Title: An investigation to ascertain whether or not time pressure influences the accuracy of final year student radiographers in abnormality detection when interpreting conventional appendicular trauma radiographs: a pilot study

Samantha Whitaker; William A Cox

University of Portsmouth

Introduction

Growing levels of service demand[1] and a shortfall in radiologists[2] are leading to increasing pressures on diagnostic imaging departments and these trends are expected to continue[2]. Additionally, the Care Quality Commission (CQC) recently reported that while several trusts demonstrate reporting backlogs, not all trusts are utilising reporting radiographers to alleviate the workload[3, 4], therefore increasing the pressure on radiologists with a larger workload than needed. Despite these pressures, it is expected, under the Royal College of Radiologists' (RCR) Standards[5], that reports are produced in a timely fashion, to support the speedy diagnosis of expected and unexpected findings and influence management[5]. However, as well as being timely, it is also important that reports are accurate as the accurate interpretation of images will ultimately provide safe and optimal care pathways for patients by answering a clinical question[5].

Given these potentially conflicting factors, it is important to consider how time pressures may affect accuracy in image interpretation. This study has been conducted to identify a link between time pressure and interpretation accuracy that may be relatable to the decision making processing of reporting practitioners. There is a rich background of theoretical work which considers the effects of time pressure on decision making including the dual systems theory, Yerkes-Dodson law and Gestalt theory. The dual systems theory is underpinned by system I (a fast, automatic cognitive process) and system II (a slower, conscious process)[6]. System I uses key information, whereas system II provides reasoning in formulating a decision[7, 8, 9]. Added time pressure has been shown not only to reduce the amount of evidence accumulated, but also to lower decision boundaries[10]. The decision boundaries in radiology may manifest in the form of satisfaction of search[11]. The Yerkes-Dodson law provides the concept behind an optimum level of stress for efficient working. Working under stress levels either side of this optimal stress can have an adverse impact on performance[12, 13]. The Gestalt law of visual perception, on the other hand, considers the potential for misperception of visual appearances due to one or more of the underpinning principles, for example; closure, proximity, symmetry, similarity and continuity[14]. It has also been recognised in literature that detectability decreases with severely restricted viewing times, but also with unlimited viewing times, overestimating with positive observations[15].

Effects of Time Pressure on Decision Making

Numerous studies suggest that there is a negative correlation between speed and accuracy and that time pressure increases risk taking[7, 8, 10, 16]. These studies performed computerised tests where participants needed to make decisions under

different levels of time pressure. However, Dambacher and Hubner's study used only sixteen participants; this sample size would not necessarily reflect a population's ability. The other three studies had a larger sample size of either 72 or 101 participants, making their results more reliable in reflecting the population. This research offers the possibility that pressure to produce timely reports may be influencing the decision making process of reporters and therefore, image interpretation accuracy. However, little work considers the above as a factor in terms of report writing which may be due to the varied responsibilities of reporters making it difficult to monitor work performance.

Studies Assessing the Impact of Time Pressure on Image Interpretation

A dated study, using only four observers, demonstrated flash viewing to be largely effective in identifying obvious lesions, however, the difference in detectability was exaggerated; when allowed unlimited time, a substantial portion of subtle lesions were missed[17]. A study by Edwards et al[18] asked fourteen radiologists to report on ninety images in 3 batches of 30. Each batch was reviewed at a different speed. The results showed that the number of false positives decreased when under pressure. However, although nearly all radiologists coped with reporting in half of their original time, they reported that they would not be able to continue under such pressure[18]; working under prolonged pressure can result in visual fatigue and cognitive overload[19]. Finally, a more recent study revealed that four out of the five radiologists that took part had more major misses when reporting at twice their original reporting speed[20]. Despite results of these studies, Muroff and Berlin[21] note that currently, validity is compromised in studies evaluating the relationship between speed and accuracy as they do not provide sufficient data for accurate and reliable results. This research was undertaken to identify the relationship between time pressure and performance in interpretation accuracy.

