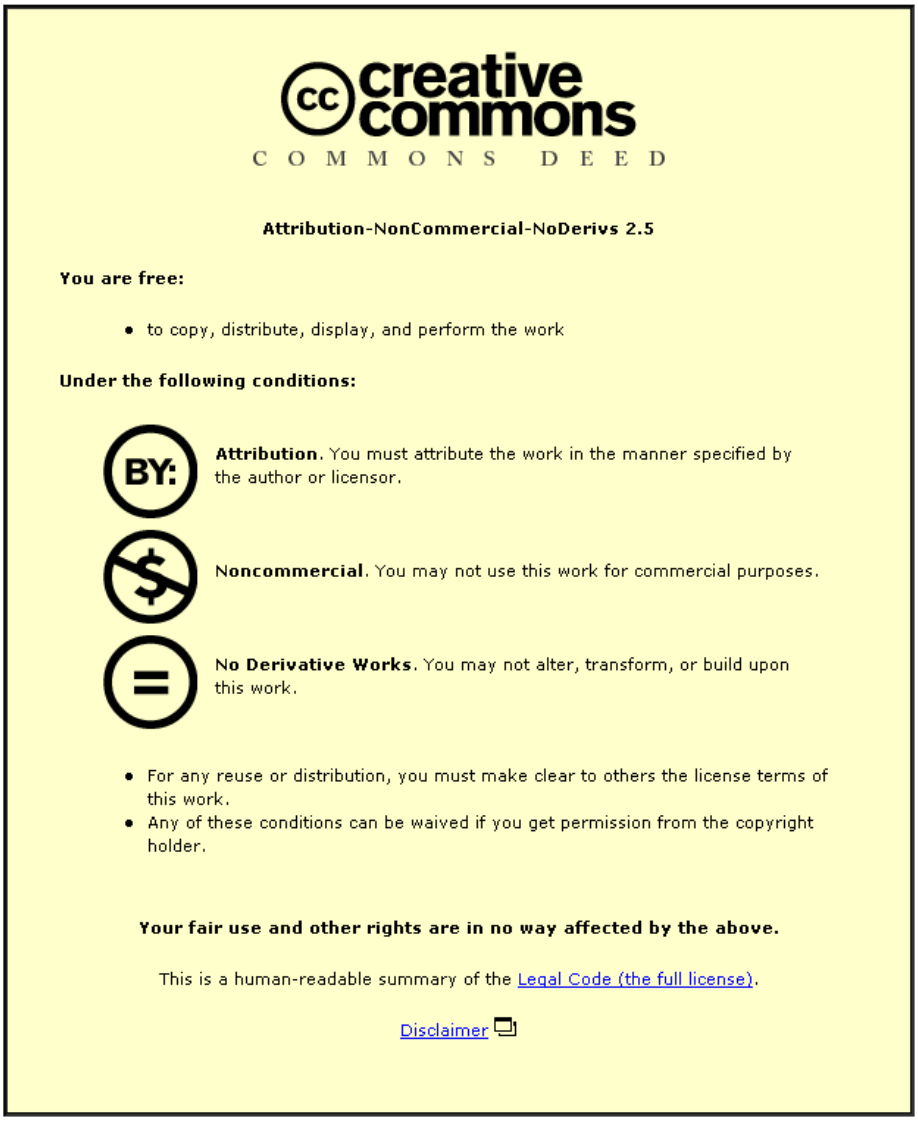

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

How much is enough? Factors affecting the optimal interpretation of breast screening mammograms.

Hazel J. Scott* and Alastair G. Gale
Applied Vision Research Centre, Loughborough University, Garendon Wing, Holywell Way, Loughborough, LE11 3TU, UK

ABSTRACT

PERFORMS (Personal Performance in Mammographic Screening), a self-assessment scheme for film-readers is undertaken as an educational tool by mammographers reading breast-screening films in the UK. The scheme has been running as a bi-annual exercise since its inception in 1991. In addition to completing the scheme each year the majority of film-readers also choose to complete a questionnaire, administered as part of the scheme, indicating key aspects of their every-day reading practice. These key aspects include, volume of cases read per week, time-on-task reading screening films, incidence and time of break periods as well as typical number of film-reading sessions per week.
Previous recommendations on best screening practice (significantly the optimum time on task) were considered in the light of these film-readers' self-reports on a current PERFORMS case set.
In addition we looked at performance accuracy of over 450 film-readers reading PERFORMS cases (60 difficult mammographic cases). Performance on measures akin to True Positive (Correct Recall Percentages) and True Negative (Correct Return to Screen Percentages) decisions were investigated. Data presented demonstrate that individual behaviours in real life screening, for the interpretation of mammographic cases, affect film-reading accuracy on a test set of mammograms for specificity and sensitivity (namely volume of cases read per week and film-reading experience). The consequences for best screening practice, in real life, are considered.

