Gaze Entropy Metrics for Mental Workload Estimation are Heterogenous During Hands-Off Level 2 Automation

Courtney M. Goodridge[1, *], Rafael C. Gonçalves[1], Ali Arabian[1], Anthony Horrobin[1], Albert Solernou[1], Yee Thung Lee[1], Yee Mun Lee[1], Ruth Madigan[1], Natasha Merat[1]

[1]Institute for Transport Studies, University of Leeds

*c.m.goodridge@leeds.ac.uk

Acknowledgements: The authors would like to thank all partners within the Hi-Drive project for their cooperation and valuable contribution. [This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006664. The sole responsibility of this publication lies with the authors. Neither the European Commission nor CINEA – in its capacity of Granting Authority – can be made responsible for any use that may be made of the information this document contains]. We would like to thank Seeing Machines for the use of their eye tracking equipment to collect data for this study and Michael Daly for software development. We would also like to thank Giancarlo Caccia Dominioni and Audrey Bruneau for their insightful comments for the design, analysis, and writing in this manuscript.

Abstract

As the level of vehicle automation increases, drivers are more likely to engage in non-driving related tasks which take their hands, eyes, and/or mind away from the driving task. Consequently, there has been increased interest in creating Driver Monitoring Systems (DMS) that are valid and reliable for detecting elements of driver state. Workload is one element of driver state that has remained elusive within the literature. Whilst there has been promising work in estimating mental workload using gaze-based metrics, the literature has placed too much emphasis on point estimate differences. Whilst these are useful for establishing whether effects exist, they ignore the inherent variability within individuals and between different drivers. The current work builds on this by using a Bayesian distributional modelling approach to quantify the within and between participants variability in Information Theoretical gaze metrics. Drivers (N = 41) undertook two experimental drives in hands-off Level 2 automation with their hands and feet away from operational controls. During both drives, their priority was to monitor the road before a critical takeover. During one drive participants had to complete a secondary cognitive task (2-back) during the hands-off Level 2 automation. Changes in Stationary Gaze Entropy and Gaze Transition Entropy were assessed for conditions with and without the 2-back to investigate whether consistent differences between workload conditions could be found across the sample. Stationary Gaze Entropy proved a reliable indicator of mental workload; 92% of the population were predicted to show a decrease when completing 2-back during hands-off Level 2 automated driving. Conversely, Gaze Transition Entropy showed substantial heterogeneity; only 66% of the population were predicted to have similar decreases. Furthermore, age was a strong predictor of the heterogeneity of the average causal effect that high mental workload had on eye movements. These results indicate that, whilst certain elements of Information Theoretic metrics can be used to estimate mental workload by DMS, future research needs to focus on the heterogeneity of these processes. Understanding

this heterogeneity has important implications toward the design of future DMS and thus the safety of drivers using automated vehicle functions. It must be ensured that metrics used to detect mental workload are valid (accurately detecting a particular driver state) as well as reliable (consistently detecting this driver state across a population).

Keywords: Distraction, workload, DMS, heterogeneity, automation, entropy

1 Introduction

The influx of automated systems in road vehicles has generated increased interest in the development of Driver Monitoring Systems (DMS). DMS refers to a collection of sensors that aim to detect whether a driver is attentive, alert, or engaged. Not only are drivers more likely to engage in non-driving related tasks (NDRTs) as vehicles transform from manual to partial driving automation (Carsten et al, 2012), but in Level 3 automation drivers are allowed to actively engage in NDRTs (SAE, 2018). This may take their hands off the wheel and eyes and mind away from the main driving task. As such, a large body of research has aimed to measure the internal states of drivers whilst using partial or conditionally automated vehicles, and how these states might change in response to NDRTs. One elusive, yet extremely relevant, driver state for informing driver readiness is workload. Workload is a general term that can be defined as the demand or difficulty that is placed upon a driver (De Waard, 1996; da Silva, 2014; Fuller, 2005; De Winter et al, 2014). *Mental workload* is more specific and has been defined as the proportion of information processing for a given task relative to an individual's processing capacity (Brookhuis & De Waard, 1993; 2000; da Silva, 2014). It should also be noted that the terms *cognitive distraction* and *cognitive load* are often used interchangeably when researchers manipulate the cognitive demand of drivers. However, there is a distinct conceptual difference; the former referring to the general removal of attention away from the driving task toward a secondary task, and the latter referring to the quantity of the cognitive resource demanded by the secondary task (Engström et al, 2017). A key aspect of mental workload is that drivers have a limited pool of cognitive resources (Wickens, 2002). Underload from the monotony of monitoring autonomous systems can result in decreased vigilance (Young & Stanton, 2002) whereas overload may occur if a driver is engaging in an NDRT and can result in sub-optimal takeover performance (Gold et al, 2015; Zeeb et al, 2016). To ensure that a driver is ready to resume control, they should ideally have moderate workload levels to reduce the likelihood of safety-critical situations (Bruggen, 2015). Hence one goal of DMS development has been to identify valid and reliable indicators of mental workload to monitor the driver during automated driving. Therefore, a specific aim of this manuscript was to investigate a family of gaze-based metrics that have shown potential in estimating mental workload in human drivers.

The dispersion of gaze has been a useful metric for measuring mental workload during manual and automated driving. Gaze dispersion is often measured as the standard deviation of raw gaze coordinates in the horizontal and vertical dimensions (Sodhi et al, 2002). During manual driving, the standard deviation of horizontal gaze reduces when the workload of the driver is increased with a secondary cognitively loading task; this phenomenon is known as visual tunneling (Reimer, 2009; Reimer et al, 2010; Wang et al, 2014). Similar effects have been observed when performing a cognitive loading secondary task during automated driving (Radlmayr et al, 2019; Wilkie et al, 2019). The sensitivity of raw gaze dispersion for detecting mental workload has proven to be a robust measure for driver monitoring systems. However, one limitation of this approach is that it does not account for the predictive nature of eye movements. Established accounts of gaze control focus on the where (spatial distribution) and the when (temporal sequence) of gaze, relative to task demands (Shiferaw et al, 2019). This is can be interpreted as being driven by bottom-up (stimulus saliency) or top-down (behavioral requirements) processes (Henderson, 2003; Shiferaw et al, 2019). However, a growing body of literature has proposed that gaze control is a system of spatial prediction (Henderson, 2017; Talter et al, 2017). Hence fixation locations are not merely instructed by top-down and bottom-up influences, but their relative contributions towards prediction and error correction when constructing an internal representation of a visual scene (Parr & Friston 2017; Spratling et al, 2017; Shiferaw et al, 2019). The brain is a prediction machine and aims to minimize error between sensory information and the internal state (Clark et al, 2013). Hence via a combination of bottom-up and top-down processes, gaze control aims to optimize visual sampling in order to make better predictions regarding the location of subsequent fixations (Parr & Friston, 2017; Spratling et al, 2017). Considering the mechanisms involved in gaze control, it can be argued that measuring differences in visual scanning behaviour during varying stages of driving may provide information on changes in the underlying processes that are influenced by increased workload (Shiferaw et al, 2019). Information Theoretic concepts such as entropy are one such method, which focus on using gaze transitions to estimate internal states.

Gaze entropy is an eye tracking metric that has shown promise for estimating mental workload and refers to the application of Information Theory to gaze data (Shiferaw et al, 2019). Within the field of Information Theory, entropy refers to the average amount of information or uncertainty for a given choice (Shannon, 1948). For a system with discrete processes, the two primary components are the source and output; the source being the total number of states that a given output can take. When applied to gaze data, there is an assumption that saccadic movements that produce fixations are outputs from a gaze control system that predicts the spatial locations of proceeding fixations (Shiferaw et al, 2019). The visual field represents all possible state spaces where a fixation could be located. To calculate the entropy of gaze fixations, fixation coordinates are divided into discrete spatial bins to generate probability distributions of a given fixation being within a given location (Shiferaw et al, 2019). The entropy value thus represents the predictability of a fixation location; a higher uncertainty (or entropy) represents a higher dispersion of gaze for a particular viewing period (Holmqvist et al, 2011). This is known as *Stationary Gaze Entropy* (H_s). Another assumption is that subsequent fixations are better predicted by current fixations via *conditional* probability rather than only *total* probability (Weiss et al, 1989; Shiferaw et al, 2019). Therefore, this provides a measure of predictability of visual scanning patterns by considering the order of fixations; this is known as *Gaze Transition Entropy* (H_t). Higher H_t is indicative of less structured, more random scanning patterns (Shiferaw et al, 2019). Because eye movements aim to optimize inference through motor action sequences (Parr & Friston, 2017), it has been proposed that there is an optimal range of H_t to efficiently sample information within the visual scene. Optimal H_t is an ideal level of complexity that balances modulation from underlying bottom-up influences with top-down prediction (Shiferaw et al, 2019). If there is an optimal range of H_t then *increased* H_t may reflect top-down interference whereby there is modulation of gaze beyond the requirements of a given task. This can manifest as highly erratic, random visual scanning. Conversely, lower than optimal H_t can result in insufficient top-down modulation thus producing insufficient visual scanning and exploration. Whilst H_t may change as a function of more visually demanding tasks or visual scenes, given an environment where these factors are experimentally controlled, H_t may change as a function of top-down engagement (Shiferaw et al, 2019).

H_s and H_t provide a quantitative assessment of visual scanning in naturalistic environments and thus have been proposed as measures that can estimate mental workload in drivers. Testing the reliability and validity of gaze entropic metrics has largely been conducted within the domain of manual driving. Schieber & Gilland (2008) found reductions in H_t as a function of secondary task load difficulty; this was further exacerbated for older drivers. The combination of older drivers having reduced visual-spatial processing resources alongside the increased demands of the secondary task resulted in this interaction effect. Schieber & Gilland (2008) proposed that metrics based on Information Theory held significant potential for monitoring driver behaviour as H_t systematically changed as a function of increased mental workload. Pillai et al (2022) implemented a similar design to investigate whether gaze entropy differentiated varying levels of cognitive load during manual driving. By calculating the signal-to-noise ratio (SNR), Pillai et al (2022) found that H_s reliably differentiated between a control task (normal driving and a detection response task) and 2-back, control and 0-back, and 0-back and 2-back conditions. Conversely, H_t could not reliably distinguish between any of these cognitive load comparisons. This suggests that it was the predictability of the dispersion of gaze, rather than gaze transitions, that was useful for estimating mental workload. One of the only experiments to study cognitive load estimation using gaze entropy during automated driving was conducted by Chen et al (2022). They investigated whether H_s changed as a function of automation level (SAE L0, L1, and L2). 3-dimensional H_s (applying the Shannon (1948) equation to coordinates in a 3-dimensional plane) negatively correlated with subjective workload during visual, auditory, or multi-modality cognitive tasks. This is indicative of gaze dispersion decreasing as a function of increased subjective workload, and thus supports similar findings of visual tunneling when cognitively loaded (Radlmayr et al, 2019; Reimer, 2009; Reimer et al, 2010; Wang et al, 2014; Wilkie et al, 2019). Chen et al (2022) concluded that H_s could be a valid indicator for visual and auditory task distractions within driver monitoring systems during partial automation.