Methods

Study Design

Participants [n=21] were split into three equal sized groups of seven and were randomly allocated their level of intended time pressure: 15 seconds (high pressure), 30 seconds (moderate pressure), or unlimited time (low pressure). A set of 30 images were presented to participants via a presentation slideshow. Each group was asked to record whether they judged the images to be normal or abnormal and to indicate location for any abnormality identified on the answer sheets provided. The pre-set timings meant images could not be manipulated; the unlimited time group (UTG) were instructed not to close the slideshow and to only progress onto the next image when satisfied with their answer to avoid a second exposure to the images. Time group sessions were conducted separately to minimise disruption upon participants finishing before those with longer time limits.

Data Collection Approach

Each answer was marked true positive (TP), when a participant correctly identified an abnormality, true negative (TN) (correctly identified the image as normal), false positive (FP) (incorrectly labelling a normal area as abnormal), or false negative (FN) (labelling an abnormality as normal). Each question number equalled one whole; with an image displaying two abnormalities, a participant would acquire a TP if both abnormalities were correctly identified, or a TP with the value of 0.5 and a FN with

the value of 0.5 if only one abnormality was identified. Results were then totalled per answer sheet, and then per time group.

Image Selection

The images selected ensured abnormalities were suitable for the level of training that had been received at that point in the students' training, however, it is recognised that this is not necessarily representative of practice. Included in the image bank was a range of extremities, including the pelvis, and paediatric examinations. Anonymised images were obtained from Radiopaedia and the university's image bank; there is a possibility that students may have seen the images previously, but they were not aware these sources were being used. There were fifteen normal and fifteen abnormal images to improve reliability behind the results of each; a number of abnormal images included multiple abnormalities. Both acute fractures and dislocations were included in the range of pathologies.

Sample

Convenience sampling was employed to recruit 21, of a potential 52 participants, from the host site; this style of sampling is a common approach, limited to a group of people most conveniently available to take part[22]. The target group was final year diagnostic radiography students due to their exposure to training on image interpretation, having passed interpretation exams in their second year.

Recruitment

The target group were emailed a participation information sheet and consent form, along with reassurance that answer sheets would be anonymous and they could withdraw from the study before it commences if they wished. Participants were informed of three potential time slots for the study which coincided with their timetables. Participants were randomly entered into time slots and emailed their allotted time.

Process

Participants were emailed the PowerPoint on the morning of their study and given the password for access when the study commenced. Participants were given brief instructions before opening the slideshow, which confirmed the given instructions. Controls were disabled to ensure the slideshow ran on timed slides with no alteration; the unlimited time group moved forward through the slides at their leisure without exiting or moving back through the slideshow.

Pilot Study

A pilot study was conducted to ensure instructions were easy to follow, images were of a reasonable size, abnormalities were visible on the monitors supplied and the timings of the slides were sufficient to apply pressure to the participants. Changes were made to the process in terms of timing and image display, based on the pilot.

Data Analysis Approach

Microsoft Excel and SPSS were used to formulate analysis of the results through graphs and ANOVA tables. The TPs, TNs, FPs and FNs were used to calculate accuracy, sensitivity and specificity across the groups, and then standard deviation

calculated for each. The ANOVA test was conducted to show any statistical significance from results, within the population.

Ethics

A favourable ethical opinion was gained for this study from the University of Portsmouth ethics committee: ref SHSSW/R/18-6.

Informed consent was obtained from every participant and the study was carried out in accordance with the Declaration of Helsinki.

Results

Figure 1 shows the total number of TPs, TNs, FPs and FNs by each time group.

