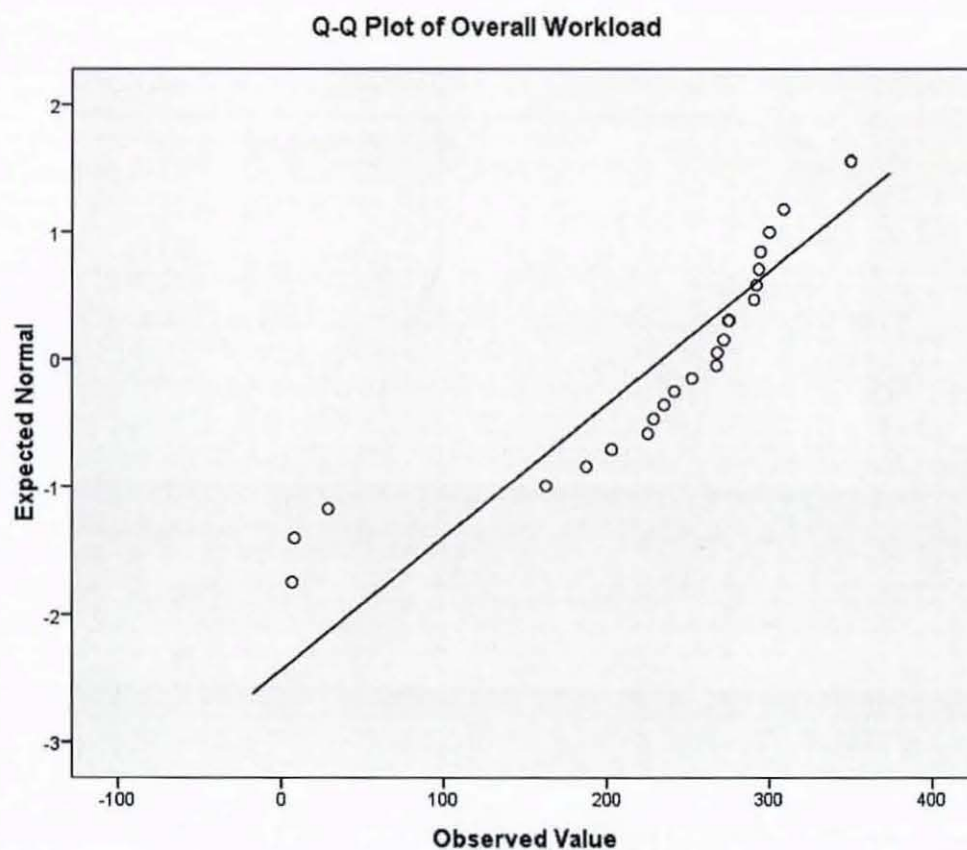


Figure A5. 2 – Q-Q plot and boxplots to assess the normality of the distribution of overall workload scores

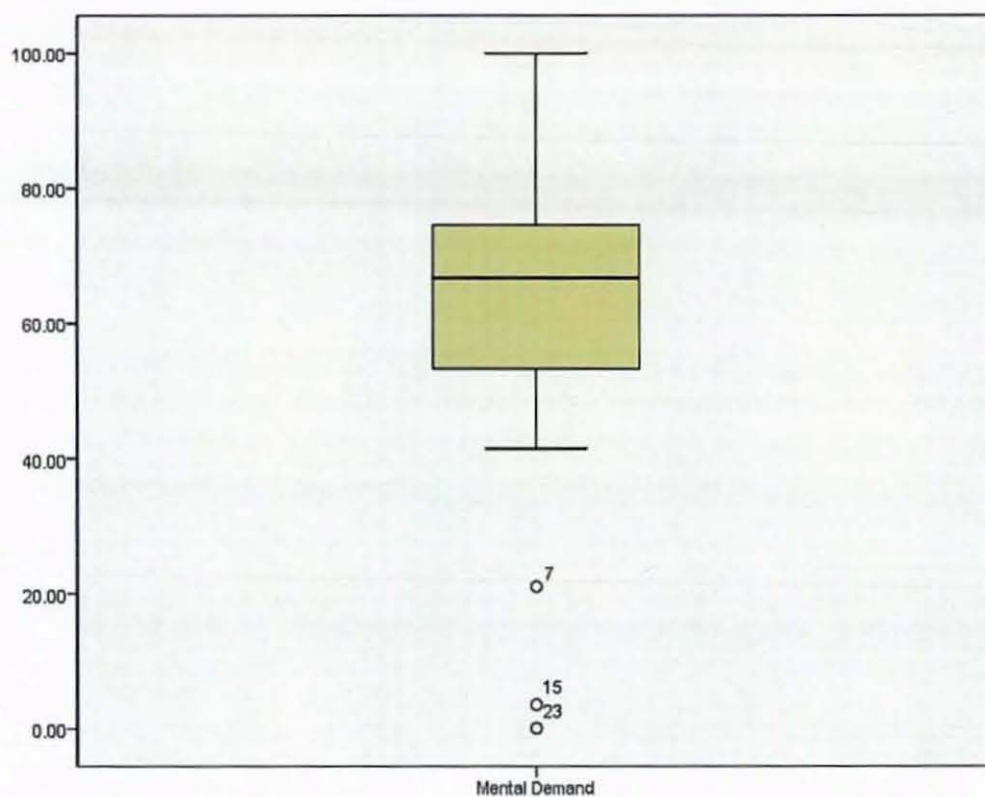
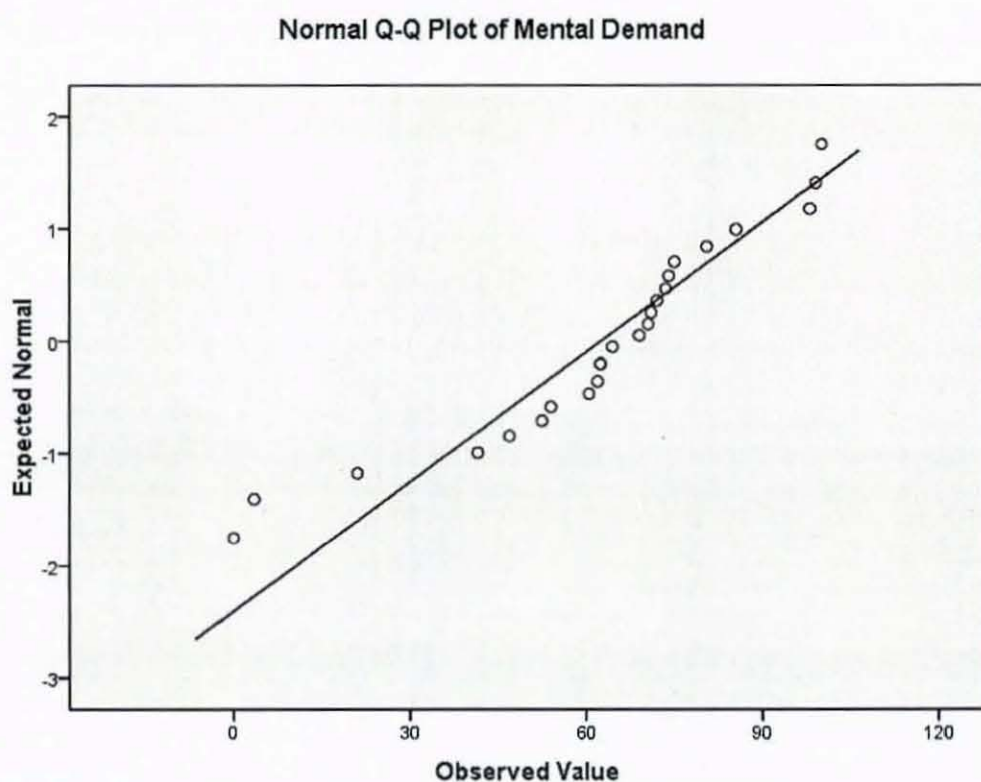


Figure A5. 3 – Q-Q plot and boxplots to assess the normality of the distribution of mental demand scores

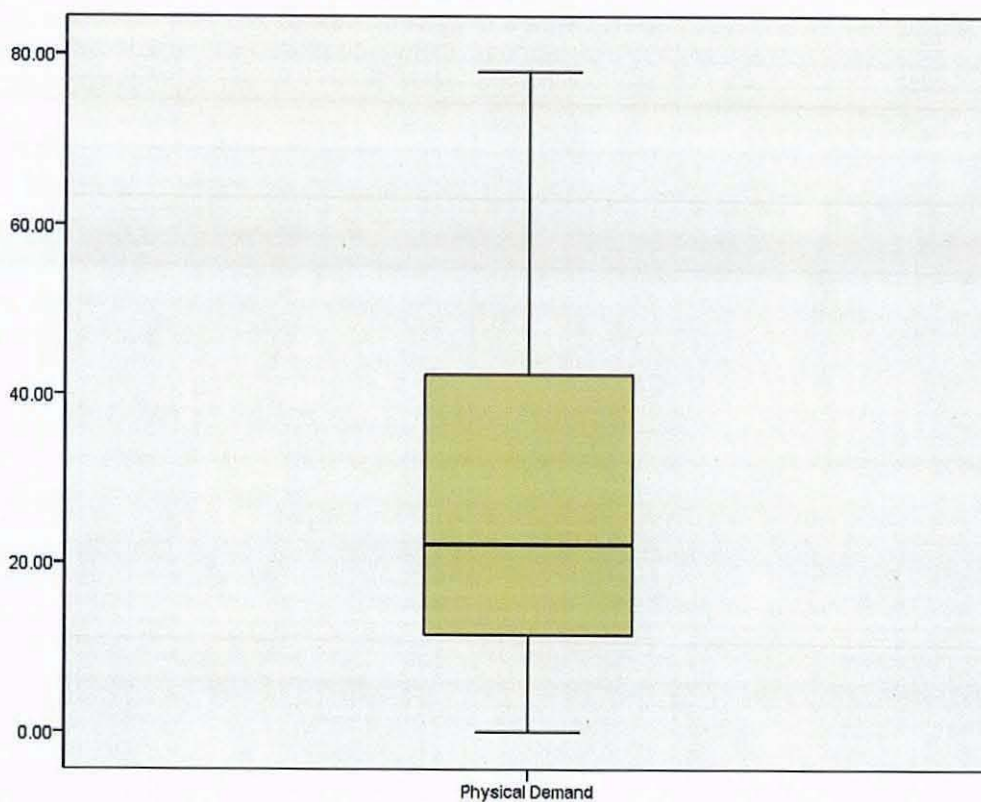
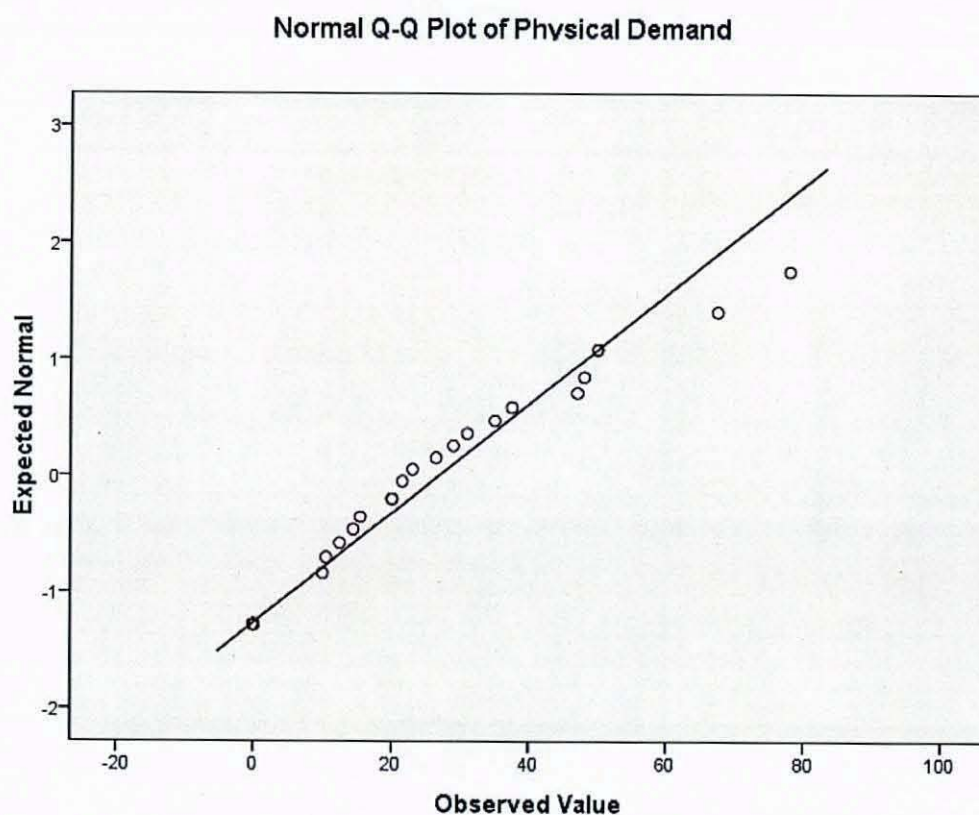


Figure A5. 4 – Q-Q plot and boxplots to assess the normality of the distribution of physical demand scores

