

Davies, Sarah and Donovan, Tim ORCID: <https://orcid.org/0000-0003-4112-861X>
(2025) Understanding human-AI interaction in medical imaging: moving towards 'appropriate reliance'. In: North-West Visual Cognition Group Seminar, 24 January 2025, Edge Hill University, UK. (Unpublished)

Downloaded from: <https://insight.cumbria.ac.uk/id/eprint/9033/>

Usage of any items from the University of Cumbria's institutional repository 'Insight' must conform to the following fair usage guidelines.

Any item and its associated metadata held in the University of Cumbria's institutional repository Insight (unless stated otherwise on the metadata record) may be copied, displayed or performed, and stored in line with the JISC fair dealing guidelines (available [here](#)) for educational and not-for-profit activities

provided that

- the authors, title and full bibliographic details of the item are cited clearly when any part of the work is referred to verbally or in the written form
- a hyperlink/URL to the original Insight record of that item is included in any citations of the work
- the content is not changed in any way
- all files required for usage of the item are kept together with the main item file.

You may not

- sell any part of an item
- refer to any part of an item without citation
- amend any item or contextualise it in a way that will impugn the creator's reputation
- remove or alter the copyright statement on an item.

The full policy can be found [here](#).

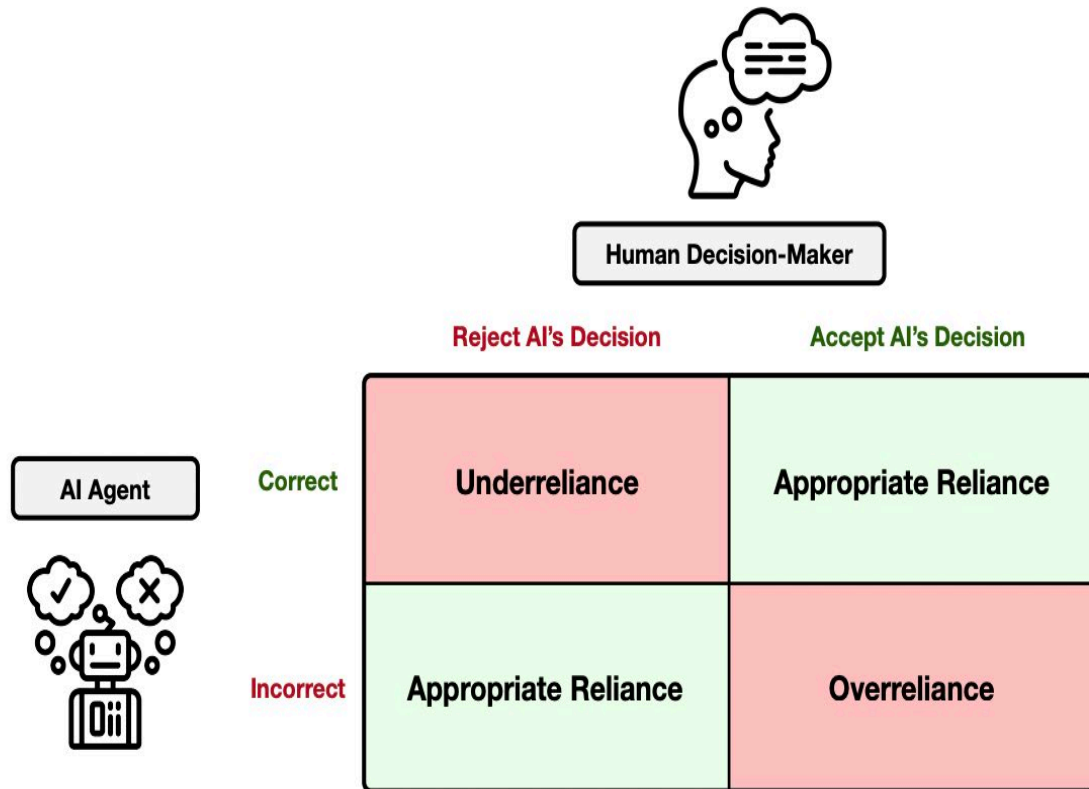
Alternatively contact the University of Cumbria Repository Editor by emailing insight@cumbria.ac.uk.



Understanding human-AI interaction in medical imaging: Moving towards 'appropriate reliance'

Sarah Davies- University of Cumbria

Tim Donovan- University of Cumbria



What is appropriate reliance and why is it important?