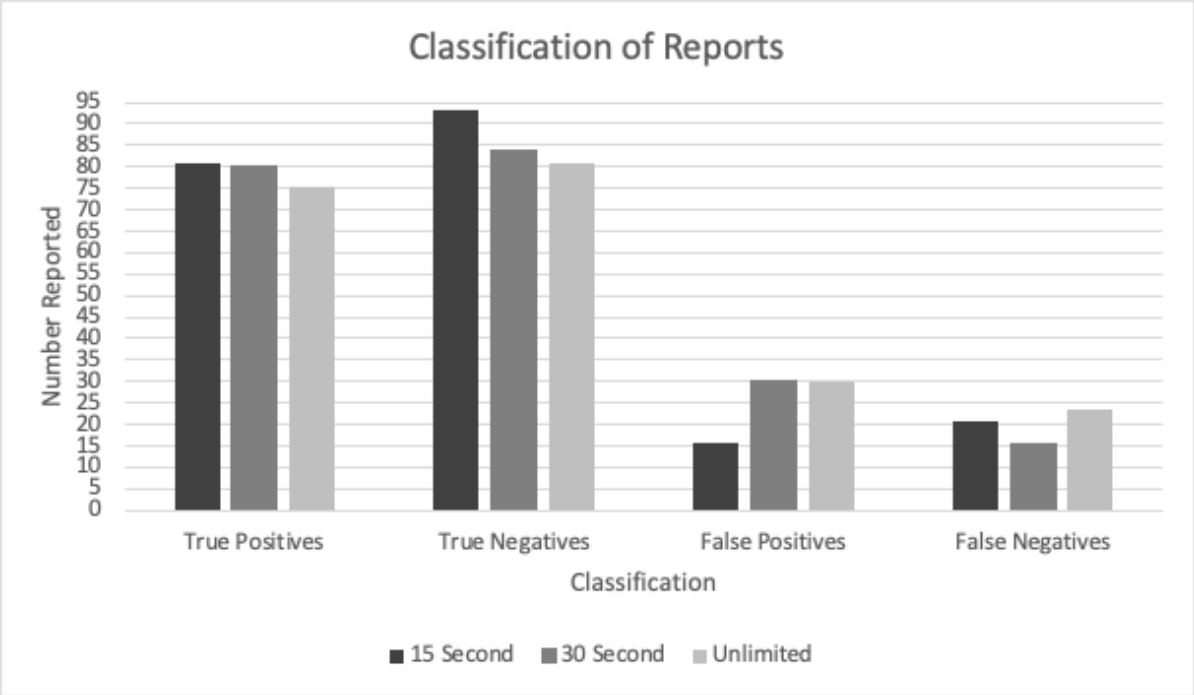

Figure 1: Classification of reports by group. NB - no statistically significant differences are demonstrated

The accuracy, sensitivity and specificity were calculated for each time group, displayed in figure 2, and table 1 also displays the standard deviation for each.

