Preregistration Form

Title

“How Effective is Reflective Search? A Time Slice Analysis of Dentistry Students’ Visual Search Strategies and Pupil Dilation during the Diagnosis of Radiographs”

Statement regarding ethics

All procedures performed in studies involving human participants were endorsed by the ethics committee of the Leibniz Knowledge Media Research Center (IWM, application LEK 2017/016) that implement recommendations of the German Psychological Association and comply with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Abstract/Study description

This study intends to analyze already collected data of dentistry students of varying semesters diagnosing Orthopantomograms (OPTs) that feature dental anomalies. We plan on sampling all data that included 10 trials of 90 s of students searching for anomalies in such OPTs. We seek to investigate fixation gaze measures across time slices, specifically in earlier and later stages of trials. Through this, we intent to compare models from related disciplines in medical diagnosis, namely the Nodine & Kundel Model (1987), with our data. We also plan on investigating the utility of pupil dilation across time slices in predicting anomaly detection, expanding on previous finding linking pupil dilation to expertise in radiograph diagnosis (Castner et al., 2020). For our time slice analysis, we define earlier and later trial stages as the first and last 30 s of trials respectively. For a more fine-grained analysis, we plan on investigating slices of 9 x 10 s individually. Since the literature does not suggest specific time slice sizes to investigate, we expect these analyses to inform future research questions. Our hypotheses for the comparison of early and late trial stages of visual search are guided by previous research investigating visual search strategies of experts in medical diagnosis: First, we expect earlier stages of visual search to be characterized by more and shorter fixations than later stages (Hypothesis 1). Secondly, we conjecture that pupil dilation is especially predictive of diagnostic performance in later trials – indicating cognitive load during a reflective search strategy (Hypothesis 2). In short, this study has important implications for future medical eye-tracking research methodology, as we expect to find time slicing to be informative about student’s visual search strategies.

Keywords

Eye Tracking, Predictive Modelling, Time Slice Analysis

Research question(s)

Our research questions are as follows:

RQ1: How does the fixation behavior of students (i.e., number of fixations and mean fixation duration) change across earlier and later stages of visual search during OPT diagnosis?

RQ2: Which role does pupil dilation and cognitive load in earlier and later stages of trials play in predicting subsequent diagnostic performance?

Study methods

We collected eye-tracking data on dentistry students' fixation gaze behavior in the context of the diagnosis of 10 Orthopantomograms (OPTs). We did so by employing RED 250 mobile eye trackers of 250Hz from SensoMotoric Instruments (SMI™). For calculating fixation metrics, we employed the default settings of the SMI software BeGaze2, classifying fixations at a threshold of 50 ms. OPTs were displayed on laptops with a screen size of 15.6 in. and a resolution of 1920 by 1080 pixels. Together with a constant testing environment, we ensured an illuminance of 30-40 lx on all laptop displays, as measured with a radiological Gossen Mavomax™ illuminance sensor.

Introduction

To standardize the experimental procedure in our data collection of dentistry students diagnosing Orthopantomograms (OPTs), we gave students 90 s each to investigate anomalies in a given OPT. However, given varying expertise between semesters in our sample, it is unlikely to assume that all students used the whole trial duration to decide on which anomalies to mark in the proceeding sections of the experiment. Therefore, we expect time slices to be informative about the visual search strategies of students.

Previous research suggests that experts' visual search strategies in medical diagnosis comprise of a quick, global impression phase at the beginning of the diagnosis (Kundel et al., 2007), while later stages might be characterized by a reflective search, as proposed by the Nodine and Kundel Model (1987). Therefore, we plan on investigating both the average fixation time and number of fixations during later and earlier stages of trials. We expect a global impression strategy in early trial stages to result in many short fixations and a reflective search strategy in later stages of trials to result in few long fixations. As the literature does not suggest specific time slice size in which different visual search phases take place, we define earlier trial stages as the first 30 s of trials of 90 s and later trial stages as the last 30 s. In a more fine-grained analysis, we will also look at time slices of 9 x 10 s individually, which is described further below.