Keywords: PERFORMS, Breast Screening, Performance, Best Practice, Radiologist, Time-on-Task, Case Volume, Task Frequency, Sensitivity.

1. INTRODUCTION

Key issues salient to the optimal interpretation of breast screening mammograms have been identified in recent years (Laming & Warren, 2000)[1]. Several recommendations were proffered including, the optimum time-on-task period that any given film-readers should allow when reading breast screening mammograms (it was recommended that film-readers should not read for more than 30 minutes at a time), that films should be double-read in inverse order and clerical checks should be in place to monitor when that detection efficiency tails off. In addition it was recommended that immediate feedback, on radiological accuracy, would improve performance. Such a mechanism exists as a self-assessment scheme in the UK which allows all participating film-readers to evaluate their performance on a test set (not real life) of mammograms. This scheme, PERFORMS (PERsonal PerFORmance in Mammographic Screening), established since 1991 is a free and anonymous exercise consisting of difficult screening cases and provides immediate and confidential feedback to all film readers on their respective performance based on a radiological "gold standard". The scheme functions as a self assessment exercise whereby mammographers have the opportunity to evaluate 120 challenging (60x2) cases bi-yearly (c.f. Gale and Walker, 1991)[2].
In addition to completing PERFORMS many film-readers also provide information, via questionnaire, on their regular mammographic reading practice. We compared these reading practices with performance on recent PERFORMS cases - with a view to elucidating which practices were a) most common and b) were associated with optimum proficiency on PERFORMS. In this way we aimed to extend the research into best screening practice by exploring typical reading practice (such as characteristics of film reading sessions) with direct comparison to individual performance characteristics - in light of previous research.

*h.scott@lboro.ac.uk; phone, 0044 1509635733; fax, 0044 1509635736; www.appliedvision.org

Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang, Berkman Sahiner, Proc. of SPIE Vol. 6515
65150F, (2007) · 1605-7422/07/$18 · doi: 10.1117/12.708159

2. METHODOLOGY

A large majority of film-readers on the UK NHSBSP, including Radiologists and Technologists, choose to take part in PERFORMS- although it is, at present, a voluntary activity. All cases for PERFORMS film sets are amalgamated, every year, nationally from Breast Screening Units throughout the UK.

Sitting PERFORMS requires individuals to enter data about each case, into a tablet PC, whereupon they receive detailed feedback via the tablet PC on all aspects of their performance (on each case) compared with an experienced panel's radiological decisions. This initial radiological standard is based on the decisions of a panel of five very experienced readers as well as, where relevant, on case pathology. Subsequently, this initial radiological opinion is replaced by a 'national radiological opinion' about each case, this is gleaned from pathology as well as from the majority opinion of all participating film-readers (over 550 in this instance) as a fairer measure of performance i.e. if any given case should be recalled or not. Therefore, for this analysis, all participants' data were calculated against this 'national opinion'.

Following the completion of 120 cases, all participants' performance, as measured against the 'national opinion', are calculated and disseminated back to individuals compared to the anonymous data of their peers. Specifically, individuals receive detailed feedback on their number of 'Correct Recall' decisions (a measure of sensitivity), 'Correct Return to Screen' decisions (a measure of specificity), percentage of correct malignancy's detected as well as ROC measures such as d' and d' for pathology.

The anonymous data of 451 UK breast-screening film-readers were compared on the most recent PERFORMS film set. In addition all individuals were invited to complete a computerised self-report detailing their most common reading practice. From this questionnaire data, we analysed elements of their regular screening practices, principally, the number and length of weekly sessions reading, time of day they normally read, details of task interruption, breaks on task, double reading practices, volume of cases read per week and years of screening experience. Accuracy on the PERFORMS set was measured by both sensitivity and specificity percentages as well as ROC d' measures.