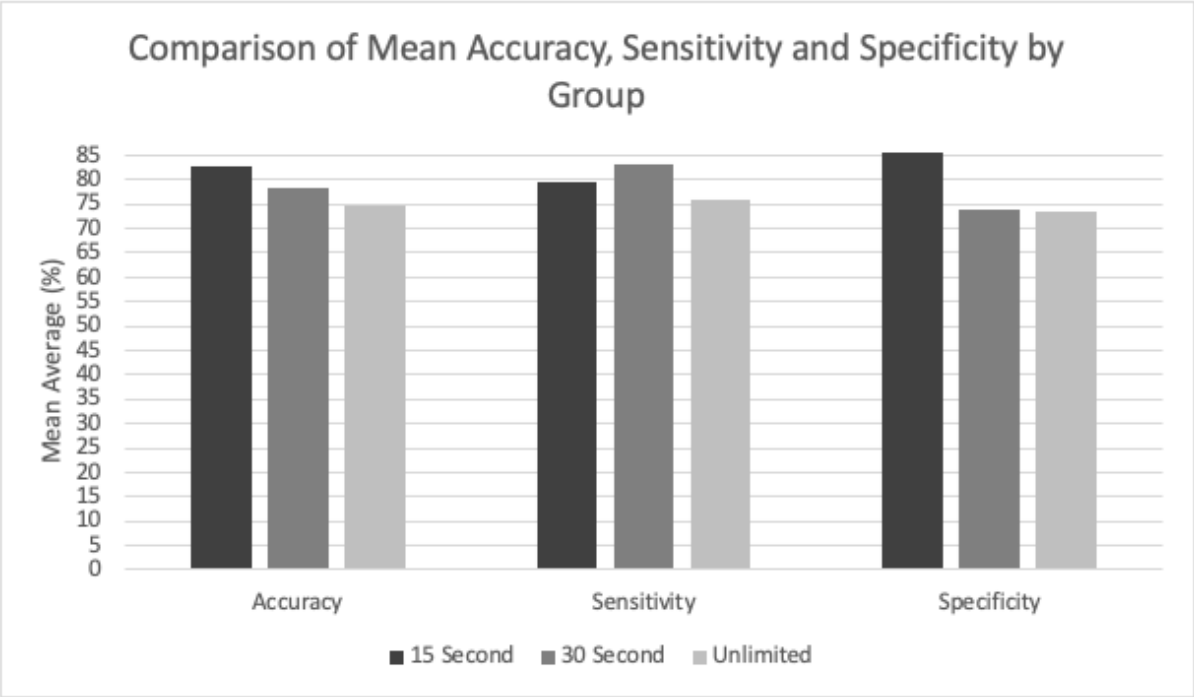

Figure 2: Mean accuracy, sensitivity and specificity by group. NB - no statistically significant differences are demonstrated

Group	Accuracy (%)	Sensitivity (%)	Specificity (%)
15 Second			
Mean Average	82.86	79.65	85.52
Standard Deviation	3.93	8.26	8.58
30 Second			
Mean Average	78.17	83.37	73.74
Standard Deviation	7.40	9.26	8.93
UTG			
Mean Average	74.52	75.83	73.29
Standard Deviation	10.79	13.16	11.54

Table 1: Comparison between time groups. NB - no statistically significant differences are demonstrated

Time comparisons within the UTG were recorded and an average time calculated per image and total for the group. Figure 3 shows the accuracy, sensitivity and specificity for each participant in the UTG, in order of completion time from the quickest to the slowest. For comparison, the 15 second group's accuracy ranged from 80%-90%, sensitivity 68.97%-89.66%, and specificity 73.33%-93.75%. The range of accuracy scores within the 30 second group were 67.22%-88.33%, sensitivity 70.37%-93.90%, and specificity 62.26%-84.85%. The completion time of each participant in the UTG, with a calculation of the average time spent per image, has also been demonstrated in table 2.