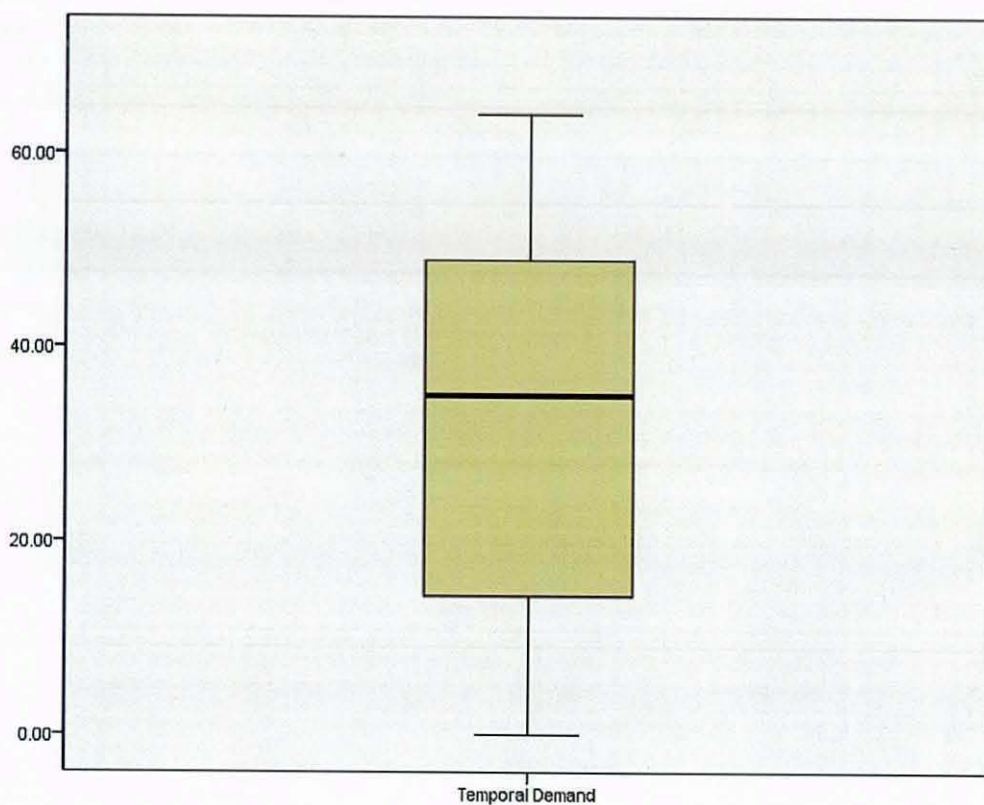
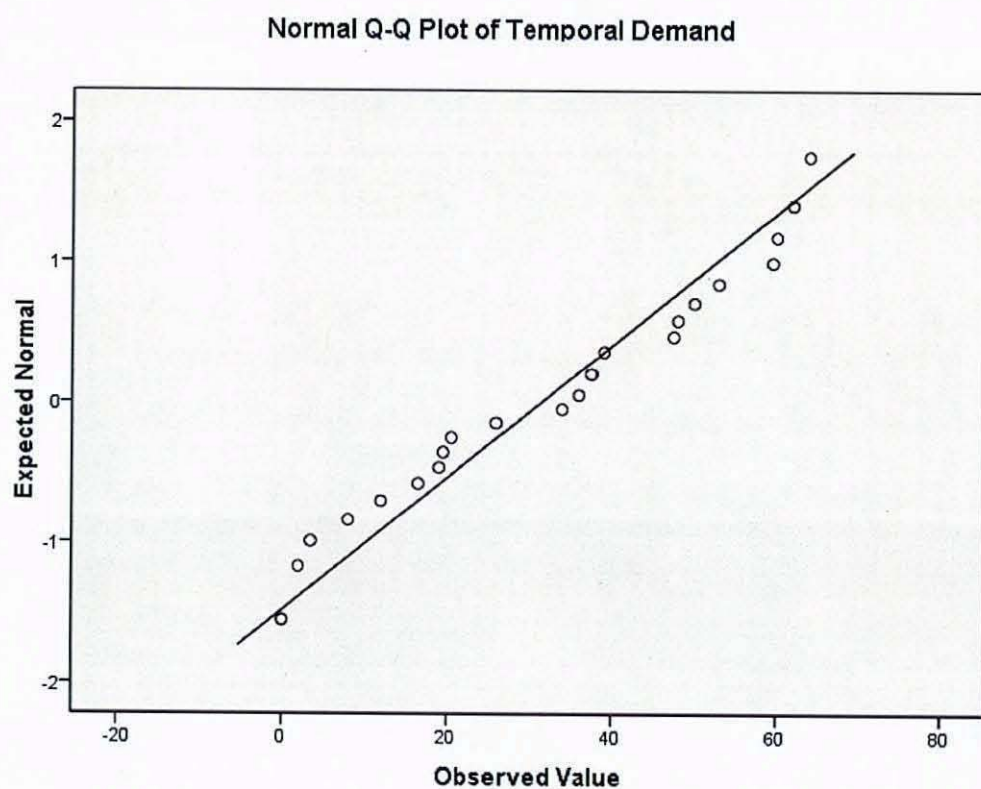


Figure A5. 5 – Q-Q plot and boxplots to assess the normality of the distribution of temporal demand scores

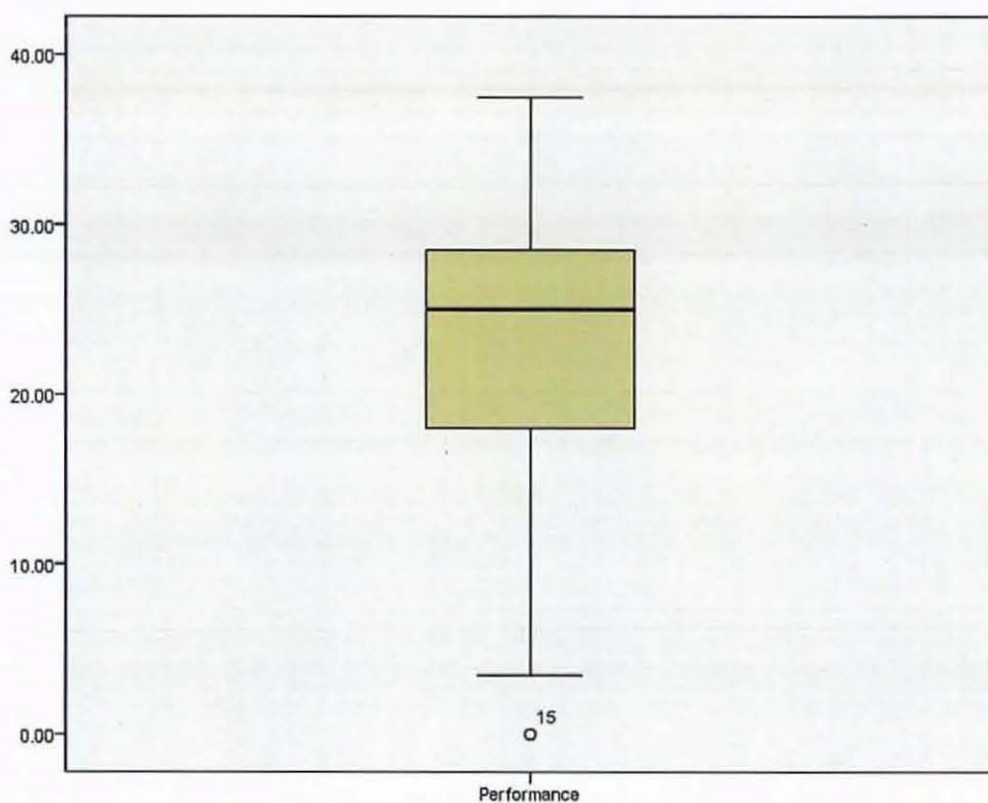
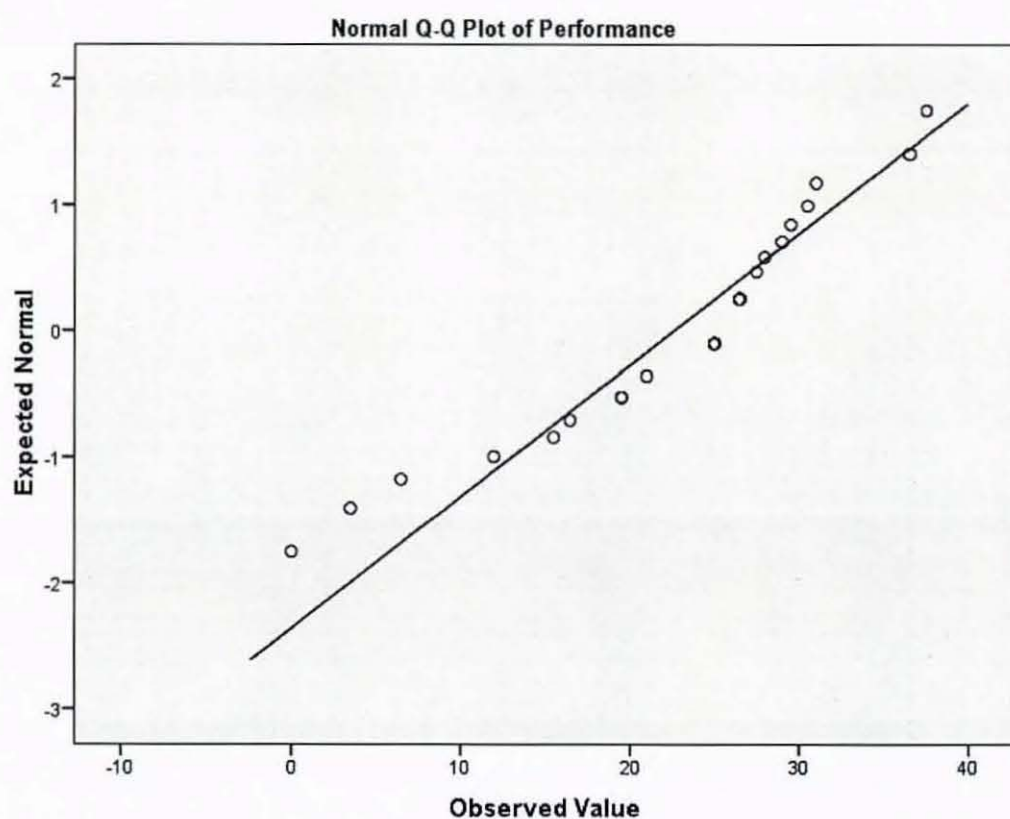


Figure A5. 6 – Q-Q plot and boxplots to assess the normality of the distribution of performance scores

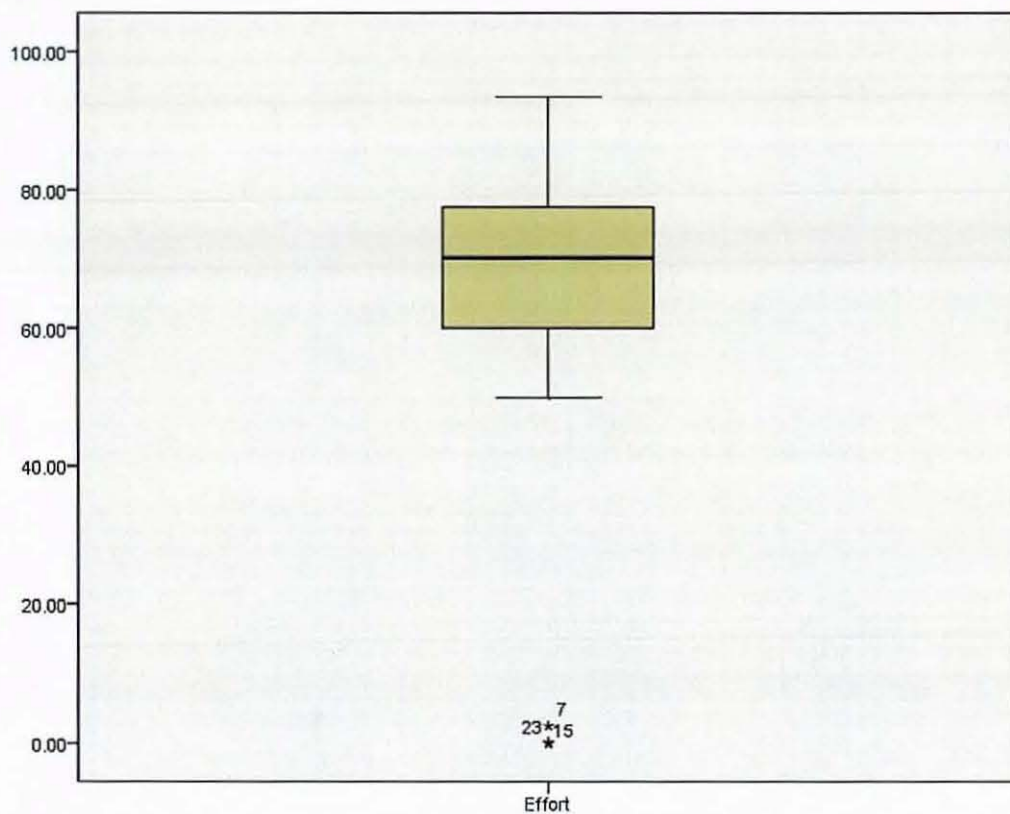
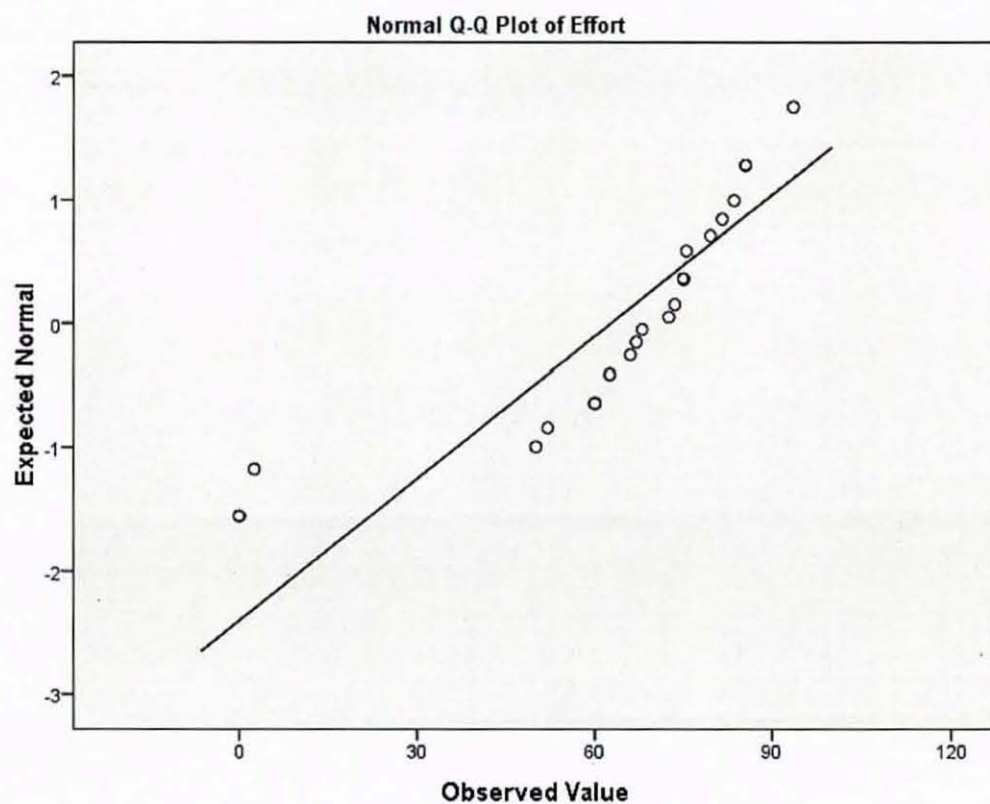