Despite evidence that gaze entropy measures can be useful for estimating mental workload, there are some limitations to this work. Chen et al (2022) utilized a desktop computer simulator where the keyboard was used for steering and pedal operations. There was also no simulated traffic or road; just a highly artificial virtual environment. Not only is this a poor replication of real driving, but the lack of stimuli within the visual scene may have produced insufficient bottom-up saliency. There was also no control condition without a secondary task, thus not allowing for any comparison of gaze entropy under normal workload situations. A wider limitation of the literature is the lack of investigation into the variation both within and between individuals. A metric that estimates mental workload must be valid (i.e., the metric systematically varies with mental workload) but it must also be reliable (i.e., the metric systematically changes in similar ways for a given population) if it is to be used in DMS within a wider population. Therefore, understanding how H_s and H_t vary is vitally important. Whilst mean differences are theoretically useful for establishing the existence of effects, they only existence in an abstract sense (Mole et al, 2020). To make applied predictions that relate to the wider population, it is vital to model and understand how a sample varies. Schieber & Gilland (2008) reported no indices of variance in H_t, thus providing no indication as to how variable H_t was when drivers were under high mental workload. Chen et al (2022) reported large individual differences in the difficulty of the spatial N-back task which may have influenced subjective ratings of mental workload alongside eye tracking metrics. However, they did not formally model these differences, or investigate whether specific individual characteristics predicted this variation. Finally, Pillai et al (2022) investigated the effects of gaze entropy by calculating SNR; a lower SNR indicates that two means are more similar. Not only is this metric focused on mean differences but averages of gaze entropy in different conditions are weighted by variance across several participants. Whilst this accounts for variation in entropy, it treats all individual differences as noise. Whilst some individual variance is undoubtedly attributed to noise in eye tracking measurement (Bottos & Balasingam, 2020; Velichkovsky et al, 1997), it is possible that individual differences could vary as function of theoretically useful variables (e.g., age, driving experience).

The aim of the current study was to investigate the feasibility of using gaze entropic metrics to estimate mental workload whilst monitoring a Level 2 automated vehicle with their hands and feet away from operational controls. Previous research has shown that eye movements change as a function of increased mental workload (Radlmayr et al, 2019; Reimer et al, 2009; Reimer et al, 2010; Wilkie et al, 2019). However, using Information Theory to study gaze metrics can go beyond understanding the spatial distribution of gaze and focus on how efficiently drivers are scanning the visual scene. Thus far, there is evidence that H_s and H_t can be used to detect driver workload (Chen et al, 2022; Pillai et al, 2022; Schieber & Gilland, 2008). However, the methodology used to make these conclusions has seemingly ignored how these variables vary within a given population. Such variance is vital, if we are to understand whether these

Information Theoretic metrics can be used by DMS to improve the safety outcomes for a wide range of users.

2 Material and methods

2.1 Participants

41 participants were recruited from a university participant pool and took part in the experiment however three had to be removed before data analysis as they either did not follow experimental instructions, or eye tracking data was not correctly captured. The remaining 38 participants (16 females, 22 males, mean age = 38.81, range = 22-65) all had normal or corrected to normal vision. All participants had a valid UK driving license (mean number of years = 17.8, range = 4-43) and were regular drivers (mean annual miles = 9355.25, range 5000-20000).

2.2 Apparatus and materials

The experiment was conducted at the University of Leeds Driving Simulator (see Figure 1). This is a motion-based driving simulator consisting of a Jaguar S-type cab encased within a 4 m spherical projection dome. The dome has a 300° field of view projection to render the driving environment. Driver controls are fully operational; pedals and steering provide haptic feedback for participants to replicate real-world driving. Longitudinal and lateral movement is also provided via a hexapod motion base and a 5 m x 5 m X-Y table. Gaze data were collected using a Seeing Machines Driver Monitoring System eye tracker sampling at 60 Hz. Subjective ratings of workload were measured via the NASA-Task Load Index (NASA-TLX). The NASA-TLX consists of 6 subscales that measure subjective ratings of mental, physical, and temporal demands as well as frustration, effort, and performance of the task (Hart, 2006).

Figure 1: University of Leeds Driving Simulator

2.3 Design

A 2 x 2 Repeated Measures design was used in this study. The two within-participant factors were event criticality and mental workload. Event criticality was manipulated by changing the time to collision at the onset of a lead vehicle braking (TTC) after a period of hands-off Level 2 automated driving. The aim of manipulating this variable was to create two levels of criticality: a “less severe” level (TTC = 5 s) that allowed participants to successfully take over without crashing, and a “severe” level that could lead to a crash if the participant was not monitoring the road correctly (TTC = 3 s). These values were chosen based on previous studies that have demonstrated that a 3 s TTC produces highly critical events, whilst a 5 s TTC provides sufficient time for takeovers (Gold et al, 2013; Mok et al, 2015; Louw & Merat, 2017). The second within-participants factor that was manipulated was mental workload. This was manipulated over two levels; a no-load condition and a high mental workload condition where participants had to complete a secondary task during the automated driving sections. To induce cognitive load, participants completed a verbal response delayed digit recall task (N-back) (Mehler et al, 2011) during the automated driving sections. The specific N-back used in the current investigation was a 2-back condition. This task was chosen because it is highly controlled, non-visual, and has been consistently shown to increase the workload of drivers during manual (Reimer, 2009; Reimer et al, 2010; Wang et al, 2014) and automated driving (Radlmayr et al, 2019; Wilkie et al, 2019).

The experiment consisted of two drives for each participant. During one drive participants completed an N-back throughout the automated period; during the other drive there was no secondary task. The order of N-back was counter-balanced across participants. Each drive lasted approximately 35 minutes and all participants drove on the same 3-lane UK motorway. Each drive consisted of 10 discrete events, each consisting of 30 s of manual driving followed by approximately 2 minutes of automated driving. After 2 minutes of automated driving, a takeover request (TOR) was delivered. Four of these events were critical: two with a TTC of 3 s, two with a TTC of 5 s. For 3 s TTCs, the lead vehicle braked suddenly and decelerated at a rate of 5.55 m/s^2, whereas for the 5 s event, the lead vehicle decelerated at 2 m/s^2. Decelerations began as soon as the takeover request (TOR) was triggered. The remaining six events were non-critical; two involved no lead vehicle, and the remaining four involved a lead vehicle that did not decelerate once the TOR was triggered. Lead vehicles appeared in front of the ego vehicle shortly before the automation was engaged. They entered the middle lane from the left-hand lane and participants were instructed to allow the lead vehicle to pull in front. Once in the middle lane, lead vehicles matched the ego-vehicle’s speed at a distance of 25 m during automation. Participants drove in the middle lane, with ambient traffic flow in the left and right lanes. Once the lead vehicle was present, the automated system engaged.

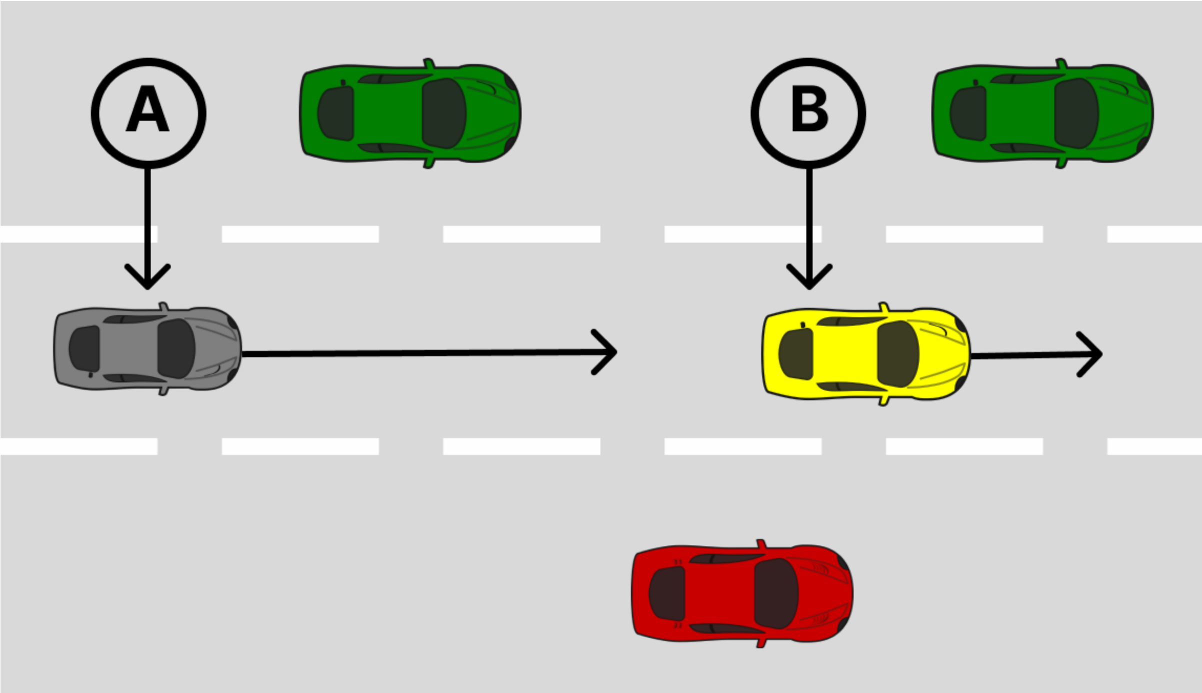

Figure 2: Schematic representation of an event. (A) represents the ego vehicle and (B) represents the lead vehicle. Lead vehicles entered from the left lane and matched the ego vehicles speed at a distance of 25 m. Following 2 minutes of automated driving, for critical trials the lead vehicle decelerated at 5.55 m/s^2 (TTC = 3 s) or 2 m/s^2 (TTC = 5 s). For non-critical trials, a TOR was delivered but the lead vehicle did not decelerate.

2.4 Procedure

Informed consent was obtained, and standardized procedural instructions were delivered. All procedures were approved by the University of Leeds Research Ethics Committee (Reference code: 2022-0353-206).

Upon arrival participants completed a number of pre-drive questionnaires (data from these questionnaires are not analysed or reported in this manuscript). Participants conducted a practice session to become familiar with all aspects of the experiment and the driving simulator dynamics. Participants were talked through the design of the Human-Machine Interface (HMI) (see Figure 3), how to disengage the automation, and completed a static N-back task. During the driving portion of the practice the 3-lane motorway contained ambient traffic. Takeovers during the practice were non-critical.

