



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/212333/>

Version: Accepted Version

Article:

Goodridge, C.M., Gonçalves, R.C., Arabian, A. et al. (2024) Gaze entropy metrics for mental workload estimation are heterogenous during hands-off level 2 automation. *Accident Analysis & Prevention*, 202. 107560. ISSN: 0001-4575

<https://doi.org/10.1016/j.aap.2024.107560>

© 2024 Published by Elsevier Ltd. This is an author produced version of an article accepted for publication in *Accident Analysis & Prevention*. Uploaded in accordance with the publisher's self-archiving policy. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



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

3 Courtney M. Goodridge^{1, *}, Rafael C. Gonçaves¹, Ali Arabian¹, Anthony Horrobin¹, Albert
4 Solernou¹, Yee Thung Lee¹, Yee Mun Lee¹, Ruth Madigan¹, Natasha Merat¹

5 ¹Institute for Transport Studies, University of Leeds

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

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

17
18
19
20
21
22
23
24
25

26 **Abstract**

27 As the level of vehicle automation increases, drivers are more likely to engage in non-driving
28 related tasks which take their hands, eyes, and/or mind away from the driving task.
29 Consequently, there has been increased interest in creating Driver Monitoring Systems (DMS)
30 that are valid and reliable for detecting elements of driver state. Workload is one element of
31 driver state that has remained elusive within the literature. Whilst there has been promising
32 work in estimating mental workload using gaze-based metrics, the literature has placed too
33 much emphasis on point estimate differences. Whilst these are useful for establishing whether
34 effects exist, they ignore the inherent variability within individuals and between different
35 drivers. The current work builds on this by using a Bayesian distributional modelling approach
36 to quantify the within and between participants variability in Information Theoretical gaze
37 metrics. Drivers (N = 38) undertook two experimental drives in hands-off Level 2 automation
38 with their hands and feet away from operational controls. During both drives, their priority was
39 to monitor the road before a critical takeover. During one drive participants had to complete a
40 secondary cognitive task (2-back) during the hands-off Level 2 automation. Changes in
41 Stationary Gaze Entropy and Gaze Transition Entropy were assessed for conditions with and
42 without the 2-back to investigate whether consistent differences between workload conditions
43 could be found across the sample. Stationary Gaze Entropy proved a reliable indicator of
44 mental workload; 92% of the population were predicted to show a decrease when completing
45 2-back during hands-off Level 2 automated driving. Conversely, Gaze Transition Entropy
46 showed substantial heterogeneity; only 66% of the population were predicted to have similar
47 decreases. Furthermore, age was a strong predictor of the heterogeneity of the average causal
48 effect that high mental workload had on eye movements. These results indicate that, whilst
49 certain elements of Information Theoretic metrics can be used to estimate mental workload by
50 DMS, future research needs to focus on the heterogeneity of these processes. Understanding

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

55 Keywords: Distraction, workload, monitoring, heterogeneity, automation, entropy

56

57

58

59

60

61

62

63

64

65

66

67

68

69 1 Introduction

70 The influx of automated systems in road vehicles has generated increased interest in the
71 development of Driver Monitoring Systems (DMS). DMS refers to a collection of sensors that
72 aim to detect whether a driver is attentive, alert, or engaged. Not only are drivers more likely
73 to engage in non-driving related tasks (NDRTs) as vehicles transform from manual to partial
74 driving automation (Carsten et al, 2012), but in Level 3 automation drivers are allowed to
75 actively engage in NDRTs (SAE, 2018). This may take their hands off the wheel and eyes and
76 mind away from the main driving task. As such, a large body of research has aimed to measure
77 the internal states of drivers whilst using partial or conditionally automated vehicles, and how
78 these states might change in response to NDRTs. One elusive, yet extremely relevant, driver
79 state for informing driver readiness is workload. Workload is a general term that can be defined
80 as the demand or difficulty that is placed upon a driver (De Waard, 1996; da Silva, 2014; Fuller,
81 2005; De Winter et al, 2014). *Mental workload* is more specific and has been defined as the
82 proportion of information processing for a given task relative to an individual's processing
83 capacity (Brookhuis & De Waard, 1993; 2000; da Silva, 2014). It should also be noted that the
84 terms *cognitive distraction* and *cognitive load* are often used interchangeably when researchers
85 manipulate the cognitive demand of drivers. However, there is a distinct conceptual difference;
86 the former referring to the general removal of attention away from the driving task toward a
87 secondary task, and the latter referring to the quantity of the cognitive resource demanded by
88 the secondary task (Engström et al, 2017). A key aspect of mental workload is that drivers have
89 a limited pool of cognitive resources (Wickens, 2002). Underload from the monotony of
90 monitoring autonomous systems can result in decreased vigilance (Young & Stanton, 2002)
91 whereas overload may occur if a driver is engaging in an NDRT and can result in sub-optimal
92 takeover performance (Gold et al, 2015; Zeeb et al, 2016). To ensure that a driver is ready to
93 resume control, they should ideally have moderate workload levels to reduce the likelihood of

94 safety-critical situations (Bruggen, 2015). Hence one goal of DMS development has been to
95 identify valid and reliable indicators of mental workload to monitor the driver during automated
96 driving. Therefore, a specific aim of this manuscript was to investigate a family of gaze-based
97 metrics that have shown potential in estimating mental workload in human drivers.

98 The dispersion of gaze has been a useful metric for measuring mental workload during manual
99 and automated driving. Gaze dispersion is often measured as the standard deviation of raw gaze
100 coordinates in the horizontal and vertical dimensions (Sodhi et al, 2002). During manual
101 driving, the standard deviation of horizontal gaze reduces when the workload of the driver is
102 increased with a secondary cognitively loading task; this phenomenon is known as visual
103 tunneling (Reimer, 2009; Reimer et al, 2010; Wang et al, 2014). Similar effects have been
104 observed when performing a cognitive loading secondary task during automated driving
105 (Radlmayr et al, 2019; Wilkie et al, 2019). The sensitivity of raw gaze dispersion for detecting
106 mental workload has proven to be a robust measure for driver monitoring systems. However,
107 one limitation of this approach is that it does not account for the predictive nature of eye
108 movements. Established accounts of gaze control focus on the where (spatial distribution) and
109 the when (temporal sequence) of gaze, relative to task demands (Shiferaw et al, 2019). This
110 can be interpreted as being driven by bottom-up or top-down processes (Shiferaw et al, 2019).
111 Bottom-up processes refer to attention that is guided by stimulus saliency of a particular image
112 or visual scene; top-down processes refer to attention that is guided by memory-based
113 knowledge and/or behavioural requirements, originating from internal visual and cognitive
114 systems (Henderson, 2003; Itti & Koch, 2001). In this sense, bottom-up processing uses “lower
115 level” input (i.e., stimulus information) to modify “higher-level” representations (i.e.,
116 integrated information in the brain), whereas top-down processing uses higher-level
117 representations to produce or modify lower-level information (Palmer, 1999; Rauss & Pourtois,
118 2013). However, a growing body of literature has proposed that gaze control is a system of

119 spatial prediction (Henderson, 2017; Talter et al, 2017). Hence fixation locations are not merely
120 instructed by top-down and bottom-up influences, but their relative contributions towards
121 prediction and error correction when constructing an internal representation of a visual scene
122 (Parr & Friston 2017; Spratling et al, 2017; Shiferaw et al, 2019). The brain aims to minimize
123 error between sensory information and the internal state (Clark et al, 2013). Hence via a
124 combination of bottom-up and top-down processes, gaze control aims to optimize visual
125 sampling in order to make better predictions regarding the location of subsequent fixations
126 (Parr & Friston, 2017; Spratling et al, 2017). Considering the mechanisms involved in gaze
127 control, it can be argued that measuring differences in visual scanning behaviour during
128 varying stages of driving may provide information on changes in the underlying processes that
129 are influenced by increased workload (Shiferaw et al, 2019). Information Theoretic concepts
130 such as entropy are one such method, which focus on using gaze transitions to estimate internal
131 states.

132 Gaze entropy is an eye tracking metric that has shown promise for estimating mental workload
133 and refers to the application of Information Theory to gaze data (Shiferaw et al, 2019). Within
134 the field of Information Theory, entropy refers to the average amount of information or
135 uncertainty for a given choice (Shannon, 1948). For a system with discrete processes, the two
136 primary components are the source and output; the source being the total number of states that
137 a given output can take. When applied to gaze data, there is an assumption that saccadic
138 movements that produce fixations are outputs from a gaze control system that predicts the
139 spatial locations of proceeding fixations (Shiferaw et al, 2019). The visual field represents all
140 possible state spaces where a fixation could be located. To calculate the entropy of gaze
141 fixations, fixation coordinates are divided into discrete spatial bins to generate probability
142 distributions of a given fixation being within a given location (Shiferaw et al, 2019). The
143 entropy value thus represents the predictability of a fixation location; a higher uncertainty (or

144 entropy) represents a higher dispersion of gaze for a particular viewing period (Holmqvist et
145 al, 2011). This is known as *Stationary Gaze Entropy* (H_s). Another assumption is that
146 subsequent fixations are better predicted by current fixations via *conditional* probability rather
147 than only *total* probability (Weiss et al, 1989; Shiferaw et al, 2019). Therefore, this provides a
148 measure of predictability of visual scanning patterns by considering the order of fixations; this
149 is known as *Gaze Transition Entropy* (H_t). Higher H_t is indicative of less structured, more
150 random scanning patterns (Shiferaw et al, 2019). Because organisms use eye movements to
151 optimize inference through motor action sequences (Parr & Friston, 2017), it has been proposed
152 that there is an optimal range of H_t to efficiently sample information within the visual scene.
153 Optimal H_t is an ideal level of complexity that balances modulation from underlying bottom-
154 up influences with top-down prediction (Shiferaw et al, 2019). If there is an optimal range of
155 H_t then *increased* H_t may reflect top-down interference whereby there is modulation of gaze
156 beyond the requirements of a given task. This can manifest as highly erratic, random visual
157 scanning for an otherwise simple road environment that contains few elements. For example,
158 a car following situation may require fairly structured gaze transitions between safety critical
159 locations (side mirrors, read-view mirror, forward headway) and thus unpredictable, random
160 transitions would be beyond requirements and less efficient for the task in hand. Conversely,
161 lower than optimal H_t can result in insufficient top-down modulation thus producing
162 insufficient visual scanning and exploration resulting in a driver potentially not attending to
163 objects within the scene such as vehicles entering the ego-vehicles lane, or pedestrians waiting
164 to cross. Whilst H_t may change as a function of more visually demanding tasks or visual scenes,
165 given an environment where these factors are experimentally controlled, H_t may change as a
166 function of top-down engagement (Shiferaw et al, 2019).