3. RESULTS

Inclusion criteria for this study were those participants (n=451) that completed not only the most recent round of PERFORMS (60 difficult cases) but who had also completed the self-report, this was a cohort of the original group of participants who completed only the PERFORMS set (approximately 550). For this analysis, overall accuracy measures based on the National Radiological Opinion (using mean scores of performance measures) were compared with all self-report data. We set out to investigate which behaviour or practice was common for breast screening readers in the UK and in addition examined any relationship with attainment on the PERFORMS set.

3.1 Normal session length

Film readers were asked to report on the most common session length, i.e. the amount of time that they would film-read before stopping. The range of session length was less than '30 minutes' to 'over 120 minutes'.

The most commonly reported session length was from '60-89 minutes' (see Figure 1) for 133 participants. This was closely followed by '30-59 minutes' (129 participants), with 'under 30 minutes' being the least reported session length (9 participants).

In order to assess if session length in real life affected performance on the PERFORMS set, a one way ANOVA was performed on the data with one IV (session length) and with DV's of the PERFORMS specificity and sensitivity measures.

There were no significant differences for measures of sensitivity, CR (Correct Recall) and Malignancies detected or for measures of specificity, CS (Correct Return to Screen) but there were significant differences for overall ROC PERFORMS measures (of d' and d' for pathology), d' [$F(4,440) = 3.4$, $p<.01$], d' pathology [$F(4,440) = 3.94$, $p<.005$]. For both these measures, participants who reported reading for shorter sessions, 30 minutes or less, performed significantly less well on the current PERFORMS set compared with those who read for longer time periods (see Figure 1a)

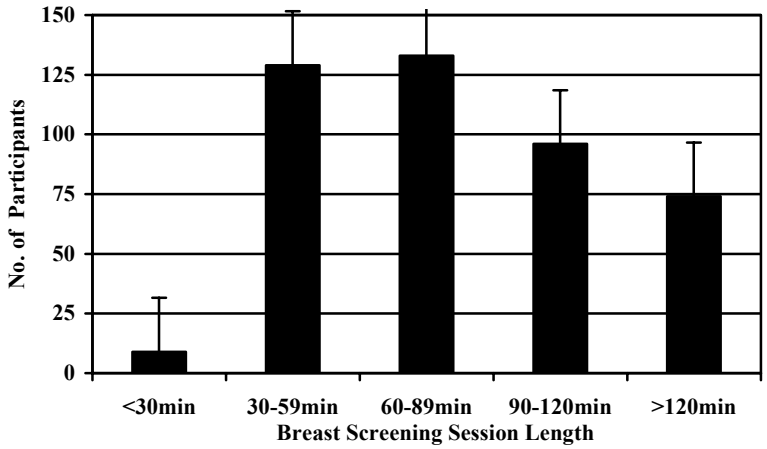

Figure 1: Reading session length in normal practice

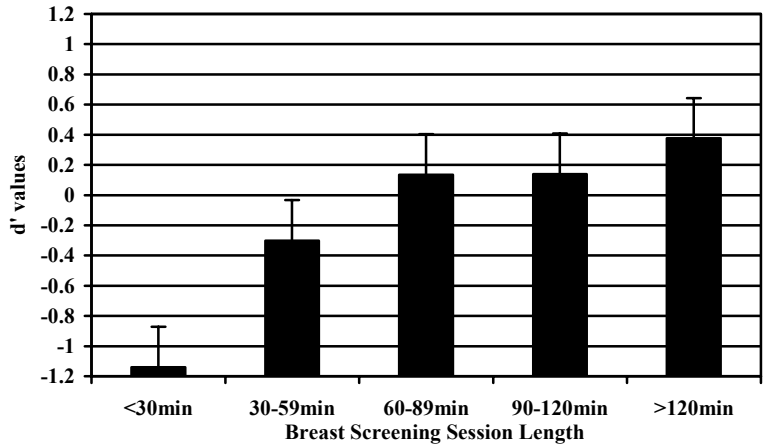

Figure 1a: Reading session length in normal practice and PERFORMS scores for d'.

3.2 Number of reading sessions per week

When reporting the number of reading sessions per week the most common was two sessions (range 1-5+sessions) with 193 participants reporting this frequency (see Figure 2). Comparison with PERFORMS results were non-significant. Therefore, reported sessions reading per week did not affect performance on the PERFORMS cases.