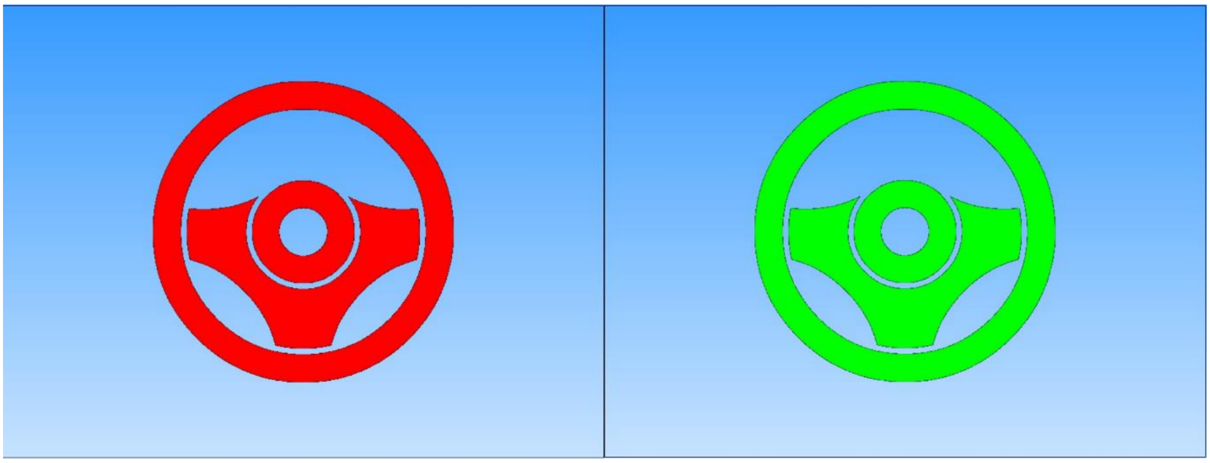

Figure 3: Icons used to indicate system status. Green steering wheels indicated the Level 2 autonomous system was activated. Red steering wheels indicated that the driver needed to take over. During manual driving, the steering wheel was greyed out. In the experiment, the red steering wheel flashed until the vehicle was back into manual driving mode.

For experimental drives, participants were instructed to enter the motorway and position themselves in the centre of the middle lane and maintain a speed of 70 MPH. After approximately 30 s of manual driving the automated system engaged automatically. This was indicated by a short auditory tone and the shifting of the steering wheel icon from grey (manual mode) to green (automation engaged) (see Figure 3). Once in automated driving mode, participants were instructed to take their hands off the wheel and feet away from the pedals and to monitor the road environment for any potential hazards. After approximately 2 minutes of automated driving, a TOR was delivered. The TOR was characterised by an auditory tone and the steering icon flashing red within the instrument cluster. Participants were instructed to take over once the TOR had been issued; this could be done by any steering input over 2°, pressing any of the pedals, or pressing a micro-switch button strapped to the steering wheel. If the driver of the ego-vehicle did not respond within 10 seconds, the automation would disengage by itself.

Following the takeover, the participant engaged in 30 s of manual driving before the automated system engaged once more. If the driver exited the middle lane during takeovers, they were instructed to return as soon as possible. There were 10 discrete events per drive and each drive lasted approximately 35 minutes. During one drive participants completed an auditory-verbal N-back task when automation was engaged, which continued until a TOR was given. Participants were instructed that a safe drive was their primary goal. After each drive, participants filled out a NASA-TLX to collect data on subjective ratings of workload. After the second experimental drive, participants completed post-drive questionnaires (data from these questionnaires is not analysed or reported in this manuscript).

2.5 Statistical modelling

The main aim of this manuscript was to investigate changes in gaze entropic eye metrics during the 2-minute automation period with and without N-back, and with and without a lead vehicle. This includes critical and non-critical trials that included a lead vehicle. Thus, data relating to the takeover and manual driving portions are not analysed within this manuscript. Data and analysis code can be found in the following link (https://github.com/courtneygoodridge/gaze_entropy_heterogenous).

2.5.1 Gaze entropy

To calculate stationary gaze entropy (H_s), the Shannon (1948) entropy equation was applied to the fixation data:

$$H_s(x) = -\sum_{i=1}^{N} p(i) log_2 p(i) \tag{1}$$

Where H_s is entropy for a given set x (time period during automation for a given condition), i is the number of state spaces or locations (in a 2-dimensional coordinate plane) of each fixation in x, N is the total number of fixations in x, and $p(i)$ is the proportion of fixations landing in a given state space. Gaze transition entropy (H_t) was calculated by applying the conditional entropy equation to 1st order Markov fixations transitions:

$$H_t(x) = -\sum_{i=1}^{N} p(i) \left[\sum_{i=1}^{N} p(i \mid j) \, log_2 p(i \mid j) \right], i \neq j \qquad (2)$$

When $p(i)$ is the stationary distribution of fixations, $p(i \mid j)$ is the probability of transitioning to state j given being currently in state i, and $i \neq j$ excludes transitions that occur within the same state space (Ellis & Stark, 1986). Fixations were split into spatial bins to apply the equations. This is the primary method of discretisation in the literature (Di Stasi et al, 2017; Krejtz et al, 2014; 2015, Raptis et al, 2017) and has been proposed as the superior method for dynamic stimuli (Shiferaw et al, 2019). For interpretability, both H_s and H_t were normalized by dividing by the maximum entropy, H_{max}. Maximum entropy is the logarithm (base 2) of all state spaces and thus represents when distributional information is at a maximum. For example, each fixation is equally spaced out within the visual scene, and each transition is completely random (Shiferaw et al, 2019). As such, H_s and H_t range from 0-1 and represent the percentage of maximum possible entropy.

2.5.2 Analytic approach

To develop human-centred driver monitoring systems that can reliably detect the mental workload of drivers, it is important to consider the distribution of driver responses rather than focusing merely on the mean. Whilst mean differences are useful for establishing the presence of effects across conditions, using mean values is limited, since it only exists in an abstract sense - no single driver can be considered “the average” (Mole et al, 2020). Furthermore, means do not contain *within* or *between individual* variability which are vital components for making real world predictions about human behaviour. Standard regression-based analyses aim to model the population mean (μ) whilst assuming that the within-participants variance (σ) is consistent. Not only is the assumption of homogeneity of variance often violated (Schielzeth et al, 2020) but there is also theoretical justification that σ might vary as a function of the manipulated variables in the experiment.

As highlighted in the Introduction, the motor coordination of eye movements aims to optimise inference (Parr & Friston, et al 2017). This implies that there is an optimal level of H_t for effective sampling of the visual scene whereby top-down processes modulate default bottom-up activation (Shiferaw et al, 2019). Whilst increases or decreases in the μ of H_t can be indicative of top-down interference or top-down modulation respectively (Shiferaw et al, 2019), the trial-by-trial variance within individuals can also be a crucial index for measuring the efficiency of visual scanning. Under the assumption that the visual scene maintains an ambient level of complexity, optimal H_t should be consistent within an individual. However, if increased mental workload results in decreases in H_t via top-down modulation, it may also affect how efficiently individuals are able to maintain optimal H_t from one trial to the next. The idea that a change in *variance* can indicate a change in a driver's internal state is not new within the driver monitoring and distraction literature. Horrey & Wickens (2007) proposed that standard statistical methods that focus on mean differences (or other measures of central tendency) are insufficient for measuring driver distraction, and that modelling large deviations in attention can reveal infrequent lapses in visual sampling control; something that can be missed when only focusing on averages. Kujala & Saarilouma (2011) found reductions in the standard deviation of fixation durations for simpler in-vehicle information systems menu deigns, thus suggesting that the variance in fixations durations could be used to assess the efficiency of visual search performance. It is thus proposed in this manuscript that a similar effect might be present for H_t , when increasing mental workload. To assess whether there are systematic changes in σ as a function of the predictor variables, the current analysis will apply distributional models. Distributional models relax the assumption of consistent σ, and allow it to be predicted by parameters as can be done when predicting μ (Bürkner, 2017).

It is also vital to quantify *between-participants variance*, as the overall aim of any analysis is to make predictions towards the population. This is particularly true for DMS, if these systems are to be reliable for establishing the state of a large and varying driver population. To model the between-participants variance, we used a multilevel modelling approach. The multilevel aspect of the model refers to the inclusion of fixed and random effects. Whilst fixed effects refer to the contribution of a predictor variable towards the average change, random effects model the variation between different participants on average, alongside how they vary in response to predictor variables (Lo & Andrews, 2015).

2.5.2.1 Model development

The population mean, μ, of all the gaze-based metrics were modelled as the linear combination of an intercept (β_0), N-back (N, β_N), presence of a lead vehicle (L, β_L), and an interaction term between these variables (NL, β_{NL}). The N-back task was parameterised as $N \in \{0, 1\}$ where $N = 1$ corresponds to the presence of the N-back during hands-off Level 2 automation. Similarly, lead vehicle was parameterised as $L \in \{0, 1\}$ where $L = 1$ corresponds to the presence of a lead vehicle during automation. The standard deviation, σ, was independently modelled as a linear combination of an intercept (α_0), N-back (α_N), presence of a lead vehicle (α_L), and an interaction (α_{NL}). Because σ cannot be negative, the $log(\sigma)$ was modelled. The distributional model structure was specified as follows:

$$Y_{ij} \sim N\left(\mu_{ij}, \sigma_{ij}\right) \tag{3}$$

$$\mu_{ij} = \left(\beta_0 + \beta_{0_j}\right) + \left(\beta_N N_i + \beta_{N_j} N_i\right) + (\beta_L L_i) + (\beta_{NL} N L_i)$$

$$log\ (\sigma_{ij}) = \left(\alpha_0 + \alpha_{0_j}\right) + \left(\alpha_N N_i + \alpha_{N_j} N_i\right) + (\alpha_L L_i)$$

$$\begin{bmatrix} \beta_{0_j} \\ \beta_{N_j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_0 \\ \beta_N \end{bmatrix}, S_\beta\right)$$

$$\begin{bmatrix} \alpha_{0_j} \\ \alpha_{N_j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \alpha_0 \\ \alpha_N \end{bmatrix}, S_\alpha\right)$$

$$S_\beta = \begin{pmatrix} \sigma^2_{\beta_{0_j}} & \rho\sigma_{\beta_{\mathrm{N}_j}}\sigma_{\beta_{0_j}} \\ \rho\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}} & \sigma^2_{\beta_{N_j}} \end{pmatrix}$$

$$S_\alpha = \begin{pmatrix} \sigma^2_{\alpha_{0_j}} & \rho\sigma_{\alpha_{\mathrm{N}_j}}\sigma_{\alpha_{0_j}} \\ \rho\sigma_{\alpha_{0_j}}\sigma_{\alpha_{N_j}} & \sigma^2_{\alpha_{N_j}} \end{pmatrix}$$

Where Y denotes the response variable, i specifies the condition of each variable, j specifies the participant, and S_β and S_α are matrices corresponding to the variance or covariance parameters.

A model was also built to investigate how N-back influenced subjective mental workload. The population mean, μ, was modelled as linear combination of an intercept (β_0) and N-back (denoted N, β_N):

$$Y_{ij} \sim N\left(\mu_{ij}, \sigma_{ij}\right) \tag{4}$$

$$\mu_{ij} = \left(\beta_0 + \beta_{0_j}\right) + \left(\beta_N N_i + \beta_{N_j} N_i\right)$$

$$\begin{bmatrix} \beta_{0_j} \\ \beta_{N_j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_0 \\ \beta_N \end{bmatrix}, S_\beta\right)$$

$$S_\beta = \begin{pmatrix} \sigma^2_{\beta_{0_j}} & \rho\sigma_{\beta_{\mathrm{N}_j}}\sigma_{\beta_{0_j}} \\ \rho\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}} & \sigma^2_{\beta_{N_j}} \end{pmatrix}$$

Where Y denotes the response variable, i specifies the condition of each variable, j specifies the participant, and S_{β} is a matrix corresponding to the variance or covariance parameters.