To further back up our findings, we will investigate the average pupil diameter during anomaly fixations in different time slices of the experimental procedure as a measure of cognitive load. We expect pupil diameter to be associated with increased odds of anomaly detection in later compared to earlier stages of trials, reflecting high cognitive load during a reflective search and decisional processes.

Additionally, we conjecture that students show fatigue during the end of the experimental session, which would indicate that the experimental procedure was to extensive for them. Since the sequence of radiographs was constant for all experimental sessions, analysis of trial length is inevitably confounded with properties of the radiographs themselves. Therefore, we plan on investigating effects of fatigue in a post-hoc analysis that could inform future research.

Problem

In a sample of dentistry students of different semesters, we would like to investigate how time slices of the diagnostic process characterize visual search strategies and predict diagnostic performance. Such time slice analysis of fixation gaze measures has largely been unaddressed in the literature and could shine a light on the visual search strategies of dentistry students.

Crucially, the Nodine and Kundel model (1987) was originally developed for describing the diagnosis of lung nodules in chest radiographs. Similarly, later studies which accrued evidence for an initial global impression strategy in experts in medical diagnosis were based on mammograph diagnosis (see for example Kundel et al., 2007). Both diagnosis situations can be characterized by a binary outcome - the objective of determining the presence or absence of any anomaly. The diagnosis of OPTs, on the other hand, is characterized by the objective of identifying all possible anomalies in an image, with the anomalies being possibly of very different nature (i.e., dental, gum or jaw anomalies). Another important difference to the described studies is that the participants in our study sample (dentistry students) are not experts. As an example, recent findings from our research, which reported that the detection of low-prevalence anomalies was connected to a general increase in coverage of OPTs after massed practice (Richter et al., 2020), suggesting that students might still amidst developing the global impression strategy that characterizes experts.

Therefore, we expect our results to inform how generalizable existing models of visual search are across our study sample, including dentistry students with varying levels of expertise and semesters.

Review of relevant scholarship

Visual search strategies and decisional processes in experts and novices:

According to the visual search model introduced by Nodine and Kundel (1987) describing the behavior of radiologists diagnosing chest radiographs, the search for anomalies can be described with reference to four phases: (a) Global impression, in which radiologists screen the radiograph for anomalies, (b) discovery search, in which radiologists selectively screen suspicious areas determined based on the global impression of the radiograph, (c) foveal verification of the detected anomalies through fixation and (d) reflective search, in which the image and its anomalies are once more screened to specifically accrue evidence for anomalies which are still not decided upon.

This initial model of visual search in medical diagnosis has leaded to a holistic processing perspective in the literature, which states that experts quickly extract information from radiographs at the beginning of the diagnosis, which then guides further visual search (Kundel et al., 2007). Backing up this account, there has been empirical evidence showing that experts, compared to novices, use fewer fixations to diagnose radiographs, while novices systematically cover the whole radiograph more thoroughly and broadly, paying attention to salient features of the image (van der Gijp et al., 2017). Still, these findings are limited to chest x-rays and call for further investigations and replication in both different domains and study samples. In the dental context, experts have been shown to apply more and longer fixations

on more subtle anomalies (Turgeon and Lam, 2016). Arguably, these differences in search strategies between different levels of expertise can be attributed to differences in prior knowledge (Eder et al., 2020).

The question arises how well the holistic processing perspective and the reflective search proposed by the Nodine and Kundel model (1987) is reflected in the data of our study sample. This account is contrasted by recent findings from our research group which linked the detection of low-prevalence anomalies to an increased coverage of OPTs after massed practice (Richter et al., 2020). This finding could indicate that students are still amidst the development of advanced visual search strategies and can still improve their diagnostic performance through increased visual focus and OPT coverage. To test these accounts, we consider investigating early and late time slices of the trial procedure to be informative, which is at the heart of this study.