167 H_s and H_t provide a quantitative assessment of visual scanning in naturalistic environments
168 and thus have been proposed as measures that can estimate mental workload in drivers. Testing

169 the reliability and validity of gaze entropic metrics has largely been conducted within the
170 domain of manual driving. Schieber & Gilland (2008) found reductions in H_t as a function of
171 secondary task load difficulty; this was further exacerbated for older drivers. The combination
172 of older drivers having reduced visual-spatial processing resources alongside the increased
173 demands of the secondary task resulted in this interaction effect. Schieber & Gilland (2008)
174 proposed that metrics based on Information Theory held significant potential for monitoring
175 driver behaviour as H_t systematically changed as a function of increased mental workload.
176 Pillai et al (2022) implemented a similar design to investigate whether gaze entropy
177 differentiated varying levels of cognitive load during manual driving. By calculating the signal-
178 to-noise ratio (SNR), Pillai et al (2022) found that H_s reliably differentiated between a control
179 task (normal driving and a detection response task) and 2-back, control and 0-back, and 0-back
180 and 2-back conditions. Conversely, H_t could not reliably distinguish between any of these
181 cognitive load comparisons. This suggests that it was the predictability of the dispersion of
182 gaze, rather than gaze transitions, that was useful for estimating mental workload. One of the
183 only experiments to study cognitive load estimation using gaze entropy during automated
184 driving was conducted by Chen et al (2022). They investigated whether H_s changed as a
185 function of automation level (SAE L0, L1, and L2). 3-dimensional H_s (applying the Shannon
186 (1948) equation to coordinates in a 3-dimensional plane) negatively correlated with subjective
187 workload during visual, auditory, or multi-modality cognitive tasks. This is indicative of gaze
188 dispersion decreasing as a function of increased subjective workload, and thus supports similar
189 findings of visual tunneling when cognitively loaded (Radlmayr et al, 2019; Reimer, 2009;
190 Reimer et al, 2010; Wang et al, 2014; Wilkie et al, 2019). Chen et al (2022) concluded that H_s
191 could be a valid indicator for visual and auditory task distractions within driver monitoring
192 systems during partial automation.

193 Despite evidence that gaze entropy measures can be useful for estimating mental workload,
194 there are some limitations to this work. Chen et al (2022) utilized a desktop computer simulator
195 where the keyboard was used for steering and pedal operations. There was also no simulated
196 traffic or road; just a highly artificial virtual environment. Not only is this a poor replication of
197 real driving, but the lack of stimuli within the visual scene may have produced insufficient
198 bottom-up saliency. There was also no control condition without a secondary task, thus not
199 allowing for any comparison of gaze entropy under normal workload situations. A wider
200 limitation of the literature is the lack of investigation into the variation both within and between
201 individuals. A metric that estimates mental workload must be valid (i.e., the metric
202 systematically varies with mental workload) but it must also be reliable (i.e., the metric
203 systematically changes in similar ways for a given population) if it is to be used in DMS within
204 a wider population. Therefore, understanding how H_s and H_t vary is vitally important. Whilst
205 mean differences are theoretically useful for establishing the existence of effects, they only
206 existence in an abstract sense (Mole et al, 2020). To make applied predictions that relate to the
207 wider population, it is vital to model and understand how a sample varies. Schieber & Gilland
208 (2008) reported no indices of variance in H_t , thus providing no indication as to how variable
209 H_t was when drivers were under high mental workload. Chen et al (2022) reported large
210 individual differences in the difficulty of the spatial N-back task which may have influenced
211 subjective ratings of mental workload alongside eye tracking metrics. However, they did not
212 formally model these differences, or investigate whether specific individual characteristics
213 predicted this variation. Finally, Pillai et al (2022) investigated the effects of gaze entropy by
214 calculating the signal to noise ratio (SNR); a lower SNR indicates that two means are more
215 similar. Not only is this metric focused on mean differences but averages of gaze entropy in
216 different conditions are weighted by variance across several participants. Whilst this accounts
217 for variation in entropy, it treats all individual differences as noise. Whilst some individual

218 variance is undoubtedly attributed to noise in eye tracking measurement (Bottos & Balasingam,
219 2020; Velichkovsky et al, 1997), it is possible that individual differences could vary as function
220 of theoretically useful variables (e.g., age, driving experience).

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

233 **2 Material and methods**

234 **2.1 Participants**

235 41 participants were recruited from a university participant pool and took part in the experiment
236 however three had to be removed before data analysis as they either did not follow experimental
237 instructions, or eye tracking data were not captured. The remaining 38 participants (16 females,
238 22 males, mean age = 38.81, range = 22-65) all had normal or corrected to normal vision. All
239 participants had a valid UK driving licence (mean number of years = 17.8, range = 4-43) and
240 were regular drivers (mean annual kilometres = 15055, range 8046-32186).

241 2.2 Apparatus and materials

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



259 Figure 1: University of Leeds Driving Simulator

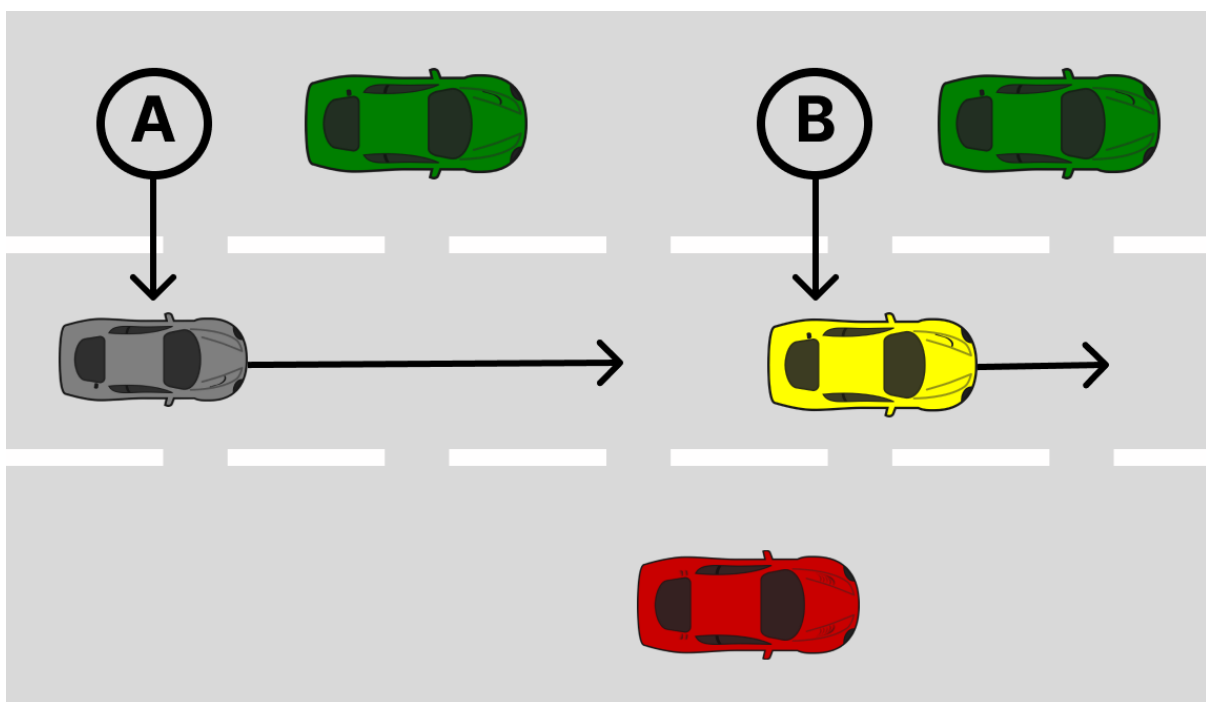
260 2.3 Design

261 A 2 x 2 Repeated Measures design was used in this study. The two within-participant factors
262 were event criticality and mental workload. Event criticality was manipulated by changing the
263 time to collision at the onset of a lead vehicle braking (TTC) after a period of hands-off Level

264 2 automated driving. The aim of manipulating this variable was to create two levels of
265 criticality: a “less severe” level (TTC = 5 s) that allowed participants to successfully take over
266 without crashing, and a “severe” level that could lead to a crash if the participant was not
267 monitoring the road correctly (TTC = 3 s). These values were chosen based on previous studies
268 that have demonstrated that a 3 s TTC produces highly critical events, whilst a 5 s TTC provides
269 sufficient time for takeovers (Gold et al, 2013; Mok et al, 2015; Louw & Merat, 2017). The
270 second within-participants factor that was manipulated was mental workload. This was
271 manipulated over two levels; a no-load condition and a high mental workload condition where
272 participants had to complete a secondary task during the automated driving sections. To induce
273 cognitive load, participants completed a verbal response delayed digit recall task (N-back)
274 (Mehler et al, 2011) during the automated driving sections. The specific N-back used in the
275 current investigation was a 2-back condition. This task was chosen because it is highly
276 controlled, non-visual, and has been consistently shown to increase the workload of drivers
277 during manual (Reimer, 2009; Reimer et al, 2010; Wang et al, 2014) and automated driving
278 (Radlmayr et al, 2019; Wilkie et al, 2019).