Figure A5. 7 – Q-Q plot and boxplots to assess the normality of the distribution of effort scores

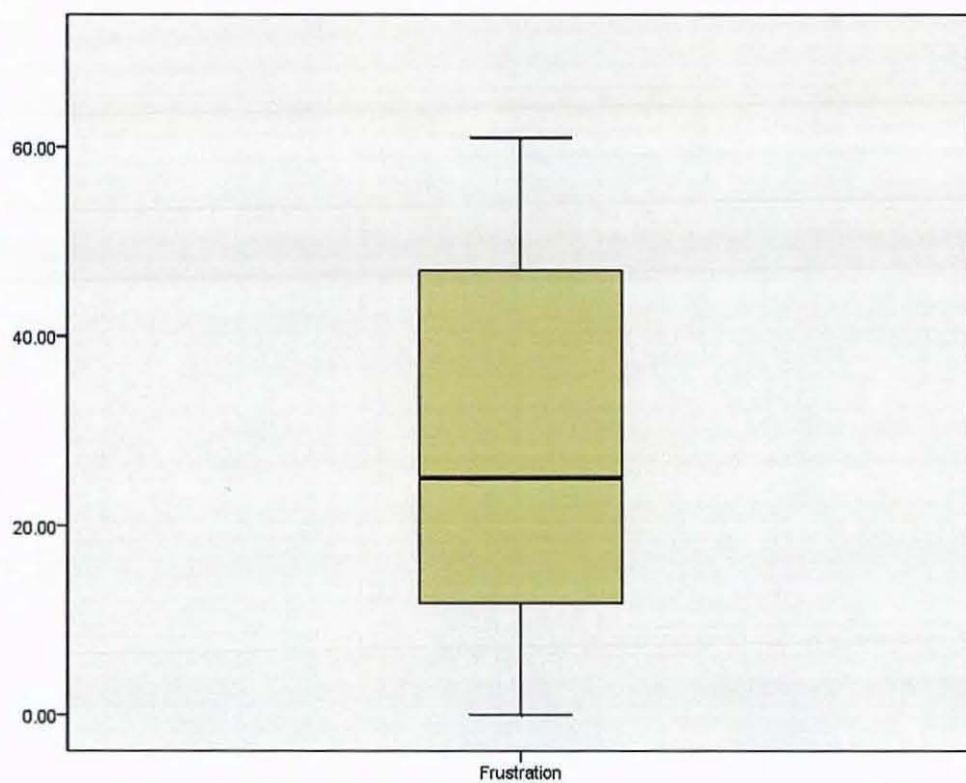
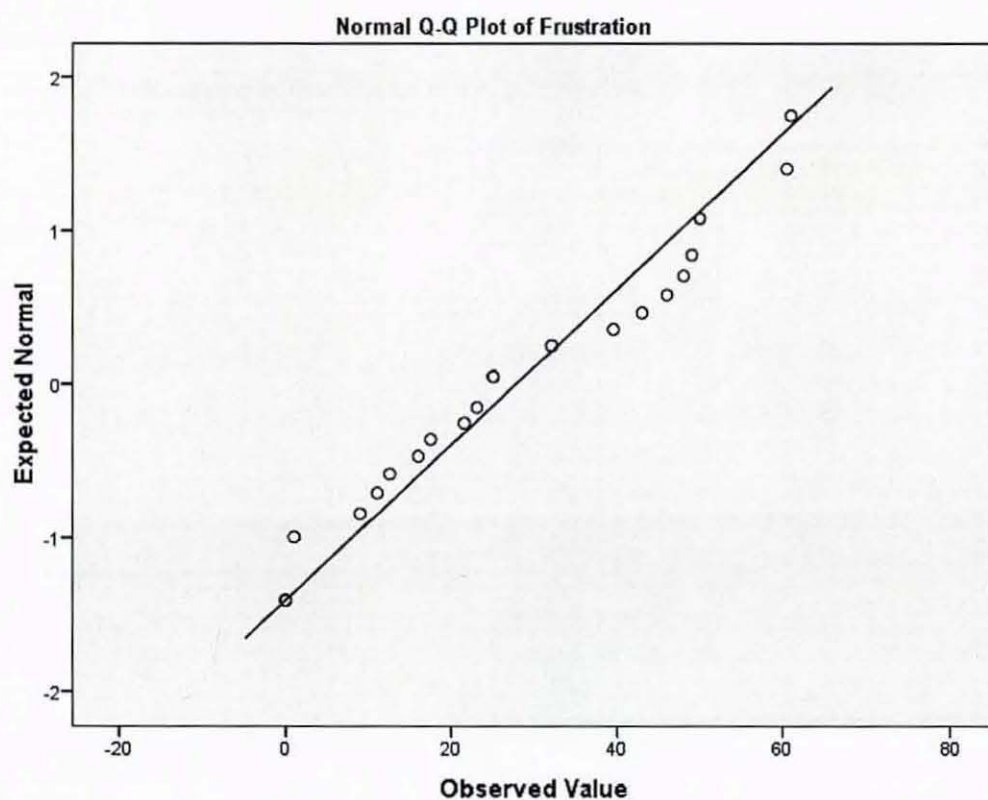


Figure A5. 8 – Q-Q plot and boxplots to assess the normality of the distribution of frustration scores

The Q-Q plots show deviations from normality for overall workload, mental demand, and effort, with the boxplots demonstrating that participant 7 was an outlier. Therefore, participant 7 was removed from the analysis and the tests for normality were conducted again.

Table A5. 3 – Tests for normality for the correlations between subscales of workload and overall workload with participant 7 removed. ^a denotes Lilliefors Significance Correction * denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Overall workload (participant 7 removed)	.137	21	.200*	.970	21	.739
Mental demand (participant 7 removed)	.145	21	.200*	.954	21	.409
Physical demand (participant 7 removed)	.128	21	.200*	.943	21	.249
Temporal demand (participant 7 removed)	.120	21	.200*	.956	21	.432
Performance (participant 7 removed)	.191	21	.045	.966	21	.640
Effort (participant 7 removed)	.103	21	.200*	.979	21	.917
Frustration (participant 7 removed)	.169	21	.122	.949	21	.325

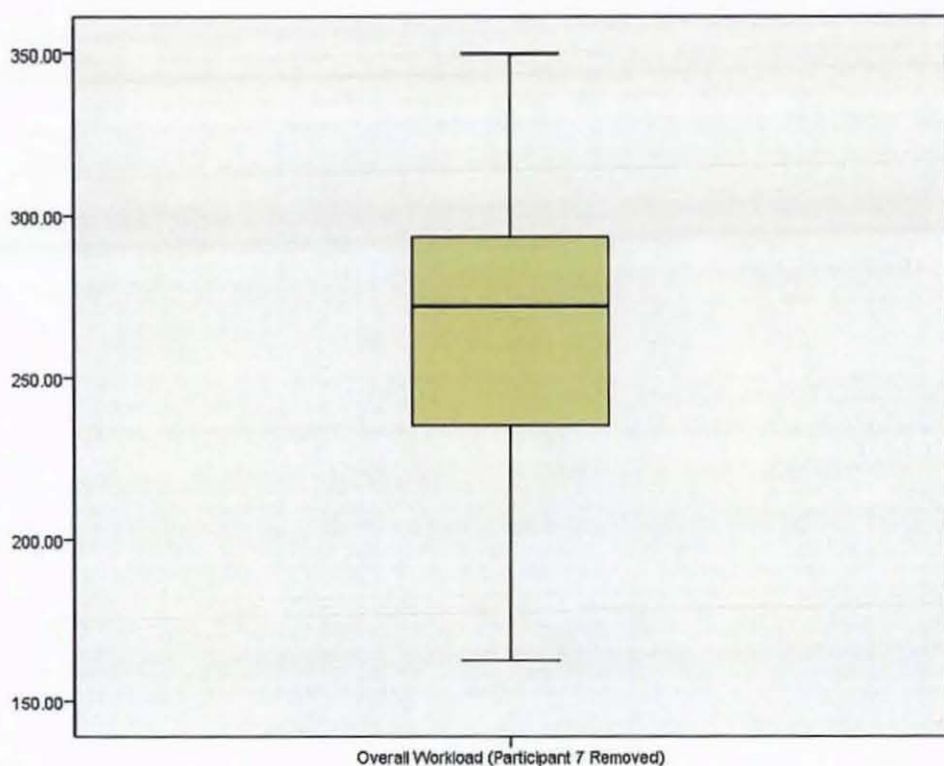
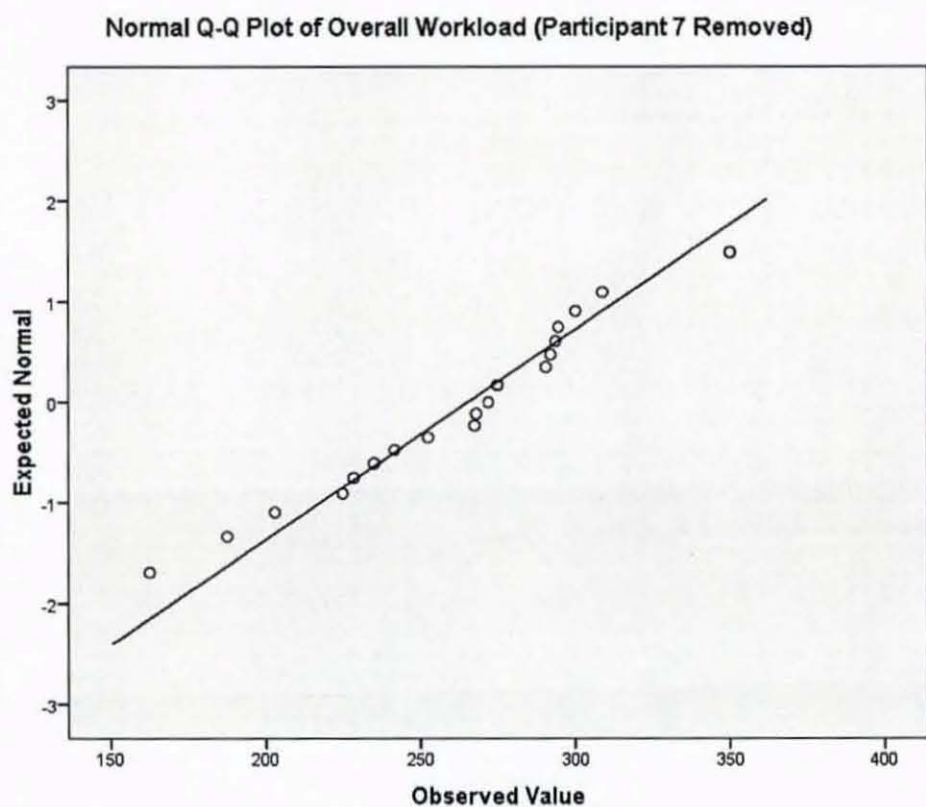


Figure A5. 9 – Q-Q plot and boxplots to assess the normality of the distribution of overall workload scores with participant 7 removed.