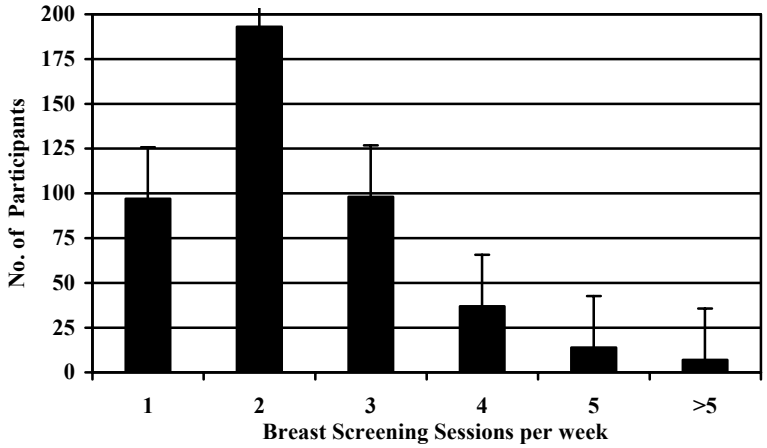

Figure 2: No. of reading sessions per week in normal practice.

3.3 Normal time of day for reading films.

The normal time of day for reading breast-screening films was most popularly between '9-12am' (n=248) and '2-6pm' (n=206). Normal time of day reading films did not relate to performance on PERFORMS (p=n.s.).

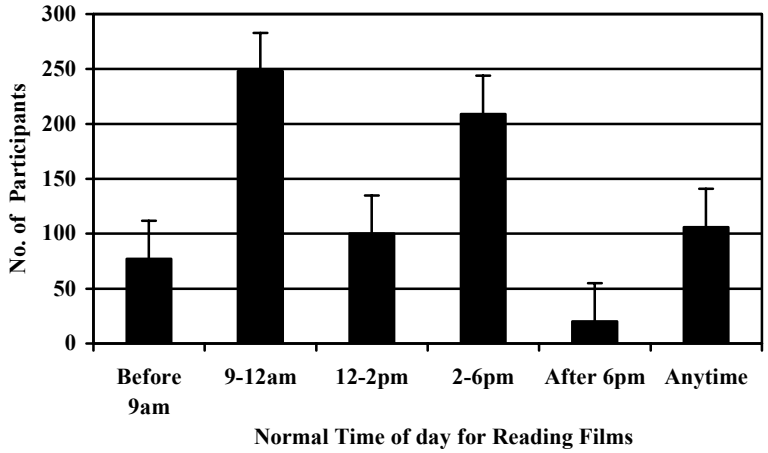

Figure 3: Normal time of day for reading films

3.4 Type of break when on task

Participants were asked to report on taking breaks whilst still on task, the type of break most commonly reported was '0-10 minutes' followed by '11-20 minutes'. Instance and type of break (see Figure 4) were not related to any of the PERFORMS measures (p=n.s.)

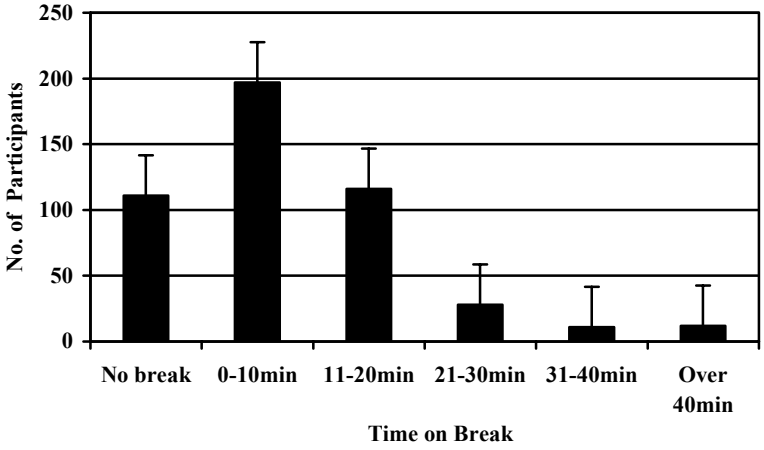

Figure 4: Instance and type of breaks whilst on task.