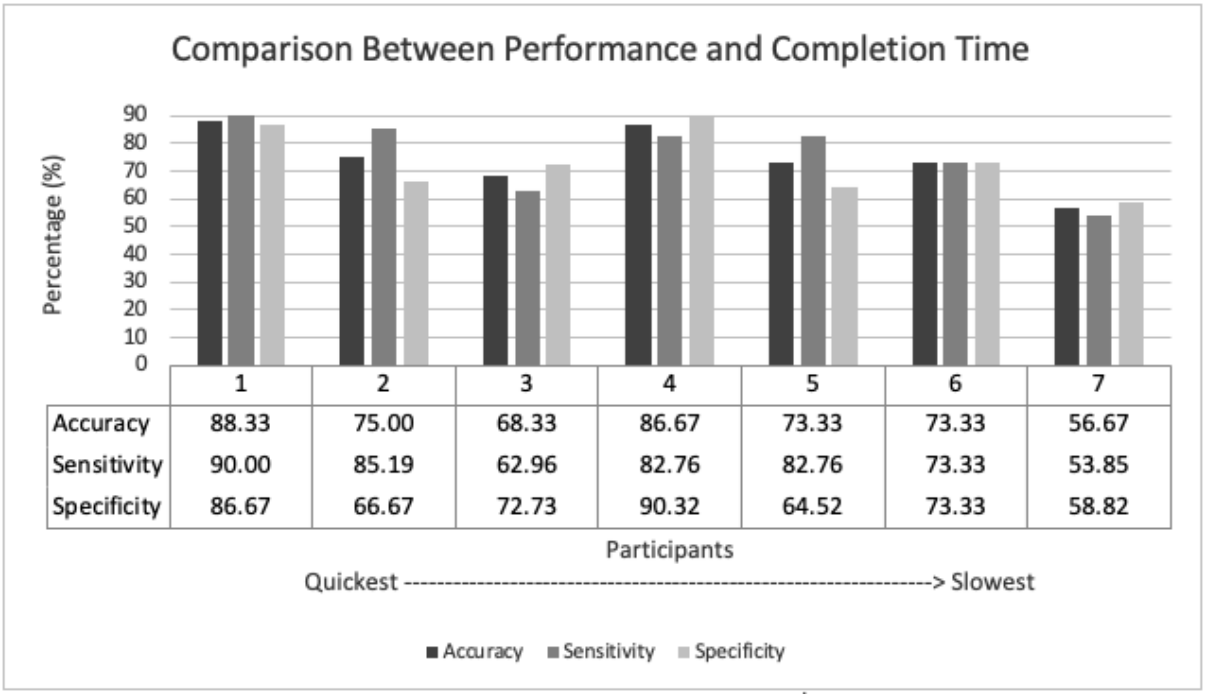

Figure 3: UTG's performance in order of completion time. NB - no statistically significant differences are demonstrated

Time	Participant							**Mean Average**
	1	2	3	4	5	6	7	
Time Taken (mins:secs)	10:36	15:30	15:44	17:32	18:21	18:39	20:01	**16:38**
Average Time per Image (secs)	21.20	31	31.47	35.07	36.70	37.30	40.03	**33.25**

Table 2: UTG's completion times. NB - no statistically significant differences are demonstrated

ROC curves were generated for each time group; the area under the curve (AUC) for the 15 second group was 0.824, 0.785 for the 30 second group, and 0.745 for the UTC. No significant results were demonstrated.

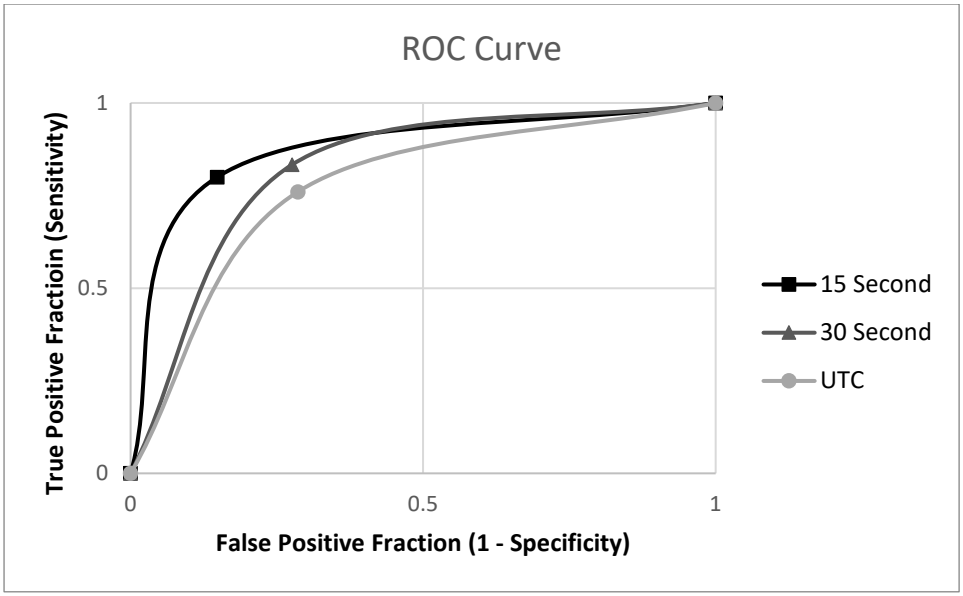

Figure 4: ROC curve: comparison of all time groups

An ANOVA test was conducted using SPSS for accuracy between all time group combinations (table 3).