Pupil dilation as a measure of cognitive load in eye tracking and visual search:

Pupillary response has been shown to be a reliable measure of cognitive load in various contexts, including memory load (Van Gerven et al., 2004) and the simulation of driving (Palinko et al., 2010). Consequently, pupil dilation has been shown to distinguish levels of expertise in the diagnosis of dental radiographs, in the context of which the pupillary response of experts, compared to novices, was systematically varying with anomaly difficulty (Castner et al., 2020). Next to this, pupil diameter increases could be linked to increased difficulty in diagnostic decision processes in medical image interpretation (Brunyé et al., 2016). Importantly, measures of pupillary response ought to be adjusted to individual baselines, as multiple contextual factors such as fatigue (Lowenstein et al., 1963) and caffeine intake (Abokyi et al., 2017) can influence pupil dilation.

Differences between previous work and the present study

- Systematic investigation of dentistry students of different semesters rather than experts compared to novices
- Systematic evaluation of time constraints in OPT diagnosis (specifically, the appropriateness of 90 s for the diagnosis of one OPT in our sample)
- Systematic comparison of fixation gaze behavior in different time slices of diagnosis
- Systematically investigating the predictive validity of pupil dilation in different time slices of diagnosis through generalized linear mixed models

Hypothesis, aims and objectives

Confirmatory analysis:

H1: The general fixation behavior during visual search in OPTs is characterized by many, short fixations in earlier stages and few, long fixations in later stages of trials

This hypothesis builds on top of the holistic processing perspective in visual search (Kundel et al., 2007) and the reflective search at the end of diagnosis proposed by the Nodine and Kundel model (1987). We expect a global impression strategy at the beginning of trials to

result in short and numerous fixations. Conversely, we conjecture that students applying longer and fewer fixations at the end of trials signals a reflective search strategy. If the search strategies proposed by Nodine and Kundel (1987) are not transferable for visual search in OPTs, we would expect that fixation gaze measures do not vary across early and late trial stages.

H2: Pupil dilation during fixations on a given anomaly in later stages of trials is positively associated with the likelihood of that anomaly being marked (and negatively associated in earlier stages)

As the pupillary response of experts was found to systematically vary with anomaly difficulty in previous studies (Castner et al., 2020), we expect increased pupil dilation in our study sample to signal elaboration of a given anomaly and increased cognitive load. Specifically, we hypothesize that increased pupil dilation during anomaly fixation in *later trial stages* to *increase* the chance that a student later marks that anomaly, which would signal higher cognitive load during a reflective search strategy. Conversely, we expect high pupil dilation during anomaly fixation in *early trial stages* to *decrease* the chance of a given anomaly being marked. This would indicate that students already experience high cognitive load during the discovery of anomalies which would signal excessive task demand.

The main effects of trial split on average fixation length and number of fixations proposed by Hypothesis 1 as well as the interaction of pupil dilation during anomaly fixations and trial split proposed in Hypothesis 2 are illustrated in Figure 1.