279 The experiment consisted of two drives for each participant. During one drive participants
280 completed an N-back throughout the automated period; during the other drive there was no
281 secondary task. The order of N-back was counter-balanced across participants. Each drive
282 lasted approximately 35 minutes and all participants drove on the same 3-lane UK motorway.
283 Each drive consisted of 10 discrete events, each consisting of 30 s of manual driving followed
284 by approximately 2 minutes of automated driving. After 2 minutes of automated driving, a
285 takeover request (TOR) was delivered. Four of these events were critical: two with a TTC of 3
286 s, two with a TTC of 5 s. For 3 s TTCs, the lead vehicle braked suddenly and decelerated at a
287 rate of 5.55 m/s^2 , whereas for the 5 s event, the lead vehicle decelerated at 2 m/s^2 . Decelerations
288 began as soon as the takeover request (TOR) was triggered. The remaining six events were

289 non-critical; two involved no lead vehicle, and the remaining four involved a lead vehicle that
290 did not decelerate once the TOR was triggered. Lead vehicles appeared in front of the ego
291 vehicle shortly before the automation was engaged. They entered the middle lane from the left-
292 hand lane and participants were instructed to allow the lead vehicle to pull in front. Once in the
293 middle lane, lead vehicles matched the ego-vehicle's speed at a distance of 25 m during
294 automation. Participants drove in the middle lane, with ambient traffic flow in the left and right
295 lanes. Once the lead vehicle was present, the automated system engaged.

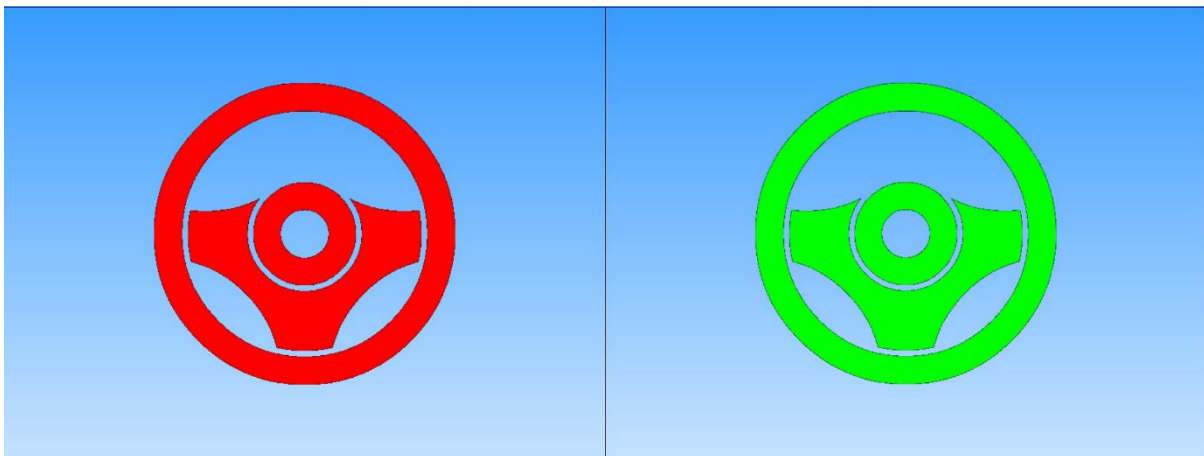


296 *Figure 2: Schematic representation of an event. (A) represents the ego vehicle and (B)*
297 *represents the lead vehicle. Lead vehicles entered from the left lane and matched the ego*
298 *vehicle's speed at a distance of 25 m. Following 2 minutes of automated driving, for critical*
299 *trials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

301 2.4 Procedure

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

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



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

316 For experimental drives, participants were instructed to enter the motorway and position
317 themselves in the centre of the middle lane and maintain a speed of 70 MPH. After
318 approximately 30 s of manual driving the automated system engaged automatically. This was

319 indicated by a short auditory tone and the shifting of the steering wheel icon from grey (manual
320 mode) to green (automation engaged) (see Figure 3). Once in automated driving mode,
321 participants were instructed to take their hands off the wheel and feet away from the pedals and
322 to monitor the road environment for any potential hazards. After approximately 2 minutes of
323 automated driving, a TOR was delivered. The TOR was characterised by an auditory tone and
324 the steering icon flashing red within the instrument cluster. Participants were instructed to take
325 over once the TOR had been issued; this could be done by any steering input over 2°, pressing
326 any of the pedals, or pressing a micro-switch button strapped to the steering wheel. During
327 piloting it became apparent that some drivers wanted to deactivate the automated system
328 without altering vehicular controls (akin to deactivating an adaptive cruise control system with
329 a button press). Hence the option for transitioning to manual driving mode via a microswitch
330 was included. If the driver of the ego-vehicle did not respond within 10 seconds, the automation
331 would disengage by itself. Following the takeover, the participant engaged in 30 s of manual
332 driving before the automated system engaged once more. If the driver exited the middle lane
333 during takeovers, they were instructed to return as soon as possible. There were 10 discrete
334 events per drive and each drive lasted approximately 35 minutes. During one drive participants
335 completed an auditory-verbal N-back task when automation was engaged, which continued
336 until a TOR was given. Participants were instructed that a safe drive was their primary goal.
337 After each drive, participants filled out a NASA-TLX to collect data on subjective ratings of
338 workload. After the second experimental drive, participants completed post-drive
339 questionnaires (data from these questionnaires is not analysed or reported in this manuscript).

340 2.5 Statistical modelling

341 The main aim of this manuscript was to investigate changes in gaze entropic eye metrics during
342 the 2-minute automation period with and without N-back, and with and without a lead vehicle.
343 This includes critical and non-critical trials that included a lead vehicle. Thus, data relating to

344 the takeover and manual driving portions are not analysed within this manuscript. Data and
345 analysis code can be found in the following link
346 (https://github.com/courtneygoodridge/gaze_entropy_heterogenous).

347 2.5.1 Gaze entropy

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

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

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

355

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

356

357 When $p(i)$ is the stationary distribution of fixations, $p(i | j)$ is the probability of transitioning
358 to state j given being currently in state i , and $i \neq j$ excludes transitions that occur within the
359 same state space (Ellis & Stark, 1986). Fixations were split into spatial bins to apply the
360 equations. This is the primary method of discretisation in the literature (Di Stasi et al, 2017;
361 Krejtz et al, 2014; 2015, Raptis et al, 2017) and has been proposed as the superior method for
362 dynamic stimuli (Shiferaw et al, 2019). For interpretability, both H_s and H_t were normalized

363 by dividing by the maximum entropy, H_{max} . Maximum entropy is the logarithm (base 2) of all
364 state spaces and thus represents when distributional information is at a maximum. For example,
365 each fixation is equally spaced out within the visual scene, and each transition is completely
366 random (Shiferaw et al, 2019). As such, H_s and H_t range from 0-1 and represent the percentage
367 of maximum possible entropy.

368 2.5.2 Analytic approach

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

380 As highlighted in the Introduction, the motor coordination of eye movements aims to optimise
381 inference (Parr & Friston, et al 2017). This implies that there is an optimal level of H_t for
382 effective sampling of the visual scene whereby top-down processes modulate default bottom-
383 up activation (Shiferaw et al, 2019). Whilst increases or decreases in the μ of H_t can be
384 indicative of top-down interference or top-down modulation respectively (Shiferaw et al,
385 2019), the trial-by-trial variance within individuals can also be a crucial index for measuring
386 the efficiency of visual scanning. Under the assumption that the visual scene maintains an
387 ambient level of complexity, optimal H_t should be consistent within an individual. However,

388 if increased mental workload results in decreases in H_t via top-down modulation, it may also
389 affect how efficiently individuals are able to maintain optimal H_t from one trial to the next.
390 The idea that a change in *variance* can indicate a change in a driver's internal state is not new
391 within the driver monitoring and distraction literature. Horrey & Wickens (2007) proposed that
392 standard statistical methods that focus on mean differences (or other measures of central
393 tendency) are insufficient for measuring driver distraction, and that modelling large deviations
394 in attention can reveal infrequent lapses in visual sampling control; something that can be
395 missed when only focusing on averages. Kujala & Saarilouma (2011) found reductions in the
396 standard deviation of fixation durations for simpler in-vehicle information systems menu
397 designs, thus suggesting that the variance in fixations durations could be used to assess the
398 efficiency of visual search performance. It is thus proposed in this manuscript that a similar
399 effect might be present for H_t , when increasing mental workload. To assess whether there are
400 systematic changes in σ as a function of the predictor variables, the current analysis will apply
401 distributional models. Distributional models relax the assumption of consistent σ , and allow it
402 to be predicted by parameters as can be done when predicting μ (Bürkner, 2017).