Dependent Variable: Accuracy						
					95% Confidence Interval	
(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
15 Second	30 Second	4.68429	4.21629	.520	-6.0764	15.4449
	Unlimited	8.33571	4.21629	.147	-2.4249	19.0964
30 Second	15 Second	-4.68429	4.21629	.520	-15.4449	6.0764
	Unlimited	3.65143	4.21629	.668	-7.1092	14.4121
Unlimited	15 Second	-8.33571	4.21629	.147	-19.0964	2.4249
	30 Second	-3.65143	4.21629	.668	-14.4121	7.1092

Table 3: ANOVA test: multiple group comparisons for accuracy. NB - no statistically significant differences are demonstrated

Discussion

ANOVA Test

The results revealed none of the combinations to be of statistical significance. However, the sample size was not large enough to power the test sufficiently to provide assurance that this accurately reflects the reality of the situation. It cannot therefore be ruled out that the results were due to chance, or due to the group at random the outlier in performance, participant 7 in the UTG, was assigned. Despite this, the results do show interesting trends that could be investigated further with a larger sample size.

Classifications

There was little variation between groups in the number of TPs, suggesting that time pressure may not have a great effect in identifying abnormalities; the variance may just be due to different levels of knowledge. On the other hand, the number of TNs decreased as time pressure decreased, suggesting that as time spent assessing an image increases, areas of uncertainty may be dwelled on with the potential to cause a misdiagnosis. However, the fact that the 30 second group had the highest level of sensitivity suggests that there is a minimum time required to fixate and consider an abnormality. It is possible that greater time pressure lowered the participants' decision boundary/quitting threshold[11] and they were, therefore, satisfied with their initial assessment. Similarly, lowering of quitting threshold may explain the resultant FP figures; the 15 second group generated almost half the number of FPs of the remaining groups. Edwards et al.'s study revealed a similar effect[18]. The Yerkes-Dodson Law[13] may explain FNs being higher either side of the optimal pressure, 30 seconds per image.

Accuracy

None of the differences between groups were statistically significant. The mean accuracy was highest in the 15 second group and lowest in the UTG; if a sufficiently powered study were to demonstrate similar results, this could suggest time pressure may be beneficial in the decision making processes underpinning image interpretation. The apparent trend between increased time pressure and accuracy could be due to the shorter time period for image assessment being the optimal pressure required for efficient performance, reflected in the Yerkes-Dodson Law[13]. This may also be seen within the UTG, as participants were aware their time will be recorded upon completion, this may have added an unintentional pressure within the group. Participant 4 may have been working at the optimum pressure for best performance, with the anomaly of participant 1, demonstrated in figure 3.
The extra time the 30 second and UTGs had would have allowed reasoning behind their decisions and more evidence accumulation[6,10], however, the decrease in accuracy suggests time is spent dwelling on areas of uncertainty the longer an image is assessed. This may provide an alternative explanation for results within the UTG, with the anomaly of participant 4, with faster participants using heuristics[6] and slower participants focussing on uncertain areas. However, it is important to recognise that, despite the apparent trend between increased time pressure and interpretation accuracy, Edwards et al.'s study revealed radiologists could not work under such pressure for long periods of time[18]. Therefore, if such cognitive processes also relate to reporting practitioners, departments should ensure plenty of breaks in an attempt to avoid visual fatigue and cognitive overload[19]. It is also important to note that the slowest observer in the UTG had the lowest performance, which may reflect competence, and a single outlier could have skewed the results due to the very limited sample size.

Sensitivity

Results could suggest that a lower level of applied pressure is beneficial in identifying pathology as the 30 second group scored highest for sensitivity, followed by the 15 second and then the UTG. The results reflect the Yerkes-Dodson Law whereby there is an optimum level of stress for best performance, yet either side of this causes performance to deteriorate[13.] However, it is important to note that the

UTG may have scored more poorly due to the impact of interruptions[23] from fellow participants leaving once they had finished. The 15 second group's performance may have been affected by a reduction in quitting threshold due to their limited time[11], or influenced visual misperception, resulting in participants overlooking abnormalities and identifying the image as normal.