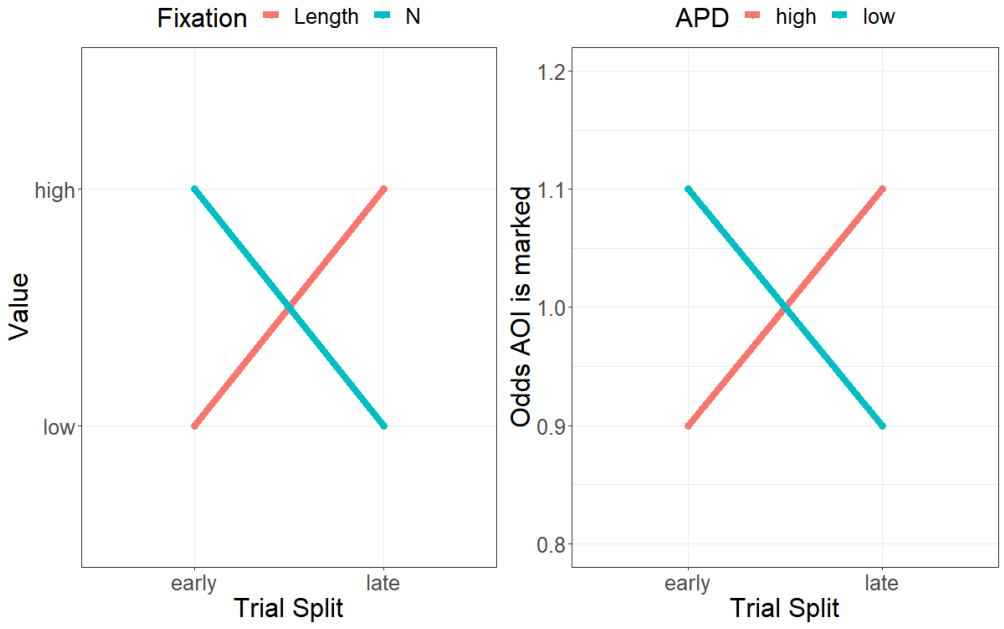

Figure 1. Illustration of the effected result patterns for Hypotheses 1 and 2, including exemplary effect sizes for the odds of a given anomaly being marked by participants. Left panel depicts the main effects of trial split on the average fixation length and number of fixations (H1) and the right panel depicts the interaction effect of average pupil diameter during anomaly fixation and trial split (APD, H2).

Post-hoc analysis:

Plotting aggregated fixation measures over all 9 time slices

Investigating the development of fixation gaze behavior over time on a more fine-grained level, we plan on investigating our data through time series plot of all 9 time slices of 10 s. This plot will feature the number of fixations and mean fixation time, ratio of fixations per second and the average pupil dilation inside of time slices. Exploring these time-series plots will also enable us to explore possible individual differences more thoroughly. As an example, one could imagine that some students might already transition in a more elaborative, reflective search strategy in the middle of trials while other students might take more time to discover anomalies initially. Such a transition might be characterized by a spike in cognitive load, indicated by increased pupil dilation during anomaly fixation. This exploration of individual differences will also help in estimating the robustness of our confirmatory analysis and its assumptions.

Gaze-likelihood analysis, testing differences between fine-grained trial time slices

Extending our exploration of time-series plots as described in the last paragraph, we will consider statistically testing differences in fixation gaze measures over time slices of 9 x 10 s through a gaze-likelihood analysis which has been applied in eye-tracking research investigating the decision-making processes in multiple choice testing (see Lindner et al., 2014, Lindner et al., 2017).

Inquiring about possible effects of session length to inform future research

We consider that students' diagnostic performance might have decayed over the course of the experimental sessions of ~ 90 minutes. As all students in our study sample were exposed to the radiographs in the same sequence, the effect of trial length on diagnostic performance is confounded with the effect of the radiographs (i.e., their difficulty). We, therefore, included this inquiry in a post-hoc analysis that could inform future research which randomizes radiographs sequences in a confirmatory setting. We plan on plotting the effect sizes of the 10 trials of experimental sessions next to each other after fitting a model just featuring the trial number as a predictor. In that way, we can estimate if there might be a systematic effect of trial position / radiograph on diagnostic performance based on our data.

Materials and Methods

Sampling Plan / Data collection / Data acquisition

Out of our data base of dentistry students diagnosing OPTs, we intend to use the largest possible data set with an equal selection and sequence of radiographs in experimental session. Specifically, this entails a set of 10 OPTs. This selection includes multiple measurement points for participants between the summer semester of 2017 and 2018.

Sample size, power and precision

Our study sample was determined through obtaining the largest possible set of data from experimental sessions with equal selection and sequence of radiographs. Applying commonly used exclusion criteria in eye-tracking research with BeGaze2 as described in the next section, our sample size amounts to N = 194 unique sessions with N = 107 unique participants.

We ran ad-hoc power-analysis based on our obtained sample size. We assumed the following minimally relevant effect sizes:

Hypothesis 1:

Predicting logarithmic mean fixation duration, we simulated linear mixed models with gaussian link function of the form:

Y = β0 + β1(trial split) + vi (semester|participant) + vj(cohort:participant)

We assumed the following model parameters:

- β1, main effect of trial split (late) on log mean fixation time: log(1.25), indicating a 25% increase in average fixation time in later stages of trials
- β0, intercept: log(1) = 0, as 1 second appeared to be a typical mean fixation time based on data from previous studies conducted in our research group
- Additionally, we obtained covariance matrices of random effects and residual variance from previous studies conducted in our research group