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

411 2.5.2.1 Model development

412 The population mean, μ , of all the gaze-based metrics were modelled as the linear combination
 413 of an intercept (β_0), N-back (N , β_N), presence of a lead vehicle (L , β_L), and an interaction term
 414 between these variables (NL , β_{NL}). The N-back task was parameterised as $N \in \{0, 1\}$ where
 415 $N = 1$ corresponds to the presence of the N-back during hands-off Level 2 automation.
 416 Similarly, lead vehicle was parameterised as $L \in \{0, 1\}$ where $L = 1$ corresponds to the
 417 presence of a lead vehicle during automation. The standard deviation, σ , was independently
 418 modelled as a linear combination of an intercept (α_0), N-back (α_N), presence of a lead vehicle
 419 (α_L), and an interaction (α_{NL}). Because σ cannot be negative, the $\log(\sigma)$ was modelled. The
 420 distributional model structure was specified as follows:

421

422

$$Y_{ij} \sim N(\mu_{ij}, \sigma_{ij}) \quad (3)$$

$$\mu_{ij} = (\beta_0 + \beta_{0j}) + (\beta_N N_i + \beta_{Nj} N_i) + (\beta_L L_i) + (\beta_{NL} N L_i)$$

$$\log(\sigma_{ij}) = (\alpha_0 + \alpha_{0j}) + (\alpha_N N_i + \alpha_{Nj} N_i) + (\alpha_L L_i)$$

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

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

$$S_\beta = \begin{pmatrix} \sigma_{\beta_{0j}}^2 & \rho \sigma_{\beta_{Nj}} \sigma_{\beta_{0j}} \\ \rho \sigma_{\beta_{0j}} \sigma_{\beta_{Nj}} & \sigma_{\beta_{Nj}}^2 \end{pmatrix}$$

$$S_\alpha = \begin{pmatrix} \sigma_{\alpha_{0j}}^2 & \rho \sigma_{\alpha_{Nj}} \sigma_{\alpha_{0j}} \\ \rho \sigma_{\alpha_{0j}} \sigma_{\alpha_{Nj}} & \sigma_{\alpha_{Nj}}^2 \end{pmatrix}$$

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

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

$$Y_{ij} \sim N(\mu_{ij}, \sigma_{ij}) \quad (4)$$

$$\mu_{ij} = (\beta_0 + \beta_{0j}) + (\beta_N N_i + \beta_{Nj} N_i)$$

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

$$S_\beta = \begin{pmatrix} \sigma_{\beta_{0j}}^2 & \rho \sigma_{\beta_{Nj}} \sigma_{\beta_{0j}} \\ \rho \sigma_{\beta_{0j}} \sigma_{\beta_{Nj}} & \sigma_{\beta_{Nj}}^2 \end{pmatrix}$$

429 Where Y denotes the response variable, i specifies the condition of each variable, j specifies
 430 the participant, and S_β is a matrix corresponding to the variance or covariance parameters.

431 2.5.2.2 Model fitting

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

440 Using a Bayesian approach, each parameter has an associated probability distribution which
441 quantifies the level of uncertainty, conditioned on the data. In this manuscript, posterior
442 distributions of parameters are described by their mean and a 95% Credible Interval (CI)
443 whereby there is a 95% probability that the true parameter value will fall; values inside this
444 density have higher credibility than those outside it (Kruschke, 2014). The *probability of*
445 *direction* (*pd*) is also reported for each fixed effect parameter. The *pd* is defined as the
446 probability that an effect is positive or negative (Makowski et al, 2019). The *pd* is strongly
447 correlated with the Frequentist p value and thus can be used as an index of an effect's existence. It
448 should be highlighted that the term "*existence*" merely relates to the consistency of an effect in one
449 direction; it makes no assumptions regarding the size, importance, relevance, or meaning of the effect
450 (Makowski et al, 2019). Hence, the reader is discouraged in making dichotomous decisions when
451 understanding whether there is an effect. Rather, they should use a combination of the *pd*, the
452 mean parameter values, and the 95% credible intervals to assess the size, direction, and
453 uncertainty of the effects.

454 **3 Results**

455 3.1 Subjective measures

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

461

462

463

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

Fixed effects		
Dependent variable:		
	<i>Mental demand</i>	<i>pd</i>
β_0	38.994 (32.656, 45.257)	100%
β_N	39.711 (32.057, 47.369)	100%
Participants	38	
Observations	76	

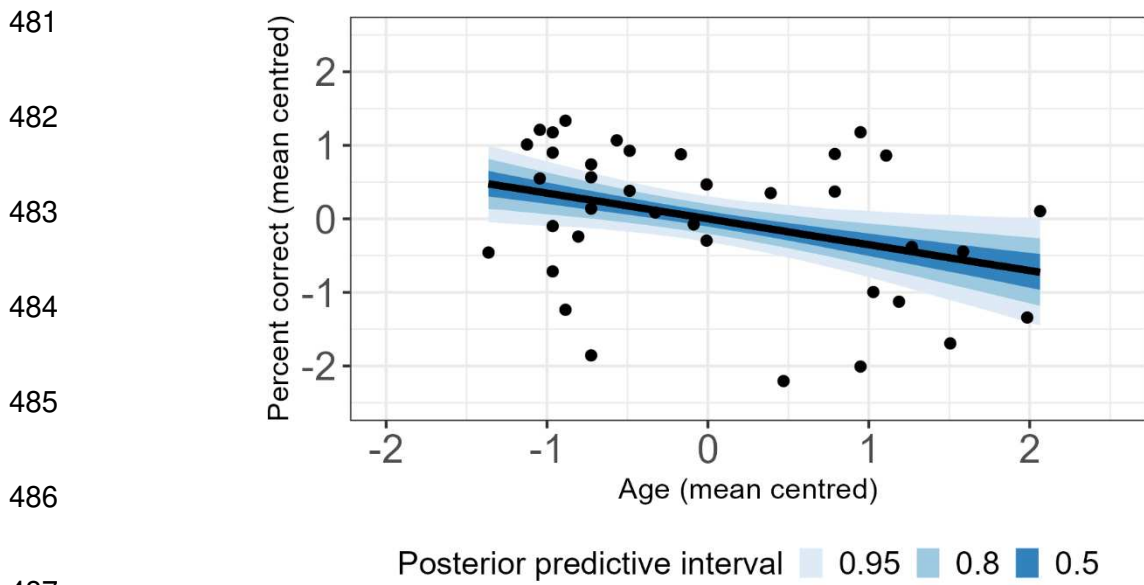
466

467 3.2 N-back performance

468 Performance data for the N-back task was only available for 37 out of 38 participants due to
 469 data loss. The average performance (percentage of correct scores) was reasonably high and
 470 homogenous across the sample ($M = 70.77$, $SD = 15.13$) however the high and low scores were
 471 quite different (range = 37.38–90.97). Previous research in manual driving had found that
 472 younger drivers had significantly better 2-back performance in comparison to older drivers
 473 (Öztürk et al, 2023). To investigate this, a univariate Bayesian correlation model was fitted on
 474 the standardised values of age and performance. The results indicate a negative correlation of
 475 $-.349$ (95% CI: $-.666$, $-.037$, $pd = 98.50\%$) suggesting that older drivers tended to have worse
 476 N-back performance. This medium effect size is slightly lower than what was been found in
 477 manual driving (Öztürk et al, 2023) although the average correlation did highlight a lot of
 478 variability; the correlation could be up to $-.666$, or as low as $-.03$ (effectively zero).

479

480



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

491 3.3 Gaze behaviours

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

496 3.3.1 Stationary Gaze Entropy (H_s)

497 3.3.1.1 Distributional parameters for H_s

498 The β_N parameter predicted an average decrease in H_s of -.141 (95% CI: -.178, -.101) when
 499 drivers completed the N-back task; equivalent to a 14 percentage point reduction in normalized
 500 H_s . The β_L parameter predicted an average decrease in H_s of -.041 (95% CI: -.058, -.022) when
 501 a lead vehicle was present during automation; equivalent to a 4 percentage point reduction. The
 502 β_{NL} parameter was estimated to be .017 suggesting that N-back reduced the difference in H_s
 503 between lead and no lead conditions by around 1.7 percentage points. However, as highlighted

504 in Table 2 there is some uncertainty for this effect; only 92% of the most probable parameters
 505 values are above 0.

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

Fixed effects		
Dependent variable:		
	H_s	pd
β_0	.474 (.428, .520)	100%
β_N	-.141 (-.178, -.101)	100%
β_L	-.041 (-.058, -.022)	100%
β_{NL}	.017 (-.006, .040)	92.77%
Participants	38	
Observations	744	

507

508 The direction of the effects for σ_{ij} of H_s are uncertain. N-back is predicted to decrease σ_{ij} by
 509 15%, however the probability that the effect is negative is only 90%. A similar pattern of results
 510 is found for the presence of the lead vehicle and the interaction effect.

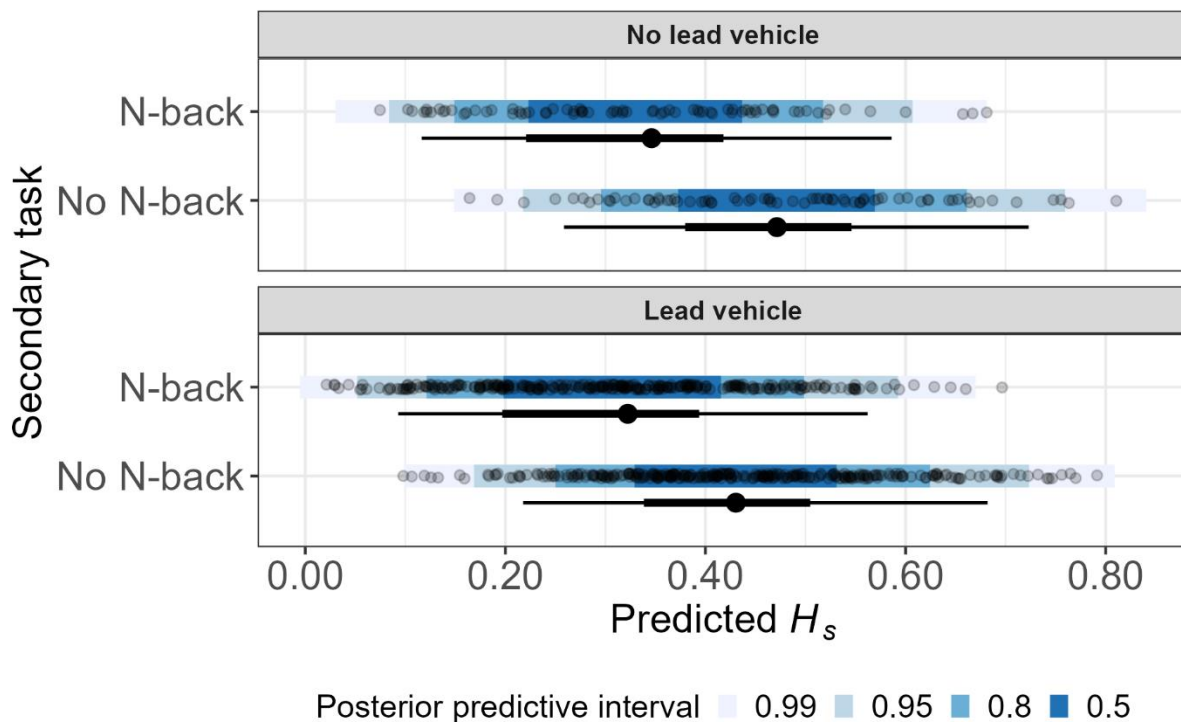
511 *Table 3: Posterior means and 95% CIs for fixed effect parameters predicting σ_{ij} of H_s*