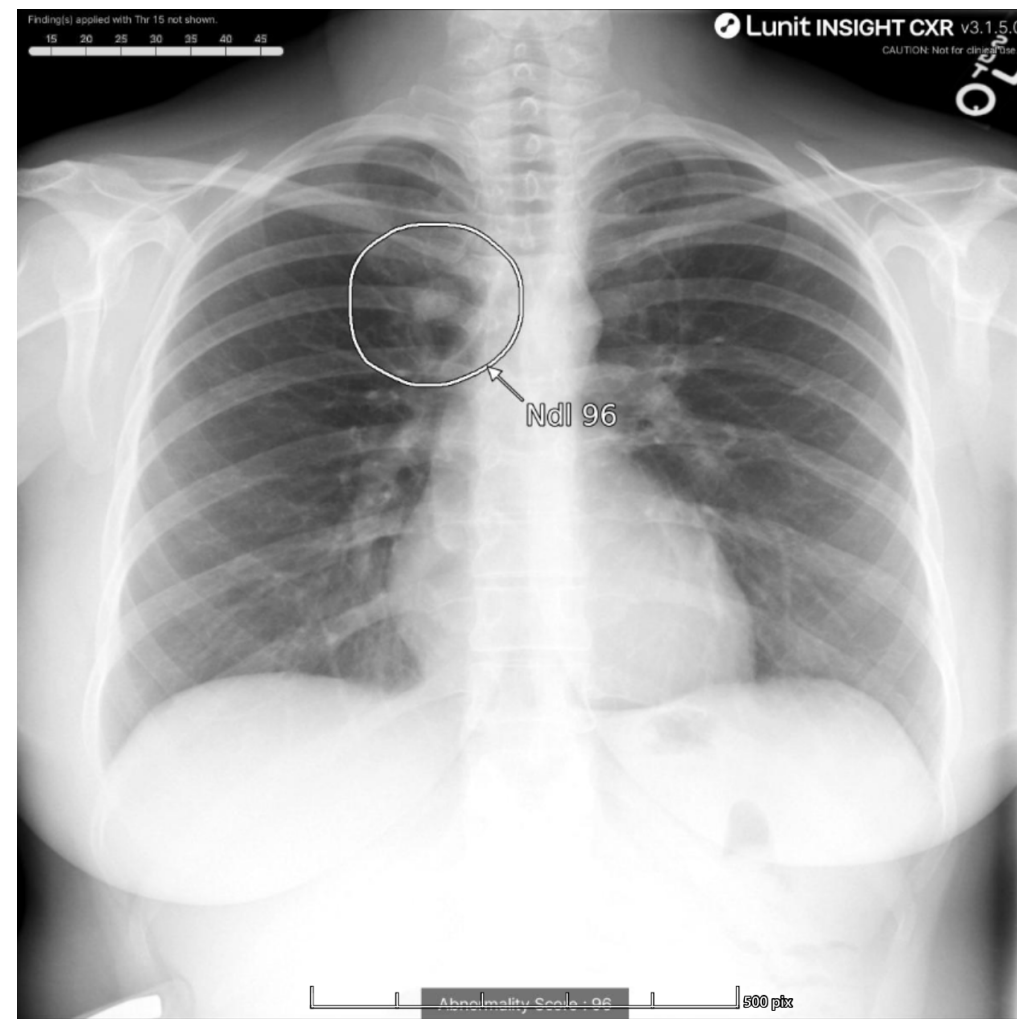
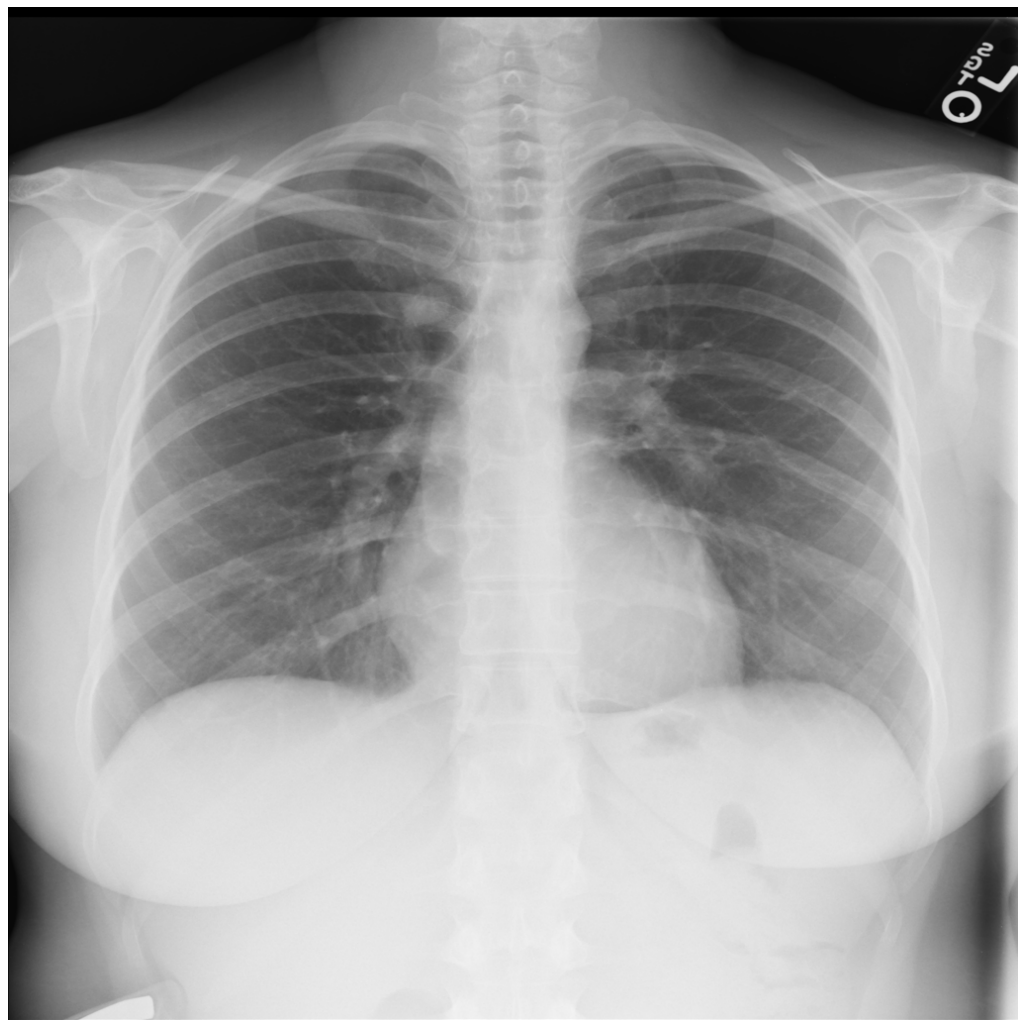
- *'the human capability to differentiate between correct and incorrect AI advice and to act upon that discrimination'* (Schemmer *et al.*, 2022)
- Over-reliance may lead to two types of error- errors of commission and/or errors of omission (Mosier and Skitka, 1996)
- Under-reliance means that performance will never change! (Reverberi *et al.*, 2022)