Specificity

Results may suggest that too much time assessing normal images has a negative impact on interpretation, as the 15 second group achieved the highest specificity. The 15 second group may have relied on the heuristics in the interpretation of normal images, in line with the dual systems theory[6], basing their decisions on key information. For example, rapid assessment of bone cortices, overlooking additional information in which the other groups may have contemplated an abnormality. The groups with more time may have progressed onto system II[6], initiating self-doubt as part of the reasoning behind the decision. The 15 second group may have also reduced time dwelling on uncertain appearances by subconsciously lowering their quitting threshold[11]. However, it is important to note that the level of knowledge and experience of reporting practitioners in practice may reduce the theory of time dwelled on areas of uncertainty, compared to participants in this study.

Standard Deviation

The standard deviation for accuracy, sensitivity and specificity was smallest in the 15 second group and largest in the UTG. The variance in completion times amongst the UTG may explain their large standard deviation score.

ROC Curve

An AUC value of 0.5 is considered as chance. However, the results from this study, being greater than 0.5, are considered acceptable results for the 30 second and UTCs and considered excellent for the 15 second group[24]. Results suggest an 82% chance the readers in the 15 second group correctly identify normal from abnormal, 79% for the 30 second group and 75% for the UTC.

Limitations

Due to voluntary participation, only 21 participants were reached; a larger sample would have benefitted a broad range of abilities amongst the final year undergraduate cohort to ensure results were reliable. Due to ethical approval, it was not possible to recruit reporting practitioners to take part. This can be seen as a disadvantage as it does not represent the abilities of reporters in practice, nonetheless, the level of knowledge required for this study was deemed suitable for the intended participants to assess abnormality detection accuracy, and identify possible changes in cognitive processes when pressure is applied. There are limitations seen within the UTC; unintentional pressure may have been added upon participants due to recording of their completion time, and as participants in the UTC finished, they may have caused disruption for those still working. Finally, as no statistical significance was demonstrated, it is possible that the results were due to chance.

Conclusion

There is a rich background of theoretical work which considers the effects of time pressure on decision making including decision making theories such as the dual systems theory and the Yerkes-Dodson law. The results may suggest a decrease in overall accuracy the longer images were assessed (accuracy at 15 seconds = 82.86, accuracy at 30 seconds = 78.17, and accuracy within unlimited time = 74.52), however, interpretation of the abnormality may be less demanding in an image bank compared with clinical practice. Also, with the small sample size the ANOVA test suggested the difference between groups within a population to be insignificant. It is recommended that future researchers use a power calculation to ensure the study is sufficiently powered. Despite this, Yerkes and Dodson's assertion that individuals have an optimum level of stress to induce their best working performance aligns with many of the results. There are implications for future research to clarify whether these findings would also hold in the clinical environment with reporting practitioners since, if they did, there would be an argument for carefully monitoring stress levels and workload in departments in order to provide the best service to patients with more accurate reports.