Fixed effects		
Dependent variable:		
	H_s	pd
α_0	-2.676 (-2.867, -2.475)	100%
α_N	-.167 (-.420, .095)	90.23%
α_L	.098 (-.103, .294)	83.35%
α_{NL}	.019 (-.265, .290)	55.13%
Participants	38	
Observations	744	

512

513 Overall, the model predicts that N-back reduces the spatial distribution of gaze. This is
 514 evidence of reduced top-down engagement when monitoring the road environment during
 515 hands-off Level 2 automated driving. This supports previous research which has shown that
 516 increased mental workload during automated driving reduces gaze dispersion (Wilkie et al,

517 2019) and suggests that H_s could be a good metric for estimating mental workload in drivers.
 518 Modelling the trial-by trial variance in H_s did not show strong effects of N-back or lead vehicle.
 519 This is highlighted in Figure 5, whereby the predictive intervals overlayed on raw data have
 520 similar ranges around their predicted means for all conditions. This suggests that *variance* in
 521 gaze dispersion from trial to trial was consistent across trials and thus changes in σ_{ij} of H_s may
 522 not be useful for detecting increased driver workload.

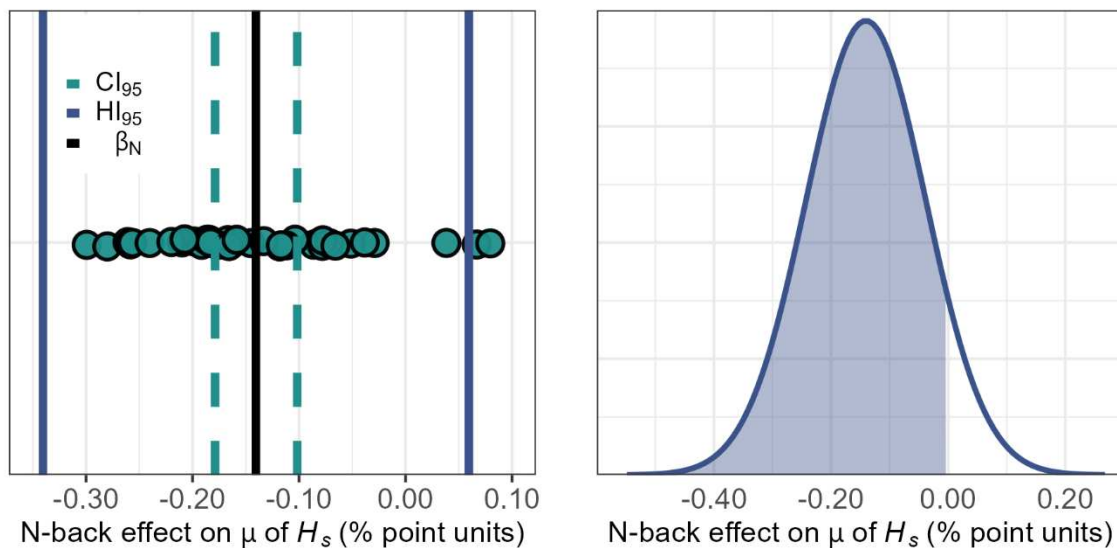


523 *Figure 5: Posterior predictive bands and posterior distribution of means plotted against raw*
 524 *data for conditions with and without a lead vehicle. The point-interval plot highlights the*
 525 *predicted mean differences between N-back/no N-back and lead/no lead vehicle alongside 50%*
 526 *and 95% credible interval bars. For both lead vehicle and N-back comparisons, the posterior*
 527 *predictive intervals are roughly of similar size highlighting the lack of evidence for N-back and*
 528 *lead vehicle affecting σ_{ij} of H_s .*

529 3.3.1.2 Heterogeneity parameters for H_s

530 Although the typical driver had reduced H_s by 14 percentage points during the N-back
531 condition, people differed in the size of this effect. Some participants had reductions as large
532 as 29 percentage points, some as a low as 3 percentage points, whereas some demonstrated
533 increases in H_s by up to 8 percentage points (see Figure 6, left panel). Despite these outlying
534 participants, the model estimates that 92% of the population are expected to have reductions in
535 H_s as a result of completing N-back during automation; the remaining 8% of the population
536 are expected to see moderate increases in H_s whilst cognitively loaded (see Figure 6, right
537 panel).

538



539 *Figure 6: Left panel: strip plot displaying the range of causal effect of N-back on H_s . The black*
540 *lines denote the average decrease in H_s (fixed effect), the blue dashed lines denote the*
541 *heterogeneity of the average casual effect of N-back (95% Credible Intervals) and the red solid*
542 *lines denote the population heterogeneity of the effect of N-back. Right panel: population*
543 *heterogeneity distribution implied by the model estimates of the mean and standard deviation.*
544 *92% of the population are predicted to demonstrate decreases in H_s when completing N-back*
545 *tasks.*

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

551 3.3.2 Gaze Transition Entropy (H_t)

552 3.3.2.1 Distributional parameters for H_t

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

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

Fixed effects		
	Dependent variable:	
	H_t	pd
β_0	.215 (.208, .222)	100%
β_N	-.021 (-.037, -.004)	99.25%
β_L	.001 (-.003, .006)	73.12%
β_{NL}	-.005 (-.012, .001)	96.08%
Participants	38	
Observations	744	

561
 562 The model also predicted differences in the σ_{ij} of H_t as a function of N-back and lead vehicle
 563 (see Table 5). The e^{α_N} parameter highlights an increase of 44% in within-participants variance

564 in H_t when completing the N-back during automation. The e^{α_L} parameter indicates that H_t
565 increased by 35% when a lead vehicle was present. The $e^{\alpha_{NL}}$ parameter suggests that the
566 difference in within-participants variance between conditions with and without a lead vehicle
567 were 23% smaller when drivers were not completing the N-back. However, there is some
568 uncertainty with this effect; the probability of the effect being above 0 is 96%.

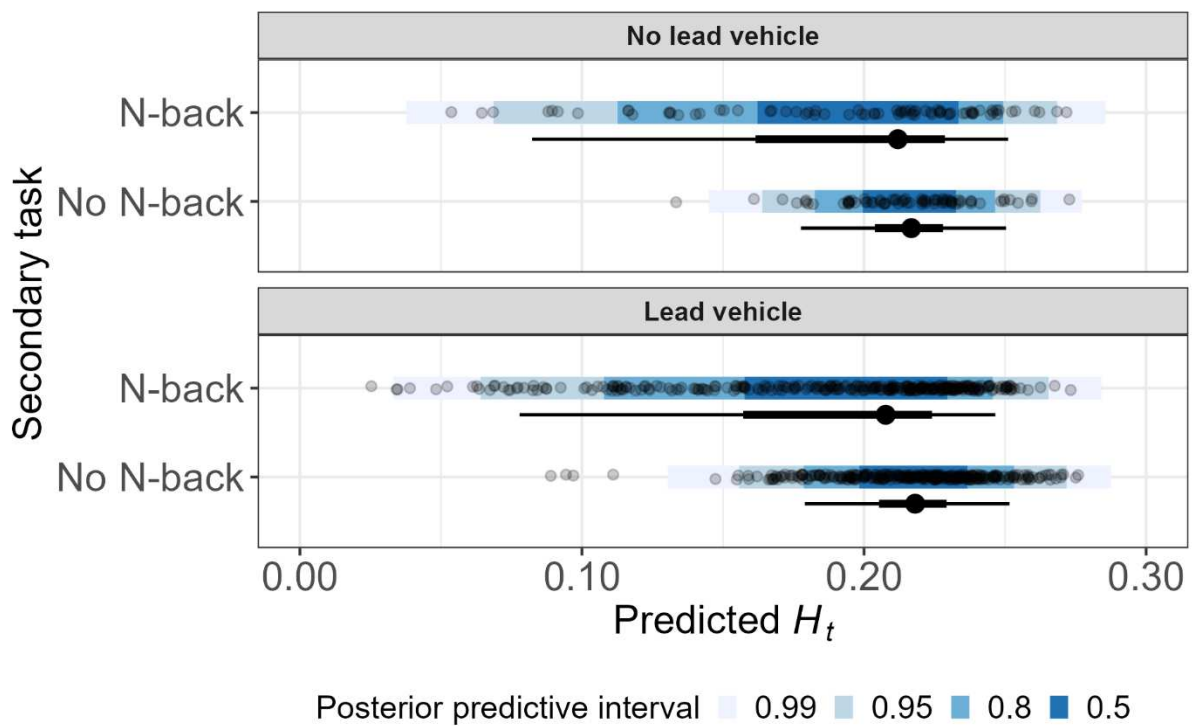
569 *Table 5: Posterior means and 95% CIs for parameters predicting the σ_{ij} of H_t*

Fixed effects		
	Dependent variable:	
	H_t	pd
α_0	-4.145 (-4.369, -3.920)	100%
α_N	.369 (.042, .696)	98.53%
α_L	.304 (.089, .524)	99.63%
α_{NL}	-.262 (-.568, .040)	95.95%
Participants	38	
Observations	744	

570
571 Model parameters highlight that completing N-back during automated driving produces
572 fixation transitions that are less erratic and more constrained within the visual scene. This
573 average decrease suggests that N-back produced top-down modulation of visual scanning
574 resulting in less complex, more constrained scanning behaviours. The concurrent reduction in
575 mean H_s and H_t as a function of N-back suggests that drivers did not perform sufficient
576 exploration of the visual scene while under high workload, and thus had reduced top-down
577 engagement whilst monitoring the automated system. This can be taken as evidence that, on
578 average, drivers during Level 2 automation who were under high workload had reduced
579 complexity of eye movements. The model also predicted *increases* in the σ_{ij} of H_t as a function
580 of N-back. The increase in σ_{ij} of H_t is highlighted in Figure 7; raw data are dispersed across a
581 broader range during N-back conditions. The systematic change in σ_{ij} as a function of N-back

582 tells us something about the relationship between visual scanning complexity and mental
 583 workload. Not only did drivers have reductions in scanning complexity, but they also failed to
 584 maintain a consistent complexity on a trial-by-trial basis. Instead, drivers demonstrated
 585 frequent fluctuations.