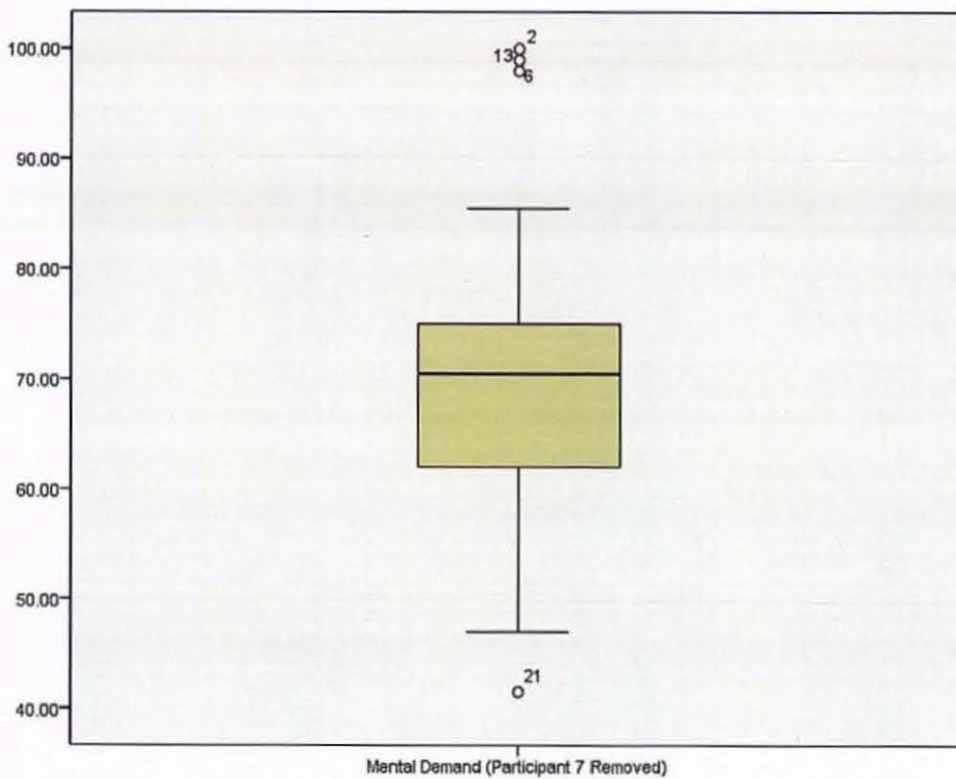
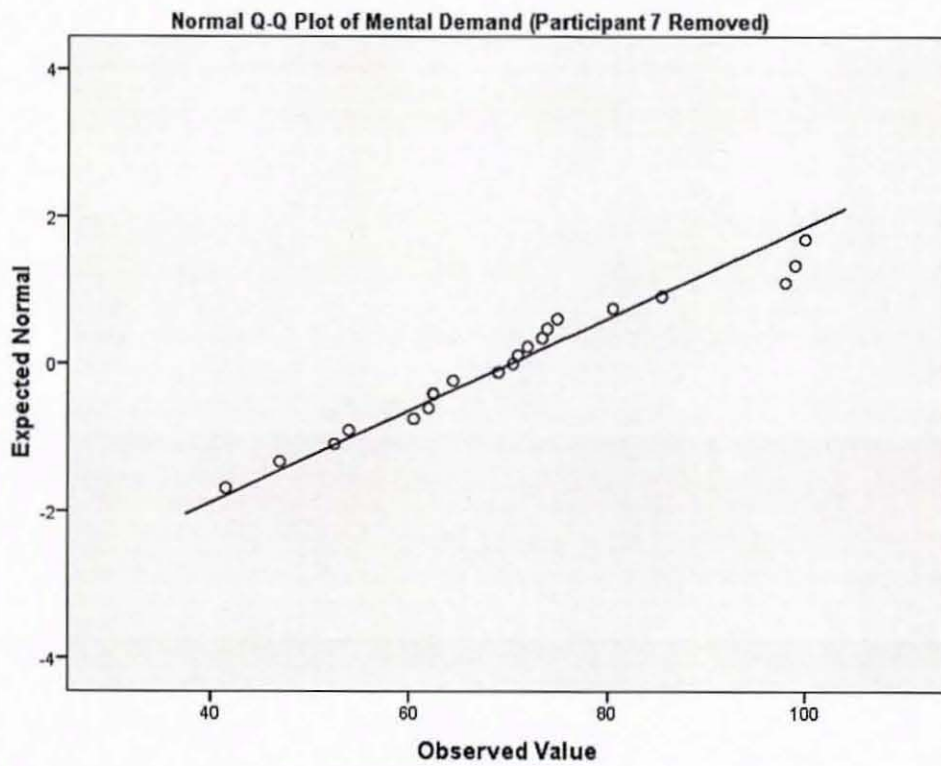


Figure A5. 10 – Q-Q plot and boxplots to assess the normality of the distribution of mental demand scores with participant 7 removed.

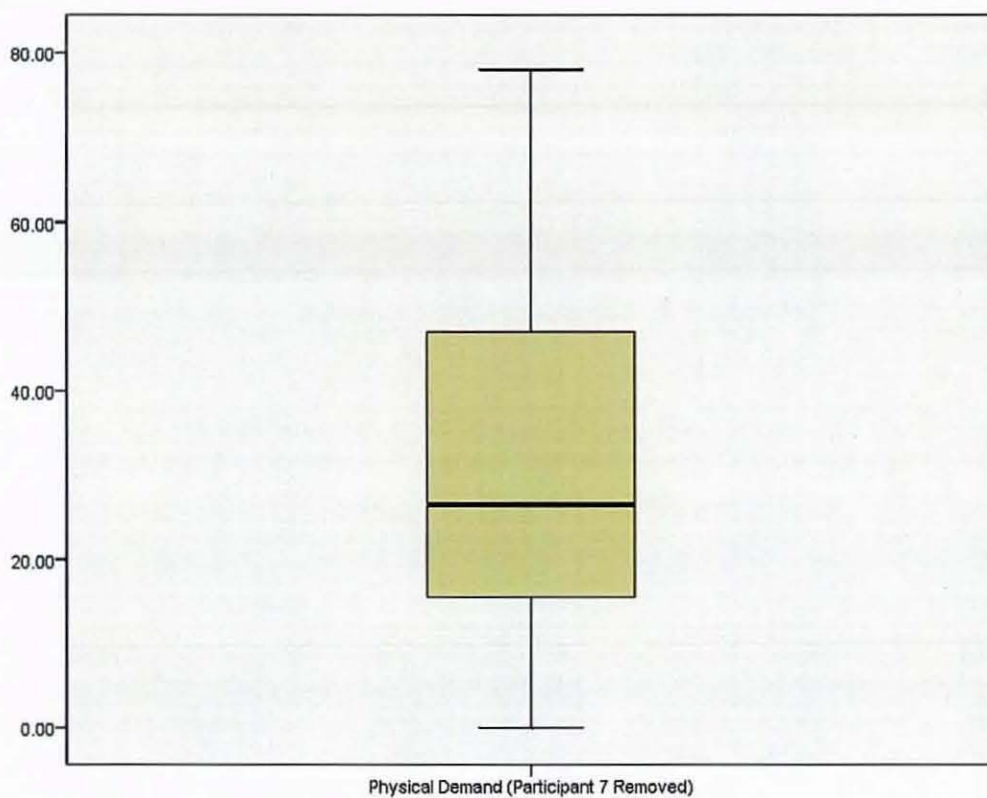
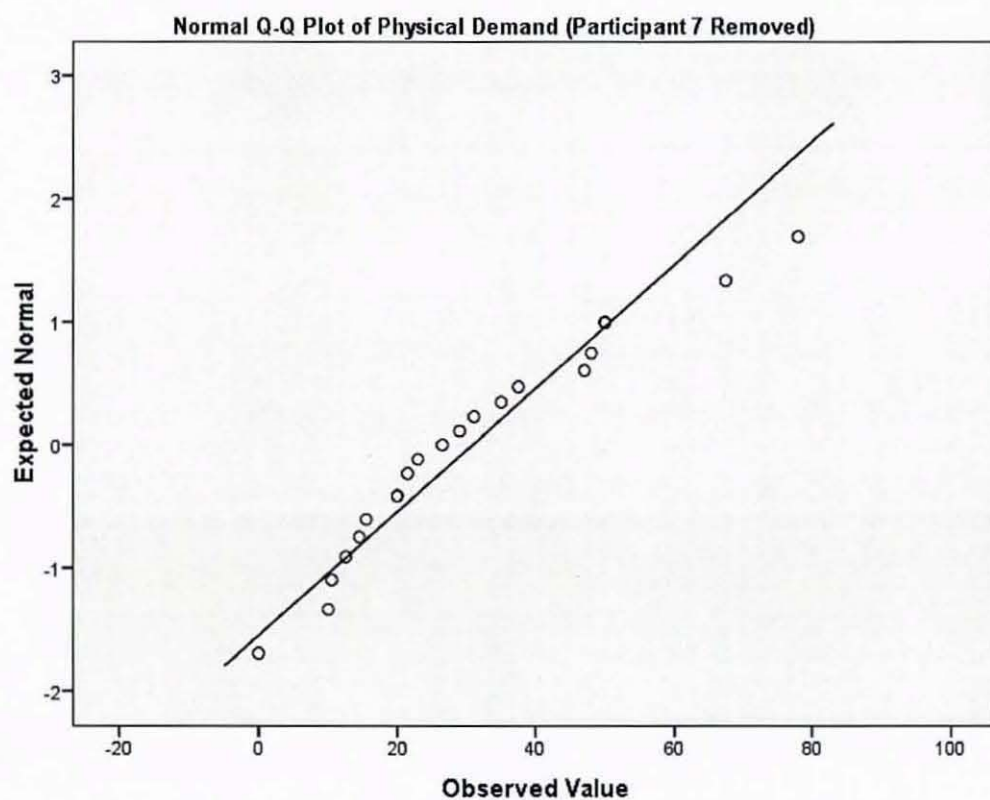


Figure A5. 11 – Q-Q plot and boxplots to assess the normality of the distribution of physical demand scores with participant 7 removed.

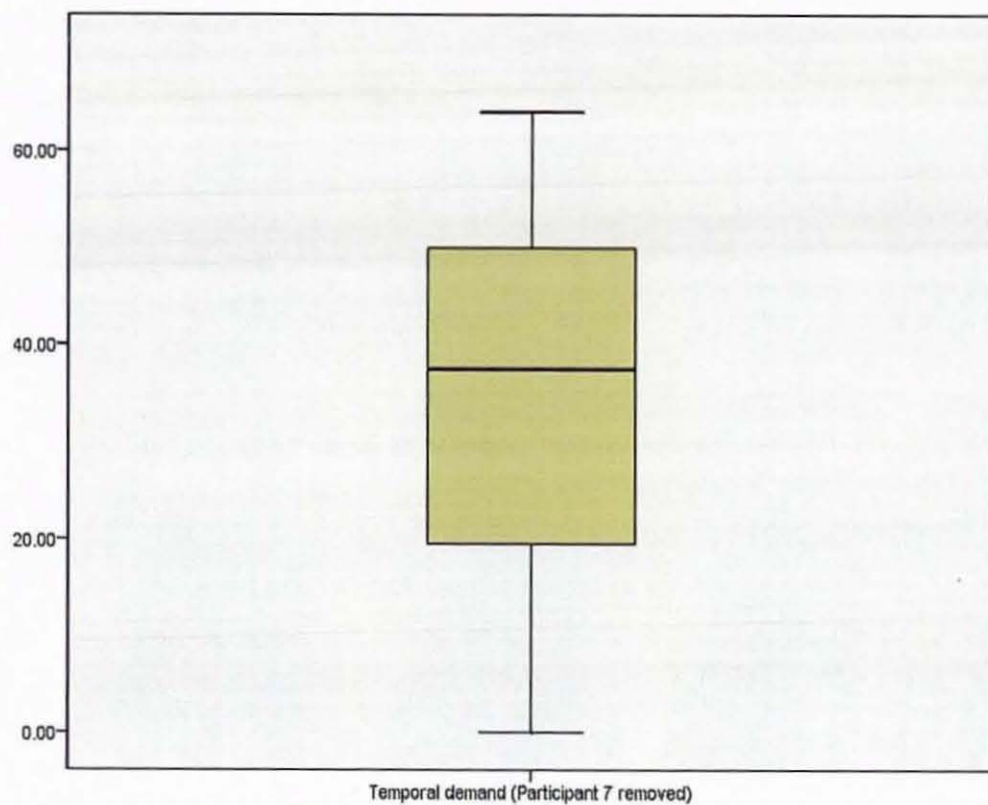
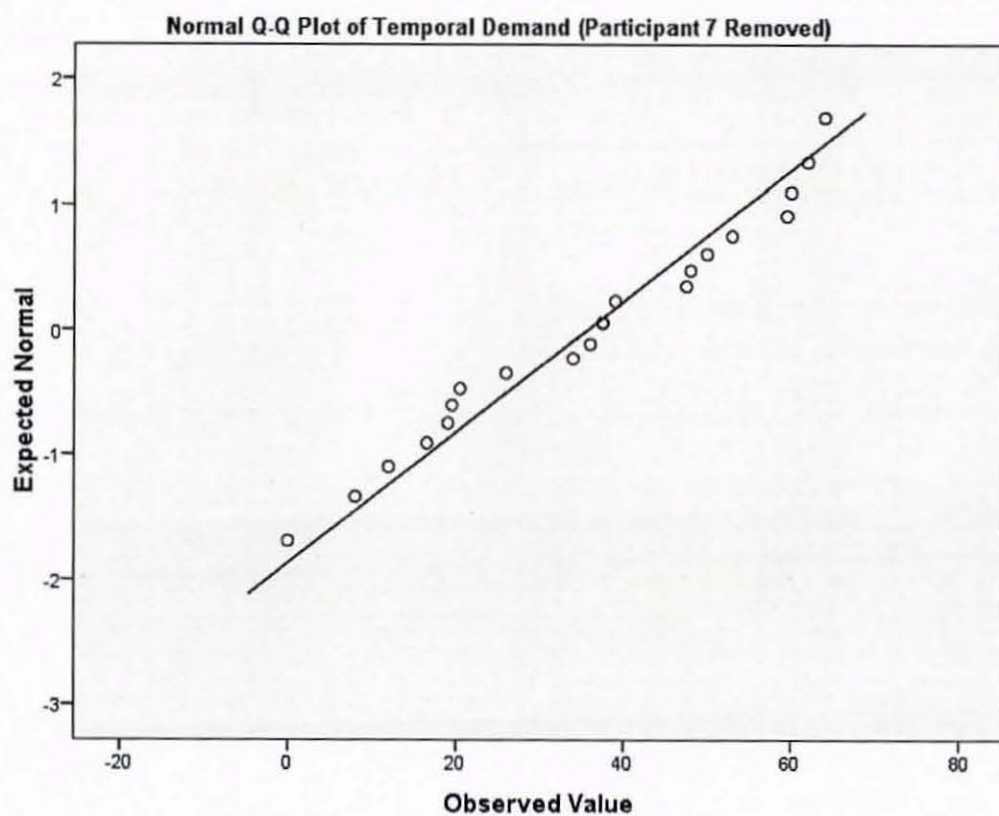


Figure A5. 12 – Q-Q plot and boxplots to assess the normality of the distribution of temporal demand scores with participant 7 removed.