2.5.2.2 Model fitting

A Bayesian approach was used in this manuscript to analyse the data. Posterior distributions were estimated using the No-U-Turn Sampler (NUTS) in the brms package in the R programming language (Bürkner, 2017). For parameters estimating mean (μ) differences between the predictor variables, informative priors were used. For distributional parameters, brms defaults were used to reflect that σ is a standard deviation and thus can only take positive values. The final models were reached by incrementally increasing model complexity. Model comparisons were made using leave-one-out cross validation and additional terms were only kept if they decreased prediction errors (Vehtari et al, 2017).

Using a Bayesian approach, each parameter has an associated probability distribution which quantifies the level of uncertainty, conditioned on the data. In this manuscript, posterior distributions of parameters are described by their mean and a 95% Credible Interval (CI) whereby there is a 95% probability that the true parameter value will fall; values inside this density have higher credibility than those outside it (Kruschke, 2014). The reader is discouraged in making dichotomous decisions when understanding whether there is an effect. Rather, they should use the mean and 95% CIs to assess size, direction, and uncertainty of an effect. Where appropriate, the *probability of direction* (pd) is also reported to illustrate what percentage of the posterior distribution is above or below 0 (Makowski et al, 2019).

3 Results

3.1 Subjective measures

To develop a ground truth regarding the cognitive loading effects of the N-back task, the mental demand facet of the NASA-TLX was compared between N-back conditions. The β_N parameter predicts that the presence of N-back during hands-off Level 2 automated driving doubled subjective scores of mental demand on average from 38.994 to 78.705. The model predicts with high certainty that N-back produced large increases in subjective mental workload.

Table 1: Posterior means and 95% CIs for fixed effect parameters predicting μ_{ij} of NASA TLX mental demand

Fixed effects	
	Dependent variable:
	Mental demand
β_0	38.994 (32.656, 45.257)
β_N	39.711 (32.057, 47.369)
Participants	38
Observations	76

3.2 N-back performance

Performance data for the N-back task was only available for 37 out of 38 participants due to data loss. The average performance was reasonably high and homogenous across the sample (M = 70.77, SD = 15.13) however the high and low scores were quite different (range = 37.38 – 90.97). Previous research in manual driving had found that younger drivers had significantly better 2-back performance in comparison to older drivers (Öztürk et al, 2023). To investigate this, a univariate Bayesian correlation model was fitted on the standardised values of age and performance. The results indicate a negative correlation of -.349 (95% CI: -.666, -.037) suggesting that older drivers tended to have worse N-back performance. This medium effect size is slightly lower than what was been found in manual driving (Öztürk et al, 2023) although the average correlation did highlight a lot of variability; the correlation could be up to -.666, or as low as -.03 (effectively zero).

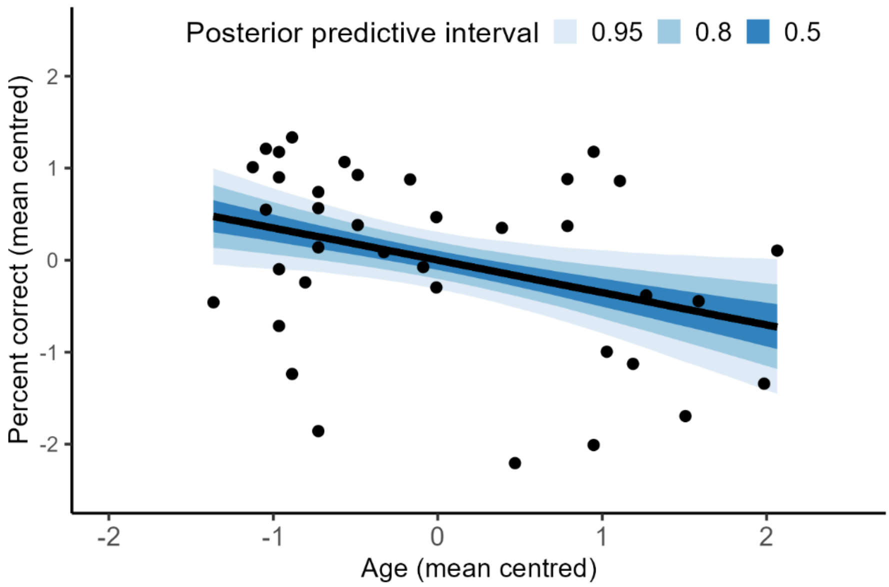

Figure 4: Correlation between age and percentage of correct 2-back responses. Values are standardized to maintain model stability. Black line represents the posterior mean surrounded by bands representing predictive intervals.

3.3 Gaze behaviours

Now that is has been established that N-back increased subjective mental workload between the different driving conditions, an investigation into differences in eye movements can be conducted to see if there were reliable differences in gaze entropic metrics as a function of N-back.

3.3.1 Stationary Gaze Entropy (H_s)

3.3.1.1 Distributional parameters for H_s

The β_N parameter predicted an average decrease in H_s of -.141 (95% CI: -.178, -.101) when drivers completed the N-back task; equivalent to a 14 percentage point reduction in normalized H_s. The β_L parameter predicted an average decrease in H_s of -.041 (95% CI: -.058, -.022) when a lead vehicle was present during automation; equivalent to a 4 percentage point reduction. The β_{NL} parameter was estimated to be .017 suggesting that N-back reduced the difference in H_s between lead and no lead conditions by around 1.7 percentage points. However, as highlighted in Figure 5 there is some uncertainty for this effect; only 92% of the most probable parameters values are above 0.

Table 2: Posterior means and 95% CIs for fixed effect parameters predicting μ_{ij} of H_s

Fixed effects

	Dependent variable:
	H_s
β_0	.474 (.428, .520)
β_N	-.141 (-.178, -.101)
β_L	-.041 (-.058, -.022)
β_{Nl}	.017 (-.006, .040)
Participants	38
Observations	744

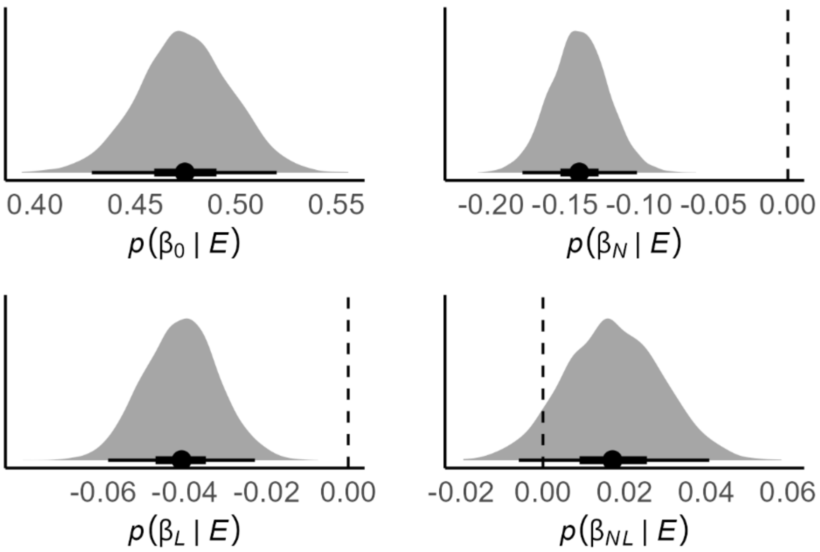

Figure 5: Posterior distributions for model parameters predicting the effect of N-back, lead vehicle, and their interaction on the μ_{ij} of H_s. N-back and lead vehicle have strong negative effects on H_t. The interaction effect is positive, but uncertain with regard to its direction. Dashed lines are presented to illustrate a null effect.

The direction of the effects for σ_{ij} of H_s are uncertain. N-back is predicted to decrease σ_{ij} by 15%, however the probability that the effect is negative is only 90%. As shown in Figure 6, a similar pattern of results is found for the presence of the lead vehicle and the interaction effect.

Table 3: Posterior means and 95% CIs for fixed effect parameters predicting σ_{ij} of H_s

	Fixed effects
	Dependent variable:
	H_s
α_0	-2.676 (-2.867, -2.475)
α_N	-.167 (-.420, .095)
α_L	.098 (-.103, .294)
α_{Nl}	.019 (-.265, .290)
Participants	38
Observations	744

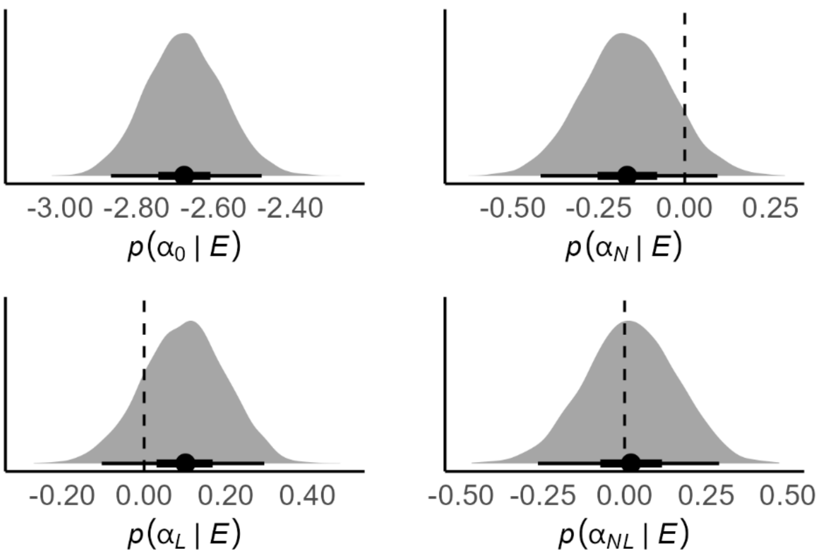

Figure 6: Posterior distribution for model parameters predicting the effect of N-back, lead vehicle, and their interaction, on σ_{ij} of H_s. The effect of N-back σ_{ij} is estimated to be negative, however there is only a 90% probability of this. The effects of lead vehicle and the interaction are estimated to be close to 0, thus highlighting high uncertainty with regard to the size and direction of their effect on the within-participants variance of H_t. Dashed lines are presented to illustrate a null effect.

Overall, the model predicts that N-back reduces the spatial distribution of gaze. This is evidence of reduced top-down engagement when monitoring the road environment during hands-off Level 2 automated driving. This supports previous research which has shown that increased mental workload during automated driving reduces gaze dispersion (Wilkie et al, 2019) and suggests that H_s could be a good metric for estimating mental workload in drivers. Modelling the trial-by trial variance in H_s did not show strong effects of N-back or lead vehicle. This is highlighted in Figure 7, whereby the predictive intervals overlayed on raw data have similar ranges around their predicted means for all conditions. This suggests that *variance* in gaze dispersion from trial to trial was consistent across trials and thus changes in σ_{ij} of H_s may not be useful for detecting increased driver workload.