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



590 *Figure 7: Posterior predictive bands and posterior distribution of means plotted against raw*
 591 *data for H_t . The point-interval plot highlights the predicted mean differences between N-*
 592 *back/no N-back and lead/no lead vehicle alongside 50% and 95% credible interval bars. It is*
 593 *evident that there are small differences in predicted means between N-back and no N-back,*
 594 *however lead vehicle seems to have no effect on mean H_t . It is also evident that σ_{ij} increases*
 595 *as a function of N-back and lead vehicle, which is highlighted by the wider predictive intervals*
 596 *and larger spread of the data.*

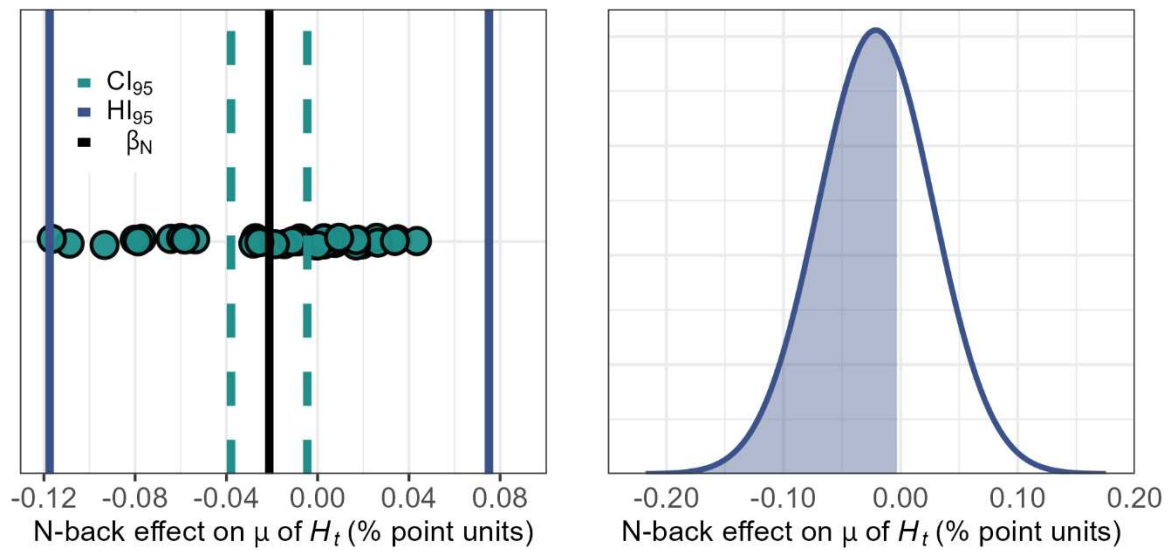
597 3.3.2.2 *Heterogeneity parameters for H_t*

598 The heterogeneity parameters of the model highlight considerable variance; the random slope
599 parameter (β_{N_j}) is almost two and a half times bigger than the average causal effect (β_N).
600 Whilst the average reduction in H_t during N-back was 2 percentage points, some people have
601 decreases in H_t of -.125 during N-back (12.5 percentage points) whereas some have *increases*
602 of up to .043 (4 percentage points) (see Figure 8, left panel). Furthermore, over 40% of the
603 sample show small-to-moderate *increases* in H_t during the N-back; a reversal of the average
604 trend. This suggests that a considerable proportion of the sample demonstrate more erratic and
605 random sampling patterns when cognitively distracted. The model predicts that only 66% of
606 the population will show an average decrease in H_t when completing the N-back during Level
607 2 automated driving (see Figure 8, right panel). The remaining 34% of the population are
608 expected to show increases in H_t , resulting in more erratic fixations transitions when
609 cognitively loaded.

610

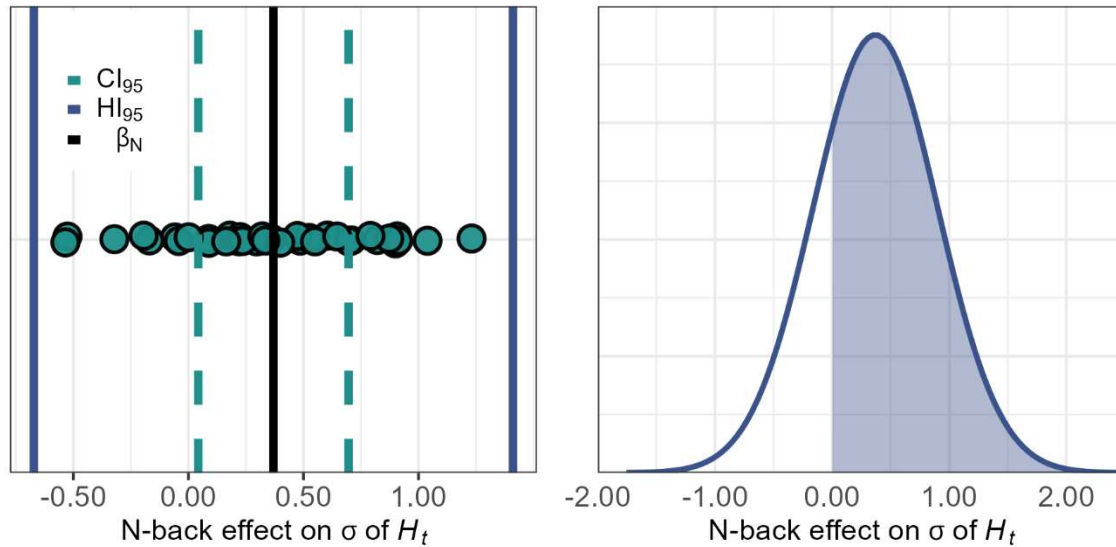
611

612



613 *Figure 8 The left panel shows a strip plot of the model predictions of the causal effect of 2-*
 614 *back on H_t . The black lines denote the average mean decrease in H_t (fixed effect), the blue*
 615 *dashed lines denote the heterogeneity of the average casual effect of N-back (95% Credible*
 616 *Intervals) and the red solid lines denote the population heterogeneity of the effect of N-back.*
 617 *The right panel shows the population heterogeneity distribution implied by the model's*
 618 *estimates of the mean and standard deviation for effect of N-back on H_t . Only 66% of the*
 619 *population are predicted to demonstrate mean decreases in H_t when completing the N-back*
 620 *task.*

621 Compare this to changes in σ_{ij} of H_t as a function of N-back. The random slope parameter
 622 predicting σ_{ij} (α_{N_j}) is only 1.5 times bigger than the average causal effect of N-back on σ_{ij}
 623 (α_N). This is further supported by looking at individual changes in σ_{ij} of H_t as a function of
 624 the N-back (see Figure 9, left panel). Whilst there is variation in the size of the effect, the
 625 direction of the effect is more consistent across the sample. This is reflected in the model
 626 predictions for the population; it predicts that 76% of the population show average increases in
 627 trial-by-trial variance when completing the N-back task during Level 2 automated driving.



629 *Figure 9: The left panel shows a strip plot of the model predictions of the causal effect of N-*
 630 *back on σ_{ij} of H_t . The black lines denote the average decrease in σ_{ij} (fixed effect), the blue*
 631 *dashed lines denote the heterogeneity of the average casual effect of N-back (95% Credible*
 632 *Intervals) and the red solid lines denote the population heterogeneity of the effect of N-back.*
 633 *The right panel shows the distribution of the individual effects of N-back on σ_{ij} of H_t in the*
 634 *population predicted by the model. 76% of the population are predicted to demonstrate*
 635 *increases in σ_{ij} of H_t when completing the N-back task.*

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

639 3.4 Understanding heterogeneity in average causal effect

640 Thus far it has been demonstrated that the mean of H_s and H_t change as a function of N-back.
 641 However, they both also demonstrate substantial variation across the sample, albeit in differing
 642 manners. H_s decreases for a majority of the sample but at varying magnitudes. Conversely, H_t
 643 decreases for only two thirds of the sample with the remaining participants showing null effects
 644 or small reversals. Whilst this is theoretically useful, it is also important to understand *why*

645 these effects are so variable. One possible explanation for entropic gaze metrics is age. Schieber
646 & Gilland (2008) found that H_t consistently decreased as secondary task load increased, and
647 these effects were exacerbated for older (67–86 years old) versus younger (19-35 years old)
648 drivers. Schieber & Gilland (2008) proposed that this could be explained by shortfalls in visual-
649 spatial resources of older drivers. A combination of loading these resources with a secondary
650 task, and the demands of visual scanning during driving, could result in diminished scanning
651 complexity under the interpretation of Wickens' (2020) Multiple Resource Theory model.
652 More recent research supports this notion, suggesting that age-related impairments of top-down
653 attentional control can exacerbate the effects that secondary cognitive tasks have on H_t
654 (Gazzaley et al, 2005; Shiferaw et al, 2019).

655 To investigate whether age-related impairments of top-down attentional control influence the
656 effect of N-back, an additional model parameter β_A specifying the effect of age and its
657 interaction with N-back was included for models of H_s and H_t . For H_s , the model predicted
658 that age accounts for 9.9% of the between-participants heterogeneity in the causal effect of N-
659 back (see Figure 10). A closer look at Figure 10 highlights that younger than average drivers
660 still had decreases in gaze dispersion during N-back, although they were slightly smaller versus
661 older than average drivers.