3.5 Interruptions whilst film-reading

The kind of interruptions film-readers reported as the most common (see Figure 5) were 'interruptions from colleagues' (n=285 participants) and the interruption least likely was 'called away to emergency' (n=22). Again no differences were found for each of these groups for any of the PERFORMS performance measures.

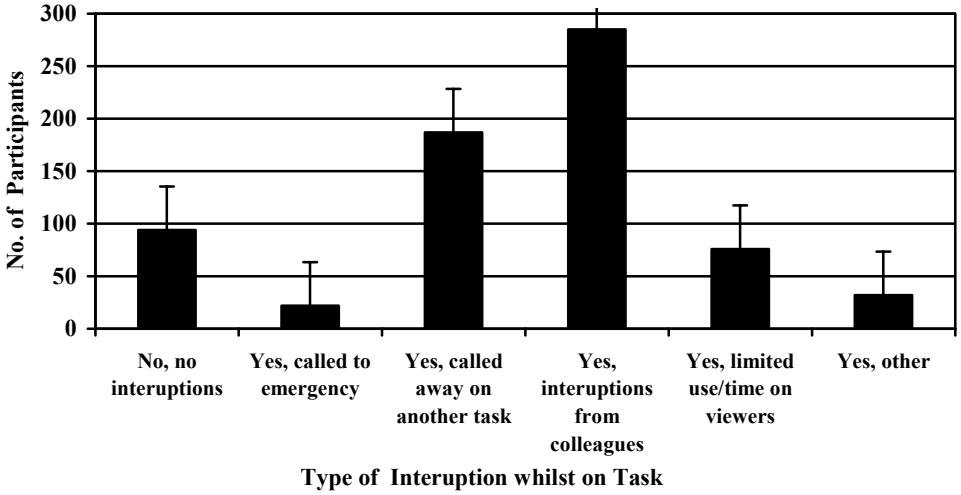

Figure 5: Instance and type of interruptions whilst on task.

3.6 Double-reading and PERFORMS

An overwhelming majority (n= 423) of participants reported that they double-read cases. This was related to sensitivity measures in PERFORMS but the number who did not double read was too small to be included in the analysis (n=2).

3.7 Film-readers individual differences- Volume

The most common volume of cases per week was between '100-199' (over 5,000 cases per year), see Figure 6. For the purposes of post-hoc analysis, the highest volume group was removed as it contained less than 5 participants. A one way analysis of variance identified significant differences for sensitivity measures, Correct Recall [$F(6,440) = 3.23$, $p<.01$] and Malignancies Detected [$F(6,440) = 3.98$, $p<.005$] as well as for d' for pathology [$F(6,440) = 2.45$, $p<.05$]. Student Newman Keuls post hoc analysis revealed that those with a higher volume (of over 400 cases per week) perform better than those who read less than 100 cases per week (see Figure 6a) on all sensitivity measures ($p<.05$),

there were no significant groups differences for d' pathology (p=n.s.). There were no significant differences for CS or for d'.

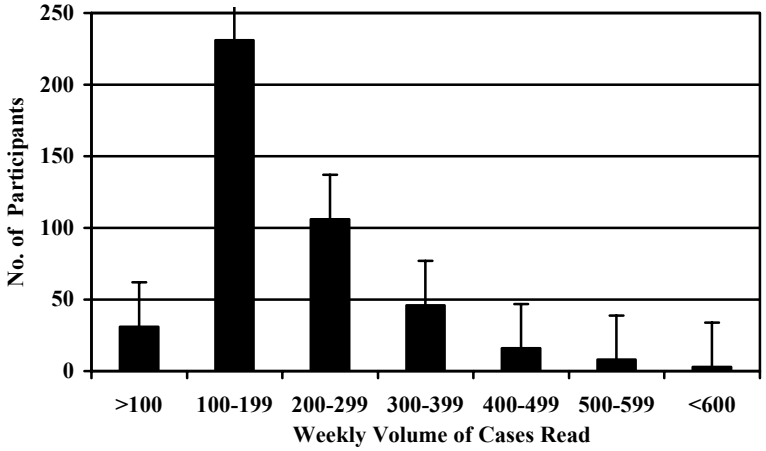

Figure 6: Typical weekly volume of cases.