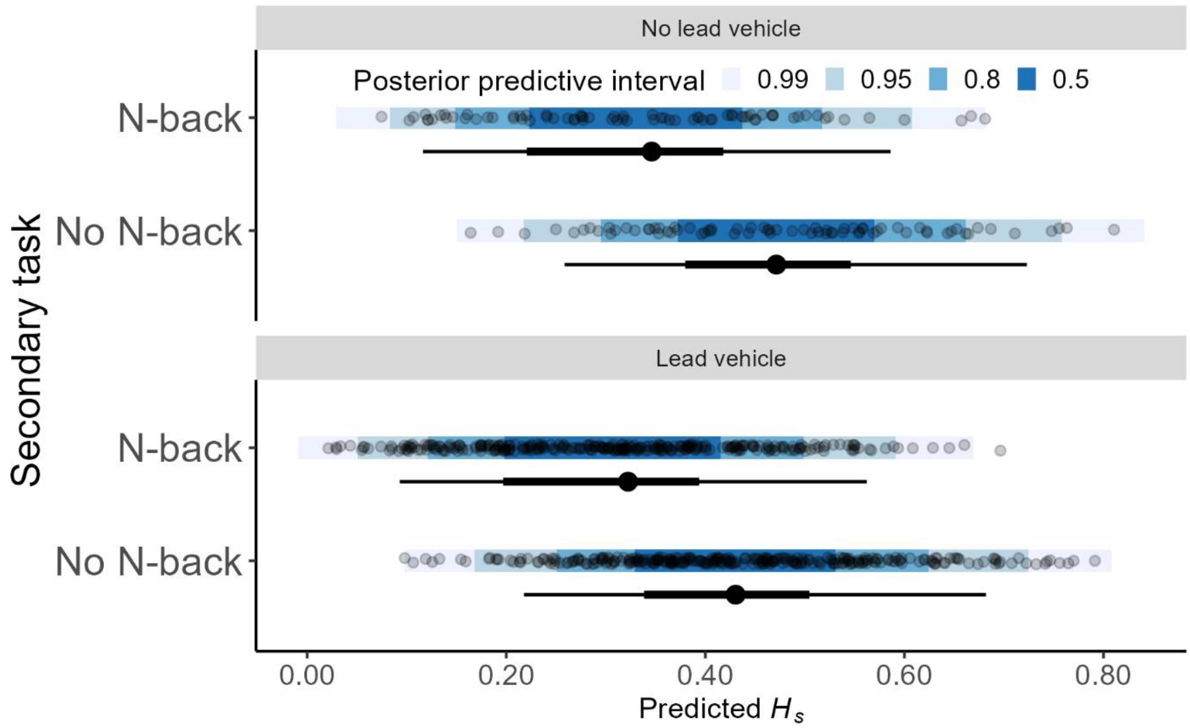

Figure 7: Posterior predictive bands and posterior distribution of means plotted against raw data for conditions with and without a lead vehicle. The point-interval plot highlights the predicted mean differences between N-back/no N-back and lead/no lead vehicle alongside 50% and 95% credible interval bars. For both lead vehicle and N-back comparisons, the posterior predictive intervals are roughly of similar size highlighting the lack of evidence for N-back and lead vehicle affecting σ_{ij} of H_s.

3.3.1.2 Heterogeneity parameters for H_s

Although the typical driver had reduced H_s by 14 percentage points during the N-back condition, people differed in the size of this effect. Some participants had reductions as large as 29 percentage points, some as a low as 3 percentage points, whereas some demonstrated *increases* in H_s by up to 8 percentage points (see Figure 8, left panel). Despite these outlying participants, the model estimates that 92% of the population are expected to have reductions in H_s as a result of completing N-back during automation; the remaining 8% of the population are expected to see moderate increases in H_s whilst cognitively loaded (see Figure 9, right panel).

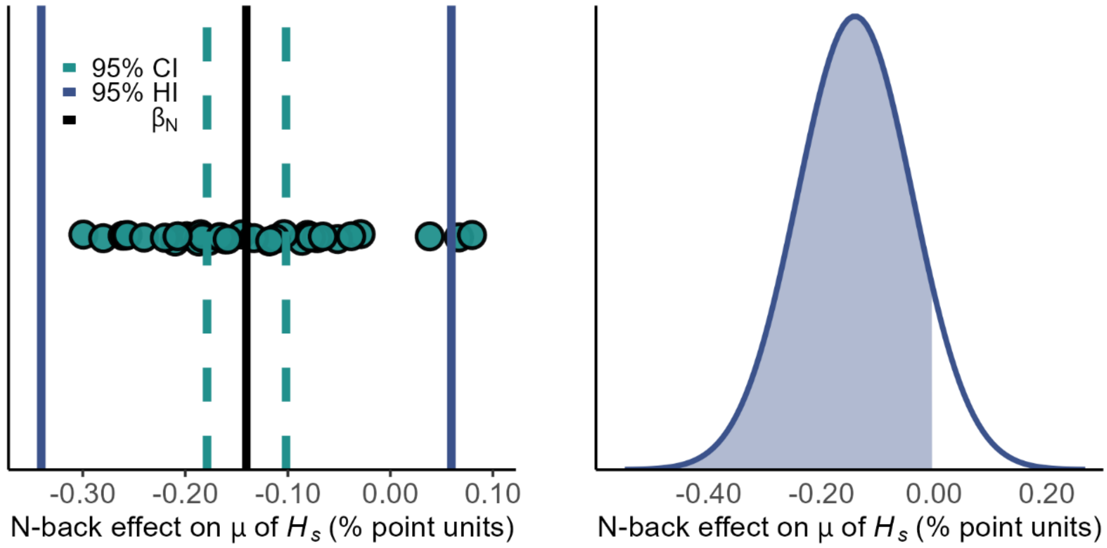

Figure 8: Left panel: strip plot displaying the range of causal effect of N-back on H_s. The black lines denote the average decrease in H_s (fixed effect), the blue dashed lines denote the heterogeneity of the average casual effect of N-back (95% Credible Intervals) and the red solid lines denote the population heterogeneity of the effect of N-back. Right panel: population heterogeneity distribution implied by the model estimates of the mean and standard deviation. 92% of the population are predicted to demonstrate decreases in H_s when completing N-back tasks.

These results suggest that H_s is a strong contender for estimating mental workload during hands-off Level 2 automated driving. Reductions in H_s during N-back are consistent across a population, with the model predicting that 92% of the population would have similar decreases under similar situations. Although the direction of this effect is consistent, the magnitude can vary drastically; up to 2.5 times larger than the average predicted from this sample.

3.3.2 Gaze Transition Entropy (H_t)

3.3.2.1 Distributional parameters for H_t

The β_N parameter predicted that the average decrease in H_t was -.021 (95% CI: -.037, -.004) when drivers were completing the N-back task during automated driving. This is equivalent to a reduction of 2 percentage points in H_t. It should be noted that the average effect could be as low as a reduction of .004 percentage points which would be effectively 0, or as high as a 3.7 percentage point reduction. The model parameters for the effect of lead vehicle and the interaction between N-back and lead vehicle were estimated as close to 0 with high certainty, thus suggesting no meaningful effect on average H_t (see Table 4).

Table 4: Posterior means and 95% CIs for parameters predicting the μ_{ij} of H_t

Fixed effects

	Dependent variable: H_t
β_0	.215 (.208, .222)
β_N	-.021 (-.037, -.004)
β_L	.001 (-.003, .006)
β_{Nl}	-.005 (-.012, .001)
Participants	38
Observations	744

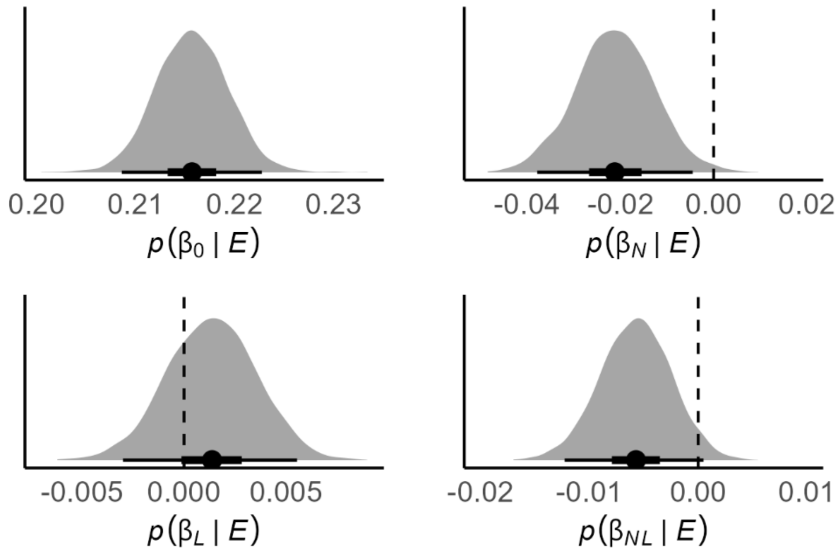

Figure 9: Posterior distribution for model parameters predicting the effect of N-back, lead vehicle, and their interaction on μ_{ij} of H_t. N-back has a small negative effect on H_t. The effects of lead vehicle and the interaction are estimated to be close to 0 with reasonably high certainty. Dashed lines are presented to illustrate a null effect.

The model also predicted differences in the σ_{ij} of H_t as a function of N-back and lead vehicle (see Table 5). The e^{α_N} parameter highlights an increase of 44% in within-participants variance in H_t when completing the N-back during automation. The e^{α_L} parameter indicates that H_t increased by 35% when a lead vehicle was present. The $e^{\alpha_{NL}}$ parameter suggests that the difference in within-participants variance between conditions with and without a lead vehicle were 23% smaller when drivers were not completing the N-back. However, there is some uncertainty with this effect; the probability of the effect being above 0 is 95% (see Figure 10).

Table 5: Posterior means and 95% CIs for parameters predicting the σ_{ij} of H_t

	Fixed effects
	Dependent variable:
	H_t
α_0	-4.145 (-4.369, -3.920)
α_N	.369 (.042, .696)
α_L	.304 (.089, .524)
α_{Nl}	-.262 (-.568, .040)
Participants	38
Observations	744

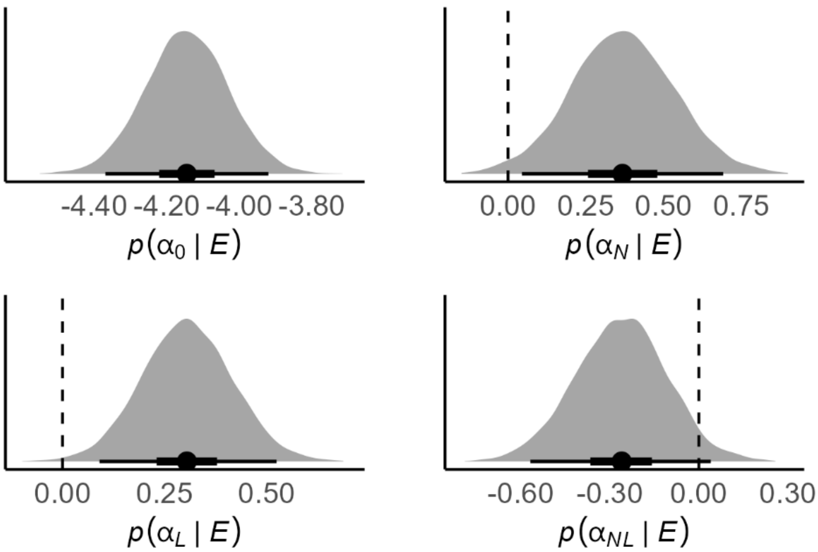

Figure 10: Posterior distribution for model parameters predicting the effect of N-back, lead vehicle, and their interaction, on σ_{ij} of H_t. N-back and lead vehicle have strong negative effects on H_t. The interaction effect is negative but slightly uncertain with regard to its direction; only 95% of the posterior distribution is above 0. Dashed lines are presented to illustrate a null effect.