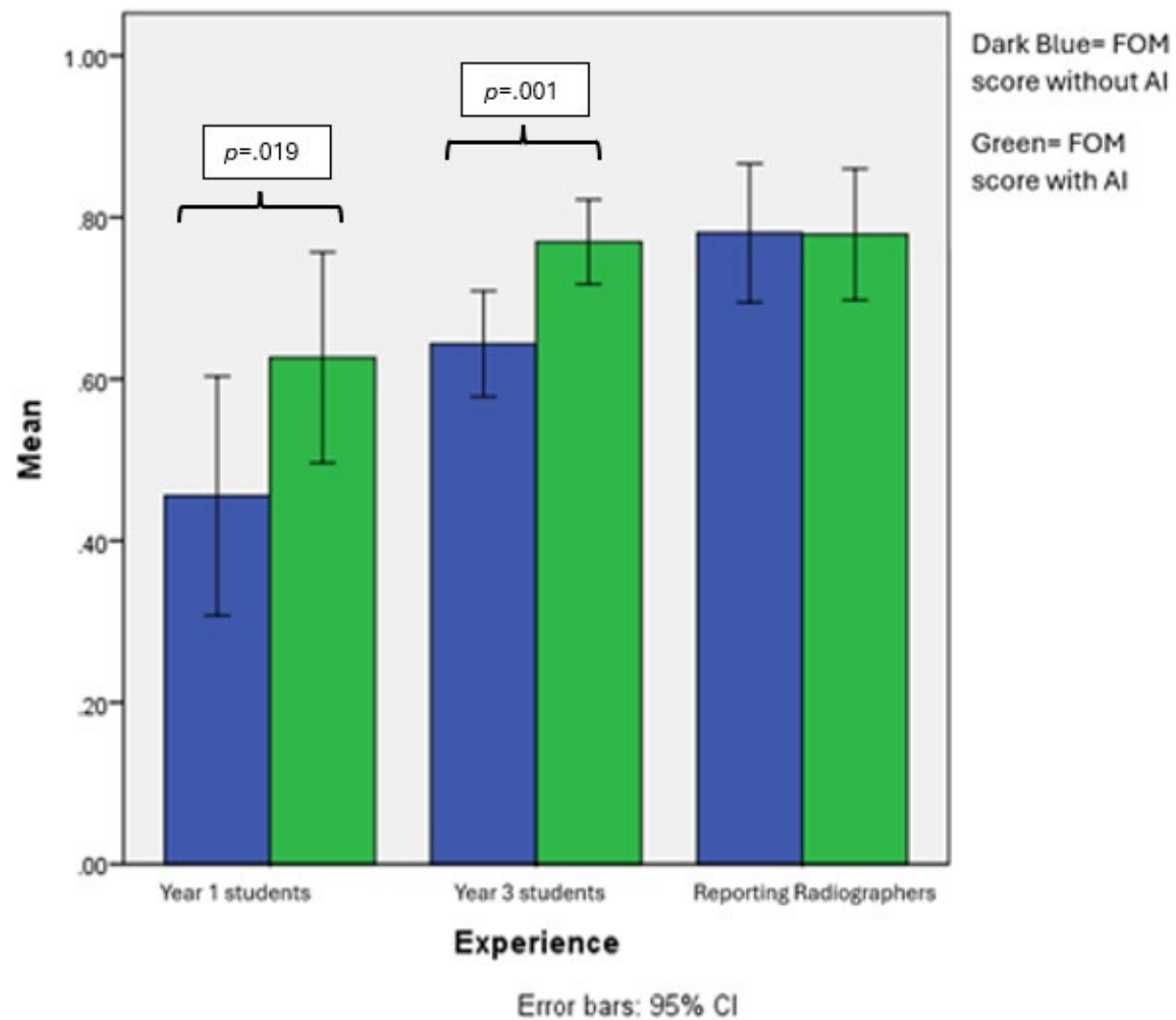


Experiment one: The impact of observer experience and attitudes towards AI technology upon observer performance with AI during a pulmonary nodule detection task: An eye tracking study