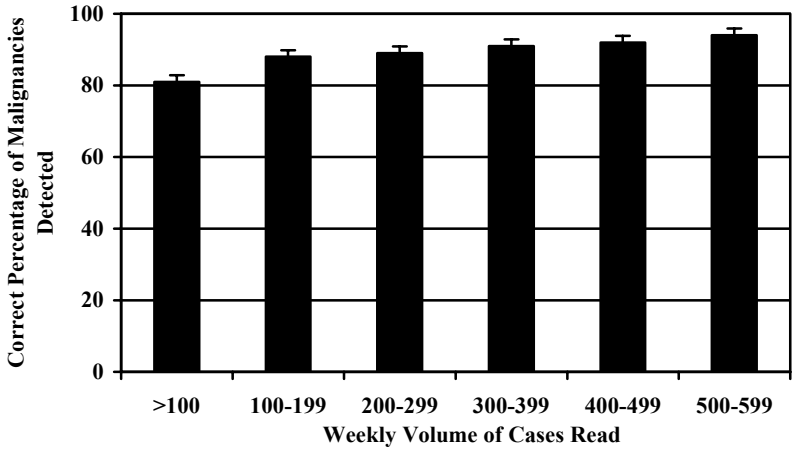

Figure 6a: Case volume and cancer detection on PERFORMS

3.8 Film-readers individual differences- Years of film-reading experience.

Years of reading experience in mammographic film reading were grouped for ease of analysis into 4 main experience groups ('1-5 years', '6-10 years', '11-15 years', and 'over 15 years), the largest group had between '1-5 years' of experience. A one-way ANOVA revealed significant differences for all PERFORMS measures ($p<.05$). However, post hoc Student Newman Keuls identified that for all sensitivity measures, experience groups of 11 years and over performed significantly better than groups of '6-10 years' ($p<.05$) which in turn performed significantly better than the lowest experience group of '1-5 years' ($p<.05$). This trend was contrary to that found for specificity measures, where the low experience groups ('1-5 years' and '6-10 years') performed significantly better than higher experienced groups of '11-15 years' ($p<.05$), see Figure 7a. For overall ROC measures Student Newman Keuls post-hoc tests showed that groups '6-10 years' and 'over 15 years' outperformed those who had been reading for '1-5 years' ($p<.05$)

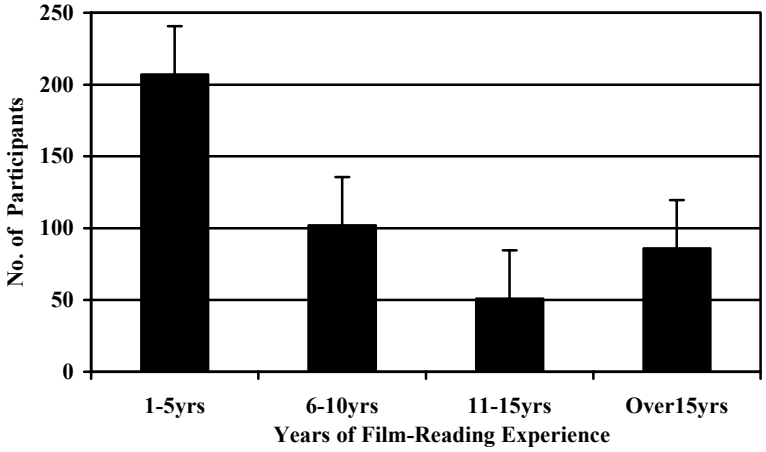

Figure 7: Years of Film-reading experience.