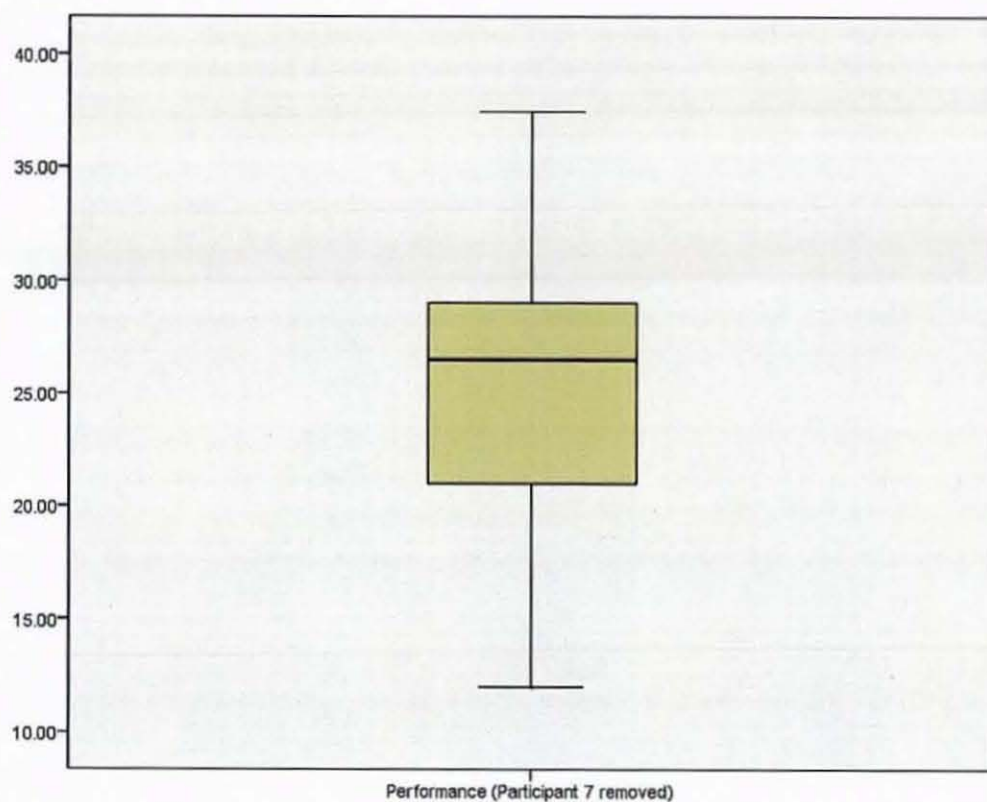
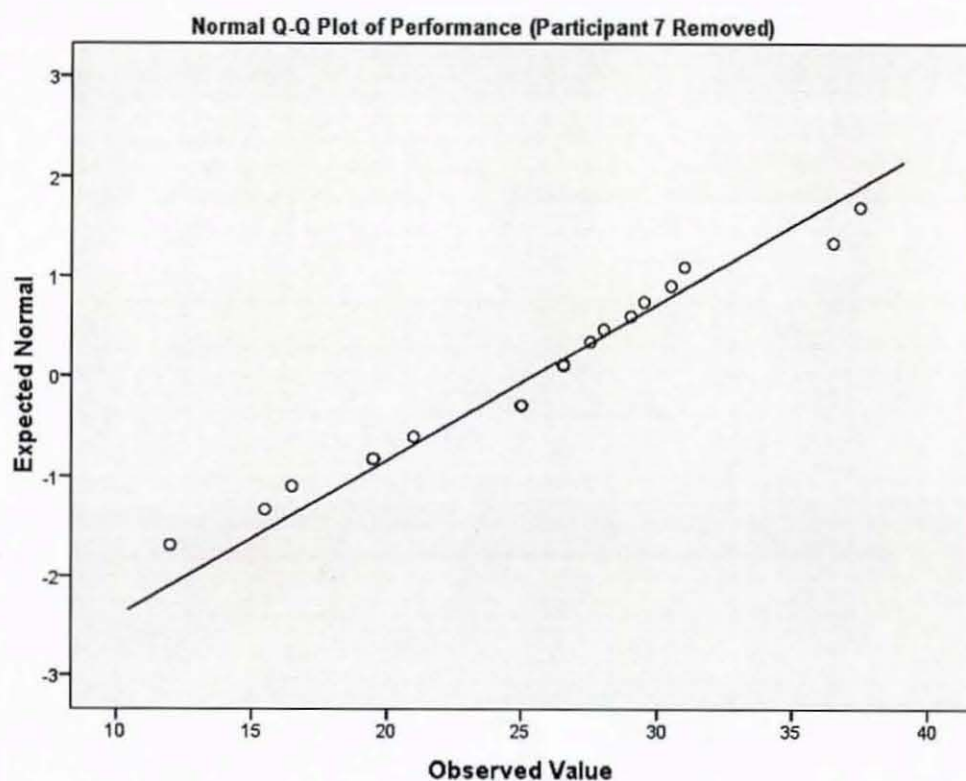


Figure A5. 13 – Q-Q plot and boxplots to assess the normality of the distribution of performance scores with participant 7 removed.

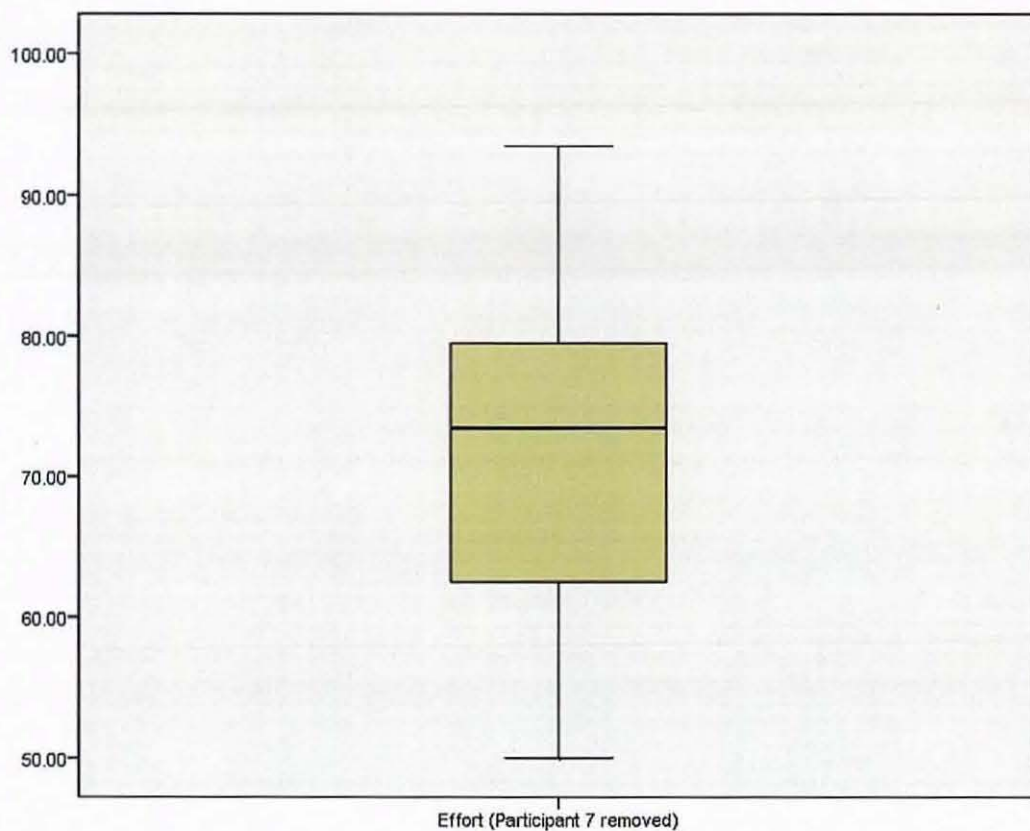
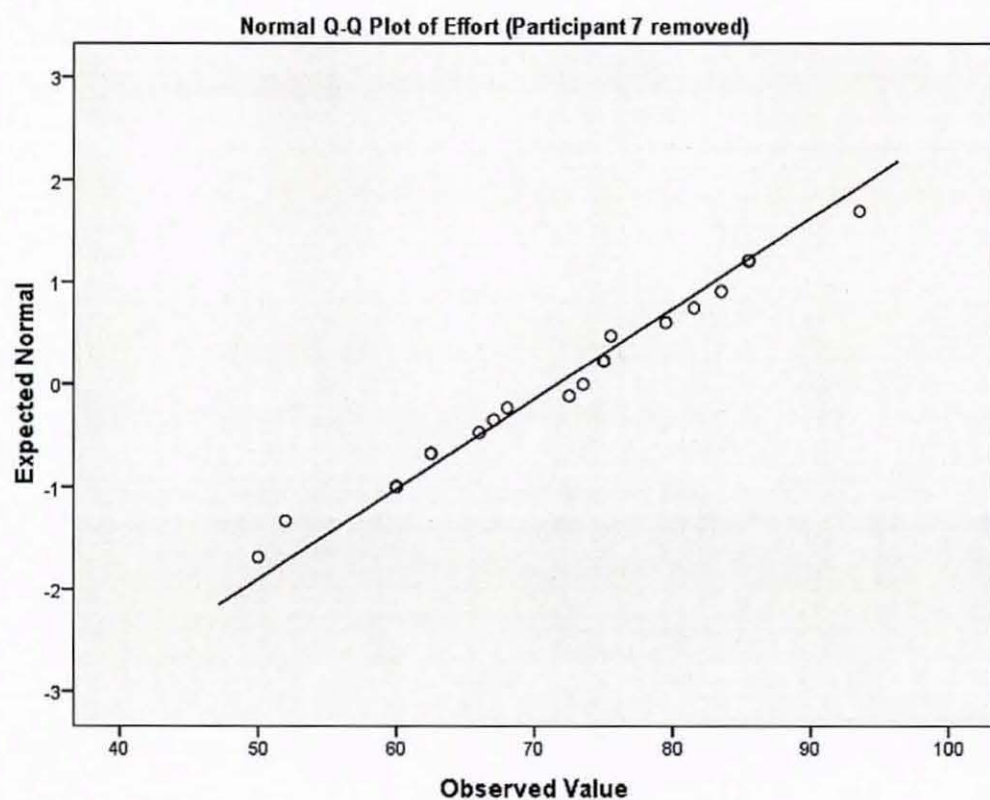


Figure A5. 14 – Q-Q plot and boxplots to assess the normality of the distribution of effort scores with participant 7 removed.

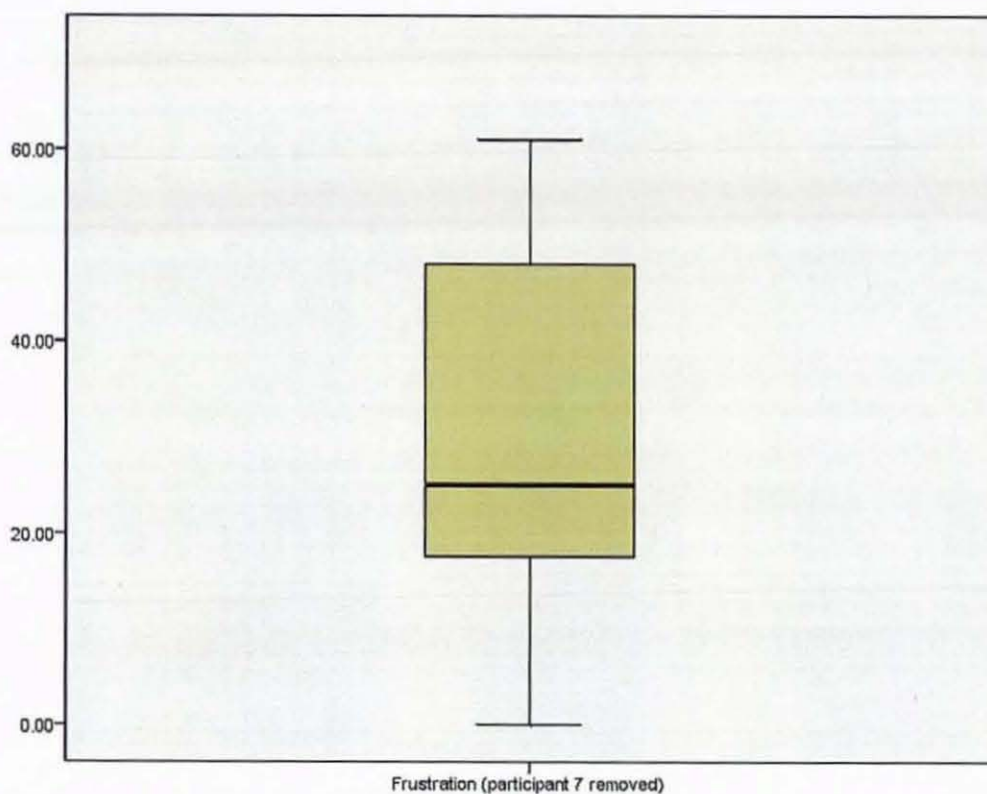
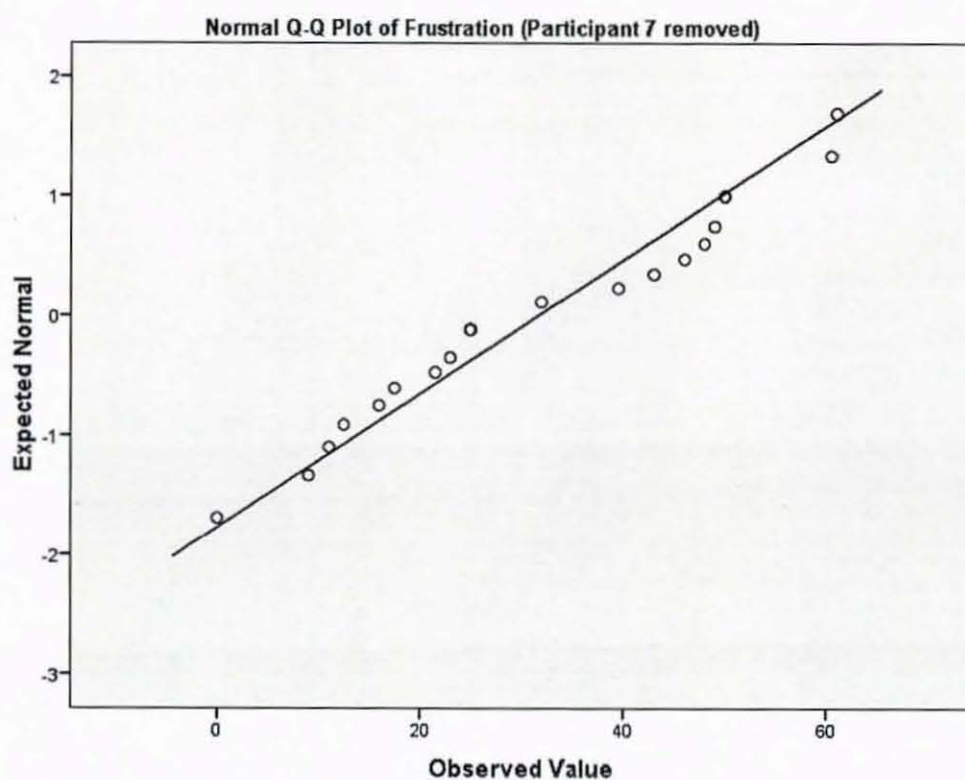