Method

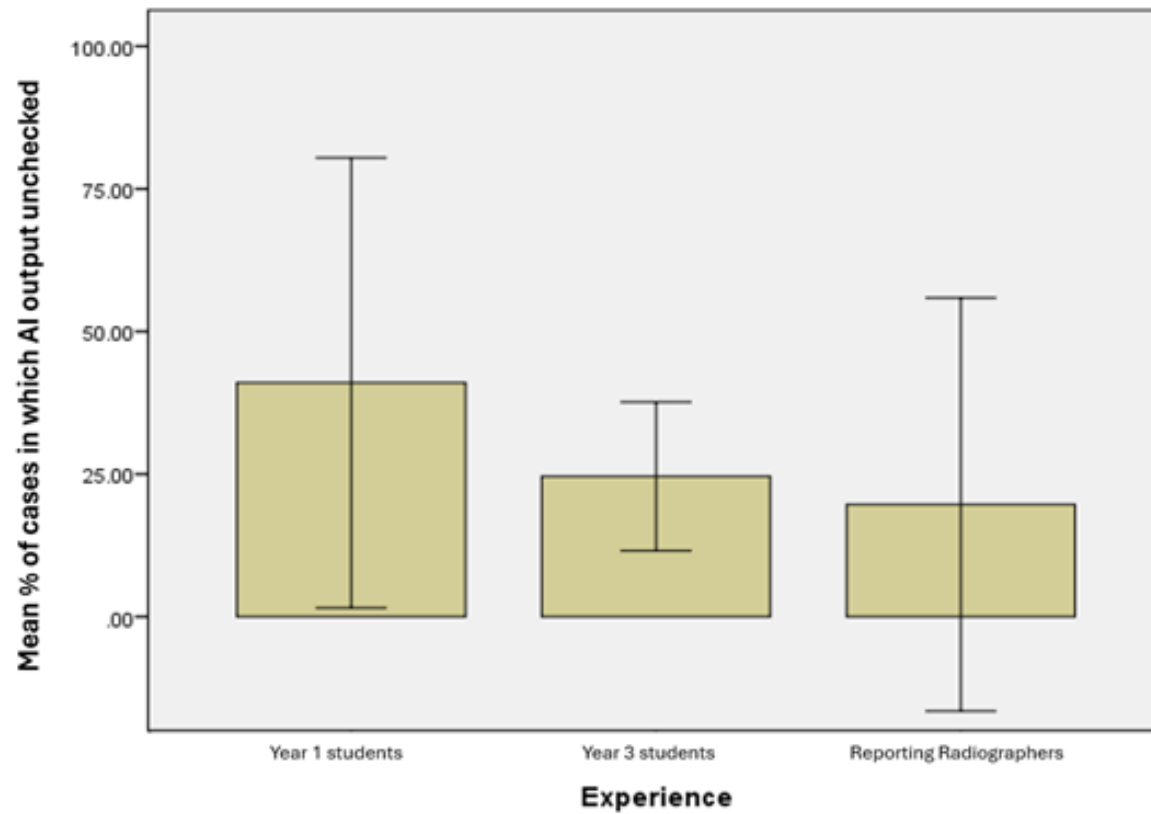
- 11 year 1 radiography students, 13 year 3 radiography students and 6 reporting radiographers
- Within-subjects design- without/with AI
- The test bank comprised of 30 CXRs at a 50% disease prevalence rate. Each diseased image contained one pulmonary nodule
- All images were analysed by Lunit Insight CXR. AI sensitivity was 87% and it produced a total of 17 FPs (Average of 0.57 FPs per case)
- Participants had to localise any suspected nodules with a mouse click and had to provide a confidence rating from 1-5 for each case (1= definitely no nodule, 5= definite pulmonary nodule)
- Participants were not given any specific instructions regarding how to use AI but were informed that they could access AI findings at any point and could make their decisions independently
- Prior to participation in the experiment, observers answered a survey to ascertain their attitude towards AI using a subset of positive and negative statements from the 'General Attitudes to Artificial Intelligence' questionnaire (Schepman and Rodway, 2020)
- RJaFroc and SPSS used for data analysis