For predicting the number of fixations, we employed generalized linear mixed models with poisson link function, which is often referred to as a log-linear model. The model featured the same model equation as above and the following effect sizes:

- β1, main effect of trial split (late) on the number of fixations: log(0.75), indicating a 25% decrease of the number of fixation in later trial stages
- β0, intercept: log(30), as 30 fixations seemed to be a reasonable baseline for the number of fixations based on data from other studies conducted in our research group
- Additionally, we obtained covariance matrices of random effects from previous studies conducted in our research group

For both models and main effects, power was estimated to be 100% based on 1000 replicas at an α level of .05.

Hypothesis 2:

Predicting anomaly detection, we simulated linear mixed models with binomial link function (logistic regression) of the form:

Y = β0 + β1β2(apd x trial split) + vi (semester|participant) + vj(cohort:participant) + vk(AOI)

- Effect of average pupil diameter in early trial stages: log(0.90), indicating 10% odds decrease in anomaly detection per additional standard deviation.
- Effect of average pupil diameter in late trial stages: log(1.10), indicating 10% odds increase in anomaly detection per additional standard deviation.
- β1β2, interaction effect (log ratio of odds ratios) of pupil diameter and trial split is therefore: log(1.10/0.90).
- (Hypothetical) resulting intercept is therefore (at 0 standard deviations in pupil dilation in early trial stages): log(1) = 0.
- Additionally, we obtained covariance matrices of random effects from previous studies conducted in our research group

Since longer fixation times on an anomaly can also signal high cognitive load (see Rayner 1998, Van Gog et al., 2009), we also estimated statistical power for detecting a second interaction of mean fixation time and trial split in the following model, assuming the same effect sizes as for the first interaction.

Y = β0 + β1β2(apd x trial split) + β3β4(mfd x trial split) + vi (semester|participant) + vj(cohort:participant) + vk(AOI)

We estimated power to be .99 for detecting a simple interaction in the least complex model and .72 for detecting a second interaction in the most complex model based on 1,000 simulations at an α level of .05.

Next to an exploratory analysis which is described below, we do not plan on any additional interim analyses nor stopping rules.

Participant characteristics

Participants were dentistry students enrolled at the University of Tübingen. They were between the sixth and tenth semester of their studies. At the times of measurement, which is between the summer semesters of 2017 and 2018 in this case, the distribution of semesters for participants in the experimental sessions is as follows (again, after applying exclusion criteria described below and with possible multiple measurements of individuals):

6th Semester	7th Semester	8th Semester	9th Semester	10th Semester
73	16	36	34	35

Procedure

In each experimental session, students were presented with ten OPTs of varying difficulty and number of anomalies. In the case of some experimental sessions, students were presented with an additional set of OPTs *after* the OPTs featured in our study sample and a 10-minute break. Before each trial, students looked at a fixation cross for 2 seconds. After an exploration phase of 90 s for each image, students were asked to mark the anomalies with a red circle through a web-based tool and without time constraints. Before the start of the experimental session, participants' quality of calibration was measured and introductory instructions were presented on screen.

Conditions and design

While there are no experimental conditions to consider in our study, our data base represents a longitudinal cohort-sequential design, such that participants can appear multiple times in our study sample. This is also true in the case of three measurement points in the sixth semester.

Variables (manipulated variables; measured variables)

Measures and covariates

Grouping Variables:

ID

Each participant was assigned a 6-character alphanumeric identifier.

Cohort

To distinguish semesters in our cohort-sequential design, each ID is nested inside one of seven cohorts.

AOI

This variable represents a unique identifier for each anomaly appearing in the experimental procedure. This variable only becomes relevant for Hypothesis 2 and is featured as a random effect in order to account for differences in anomaly difficulty. For Hypothesis 1, we sampled fixation gaze measures for all fixations including fixations on whitespace (regions excluding AOIs).

General covariates and predictors:

Semester (covariate)

To account for differences in expertise across participants, we will feature semester as a covariate in our models. Multiple measurement points in semester 6 will be treated as decimal points in the semester variable.