Figure A5. 15 – Q-Q plot and boxplots to assess the normality of the distribution of frustration scores with participant 7 removed.

Speed of Reading

A priori comparison of time taken per case at the digital and hybrid workstations. For a within subjects t test the difference between the two scores obtained for each subject at the hybrid and digital workstations should be normally distributed. Both Kolmogorov-Smirnov and Shapiro Wilk tests show no deviation from the assumption of normality. The Q-Q plots show no obvious skewing or kurtosis, and the boxplots show no outliers.

Table A5. 4 – Tests for normality for difference between the time taken per case at the digital and hybrid workstations. ^a denotes Lilliefors Significance Correction * denotes a lower bound of the true significance.

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Difference between time taken at the hybrid and digital workstation	.203	8	.200*	.961	8	.817
Difference between time taken at the hybrid and digital workstation with recalled cases excluded	.152	8	.200*	.963	8	.836

Normal Q-Q Plot of the Difference Between Time Taken at the Hybrid and Digital Workstations

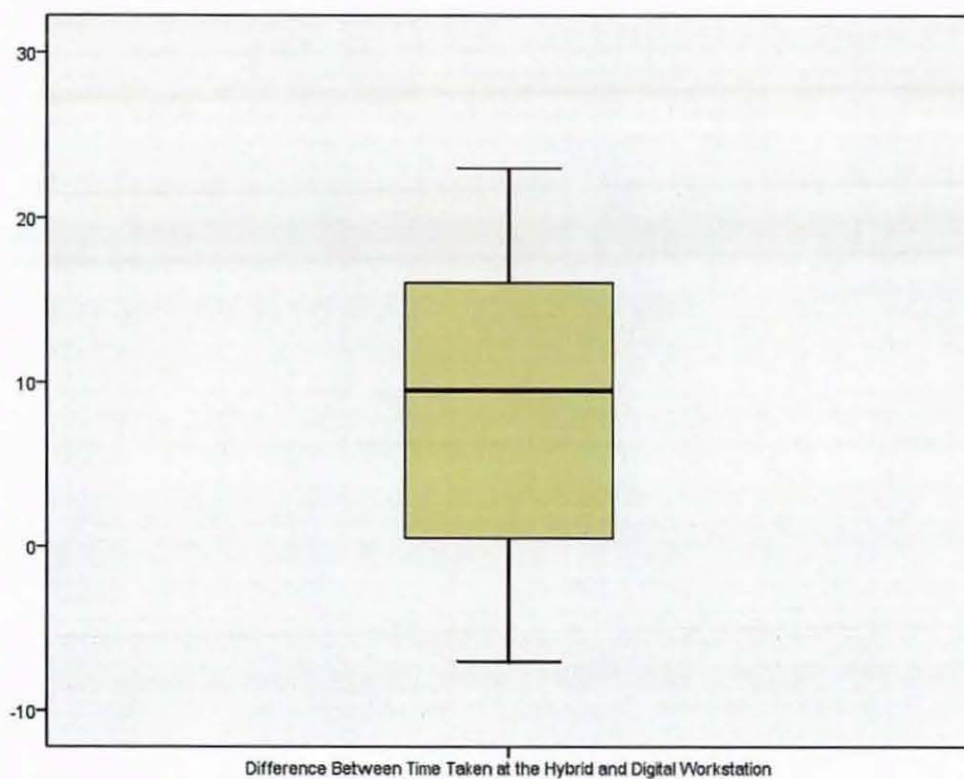
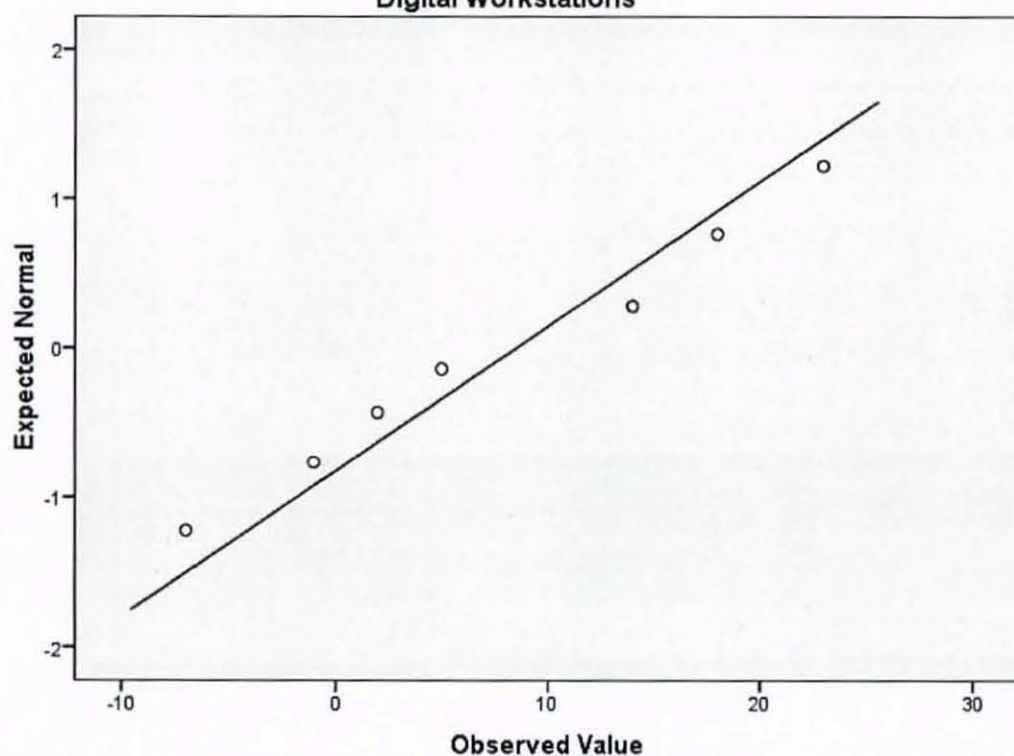


Figure A5. 16 – Q-Q plot and boxplots to assess the normality of the differences between time taken per case at the digital and hybrid workstations

Normal Q-Q Plot of the Difference Between Time Taken (Excluding Recalled Cases) at the Hybrid and Digital Workstations

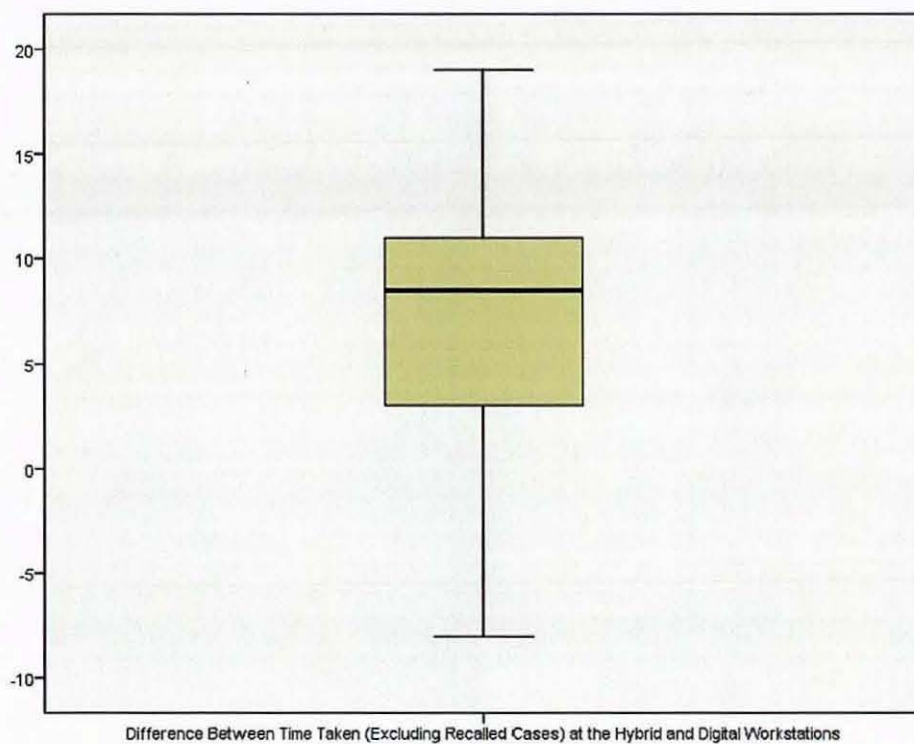
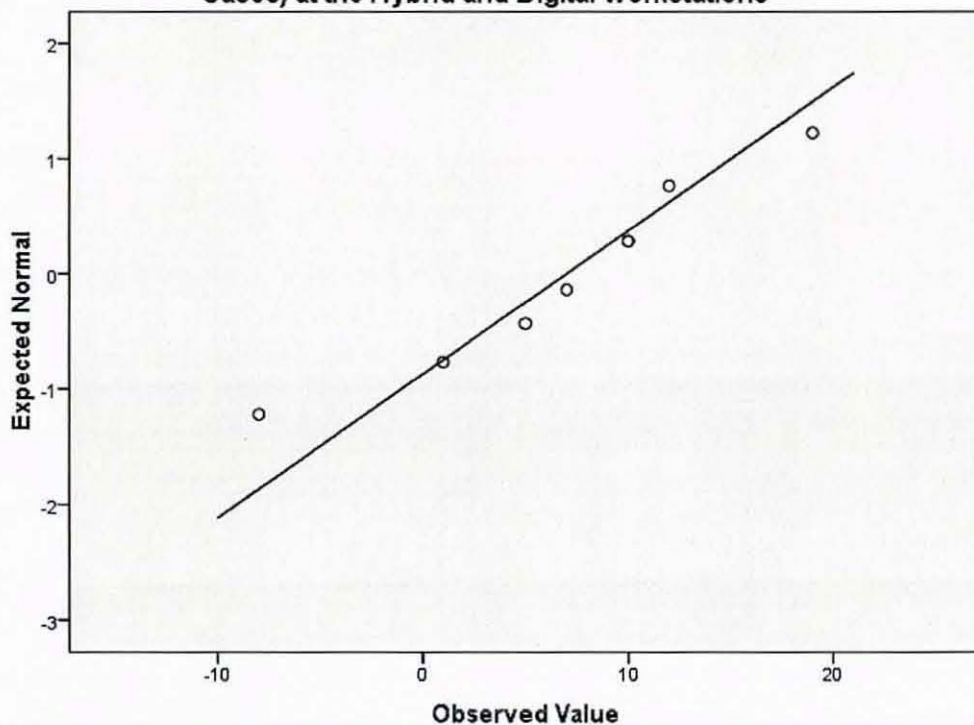


Figure A5. 17 – Q-Q plot and boxplots to assess the normality of the differences between time taken per case (excluding recalled cases) at the digital and hybrid workstations

Appendix 6 – Participant Information Sheet and Informed Consent Form

Digitisation of Prior Mammograms: Effect on Radiologist Performance

PARTICIPANT INFORMATION SHEET

The aim of this study is to measure the effect of digitising the prior mammograms on radiologist and advanced practitioner performance in reading mammograms in the NHS Breast Screening Programme

Taking part in the study will involve the following

- You will be asked to complete six hours of reading difficult test cases. Approximately half of these will be cancerous and half either normal or benign. You will be asked to locate the cancerous lesions and give a report on the probability of malignancy for each lesion you find. You will be video recorded whilst undertaking this task. The filming will be by four fixed video cameras located in unobtrusive places in the radiology reading room. Anonymised stills from the video will be used in publications only with your consent.
- For one of the six sessions your eye movements will be recorded using light-weight head-mounted eye tracking equipment.

The total time commitment for this study is 6 hours in one hour sessions. Initially three sessions will be conducted, ideally over a three week period, then a break of at least one month, then another three sessions.