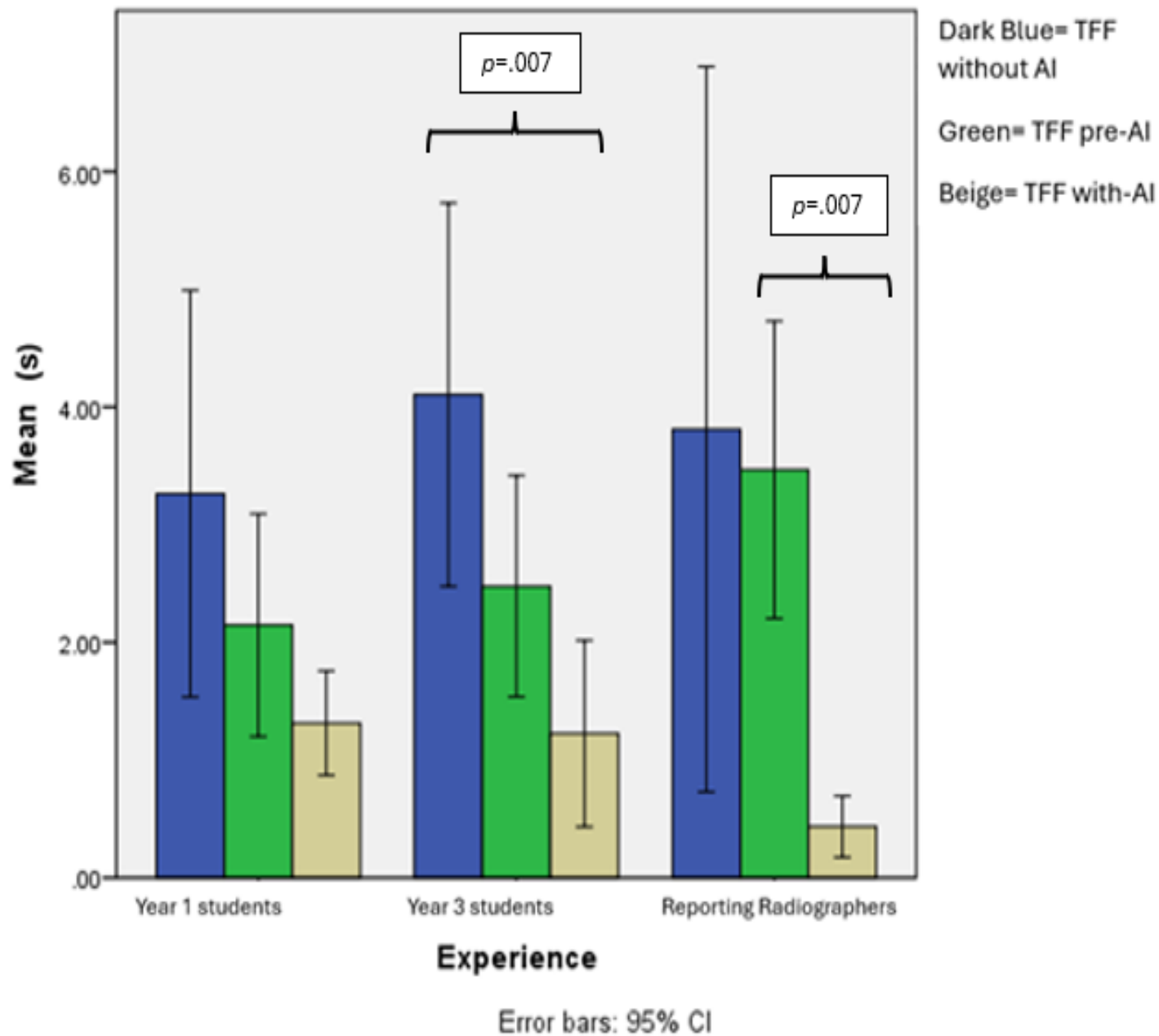
Results- Figure of Merit (FOM)

- Figure of Merit score represents the probability that a lesion (nodule) is rated higher than the highest rated non-lesion on a normal image (Chakraborty and Berbaum, 2004)
- Mean FOM without AI= 0.60, mean FOM with AI= 0.72 ($p=.001$)
- Year 1 students had the greatest increase in mean FOM with AI (+0.17, $p=.019$), no change in the performance of reporting radiographers.
- Year 3 students were the only group in which all observers experienced an increase in FOM with AI



Interaction with AI

- 63% of all participants opted to check AI output in 100% of cases
- Trend towards increased engagement with AI output with increasing levels of experience. However, the difference between groups was non-significant ($p=.167$)
- Not checking AI had limited impact upon performance

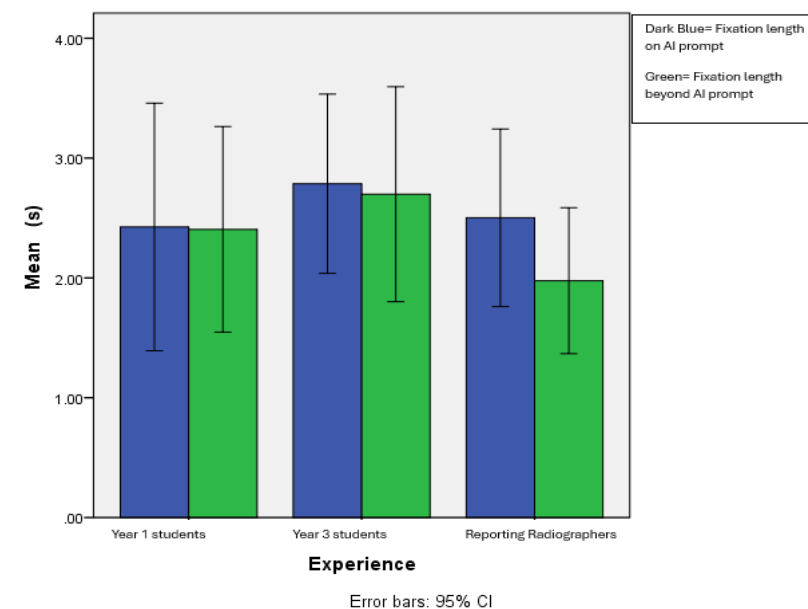
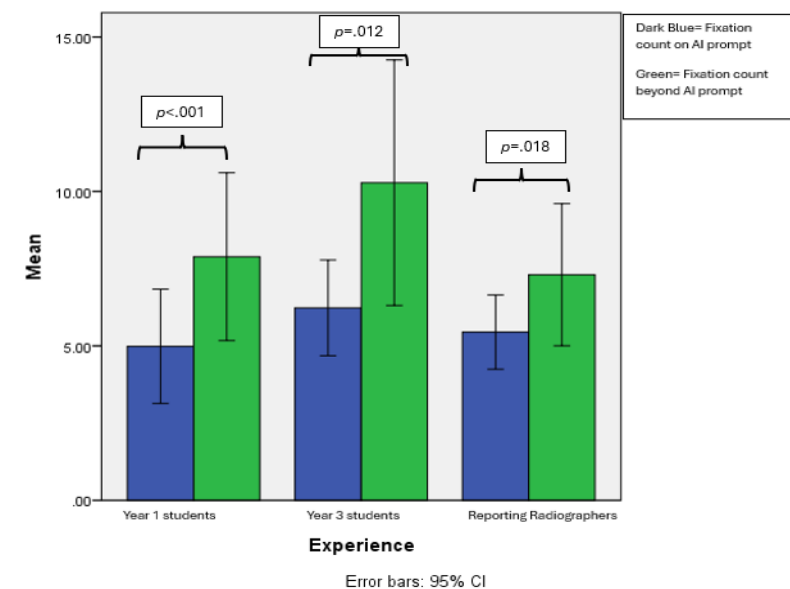