Model parameters highlight that completing N-back during automated driving produces fixation transitions that are less erratic and more constrained within the visual scene. This average decrease suggests that N-back produced top-down modulation of visual scanning resulting in less complex, more constrained scanning behaviours. The concurrent reduction in mean H_s and H_t as a function of N-back suggests that drivers did not perform sufficient exploration of the visual scene while under high workload, and thus had reduced top-down engagement whilst monitoring the automated system. This can be taken as evidence that, on average, drivers during Level 2 automation who were under high workload had reduced complexity of eye movements. The model also predicted *increases* in the σ_{ij} of H_t as a function of N-back. The increase in σ_{ij} of H_t is highlighted in Figure 11; raw data are dispersed across a broader range during N-back conditions. The systematic change in σ_{ij} as a function of N-back tells us something about the relationship between visual scanning complexity and mental workload. Not only did drivers have reductions in scanning complexity, but they also failed to maintain a consistent complexity on a trial-by-trial basis. Instead, drivers demonstrated frequent fluctuations.

The presence of a lead vehicle had no meaningful effect on mean H_t. However, σ_{ij} did increased by 35% in the presence of a lead vehicle. This suggests that when following a lead vehicle, drivers struggled to maintain their scanning complexity within an optimal range; instead, their trial-by-trial variance in H_t was high.

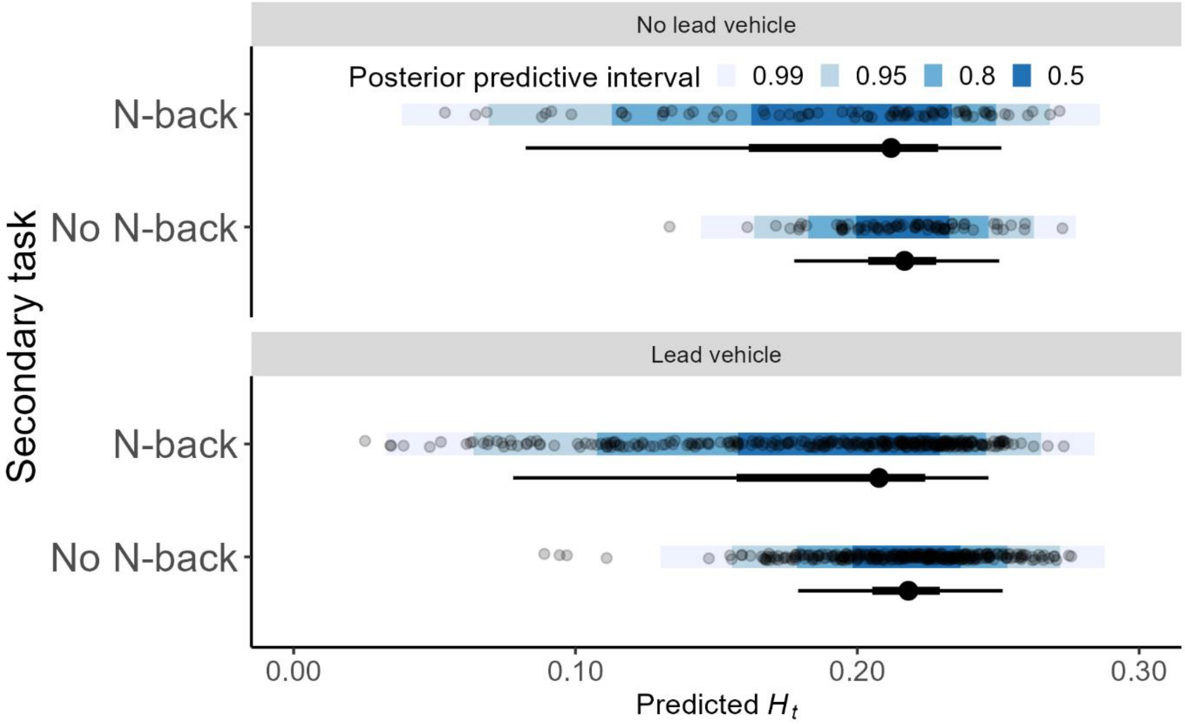

Figure 11: Posterior predictive bands and posterior distribution of means plotted against raw data for H_t. The point-interval plot highlights the predicted mean differences between N-back/no N-back and lead/no lead vehicle alongside 50% and 95% credible interval bars. It is evident that there are small differences in predicted means between N-back and no N-back, however lead vehicle seems to have no effect on mean H_t. It is also evident that σ_{ij} increases as a function of N-back and lead vehicle, which is highlighted by the wider predictive intervals and larger spread of the data.

3.3.2.2 Heterogeneity parameters for H_t

The heterogeneity parameters of the model highlight considerable variance; the random slope parameter (β_{N_j}) is almost two and a half times bigger than the average causal effect (β_N). Whilst the average reduction in H_t during N-back was 2 percentage points, some people have decreases in H_t of -.125 during N-back (12.5 percentage points) whereas some have *increases* of up to .043 (4 percentage points) (see Figure 12, left panel). Furthermore, over 40% of the sample show small-to-moderate *increases* in H_t during the N-back; a reversal of the average trend. This suggests that a considerable proportion of the sample demonstrate more erratic and random sampling patterns when cognitively distracted. The model predicts that only 66% of the population will show an average decrease in H_t when completing the N-back during Level 2 automated driving (see Figure 12, right panel). The remaining 34% of the population are expected to show increases in H_t, resulting in more erratic fixations transitions when cognitively loaded.

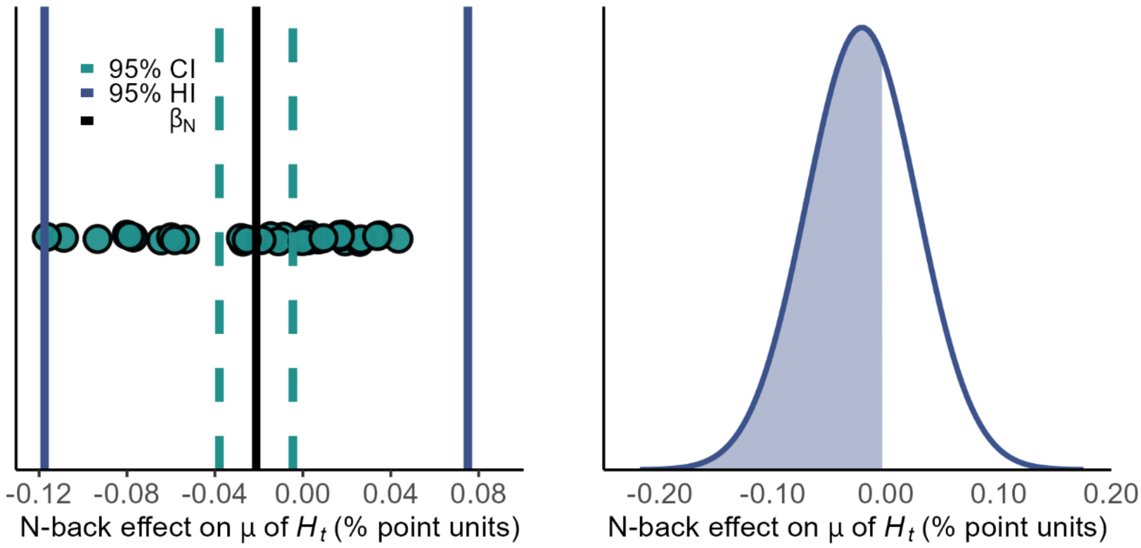

Figure 12 The left panel shows a strip plot of the model predictions of the causal effect of 2-back on H_t. The black lines denote the average mean decrease in H_t (fixed effect), the blue dashed lines denote the heterogeneity of the average casual effect of N-back (95% Credible Intervals) and the red solid lines *denote the population heterogeneity of the effect of N-back. The right panel shows the population heterogeneity distribution implied by the model's estimates of the mean and standard deviation for effect of N-back on H_t. Only 66% of the population are predicted to demonstrate mean decreases in H_t when completing the N-back task.*

Compare this to changes in σ_{ij} of H_t as a function of N-back. The random slope parameter predicting σ_{ij} (α_{N_j}) is only 1.5 times bigger than the average causal effect of N-back on σ_{ij}

(α_N). This is further supported by looking at individual changes in σ_{ij} of H_t as a function of the N-back (see Figure 13, left panel). Whilst there is variation in the size of the effect, the direction of the effect is more consistent across the sample. This is reflected in the model predictions for the population; it predicts that 76% of the population show average increases in trial-by-trial variance when completing the N-back task during Level 2 automated driving.

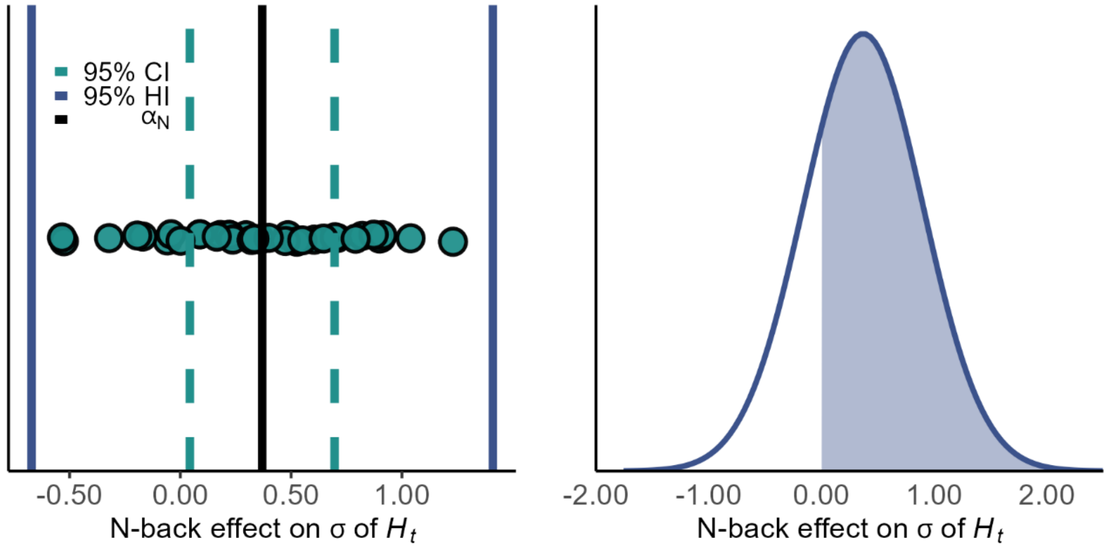

Figure 13: The left panel shows a strip plot of the model predictions of the causal effect of N-back on σ_{ij} of H_t. The black lines denote the average decrease in σ_{ij} (fixed effect), the blue dashed lines denote the heterogeneity of the average casual effect of N-back (95% Credible Intervals) and the red solid lines *denote the population heterogeneity of the effect of N-back. The right panel shows the distribution of the individual effects of N-back on σ_{ij} of H_t in the population predicted by the model. 76% of the population are predicted to demonstrate increases in σ_{ij} of H_t when completing the N-back task.*

These findings provide further credence to the assessment of H_t made in the previous section. Both μ_{ij} and σ_{ij} of H_t change as a function of N-back. However, changes in σ_{ij} are predicted to be more consistent across the population.