All data will be anonymised immediately after data collection, and will be reported in such a way to ensure complete confidentiality. The investigators are not interested in individual performance, only on collective group performance under the two conditions. Under no circumstances will details of an individual's performance be divulged to anyone other than the participant themselves.

After participation in the study, if you would like details of your performance to be supplied confidentially, then this can be requested via email to s.phillips2@lboro.ac.uk. A report detailing your answers under each condition alongside the 'correct' answers would then be provided via email, so other colleagues and participants are not aware of the request. This data will only be supplied upon request.

You have the right to withdraw from this study at any stage for any reason, and that I will not be required to explain your reasons for withdrawing.

Researcher contact details
Sian Taylor-Phillips
Tel 07725000262
Email s.phillips2@lboro.ac.uk

Project Supervisor: Prof Alastair Gale
Tel 01509635703
Email a.g.gale@lboro.ac.uk

Local Collaborator:
Dr Alison Duncan

Local Independent Contact Point
Research and Development Office
02476 966197
02476 966202

Digitisation of Prior Mammograms: Effect on Radiologist Performance

INFORMED CONSENT FORM

(to be completed after Participant Information Sheet has been read)

The purpose and details of this study have been explained to me. I understand that this study is designed to further scientific knowledge and that all procedures have been approved by the Loughborough University Ethical Advisory Committee, and the Caldecott Guardian at Coventry Hospital.

- I have read and understood the information sheet and this consent form.
- I have had an opportunity to ask questions about my participation.
- I understand that I am under no obligation to take part in the study.
- I understand that I have the right to withdraw from this study at any stage for any reason, and that I will not be required to explain my reasons for withdrawing.
- I understand that all the information I provide will be treated in strict confidence.
- I understand that all data including video and eye tracking data will be anonymised.
- I agree to participate in this study.

Your name

Your signature

Signature of investigator

Date

Appendix 7 – Participant Instructions

Digitisation of Prior Mammograms: Effect on Radiologist Performance

Instructions for Participants:

Thank you for agreeing to participate in this experiment. There are a total of six sessions for each participant, each of which lasts approximately an hour. This study measures the difference in performance under two conditions, and therefore individual performance is not of interest to the experimenters. However if at the end of the study if you wish to review your performance then details of the cases you correctly and incorrectly classified can be given confidentially.

This experiment is a simulation of reading mammograms as a part of the NHS Breast Screening Programme. There are approximately half malignant cases and half non-malignant over the six sessions, however this split may vary by session. For each case please examine the mammograms, and then mark the locations of any lesions with a cross on the data recording sheet. Please mark as many lesions as you can see per case, or none at all. Then number each lesion and rate the probability of malignancy. If there are more than three lesions in a particular case then just rate the most suspicious three. It is very important that you err on the side of marking too many rather than too few lesions. If there are any indications of possible malignancy then please mark the lesion. Finally state whether you would recall the case or return to screen if you were reading it for the NHS Breast Screening Programme.

There are three practice cases so that you can get used to the method of reporting, and raise any questions with the experimenter. There are 54 cases to review per session, which is expected to last an hour or less. For half of these the prior mammograms will be in film format, and for the other half the prior mammograms will be digitised

It is important to the study design to complete the first three sessions on the same day of the week and at the same time of day for three consecutive weeks where possible, and to minimise interruptions during the reading sessions.

Once again thank you for your participation.

Digitisation of Prior Mammograms: Effect on Radiologist Performance

Reminder of Instructions:

- Mark a lesion if there is **any** indication of possible malignancy
- Number the lesions where there is more than one for a case
- Always report whether you would recall or return to screen if you read the case in the NHS Breast Screening Programme
- There are 54 cases to review per session, which is expected to last an hour or less.

Appendix 8 – Examples of Data Recording Sheets for Performance Experiment

Case 159 – Malignant spiculate mass on right breast

**Case 66 – Malignant ill defined mass and suspicious calcifications on
left breast**

**Case 98- Not malignant. Well defined mass on right breast (cyst) which
has not changed size since previous mammograms. This case went to
arbitration in the breast screening programme but was not recalled.**

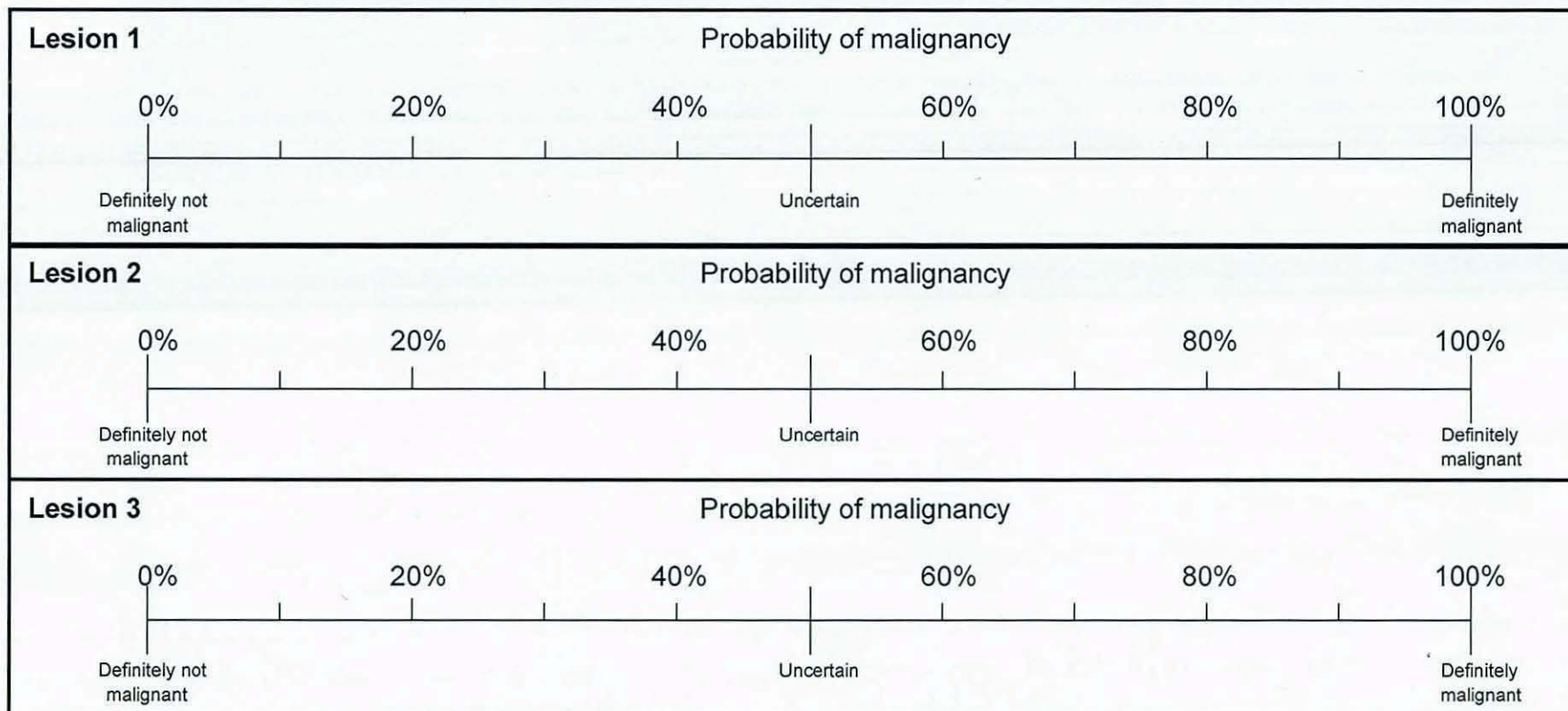
**Case 86 – Not malignant. Spiculated mass in left breast. This case was
recalled and biopsied in the breast screening programme.**

**Case 21 – Not malignant. Architectural Distortion in left breast.
This case was recalled for further tests in the breast screening
programme but not biopsied.**