Time slice (predictor)

We plan on slicing trials of OPT diagnosis through self-programmed algorithms in R (R Core Team, 2020). Fixation gaze measures will then be aggregated within these time slices. For our confirmatory analysis, we will feature a binary factor with the levels ”early” and “late”, representing the first and last 30 s of trials, with the first time slice as the reference category. Similarly, we will slice trials in sections of 9 x 10 s and experimental sessions in 10 x 10 units of OPTs (read: trials) in our post-hoc analyses.

Trial Level Variables (Hypothesis 1):

Mean Fixation duration (dependent variable, Hypothesis 1)

The mean fixation duration represents the average length of fixations inside of a given time slice in seconds. This variable appears to be approximately normally distributed when converted to log scale.

Number of fixations (dependent variable, Hypothesis 1)

This variable denotes how often a fixation (defined as resting gaze for more than 50 ms) occurs inside of a given time slice. Since the number of fixations represents count data, it will be modeled through a poisson distribution, frequently referred to as log-linear modelling.

AOI level variables (Hypothesis 2):

Pupil Dilation (predictor)

Pupil dilation will be taken from BeGaze2, which calculates the average pupil dilation diameter in mm during a fixation. For taking individual and contextual differences in pupil dilation into account, we will calculate a baseline which is a fixation cross before each trial. The average pupil diameter measured during fixations on these fixation crosses will be subtracted from each average pupil dilation diameter during a given fixation in the corresponding trials. As pupil dilation appears to be approximately normally distributed, we will center pupil dilation around a mean of 0 and scale to a standard deviation of 1 in order be able to interpret model parameters more intuitively.

Diagnostic Performance (dependent variable, Hypothesis 2)

For each anomaly, diagnostic performance is coded as a binary variable with 1 representing that the participant marked the anomaly in the marking phase and 0 if not.

Unit of analysis

There are two main resolutions of analysis to be differentiated in this study.

For Hypothesis 1, testing the general fixation behavior of participants, the unit of analysis is a time slice inside of trials (OPTs).

For Hypothesis 2, testing the role of cognitive load across time slices on anomaly detection, the unit of analysis is one anomaly inside of a given time slice. Notably, this implies that data points are only included in our study sample if participants have fixated a given AOI in the given time slice at least once.

additional operational definitions

None

Scales

None

Analysis Plan

As outlined in our power analysis, we will fit three types of generalized linear mixed models.

For Hypothesis 1, we will test the main effect of trial split (early, late) in predicting log mean fixation duration (gaussian link function) and number of fixations (poisson link function) during visual search:

$Y = \beta 0 + \beta 1(\text{trial split}) + vi(\text{semester}|\text{participant}) + vj(\text{cohort:participant})$

Next to semester as a covariate within participants, the model equation includes random effects for participants. This random effect fits individual intercepts for instances of participants and accounts for differences between them.

For Hypothesis 2, we will test the interaction effect of average pupil diameter during anomaly fixation and trial split (early, late) on corresponding anomaly detection:

$Y = \beta 0 + \beta 1\beta 2(\text{apd x trial split}) + vi(\text{semester}|\text{participant}) + vj(\text{cohort:participant}) + vk(\text{AOI})$

With the target variable (log-odds in logistic regression) as the natural logarithm to the ratio of the probabilities of a given anomaly to be correctly marked or not marked:

$Y = \log[P(\text{AOI marked}=1)/P(\text{AOI marked}=0)]$

Compared to the other models on trial level, this model additionally entails a random effect for the AOIs, fitting individual intercepts for each instance in order to control for varying anomaly difficulty.

As another control measure for cognitive load, we will also consider fitting a more complex model including an additional interaction term of mean fixation duration and trial split:

$Y = \beta 0 + \beta 1\beta 2(\text{apd x trial split}) + \beta 3\beta 4(\text{mfd x trial split}) + vi(\text{semester}|\text{participant}) + vj(\text{cohort:participant}) + vk(\text{AOI})$

We will compare the utility of both models through a likelihood ratio test and model information criteria (i.e., AIC).