3.4 Understanding heterogeneity in average causal effect

Thus far is has been demonstrated that the mean of H_s and H_t change as a function of N-back. However, they both also demonstrate substantial variation across the sample, albeit in differing manners. H_s decreases for a majority of the sample but at varying magnitudes. Conversely, H_t decreases for only two thirds of the sample with the remaining participants showing null effects or small reversals. Whilst this is theoretically useful, it is also important to understand *why* these effects are so variable. One possible explanation for entropic gaze metrics is age. Schieber & Gilland (2008) found that H_t consistently decreased as secondary task load increased, and these effects were exacerbated for older (67–86 years old) versus younger (19-35 years old) drivers. Schieber & Gilland (2008) proposed that this could be explained by shortfalls in visual-spatial resources of older drivers. A combination of loading these resources with a secondary task, and the demands of visual scanning during driving, could result in diminished scanning complexity under the interpretation of Wickens' (2020) Multiple Resource Theory model. More recent research supports this notion, suggesting that age-related impairments of top-down attentional control can exacerbate the effects that secondary cognitive tasks have on H_t (Gazzaley et al, 2005; Shiferaw et al, 2019).

To investigate whether age-related impairments of top-down attentional control influence the effect of N-back, an additional model parameter β_A specifying the effect of age and its interaction with N-back was included for models of H_s and H_t. For H_s, the model predicted that age accounts for 9.9% of the between-participants heterogeneity in the causal effect of N-back (see Figure 14). A closer look at Figure 15 highlights that younger than average drivers still had decreases in gaze dispersion during N-back, although they were slightly smaller versus older than average drivers.

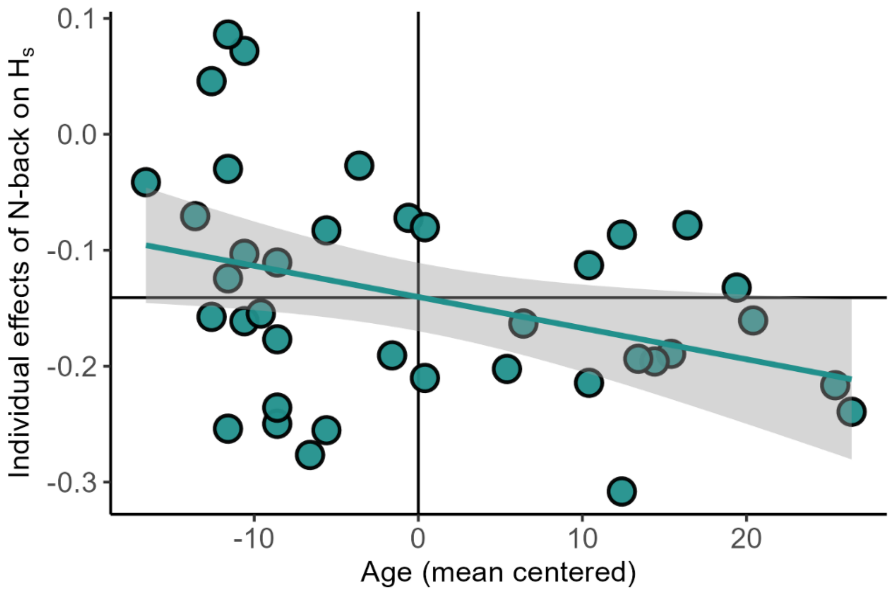

Figure 14: Individual effects of N-back on H_s plotted against mean centred age. X axis vertical line denotes mean age, y axis horizontal line denotes the average effect of N-back. All people in the sample show decreases in gaze dispersion due to N-back. However, this effect is more prominent for older than average people.

As for H_t, the model predicts that driver age accounts for 19% of between-participants heterogeneity in the causal effect of N-back. This suggests that age had a larger impact on how N-back effected H_t in comparison to how it impacted H_s. Furthermore, how the between-participants variance manifested was different. Younger than average drivers tended to show null effects or even small reversals of the average causal effect, whereas older drivers observed large reductions in H_t attributed to the effect of the N-back task (see Figure 15).

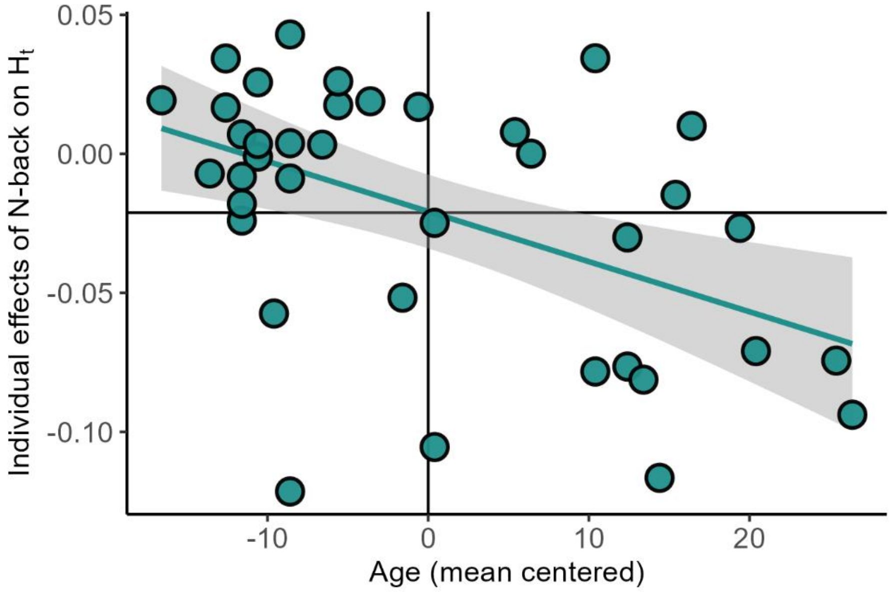

Figure 15: Individual effects of N-back H_t on plotted again mean centred age. X axis vertical line denotes mean age, y axis horizontal line denotes the average effect of N-back. Young than average people appear to have almost no effect of N-back on H_t, with even some slight reversals. Conversely, older than average people tend to have stronger than average effects of N-back on H_t.

4 Discussion

The aim of this study was to investigate whether gaze metrics based on Information Theory could be used to estimate mental workload during hands-off Level 2 automated driving. Drivers had to monitor a road environment before taking over during critical and non-critical situations. The data presented in this manuscript focused on whether changes in eye movements during automated driving were associated with changes in mental workload. The observed data revealed that H_s was a reliable indicator of mental workload; the model predicted that 92% of the population would have decreases in H_s when completing the N-back task. Despite this, there was substantial variability in the size of the effect, with some people predicted to exhibit effects more than double the size of the average causal effect. Conversely, in contrast to previous work (Schieber & Gilland, 2008) H_t was found to be much less reliable for detecting mental workload. Although the model predicted average reductions in gaze transition complexity for high workload conditions, only 66% of the population would exhibit similar decreases in H_t during N-back. Participant age appeared to be a strong predictor for how N-back influenced gaze entropic measures, accounting for 9.5% and 19% of the between-participants heterogeneity in the causal effect of N-back on H_s and H_t, respectively.

The current manuscript supports previous work that gaze dispersion reduces when mental workload increases (Reimer et al, 2009; 2010; Louw & Merat, 2017; Wilkie et al, 2019). The analysis also aligns with previous work that gaze complexity decreases under high mental workload (Schieber & Gilland, 2008). However, the analytic approach employed in this paper improves upon previous work by explicitly modelling and quantifying a key assumption of human behaviour; that people are inherently heterogenous. To build theories of psychological processes that inform eye movements during partial and conditional automated driving, it is advisable to take into account the heterogeneity of the sample (Bogler et al, 2019). This is especially vital when heterogeneity is sufficient such that null effects or reversals are observed in the data (Bogler et al, 2019). In the current manuscript, this was observed for H_t as a function of N-back. Under the assumption that this variance is not due to poor experimental control, such theories will need to include subpopulations that differ in causal processes. One previous attempt at this approach was by Reimer et al (2009) who considered the pattern of visual tunneling under high mental workload in the population by computing change scores from pre-task periods of gaze dispersion for each individual. Although this identifies whether individuals in the sample follow average trends, it does not generate a population distribution of the effects of mental workload on eye movements. Instead, the current manuscript constructed a population heterogeneity distribution implied by the models estimate of the population mean (μ) and standard deviation (σ) for each gaze entropic metric.

The effect of N-back on H_s and H_t differed as function of age, albeit in slightly different ways. For H_s, a majority of the sample showed reductions in the spatial distribution of gaze as a function of N-back; this reduction was weaker for younger than average participants. Conversely, for H_t there was no effect of N-back for the younger than average sample. There were even small *increases* in gaze complexity when completing the N-back. The older than average drivers showed a strong decrease in gaze transition complexity. It is important to note that age had minimal effects on H_s and H_t directly; rather, age influenced how much N-back affected gaze. In this sense, the current findings support previous work that report the lack of a direct effect of age on gaze centralization (Reimer et al, 2010; 2012). One explanation for the indirect effect of age on gaze entropy could be due to a healthy age-related cognitive decline. Top-down modulation underlies selective attention by suppressing the neural activity associated with the interference of task irrelevant representations (Gazzaley et al, 2005; Ploner et al, 2001). In the context of gaze control, top-down modulation also allows for efficient sampling of the environment by overriding bottom-up input, thus allowing drivers to efficiently monitor dynamic scenes (Shiferaw et al, 2019). However, research has found that older populations struggle to suppress task irrelevant information (Gazzaley et al, 2005). Consequently, this combination leads to a reduction in scanning complexity due to the interference of the N-back task, in combination with already weakened top-down selective attention processes of older than average participants.

In terms of their implications, these results can provide DMS designers with some important principles for using the correct metrics for detecting mental workload. A key aspect to be considered is that driver demographics should be taken into account when using DMS to establish driver state in vehicles. This manuscript clearly demonstrates that age influenced the extent to which N-back changed gaze-based metrics. As such, if DMS were to use H_s as an indicator of mental workload, differing thresholds might be necessary for drivers of different ages. For example, it might be necessary for a smaller threshold in the reduction of spatial dispersion for younger drivers as their gaze might be less effected by N-back, even though they might be experiencing high levels of mental workload, which could, in turn, impair their takeover performance. Another element to for DMS engineers to consider is which parameter of the gaze metric distribution should be used to establish a change in driver state. The current state of the art assumes that changes in central tendency should be used (e.g. a change in mean H_t establishes that N-back induces high mental workload). However, the current findings suggest that changes in *variance* may be more reliable. Increases in the trial-by-trial variance of H_t were predicted to be found in 76% of the population during high mental workload; only 66% of the population were predicted to follow average trends regarding a change in mean H_t. This suggests that changes in the variance of gaze complexity were more reliable than changes in the mean. High trial-by-trial variance during N-back suggests that drivers had frequent fluctuations in the complexity of their gaze from one trial to the next. Rather than finding an optimal level of gaze transitions that were suitable for all trials, the randomness of the transitions changed frequently. It has been proposed that the motor controls involved in eye movements aim to optimize inference (Parr & Friston, 2017) which implies that there are optimal levels of H_t to sample the environment efficiently (Shiferaw et al, 2019). Hence the results in the current manuscript suggest that high mental workload disrupts this eye movement optimization, resulting in variable, inefficient monitoring of the driving environment. The utilization of variance as an indicator for mental workload supports results from research within the visual distraction domain. These show, for example, that presentation of information by certain in-vehicle information systems reduces variations in fixation durations, supporting more efficient information processing (Horrey & Wickens, 2007; Kujala & Saarilouma, 2011). A similar suggestion is made here; without N-back trial-by-trial variance is small suggesting drivers establish and optimal H_t that allows them to efficiently sample the road. As mental workload increases, so does the variance in H_t, which is proposed as an indicator for visual scanning inefficiency. These findings suggest more research is needed to understand whether different parameters of response distributions can be used as indicators of mental workload.