Eye tracking metrics- Time to First Fixation (TFF)

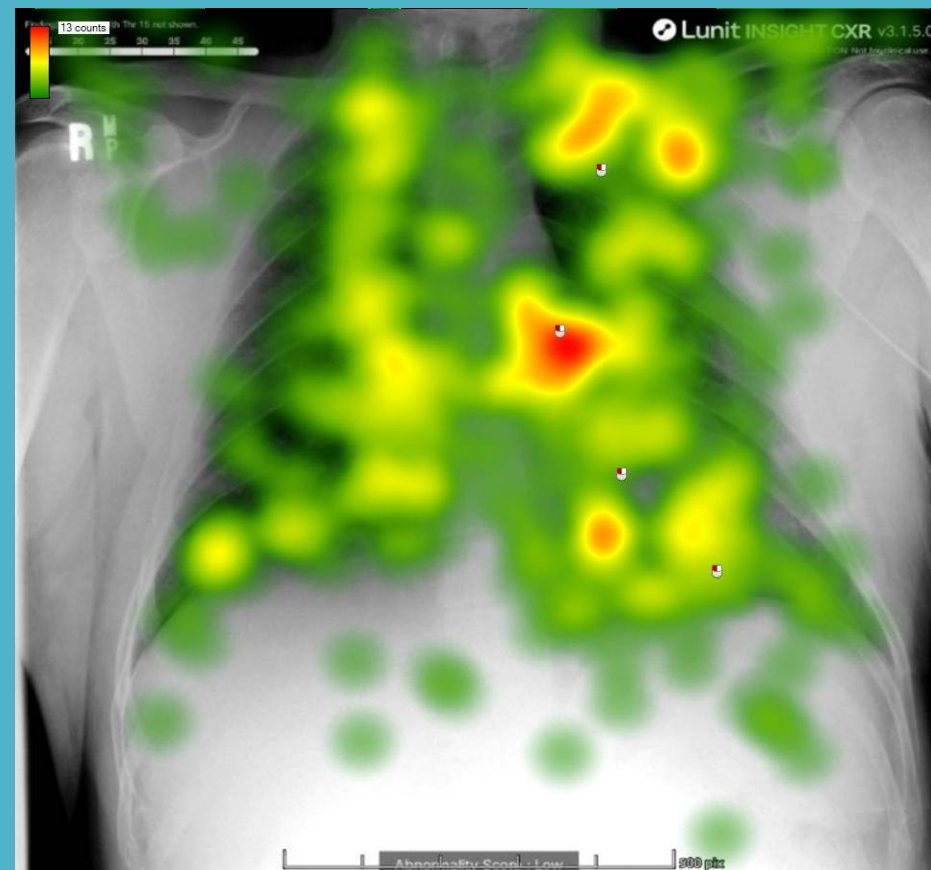
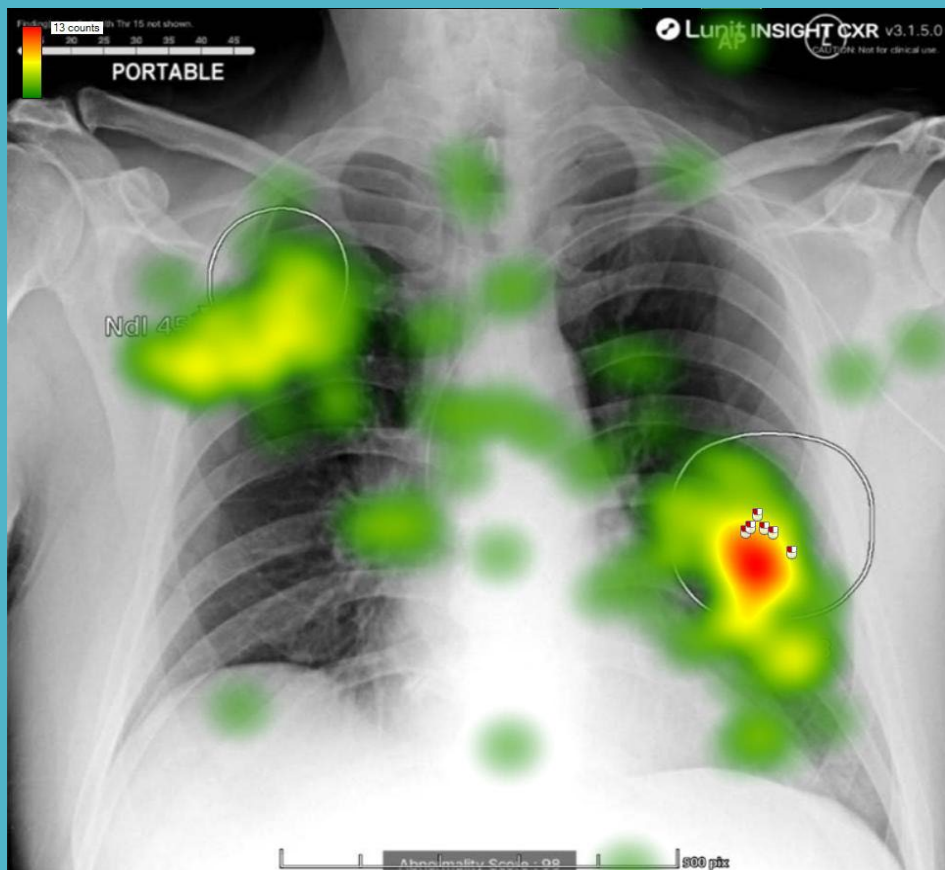
- TFF- time from image onset to first fixation on nodule
- Mean TFF in the with-AI condition was significantly shorter than both the pre-AI and without-AI conditions (both $p<.001$)