References

1. [dataset] NHS England. *Diagnostic imaging dataset annual statistical release.* 2017. Available from: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2018/11/Annual-Statistical-Release-2017-18-PDF-1.6MB-1.pdf [accessed 28 June 2019]
2. Royal College of Radiologists. *Clinical radiology UK workforce census 2017 report.* London: The Royal College of Radiologists, 2018.
3. Care Quality Commission. *Radiology review: a national review of radiology reporting within the NHS in England.* 2018. Available from: https://www.cqc.org.uk/publications/themed-work/radiology-review [accessed 20 October 2019]
4. NHS Improvement. *Seven day hospital services learning and sharing event: summary of discussion.* 2017. Available from: https://improvement.nhs.uk/resources/seven-day-services-learning-and-sharing-event-challenges-and-solutions/ [accessed 30 June 2019]
5. Royal College of Radiologists. *Standards for interpretation and reporting of imaging investigations, second edition.* London: The Royal College of Radiologists, 2018.
6. Lindner F, Rose J. *No need for more time: intertemporal allocation decisions under time pressure*. Journal of Economic Psychology 2017;60:53-70, https://dx.doi.org/10.1016/j.joep.2016.12.004
7. Hu Y, Wang D, Pang K, Xu G, Guo J. *The effect of emotion and time pressure on risk decision-making*. Journal of Risk Research 2014;18(5):637-650, https://dx.doi.org/10.1080/13669877.2014.910688
8. Madan CR, Spetch ML, Ludvig EA. *Rapid makes risky: time pressure increases risk seeking in decisions from experience.* Journal of Cognitive Psychology 2015;27(8):921-928, https://doi.org/10.1080/20445911.2015.1055274
9. Bobadilla-Suarez S, Love BC. *Fast or frugal, but not both: decision heuristics under time pressure*. Journal of Experimental Psychology: Learning, Memory and Cognition 2018;44(1):24-33, https://dx.doi.org/10.1037/xlm0000419.supp
10. Dambacher D, Hubner R. Time pressure affects the efficiency of perceptual processing in decisions under conflict. *Psychological Research 2014;79*:83-94, https://dx.doi.org/10.1007/s00426-014-0542-z
11. Wolfe JM, Van Wert MJ. Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology 2010;20*(2):121-124, https://dx.doi.org/10.1016/j.cub.2009.11.066
12. Salehi B, Cordero MI, Sandi C*. Learning under stress: the inverted-ushape function revisited*. Learning & Memory 2016;17:522-530, https://dx.doi.org/10.1101/lm.1914110
13. Yerkes RM, Dodson JD. *The relation of strength of stimulus to rapidity of habit-formation*. Journal of Comparative Neurology and Psychology 1908;18(5):459-482, https://dx.doi.org/10.1002/cne.920180503
14. Koontz NA, Gunderman RB. *Gestalt theory: implications for radiology education.* American Journal of Roentgenology 2008;190(5):1156-1160, https://dx.doi.org/10.2214/ajr.07.3268
15. Berlin L. *Liability of interpreting too many radiographs*. AJR 2000;175:17-22, https://dx.doi.org/10.2214/ajr.175.1.1750017

16. Young DL, Goodie AS, Hall DB, Wu E. *Decision making under time pressure, modelled in a prospect theory framework*. Organizational Behaviour and Human Decision Processes 2012;118:179-188, https://dx.doi.org/10.1016/j.obhdp.2012.03.005
17. Oestmann JW, Greene R, Kushner DC, Bourgouin PM, Linetsky L, Llewellyn HJ. *Lung lesions: correlation between viewing time and detection.* Radiology 1988;166(2):451-3, https://dx.doi.org/10.1148/radiology.166.2.3336720
18. Edwards AJ, Ricketts C, Dubbins PA, Roobottom CA, Wells IP. *The effect of reporting speed on plain film reporting errors.* Clinical Radiology *2003;58*(12):971-979, https://dx.doi.org/10.1016/s0009-9260(03)00289-7
19. Khan S, Hedges WP. *What is the relation between number of sessions worked and productivity of radiologists: a pilot study?* Journal of Digital Imaging 2015;29(2):165-174, https://dx.doi.org/10.1007/s10278-015-9825-1
20. Sokolovskaya E, Shinde T, Ruchman RB, Kwak AJ, Lu S, Shariff YK, et al. *The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: a pilot study.* JACR 2015;12(7):683-8, https://dx.doi.org/10.1016/j.jacr.2015.03.040
21. Muroff LR, Berlin L. *Speed versus interpretation accuracy: current thoughts and literature review.* AJR 2019;213:490-492, https://dx.doi.org/10.2214/ajr.19.21290
22. Walliman N. *Research methods: the basics*. 2nd ed. Oxon: Routledge; 2018.
23. Williams LH, Drew T. *Distraction in diagnostic radiology: how is search through volumetric medical images affected by interruptions?* Cognitive Research: Principles and Implications 2017;2(12):1-11, https://dx.doi.org/10.1186/s41235-017-0050-y
24. Mandrekar JN. *Receiver operating characteristic curve in diagnostic test assessment.* Journal of Thoracic Oncology 2010;5(9):1315-16, https://doi.org/10.1097/JTO.0b013e3181ec173d