Another interesting result from this study was the effect of lead vehicle presence. There was a small but consistent decrease in the spatial distribution of gaze for trials with lead vehicles. This supports previous research that drivers reduce the spread of their gaze and reallocate attention towards lead vehicles (Crundall et al, 2004). A key difference is that Crundall et al (2004) observed reductions in gaze dispersion only when instructing drivers to follow a lead vehicle during manual driving. Conversely, participants in the current study were instructed to monitor the entire road environment for hazards. Despite this request, the lead vehicle was clearly a salient object within the road environment and thus likely attracted drivers' attention. This may pose a problem for DMS in the real world, given that gaze dispersion has been shown to decrease in the presence of a lead vehicle, irrespective of increasing mental workload. Therefore, DMS will need to ensure that it can distinguish between drivers attending towards vehicles on the road ahead, and those under increased mental workload. It should be noted that the average reduction in gaze dispersion was much smaller for lead vehicles versus N-back conditions, however this still will not disentangle drivers who had smaller reductions in gaze dispersion during N-back conditions.

One limitation of the current work is that these model predictions need to be validated on a wider range of datasets. A statistical model is only as good as the data used to fit it. Whilst age ranges and gender balance were representative in the current sample, they mostly represented white, British drivers in the north of England. As such, whether their behaviours translate well to drivers from different cultures remains to be seen. Another limitation with the current work is the use of a Gaussian distribution as the likelihood for the modelling. Whilst the data were approximated by a Gaussian distribution, and the posterior predictive checks appear to fit the data well, normalized H_s and H_t are technically continuous variables bounded between 0 and 1. Conversely, any value is possible for a Gaussian distribution. The Beta distribution is a candidate that might be better suited for modelling these types of variables (Paolino, 2001; Ferrari & Cribari-Neto, 2004). Whilst a comparison of Gaussian and Beta likelihoods on clinical data highlighted that the estimates were very similar (Kurz, 2023) the Beta distribution is a better conceptual fit and produced slightly more precise estimates. Future research may compare these methods to investigate any differences in the context of gaze metrics.

In conclusion, Information Theoretic eye-based metrics have shown some promise in identifying increased mental workload in drivers engaging in an N-back task during hands-off Level 2 automated driving. Both H_s (Pillai et al, 2022) and H_t (Schieber & Gilland, 2008) were found to decrease as a function of increasing task load. However, the current research suggests that this assessment is incomplete. Whilst the average trends are consistent with previous research, there is substantial variance in how eye movements change as a function of task load across a population. For future DMS systems that apply to a multitude of drivers, this variance needs to be properly measured and quantified. One potential source of this heterogeneity is age, and thus DMS designers should consider how their input metrics are influenced by differing demographic variables.

References

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601.

Bottos, S., & Balasingam, B. (2020). Tracking the progression of reading using eye-gaze point measurements and hidden markov models. *IEEE Transactions on Instrumentation and Measurement*, *69*(10), 7857-7868.

Brookhuis, K. A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics*, *36*(9), 1099-1110.

Brookhuis, K. A., & de Waard, D. (2000). Assessment of drivers' workload: Performance and subjective and physiological indexes. In *Stress, workload, and fatigue* (pp. 321-333). CRC press.

Bruggen, A. (2015). An empirical investigation of the relationship between workload and performance. *Management Decision*, *53*(10), 2377-2389.

Bürkner, P. C. (2017). Advanced Bayesian multilevel modeling with the R package brms. *arXiv preprint arXiv:1705.11123*.

Carsten, O., Lai, F. C., Barnard, Y., Jamson, A. H., & Merat, N. (2012). Control task substitution in semiautomated driving: Does it matter what aspects are automated?. *Human factors*, *54*(5), 747-761.

Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving. *Transportation research part F: traffic psychology and behaviour*, *89*, 109-128.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

Crundall, D., Shenton, C., & Underwood, G. (2004). Eye movements during intentional car following. *Perception*, *33*(8), 975-986.

da Silva, F. P. (2014). Mental workload, task demand and driving performance: what relation?. *Procedia-Social and Behavioral Sciences*, *162*, 310-319.

De Waard, D., & Brookhuis, K. A. (1996). The measurement of drivers' mental workload.

De Winter, J. C., Happee, R., Martens, M. H., & Stanton, N. A. (2014). Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation research part F: traffic psychology and behaviour*, *27*, 196-217.

Di Stasi, L. L., Díaz-Piedra, C., Ruiz-Rabelo, J. F., Rieiro, H., Carrion, J. M. S., & Catena, A. (2017). Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based indices. *Applied Ergonomics*, *65*, 168-174.

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human factors*, *28*(4), 421-438.

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human factors*, *59*(5), 734-764.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, *31*(7), 799-815.

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident analysis & prevention*, *37*(3), 461-472.

Gazzaley, A., Cooney, J. W., Rissman, J., & D'esposito, M. (2005). Top-down suppression deficit underlies working memory impairment in normal aging. *Nature neuroscience*, *8*(10), 1298-1300.

Gold, C., Berisha, I., & Bengler, K. (2015, September). Utilization of drivetime–performing non-driving related tasks while driving highly automated. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, No. 1, pp. 1666-1670). Sage CA: Los Angeles, CA: SAGE Publications.

Gold, C., Damböck, D., Bengler, K., & Lorenz, L. (2013). Partially automated driving as a fallback level of high automation. In *6. tagung fahrerassistenzsysteme*.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*, *7*(11), 498-504.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in cognitive sciences*, *21*(1), 15-23.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Horrey, W. J., & Wickens, C. D. (2007). In-vehicle glance duration: distributions, tails, and model of crash risk. *Transportation research record*, *2018*(1), 22-28.

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., ... & Villalobos, N. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception (TAP)*, *13*(1), 1-20.

Krejtz, K., Szmidt, T., Duchowski, A. T., & Krejtz, I. (2014, March). Entropy-based statistical analysis of eye movement transitions. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 159-166).

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Kujala, T., & Saariluoma, P. (2011). Effects of menu structure and touch screen scrolling style on the variability of glance durations during in-vehicle visual search tasks. *Ergonomics*, *54*(8), 716-732.

Kurz, S. (2023, June, 25). Causal inference with beta regression. (https://solomonkurz.netlify.app/blog/2023-06-25-causal-inference-with-beta-regression/#ref-paolino2001maximum)

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in psychology*, *6*, 1171.

Louw, T., & Merat, N. (2017). Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation. *Transportation Research Part C: Emerging Technologies*, *76*, 35-50.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

Mehler, B., Reimer, B., & Dusek, J. A. (2011). MIT AgeLab delayed digit recall task (n-back). *Cambridge, MA: Massachusetts Institute of Technology*, *17*.

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015, September). Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. In *2015 IEEE 18th international conference on intelligent transportation systems* (pp. 2458-2464). IEEE.

Mole, C., Pekkanen, J., Sheppard, W., Louw, T., Romano, R., Merat, N., ... & Wilkie, R. (2020). Predicting takeover response to silent automated vehicle failures. *Plos one*, *15*(11), e0242825.

Öztürk, İ., Merat, N., Rowe, R., & Fotios, S. (2023). The effect of cognitive load on Detection-Response Task (DRT) performance during day-and night-time driving: A driving simulator study with young and older drivers. *Transportation research part F: traffic psychology and behaviour*, *97*, 155-169.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, *9*(4), 325-346.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific reports*, *7*(1), 14678.

Pillai, P., Balasingam, B., Kim, Y. H., Lee, C., & Biondi, F. (2022). Eye-gaze metrics for cognitive load detection on a driving simulator. *IEEE/ASME Transactions on Mechatronics*, *27*(4), 2134-2141.

Ploner, C. J., Ostendorf, F., Brandt, S. A., Gaymard, B. M., Rivaud-Péchoux, S., Ploner, M., ... & Pierrot-Deseilligny, C. (2001). Behavioural relevance modulates access to spatial working memory in humans. *European Journal of Neuroscience*, *13*(2), 357-363.

Radlmayr, J., Fischer, F. M., & Bengler, K. (2019). The influence of non-driving related tasks on driver availability in the context of conditionally automated driving. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20* (pp. 295-304). Springer International Publishing.

Raptis, G. E., Fidas, C. A., & Avouris, N. M. (2017, May). On implicit elicitation of cognitive strategies using gaze transition entropies in pattern recognition tasks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1993-2000).

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record*, *2138*(1), 13-19.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2010, September). The impact of systematic variation of cognitive demand on drivers' visual attention across multiple age groups. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 24, pp. 2052-2055). Sage CA: Los Angeles, CA: SAGE Publications.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups. *Human factors*, *54*(3), 454-468.

Schieber, F., & Gilland, J. (2008, September). Visual entropy metric reveals differences in drivers' eye gaze complexity across variations in age and subsidiary task load. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, No. 23, pp. 1883-1887). Sage CA: Los Angeles, CA: SAGE Publications.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, *11*(9), 1141-1152.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shiferaw, B., Downey, L., & Crewther, D. (2019). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, *96*, 353-366.

Sodhi, M., Reimer, B., & Llamazares, I. (2002). Glance analysis of driver eye movements to evaluate distraction. *Behavior Research Methods, Instruments, & Computers*, *34*, 529-538.

Spratling, M. W. (2017). A predictive coding model of gaze shifts and the underlying neurophysiology. *Visual Cognition*, *25*(7-8), 770-801.

Standard, S. (2018). J3016B: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles-SAE International.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological review*, *124*(3), 267.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413-1432.

Velichkovsky, B., Sprenger, A., & Unema, P. (1997). Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem". In *Human-Computer Interaction INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–18th July 1997, Sydney, Australia* (pp. 509-516). Springer US.

Wang, Y., Reimer, B., Dobres, J., & Mehler, B. (2014). The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transportation research part F: traffic psychology and behaviour*, *26*, 227-237.

Weiss, R. S., Remington, R., & Ellis, S. R. (1989). Sampling distributions of the entropy in visual scanning. *Behavior Research Methods, Instruments, & Computers*, *21*(3), 348-352.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159-177.

Wickens, C. D. (2020). Processing resources and attention. In *Multiple task performance* (pp. 3-34). CRC Press.

Wilkie, R., Mole, C., Giles, O., Merat, N., Romano, R., & Makkula, G. (2019, June). Cognitive load during automation affects gaze behaviours and transitions to manual steering control. In *Driving Assessment Conference* (Vol. 10, No. 2019). University of Iowa.

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human factors*, *44*(3), 365-375.

Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident analysis & prevention*, *92*, 230-239.