Case 159

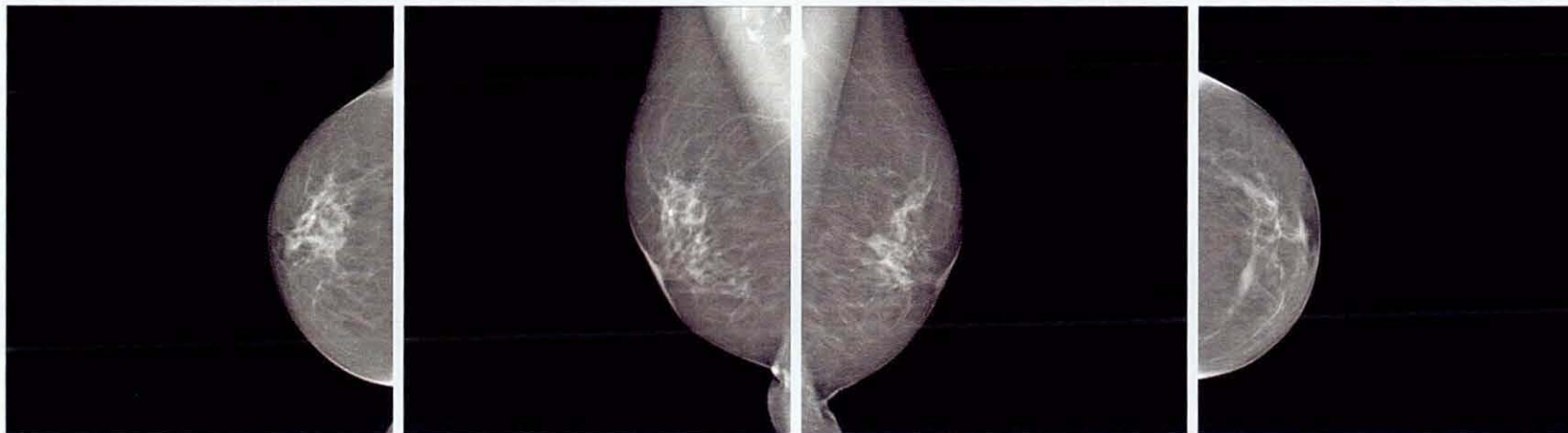


293

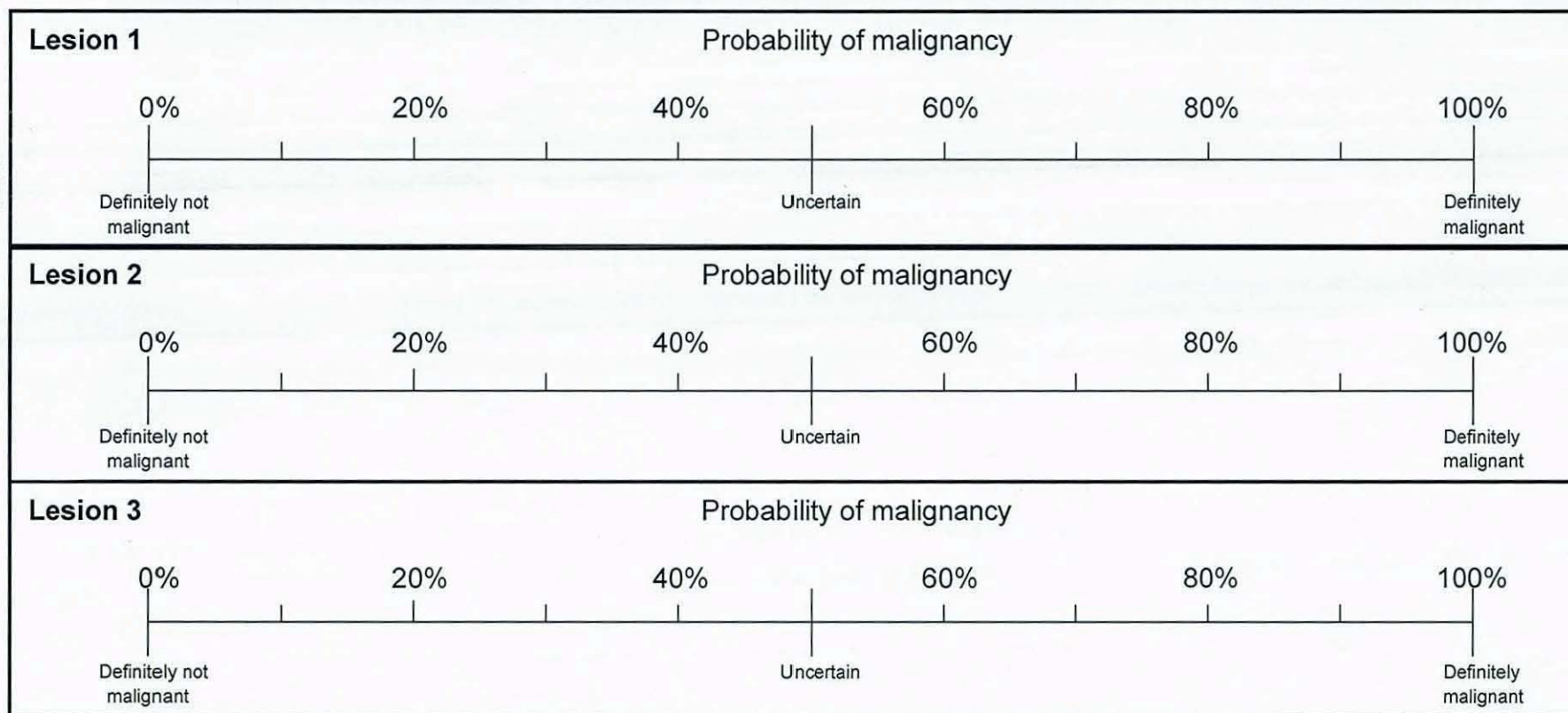


Return to Screening / Recall

Case 66



294

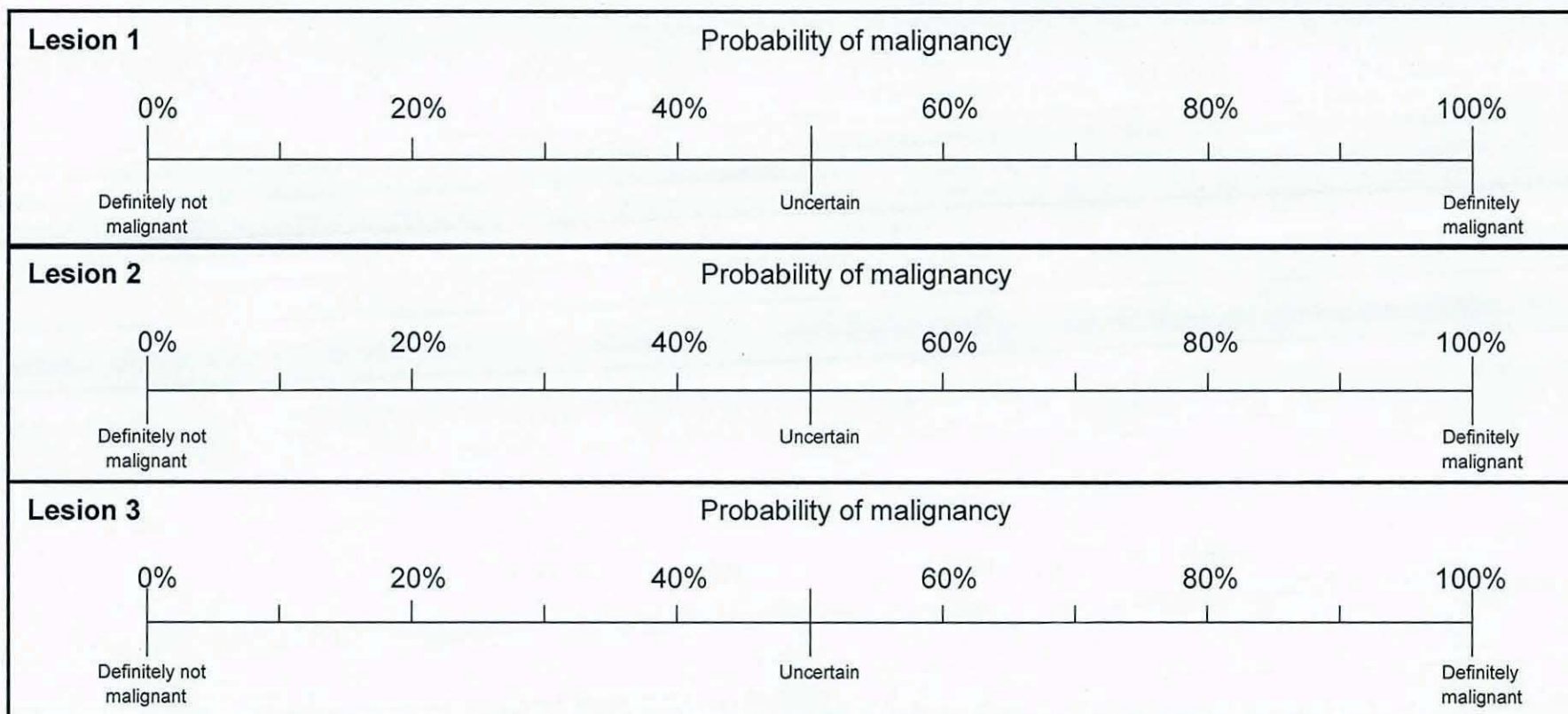


Return to Screening / Recall

Case
98

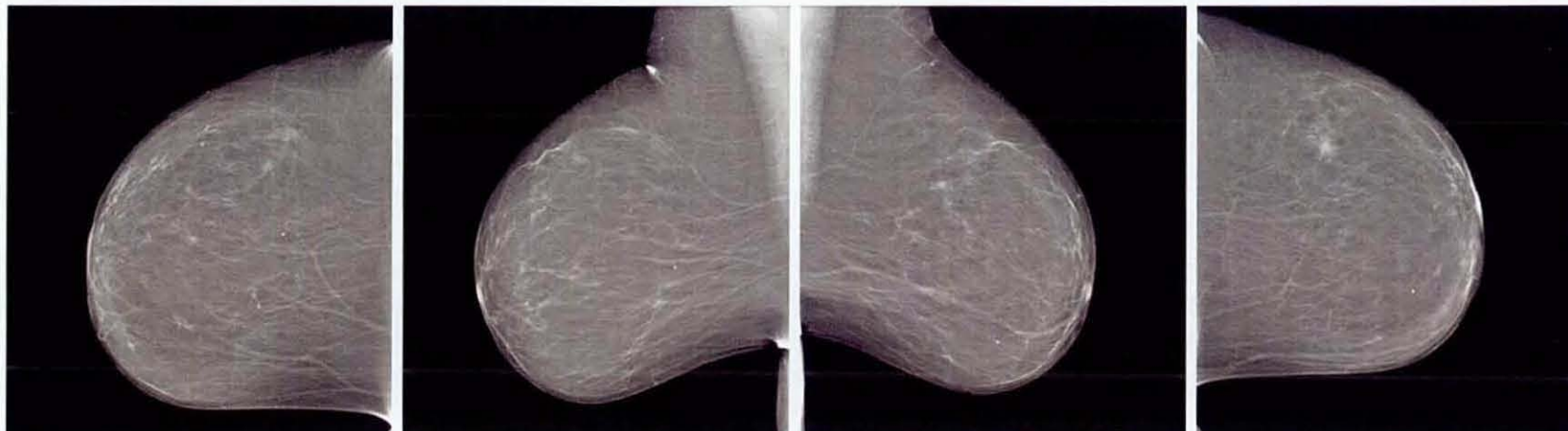


295



Return to Screening / Recall

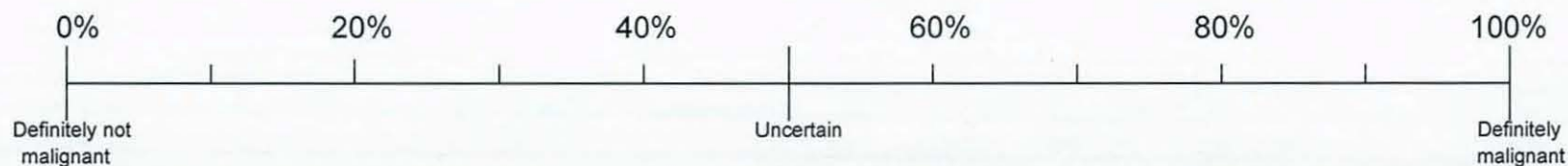
Case
86



296

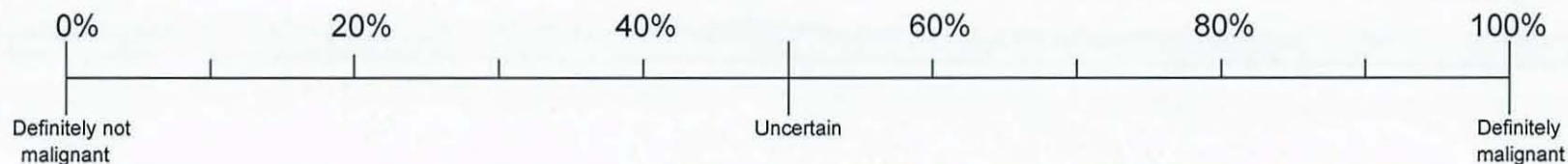
Lesion 1

Probability of malignancy



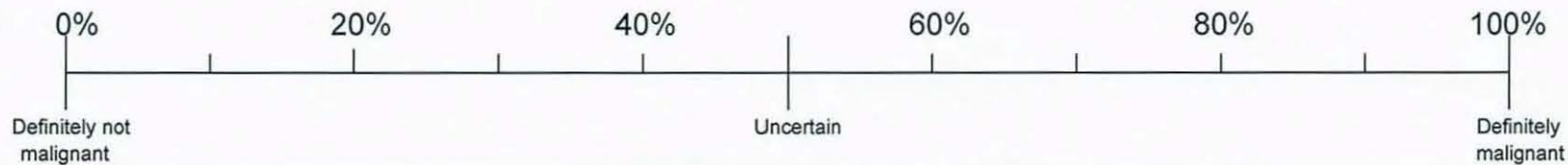
Lesion 2

Probability of malignancy



Lesion 3

Probability of malignancy

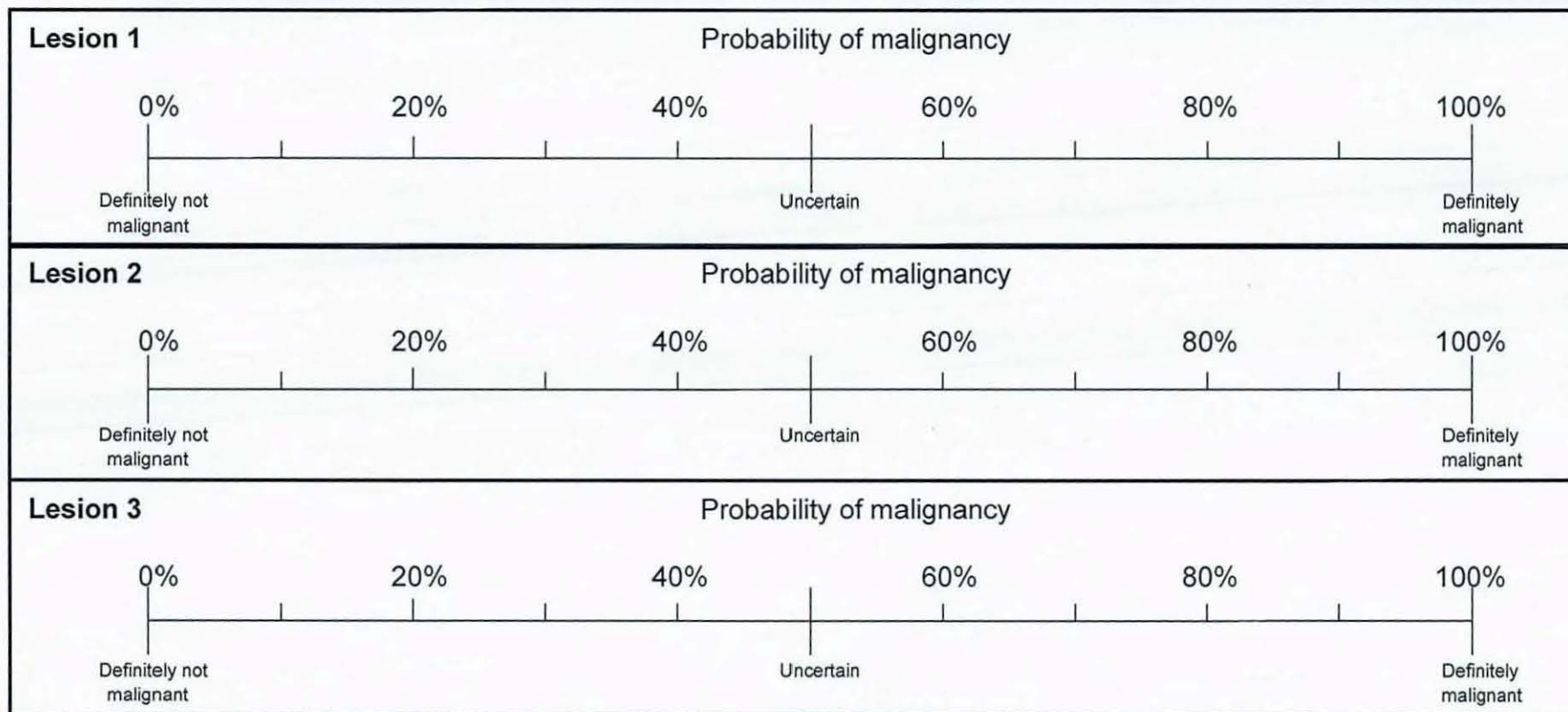


Return to Screening / Recall

Case 21



297



Return to Screening / Recall

Appendix 9 - Publications by Sian Taylor-Phillips

Taylor-Phillips S., Wallis, M.G., Gale, A G., Should previous mammograms be digitised in the transition to digital mammography?, *European Radiology*, 2009, 19 (8), pp 1890-1896.

Taylor-Phillips, S., Wallis, M.G., Duncan, A., Gale, A.G., The Effect of Digitising Film Prior Mammograms on Radiologists' Performance in Breast Screening: A JAFROC Study, in *Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, 2009, Sahiner, B., and Manning, D.J. (eds), Proceedings of SPIE, 7263, (SPIE, Bellingham, WA, 2006), 7263-36.

Taylor-Phillips, S., Wallis, M.G., Gale, A G , Should prior film mammograms be digitised during the transition to digital mammography? *Breast Cancer Research*, 2008, 10(3), pp25.

Radiologist Performance in the 21st Century: The impact of Workstation Type and MSD's, Taylor-Phillips, S., Gale, A G , Wallis, M.G., *Contemporary Ergonomics*, 2008, pp 363-368.

Taylor-Phillips, S.; Wallis, M. G.; Gale, A. G., Mammography workstation design: effect on mammographer behaviour and the risk of musculoskeletal disorders, in *Medical Imaging: Image Perception, Observer Performance, and Technology Assessment 2008*, Sahiner, B., and Manning, D.J. (eds), Proceedings of SPIE, 6917, (SPIE, Bellingham, WA, 2006), 6917-51.

Taylor-Phillips, S.; Wallis, M G.; Duncan, A , Gale, A. G., Performance in Digital Mammography with and without Film Prior Mammograms, *Breast Cancer Research* 2009, **11**(Suppl 2),pp15