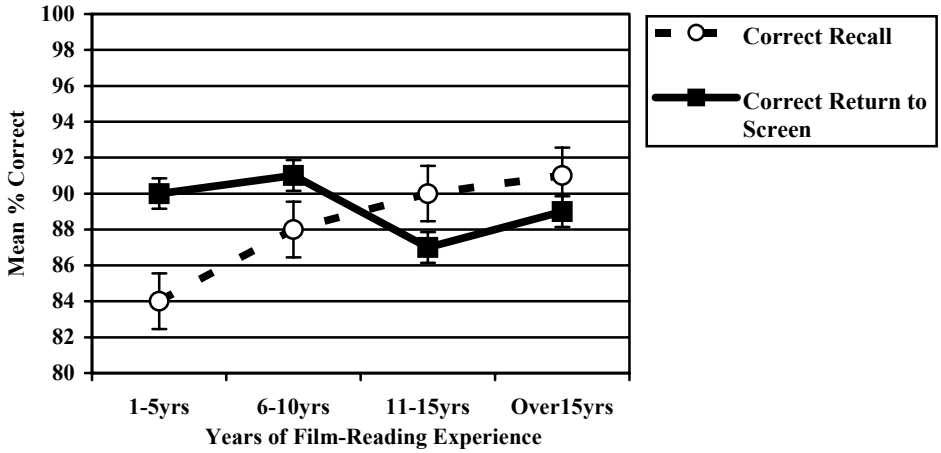

Figure 7a: Sensitivity and specificity on PERFORMS by years of film-reading experience

3.9 Typical readers' profile and relation to attainment on PERFORMS

A common profile for the 451 participants who completed the self-report was as follows:-

Table 1: Self -report profile and relation performance on PERFORMS

Measure	Most Typical Practice	Related to PERFORMS?
Normal Session Length*	60-89minutes	Yes, d' and d' pathology only.
No. of Sessions	2 per week	No
Time of Day	9-12am	No
Type of Break	0-10mins	No
Double Reading	Yes	NA
Volume of Cases*	100-199 per week	Yes, all Sensitivity measures
Years reading*	1-5years	Yes, all PERFORMS measures.

* Sign. differences at the p<.05 level.

Reading behaviour related to PERFORMS results were, session length, volume and experience. However, when all three were entered into a multiple regression analysis (enter method) a weak but significant model emerged in which volume was not a significant predictor of any performance measures over years of reading experience where session length was also entered (see Table 2). Contrastingly, using the same model, years of reading experience was a significant predictor of all sensitivity, specificity and ROC measures. In addition, normal length of session was a significant predictor of malignancies detected, d' and d' pathology.

Table 2: Multiple-regression results for all PERFORMS measures

Using the enter method a weak but significant model emerged for each of the performance measures.

Measure	Model Significance and Adjusted R^2		Predictor variables	Beta	P
Correct Recall	($F_{3,432}$ = 15.684,p<.0005)	Adj. R^2=.092	Years Experience	.302	p<.0005
			(Volume and Session Length, p=n.s in this model)		
Correct Return to Screen	($F_{3,432}$ = 2.731,p<.05)	Adj. R^2=.012	Years Experience	.-132	p<.025
			(Volume and Session Length, p=n.s, in this model)		
Malignancies Detected	($F_{3,432}$ = 19.041,p<.0005)	Adj. R^2=.111	Years Experience	.323	p<.0005
			Session Length	.106	p<.025
			(Volume, p=n.s in this model)		
d'	($F_{3,432}$ = 5.559,p<.001)	Adj. R^2=.030	Years Experience	.147	p<.005
			Session Length	.106	p<.05
			(Volume, p=n.s in this model)		
d' for Pathology	($F_{3,432}$ = 7.821,p<.0005)	Adj. R^2=.045	Years Experience	.147	p<.005
			Session Length	.106	p<.05
			(Volume, p=n.s in this model)		

4. DISCUSSION

To the question of best screening practice, in terms of screening practice that is common (from participants' self-reports) individuals are reading in sessions twice as long as the Laming and Warren (2000) report recommends (30 minutes) at a time. Furthermore, from overall d' values, those individuals that report typically reading for under 30 minutes continuously, in real life screening, performed less well here than those who normally read for longer. This may suggest that as well as being a recommended maximum time for film-reading sessions, there may also be a minimum. However, the number of participants in the under 30 minutes group was very small (n=9) so this may well account for the variability in performance compared to the longer time sessions. Furthermore, there were no significant differences between the other session time periods.

The self-reports illustrate a likely 'profile' of typical reading behaviour, namely; reading in sessions lasting just over an hour, reading mammograms in two weekly sessions, reading was frequently completed in the morning between 9-12pm with breaks of approximately 10 minutes. The type of interruption most prevalent during reading sessions was reported as interruptions from colleagues. Double reading was overwhelmingly adhered to in the current sample and individuals read on average between 100-199 cases per week (annual volume of 5,200 cases). This is in accordance with standard UK NHS recommendations (that individuals read over 5,000 cases per year). In terms of years of reading experience, the largest group currently reported having between 1-5 years experience, followed by the most experienced group (of reading for over 15 years).