662

663

664

665

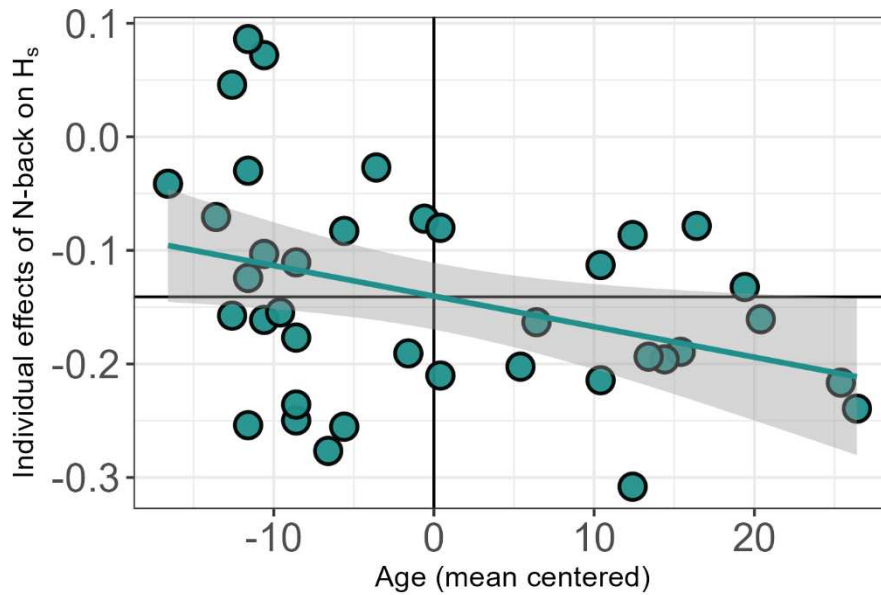
666

667

668

669

670



681 *Figure 10: Individual effects of N-back on H_s plotted against mean centred age. X axis vertical*
 682 *line denotes mean age, y axis horizontal line denotes the average effect of N-back. All people*
 683 *in the sample show decreases in gaze dispersion due to N-back. However, this effect is more*
 684 *prominent for older than average people.*

685 As for H_t , the model predicts that driver age accounts for 19% of between-participants
 686 heterogeneity in the effect of N-back. This suggests that age had a larger impact on how N-
 687 back effected H_t in comparison to how it impacted H_s . Furthermore, how the between-
 688 participants variance manifested was different. Younger than average drivers tended to show
 689 null effects or even small reversals of the average causal effect, whereas older drivers observed
 690 large reductions in H_t attributed to the effect of the N-back task (see Figure 11).

691
 692
 693
 694
 695
 696
 697
 698

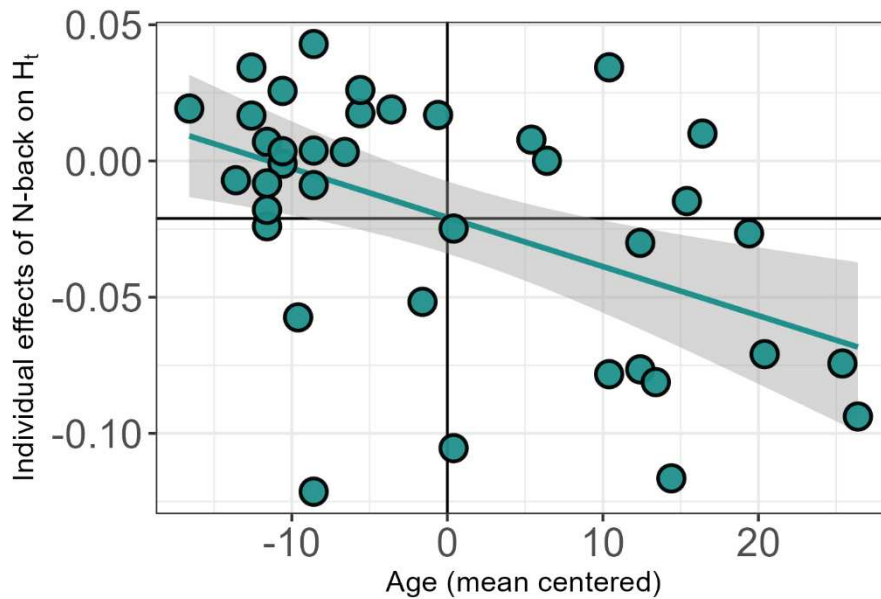


Figure 11: Individual effects of N-back H_t on plotted again mean centred age. X axis vertical line denotes mean age, y axis horizontal line denotes the average effect of N-back. Younger than average people appear to have almost no effect of N-back on H_t , with even some slight reversals. Conversely, 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

725 mental workload. Although the model predicted average reductions in gaze transition
726 complexity for high workload conditions, only 66% of the population would exhibit similar
727 decreases in H_t during N-back. Participant age appeared to be a strong predictor for how N-
728 back influenced gaze entropic measures, accounting for 9.5% and 19% of the between-
729 participants heterogeneity in the causal effect of N-back on H_s and H_t , respectively.

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

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

768 In terms of their implications, these results can provide DMS designers with some important
769 principles for using the correct metrics for detecting mental workload. A key aspect to be
770 considered is that driver demographics should be taken into account when using DMS to
771 establish driver state in vehicles. This analysis demonstrates that age was associated with the
772 extent to which N-back changed gaze-based metrics. As such, if DMS were to use H_s as an
773 indicator of mental workload, differing thresholds might be necessary for drivers of different

774 ages. For example, it might be necessary for a smaller threshold in the reduction of spatial
775 dispersion for younger drivers as their gaze might be less effected by N-back, even though they
776 might be experiencing high levels of mental workload, which could, in turn, impair their
777 takeover performance. Another element to for DMS engineers to consider is which parameter
778 of the gaze metric distribution should be used to establish a change in driver state. The current
779 state of the art assumes that changes in central tendency should be used (e.g. a change in mean
780 H_t establishes that N-back induces high mental workload). However, the current findings
781 suggest that changes in *variance* may be more reliable. Increases in the trial-by-trial variance
782 of H_t were predicted to be found in 76% of the population during high mental workload; only
783 66% of the population were predicted to follow average trends regarding a change in mean H_t .
784 This suggests that changes in the variance of gaze complexity were more reliable than changes
785 in the mean. High trial-by-trial variance during N-back suggests that drivers had frequent
786 fluctuations in the complexity of their gaze from one trial to the next. Rather than finding an
787 optimal level of gaze transitions that were suitable for all trials, the randomness of the
788 transitions changed frequently. It has been proposed that the motor controls involved in eye
789 movements aim to optimize inference (Parr & Friston, 2017) which implies that there are
790 optimal levels of H_t to sample the environment efficiently (Shiferaw et al, 2019). Hence the
791 results in the current manuscript suggest that high mental workload disrupts this eye movement
792 optimization, resulting in variable, inefficient monitoring of the driving environment. The
793 utilization of variance as an indicator for mental workload supports results from research within
794 the visual distraction domain. These show, for example, that presentation of information by
795 certain in-vehicle information systems reduces variations in fixation durations, supporting
796 more efficient information processing (Horrey & Wickens, 2007; Kujala & Saarilouma, 2011).
797 A similar suggestion is made here; without N-back trial-by-trial variance is small suggesting
798 drivers establish and optimal H_t that allows them to efficiently sample the road. As mental

799 workload increases, so does the variance in H_t , which is proposed as an indicator for visual
800 scanning inefficiency. These findings suggest more research is needed to understand whether
801 different parameters of response distributions can be used as indicators of mental workload.

802 Another interesting result from this study was the effect of lead vehicle presence. There was a
803 small but consistent decrease in the spatial distribution of gaze for trials with lead vehicles.
804 This supports previous research that drivers reduce the spread of their gaze and reallocate
805 attention towards lead vehicles (Crundall et al, 2004). A key difference is that Crundall et al
806 (2004) observed reductions in gaze dispersion only when instructing drivers to follow a lead
807 vehicle during manual driving. Conversely, participants in the current study were instructed to
808 monitor the entire road environment for hazards. Despite this request, the lead vehicle was
809 clearly a salient object within the road environment and thus likely attracted drivers' attention.
810 This may pose a problem for DMS in the real world, given that gaze dispersion has been shown
811 to decrease in the presence of a lead vehicle, irrespective of increasing mental workload.
812 Therefore, DMS will need to ensure that it can distinguish between drivers attending towards
813 vehicles on the road ahead, and those under increased mental workload. It should be noted that
814 the average reduction in gaze dispersion was much smaller for lead vehicles versus N-back
815 conditions, however this still will not disentangle drivers who had smaller reductions in gaze
816 dispersion during N-back conditions.

817 Another thing to highlight is the artificial nature of the N-back task as a method for inducing
818 mental workload. Human Factors researchers have used a range of tasks to induce non-visual
819 mental workload during manual and automated driving experiments. One of the most common
820 is the N-back task because is easy to control and systematically manipulate the level of mental
821 workload (Reimer, 2009; Reimer et al, 2010; Wang et al, 2014) and to quantify performance
822 (Goodridge et al, 2023). However, other tasks have been used such as the Sustained Attention
823 Response Task (SART) (Hawkins et al, 2014) and the Paced Auditory Serial Addition Task

824 (PASAT) (Thompson et al, 2012). What these tasks share is that they all involve drivers having
825 to maintain digits (or sometimes letters in the case of N-back and the SART) in their working
826 memory, thus loading these cognitive resources. Although these tasks allow for tight control
827 of mental workload manipulations, in the real world, it is unlikely that drivers will practice
828 keeping letters and numbers in their memory. Some have proposed that more natural tasks
829 should be used to better reflect the sorts of NDRTs that real drivers will complete during higher
830 levels of automation (Goodridge et al, 2023). Whilst some studies have used more naturalistic
831 hands-free phone conversations (Recarte & Nunes, 2003; Treffner & Barrett, 2004), it is harder
832 to experimentally control and thus test, the effects that these conversations might have on gaze
833 behaviours. The Twenty Questions Task (TQT) has been used which is still conversational
834 (and thus more realistic) but has elements of experimental control regarding the target word a
835 driver must reach (Merat et al, 2012). Future research should compare artificial and natural
836 modes of mental workload induction to investigate whether they have differing effects on eye
837 movements that could be used by DMS.