Attention allocation- images with AI prompts

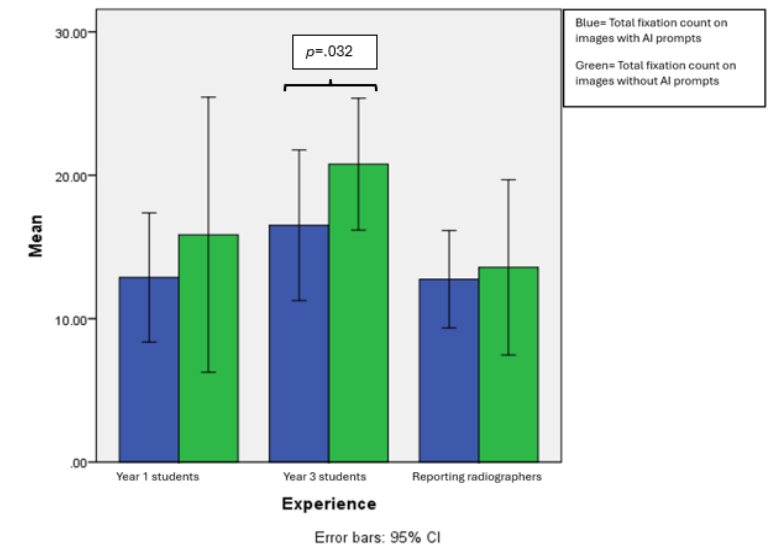
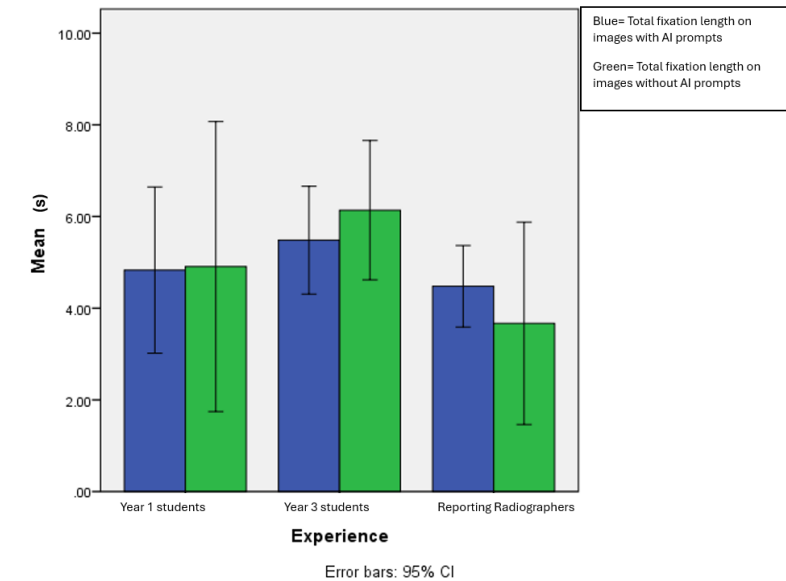
- Fixation count beyond AI prompt was significantly higher than fixation count on AI prompt (+3.19, $p<.001$)
- Fixation length on AI prompt was slightly higher than fixation length beyond AI prompt (+0.15s, $p=.454$)