Post-hoc Analysis:

Analogous to the models defined above, we will aggregate each fixation gaze measure in 9 time slices of 10 s in order to explore their progression inside of trials on a more fine-grained level. As mentioned, we consider statistically testing differences in fixation gaze measures over time through a gaze-likelihood analysis which has been applied in eye-tracking research investigating the decision-making processes in multiple choice testing (see Lindner et al., 2014, Lindner et al., 2017). At least, we will explore these differences through time-series plots, possibly on the level of individual participants or participant clusters.

Finally, we will consider fitting the following model on AOI level in order to explore possible influences of trials on anomaly detection.

$Y = \beta0 + \beta1(trial) + vi(semester|participant) + vj(cohort:participant) + vk(AOI)$

Plotting the effect sizes of all 10 levels of the trial predictor might reveal a trend indicating fatigue and excessive demand over the course of the experimental procedure.

Preprocessing

Data Inclusions/exclusion criteria

Experimental sessions will be excluded from the study sample if calibration indicates a deviation of more than .6 visual degrees for *either* the x or y axis. Additionally, sessions will be excluded if the overall tracking ratio falls below 80%. Lastly, we will exclude one experimental session from data analysis which accrued trials of longer than 90 s most likely due to a technical error.

Transformations

We will log-transform mean fixation duration for predictive modelling, as that variable appears approximately normally distributed when log transformed.

As mentioned, we will standardize (baselined) average pupil dilation to a mean of 0 and a standard deviation of 1 as this variable appears to be approximately normally distributed. This transformation will help making meaningful interpretations of the estimated effect size of pupil dilation on diagnostic performance.

Tests (Sequential analyses, T-Test, ANOVA, MANOVA, ANCOVA, Pearson correlation, Regression, ...)

All statistical tests are based on generalized linear mixed models which will be fitted using the package lme4 (Bates et al., 2015) in R (R Core Team, 2020). For all tests, α will be set to .05.

Hypothesis 1:

First, we will test and interpret the main effect of trial split in the following model with gaussian link function predicting log mean fixation duration:

$Y = \beta0 + \beta1(trial\ split) + vi(semester|participant) + vj(cohort:participant)$

The predictor trial split with the levels “early” and “late” will be featured in effect coding such that $\beta 1$ denotes the additionally predicted (log) mean fixation time in later stages of trials. In our power analysis, we proposed a minimally relevant effect of log(1.25), which means that the model predicts a 25% increase in mean fixation time for late trial stages.

For investigating the number of fixations across trial stages, we will test and interpret the main effect of trial split in the same model equation, but with poisson link function and the number

of fixations as target variable. In this case, β 1 denotes to effect of later trial stages on the log number of fixations. In our power analysis, we proposed a minimally relevant effect of log(0.75), which means that the model predicts a 25% decrease in the number of fixations for late trial stages.

Hypothesis 2:

Predicting anomaly detection on an AOI level, we will test the effect of pupil dilation across early and late trial stages on diagnostic performance through generalized linear mixed model with binomial link function. To account for differences in anomaly difficulty, we also feature the anomaly itself as a random effect in the following model:

Y = β0 + β1β2(apd x trial split) + vi(semester|participant) + vj(cohort:participant) + vk(AOI)

The exponential effect sizes of the model predictors will directly be interpreted in terms of their contribution to the estimated likelihood of a given anomaly being marked. As an example, if the main effect of pupil dilation will be estimated at β = 0.10 then exp(β) ~ 1.10 which means that a positive change in one standard deviation in pupil dilation corresponds to a 10% increase in the odds of a given anomaly being marked given all other predictors remain constant.