Although such questionnaire/self-reports are useful in suggesting an overall common reader profile for the majority of breast screening readers who completed the PERFORMS scheme, perhaps more importantly, we investigated if performance on PERFORMS was affected by any reported aspect of real-life screening practice. As previously mentioned, normal length of reading session time was significant, but no other element relating to real life screening

practice had any effect on PERFORMS results, with the exception of volume of cases read and years of screening experience.

Volume of cases read had a significant effect on all sensitivity measures, whereby performance improved with higher volumes of weekly cases (significantly those participants who read over 400 cases per week were better than those who read up to 100 cases per week). These data echo our previous findings on Technologists' case volume and real-life factors[3], which demonstrated that volume drives sensitivity on PERFORMS. The study highlighted that those who read a greater number of cases detected significantly more percentages of malignant cases and had a higher percentage of Correct Recall without affecting specificity scores. In addition, these results also support Esserman, Cowley, Eberle, Kirkpatrick, Chang, Berbaum and Gale's (2002)[4] findings, which revealed that a high volume of cases in real life mirrored better overall performance on the PERFORMS scheme, although they found this for *both* sensitivity and specificity measures.

Moreover, Esserman et al. (2002) did not specifically test years of experience (which they argued is both a combination of volume of cases and amount of time as a Mammographer) however, they did concede that this may be akin to the relationship between volume and performance. Indeed, this is what was found in the present study. Years of reading experience (collapsed into four main groupings) were significantly related to all PERFORMS measures. These data showed different patterns for sensitivity measures (where increased years of experience had significantly beneficial effect on performance) compared to specificity measures where those groups with less years of experience showed significantly better performance. This suggests that years of experience, like case volume, drives sensitivity rather than specificity measures (which conflicts somewhat with what we have found for years of experience previously for Technologists[3]).

Multiple regression analysis showed that, of the three factors that had a significant effect on performance, namely; volume, experience and time on session, only years of reading experience was significant for all predictive models (see Table 2). Volume of cases was non-significant in all cases and length of session time was significant only for ROC measures and malignancies detected. Although these models were of weak (but significant) predictive value they served to possibly illuminate the significance of years of experience over volume of cases read.

Pinpointing which screening practices were optimal, from these data, was less than obvious. Main factors affecting performance were shown to be volume, experience and session time. Years of screening experience, rather than any particular reading habit or 'style', affected performance on the present PERFORMS case set. However, one limitation of the current study is that it compared reported reading styles in real life with performance on PERFORMS, real life reading practices may not necessarily reflect how individuals read PERFORMS at any given time. It was suggested that a further study examining factors relating to how participants actually took part in the PERFORMS scheme (e.g. time on task, incidence of breaks, no of reading sessions etc) would further elucidate which reading practices best characterise optimum radiological performance.

5. CONCLUSIONS

The results indicated specific trends in reading practices in terms of volume of cases read, session length and years of reading experience. These were related to attainment on the PERFORMS scheme for the current year.

It was concluded that there were definite trends distinguishing everyday screening practice and these were related to previous work on best screening practice. In order to provide a bench mark for how these behaviours bear upon performance it is suggested that a further study relating actual PERFORMS reading behaviour to individual accuracy on a test set of mammograms was warranted.

ACKNOWLEDGEMENTS

This work is supported by the UK National Health Service Breast Screening Programme.

REFERENCES

1. Laming, D. & Warren, R. Improving the detection of cancer in the screening of mammograms. Journal of Medical Screening, 2000:7:24-30.
2. Gale A.G. & Walker G.E. Design for performance: quality assessment in a national breast screening programme. In E. Lovesay (Ed.) Ergonomics - design for performance 1991, Taylor & Francis, London.
3. Scott, H.J., Gale, A.G., Wooding, D.S.: Breast-Screening Technologists: does real life case volume affect performance?, In: Medical Imaging 2004: Image perception and performance. Miguel P. Eckstein & D.P. Chakraborty (Ed.). Proceedings of SPIE Vol. 5372, 399-406.
4. Esserman L., Cowley H., Eberle C., Kirkpatrick A., Chang S., Berbaum K., & Gale A.G.: Improving the Accuracy of Mammography: Volume and Outcome Relationships. Journal of the National Cancer Institute, 2002, Vol. 94, No. 5, 369-375, March 6