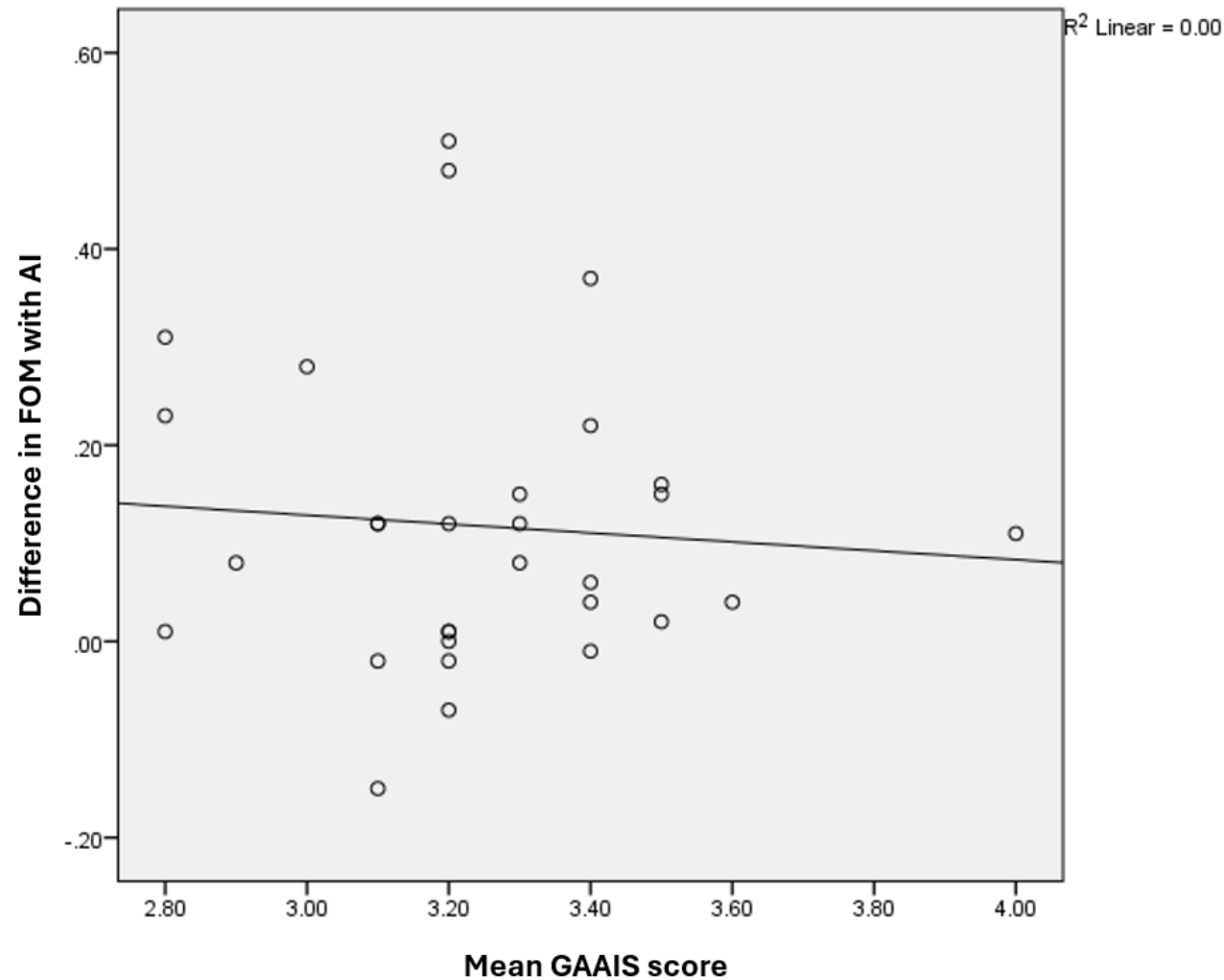
Images with prompts vs images without (all participants)



- Fixation count on images without prompts was higher than fixation count on images with prompts (+3.1, $p=.059$)
- For Year 3 students, the difference was statistically significant (+4.25, $p=.032$)
- Fixation length on images without prompts was higher than fixation length on images with prompts (+0.15s, $p=.954$)
- Observers less susceptible to errors of omission?



Attitudes towards AI



- Year 3 students held the most positive attitude towards AI technology. However, Kruskal-Wallis test revealed that the difference between group GAAIS scores was non-significant ($p=.070$)
- Spearman's rank correlation test revealed that there was no relationship between mean GAAIS score and the difference in FOM with AI ($r_s=-0.11$, $p=.954$)



Experiment 2- Confirmation bias in observer performance during an AI-assisted pulmonary nodule detection task: An eye tracking study



Method

- Within subjects design- confirmation bias trial and control
- Sample will consist of Year 3 radiography students, radiographers and reporting radiographers
- Confirmation bias trial- initial 10 cases will be high human-AI agreement cases based on the findings of experiment 1
- Participants will be required to localise any suspicious nodules with a mouse click and to provide a confidence rating from 1-5 on both the initial image and AI image

Is there a statistically significant difference in agreement fraction/switch fraction between trials?

Is there a statistically significant difference in FOM between trials?

Is there any difference in visual search behaviour across the two trials? Eg fixation count, fixation length on AI image



Participant	AF- CB Trial	AF-Control Trial	SF- CB Trial	SF- Control Trial
11	0.87	0.6	0.31	0.17
12	0.73	0.73	0.55	0.45
13	0.8	0.8	0.5	0.58
14	0.67	0.7	0.15	0.1
15	0.67	0.67	0.35	0.45
16	0.63	0.5	0.05	0.13
19	0.63	0.87	0.37	0.42
20	0.63	0.7	0.42	0.38
21	0.8	0.6	0.5	0.33
3	0.6	0.63	0.39	0.21
4	0.7	0.73	0.1	0.23
5	0.57	0.33	0.41	0.3
6	0.47	0.43	0.21	0.15
8	0.57	0.67	0.18	0.25
18	0.67	0.73	0.4	0.32
1	0.83	0.73	0.44	0.36
10	0.73	0.7	0.23	0.29
Average	0.68	0.65	0.33	0.30

Preliminary results- Year 3 students

Across the first 10 cases:

CB trial AF= M= 0.91 (SD= 0.10)

Control trial AF= M= 0.45 (SD= 0.17)

Paired samples t-test= $t(16)= 8.79, p<.001$

CB trial SF= M= 0.10 (SD= 0.08)

Control trial SF= M= 0.58 (SD= 0.27)

Paired samples t-text= $t(16)= -7.18, p<.001$

Thanks for
listening, any
questions?