Interpreting the interaction effect size of pupil dilation and trial split (early, late), it helps to conceptualize the following two model equations which result from a data point measured in either earlier or later stages of trials:

For the early trial split, the following model equation holds:

Y = β 0 + β 1(pupil dilation) + vi(semester|participant) + vj(cohort:participant) + vk(AOI)

For the late trial split, the resulting model equation is:

Y = (β 0 + β 2) + (β 1 + β 3) (pupil dilation) + vi(semester|participant) + vj(cohort:participant) + vk(AOI)

In both equations, β 1 corresponds to the effect of pupil dilation in the early trial split. On the other hand, β 3 in the latter model corresponds to the *additional* effect of pupil dilation in the late trial split (which is also what R reports as the interaction effect size). Exponentiating this interaction term corresponds to the *ratio of odds ratios* of the effects of pupil dilation in the early split [exp(β 1)] and the late split [exp(β 1 + β 3)]. As an example, if we expect the effect of pupil dilation in the early split to be β 1 = -0.11 which is ~ 10% of an odds decrease for an additional standard deviation of pupil dilation, and (β 1 + β 3) = 0.10 ~ 10% odds increase for an additional standard deviation in the late split, then the ratio of odds ratios is exp(0.10) / exp(-0.11) ~ 1.23. Log(1.23) ~ 0.21 then conversely yields the interaction model term β 3 since (β 1 + β 3) = (-0.11 + 0.21) = 0.10.

References

Abokyi, S., Owusu-Mensah, J., & Osei, K. A. (2017). Caffeine intake is associated with pupil dilation and enhanced accommodation. *Eye*, *31*(4), 615-619. https://doi.org/10.1038/eye.2016.288

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,* 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Brunyé, T. T., Eddy, M. D., Mercan, E., Allison, K. H., Weaver, D. L., & Elmore, J. G. (2016). Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation. *BMC medical informatics and decision making*, *16*(1), 77. https://doi.org/10.1186/s12911-016-0322-3

Castner, N., Appel, T., Eder, T., Richter, J., Scheiter, K., Keutel, C., ... & Kasneci, E. (2020). Pupil diameter differentiates expertise in dental radiography visual search. *PloS one*, *15*(5). https://doi.org/10.1371/journal.pone.0223941

Eder, T. F., Richter, J., Scheiter, K., Keutel, C., Castner, N., Kasneci, E., & Huettig, F. (2020). How to support dental students in reading radiographs: effects of a gaze-based compare-and-contrast intervention. *Advances in Health Sciences Education.* https://doi.org/10.1007/s10459-020-09975-w

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*, *242*(2), 396-402. https://doi.org/10.1148/radiol.2422051997

Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and instruction*, *47*, 91-102. https://doi.org/10.1016/j.learninstruc.2016.10.007

Lindner, M. A., Eitel, A., Thoma, G. B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology*, *28*(5), 738-752. https://doi.org/10.1002/acp.3060

Lowenstein, O., Feinberg, R., & Loewenfeld, I. E. (1963). Pupillary movements during acute and chronic fatigue: A new test for the objective evaluation of tiredness. *Investigative Ophthalmology & Visual Science*, *2*(2), 138-157.

Nodine, C. F., & Kundel, H. L. (1987). The cognitive side of visual search in radiology. In *Eye movements from physiology to cognition* (pp. 573-582). Elsevier. https://doi.org/10.1016/B978-0-444-70113-8.50081-3

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010, March). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 141-144). https://doi.org/10.1145/1743666.1743701

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Richter, J., Scheiter, K., Eder, T. F., Huettig, F., & Keutel, C. (2020). How massed practice improves visual expertise in reading panoramic radiographs in dental students: An eye tracking study. *Plos one*, *15*(12). https://doi.org/10.1371/journal.pone.0243060

Turgeon, D. P., & Lam, E. W. (2016). Influence of experience and training on dental students' examination performance regarding panoramic images. *Journal of Dental Education*, *80*(2), 156-164. https://doi.org/10.1002/j.0022-0337.2016.80.2.tb06071.x

Van Der Gijp, A., Ravesloot, C. J., Jarodzka, H., Van der Schaaf, M. F., Van der Schaaf, I. C., van Schaik, J. P., & Ten Cate, T. J. (2017). How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, *22*(3), 765-787. https://doi.org/10.1007/s10459-016-9698-1

Van Gerven, P. W., Paas, F., Van Merriënboer, J. J., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167-174. https://doi.org/10.1111/j.1469-8986.2003.00148.x

Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. Computers in Human Behavior, 25(2), 325-331. https://doi.org/10.1016/j.chb.2008.12.021