838 One limitation of the current work is that these model predictions need to be validated on a
839 wider range of datasets. A statistical model is only as good as the data used to fit it. Whilst age
840 ranges and gender balance were representative in the current sample, they mostly represented
841 white, British drivers in the north of England. As such, whether their behaviours translate well
842 to drivers from different cultures remains to be seen. Another limitation with the current work
843 is the use of a Gaussian distribution as the likelihood for the modelling. Whilst the data were
844 approximated by a Gaussian distribution, and the posterior predictive checks appear to fit the
845 data well, normalized H_s and H_t are technically continuous variables bounded between 0 and
846 1. Conversely, any value is possible for a Gaussian distribution. The Beta distribution is a
847 candidate that might be better suited for modelling these types of variables (Paolino, 2001;
848 Ferrari & Cribari-Neto, 2004). Whilst a comparison of Gaussian and Beta likelihoods on

849 clinical data highlighted that the estimates were very similar (Kurz, 2023) the Beta distribution
850 is a better conceptual fit and produced slightly more precise estimates. Future research may
851 compare these methods to investigate any differences in the context of gaze metrics.

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

862

863

864

865

866

867

868

869

870

871 **References**

- 872 Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in
873 psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601.
- 874 Bottos, S., & Balasingam, B. (2020). Tracking the progression of reading using eye-gaze point
875 measurements and hidden markov models. *IEEE Transactions on Instrumentation and*
876 *Measurement*, *69*(10), 7857-7868.
- 877 Brookhuis, K. A., & De Waard, D. (1993). The use of psychophysiology to assess driver
878 status. *Ergonomics*, *36*(9), 1099-1110.
- 879 Brookhuis, K. A., & de Waard, D. (2000). Assessment of drivers' workload: Performance and
880 subjective and physiological indexes. In *Stress, workload, and fatigue* (pp. 321-333). CRC
881 press.
- 882 Bruggen, A. (2015). An empirical investigation of the relationship between workload and
883 performance. *Management Decision*, *53*(10), 2377-2389.
- 884 Bürkner, P. C. (2017). Advanced Bayesian multilevel modeling with the R package
885 brms. *arXiv preprint arXiv:1705.11123*.
- 886 Carsten, O., Lai, F. C., Barnard, Y., Jamson, A. H., & Merat, N. (2012). Control task
887 substitution in semiautomated driving: Does it matter what aspects are automated?. *Human*
888 *factors*, *54*(5), 747-761.
- 889 Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing eye-tracking metrics of mental
890 workload caused by NDRTs in semi-autonomous driving. *Transportation research part F:*
891 *traffic psychology and behaviour*, *89*, 109-128.
- 892 Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive
893 science. *Behavioral and brain sciences*, *36*(3), 181-204.

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

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

898 De Waard, D. (1996). The measurement of drivers' mental workload. PhD thesis, University
899 of Groningen, Traffic Research Centre.

900 De Winter, J. C., Happee, R., Martens, M. H., & Stanton, N. A. (2014). Effects of adaptive
901 cruise control and highly automated driving on workload and situation awareness: A review of
902 the empirical evidence. *Transportation research part F: traffic psychology and behaviour*, 27,
903 196-217.

904 Di Stasi, L. L., Díaz-Piedra, C., Ruiz-Rabelo, J. F., Rieiro, H., Carrion, J. M. S., & Catena, A.
905 (2017). Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based
906 indices. *Applied Ergonomics*, 65, 168-174.

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

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

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

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

915 Gazzaley, A., Cooney, J. W., Rissman, J., & D'esposito, M. (2005). Top-down suppression
916 deficit underlies working memory impairment in normal aging. *Nature neuroscience*, 8(10),
917 1298-1300.

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

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

924 Goodridge, C. M., Gonçalves, R. C., & Öztürk, İ. (2023, September). What do we mean by
925 cognitive load? Towards more accurate definition of the term for better identification by driver
926 monitoring systems. In *Adjunct Proceedings of the 15th International Conference on*
927 *Automotive User Interfaces and Interactive Vehicular Applications* (pp. 256-259).

928 Hawkins, G. E., Mittner, M., Forstmann, B. U., & Heathcote, A. (2019). Modeling distracted
929 performance. *Cognitive psychology*, 112, 48-80.

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

932 Henderson, J. M. (2017). Gaze control as prediction. *Trends in cognitive sciences*, 21(1), 15-
933 23.

934 Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J.
935 (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

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

938 Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews*
939 *neuroscience*, 2(3), 194-203.

940 Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., ... & Villalobos,
941 N. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception (TAP)*, 13(1), 1-
942 20.

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

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

947 Kujala, T., & Saariluoma, P. (2011). Effects of menu structure and touch screen scrolling style
948 on the variability of glance durations during in-vehicle visual search tasks. *Ergonomics*, 54(8),
949 716-732.

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

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

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

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

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

963 Merat, N., Jamson, A. H., Lai, F. C., & Carsten, O. (2012). Highly automated driving,
964 secondary task performance, and driver state. *Human factors, 54(5), 762-771.*

965 Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015, September).
966 Emergency, automation off: Unstructured transition timing for distracted drivers of automated
967 vehicles. In *2015 IEEE 18th international conference on intelligent transportation systems* (pp.
968 2458-2464). IEEE.

969 Mole, C., Pekkanen, J., Sheppard, W., Louw, T., Romano, R., Merat, N., ... & Wilkie, R.
970 (2020). Predicting takeover response to silent automated vehicle failures. *Plos one, 15(11),*
971 *e0242825.*

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

976 Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* MIT press.

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

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

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

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

987 Radlmayr, J., Fischer, F. M., & Bengler, K. (2019). The influence of non-driving related tasks
988 on driver availability in the context of conditionally automated driving. In *Proceedings of the*
989 *20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport*
990 *Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20* (pp.
991 295-304). Springer International Publishing.

992 Raptis, G. E., Fidas, C. A., & Avouris, N. M. (2017, May). On implicit elicitation of cognitive
993 strategies using gaze transition entropies in pattern recognition tasks. In *Proceedings of the*
994 *2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1993-
995 2000).

996 Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive
997 coding?. *Frontiers in psychology*, *4*, 276.

998 Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual
999 search, discrimination, and decision making. *Journal of experimental psychology:*
1000 *Applied*, *9*(2), 119.

1001 Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual
1002 tunneling. *Transportation Research Record*, *2138*(1), 13-19.

1003 Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2010, September). The impact of
1004 systematic variation of cognitive demand on drivers' visual attention across multiple age
1005 groups. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol.
1006 54, No. 24, pp. 2052-2055). Sage CA: Los Angeles, CA: SAGE Publications.

1007 Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of
1008 variations in short-term memory demands on drivers' visual attention and driving performance
1009 across three age groups. *Human factors*, 54(3), 454-468.

1010 Schieber, F., & Gilland, J. (2008, September). Visual entropy metric reveals differences in
1011 drivers' eye gaze complexity across variations in age and subsidiary task load. In *Proceedings*
1012 *of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, No. 23, pp. 1883-
1013 1887). Sage CA: Los Angeles, CA: SAGE Publications.

1014 Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Alaguela, H., Teplitsky, C.,
1015 ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of
1016 distributional assumptions. *Methods in ecology and evolution*, 11(9), 1141-1152.

1017 Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical*
1018 *journal*, 27(3), 379-423.

1019 Shiferaw, B., Downey, L., & Crewther, D. (2019). A review of gaze entropy as a measure of
1020 visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, 96, 353-366.

1021 Sodhi, M., Reimer, B., & Llamazares, I. (2002). Glance analysis of driver eye movements to
1022 evaluate distraction. *Behavior Research Methods, Instruments, & Computers*, 34, 529-538.

1023 Spratling, M. W. (2017). A predictive coding model of gaze shifts and the underlying
1024 neurophysiology. *Visual Cognition*, 25(7-8), 770-801.

1025 SAE. (2018). J3016B: Taxonomy and Definitions for Terms Related to Driving Automation
1026 Systems for On-Road Motor Vehicles-SAE International.

1027 Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. (2017). LATEST: A model of saccadic
1028 decisions in space and time. *Psychological review*, 124(3), 267.

1029 Thompson, K. R., Johnson, A. M., Emerson, J. L., Dawson, J. D., Boer, E. R., & Rizzo, M.
1030 (2012). Distracted driving in elderly and middle-aged drivers. *Accident Analysis &*
1031 *Prevention, 45*, 711-717.

1032 Treffner, P. J., & Barrett, R. (2004). Hands-free mobile phone speech while driving degrades
1033 coordination and control. *Transportation Research Part F: Traffic Psychology and*
1034 *Behaviour, 7*(4-5), 229-246.

1035 Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-
1036 one-out cross-validation and WAIC. *Statistics and computing, 27*, 1413-1432.

1037 Velichkovsky, B., Sprenger, A., & Unema, P. (1997). Towards gaze-mediated interaction:
1038 Collecting solutions of the “Midas touch problem”. In *Human-Computer Interaction*
1039 *INTERACT'97: IFIP TC13 International Conference on Human-Computer Interaction, 14th–*
1040 *18th July 1997, Sydney, Australia* (pp. 509-516). Springer US.

1041 Wang, Y., Reimer, B., Dobres, J., & Mehler, B. (2014). The sensitivity of different
1042 methodologies for characterizing drivers' gaze concentration under increased cognitive
1043 demand. *Transportation research part F: traffic psychology and behaviour, 26*, 227-237.

1044 Weiss, R. S., Remington, R., & Ellis, S. R. (1989). Sampling distributions of the entropy in
1045 visual scanning. *Behavior Research Methods, Instruments, & Computers, 21*(3), 348-352.

1046 Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in*
1047 *ergonomics science, 3*(2), 159-177.

1048 Wickens, C. D. (2020). Processing resources and attention. In *Multiple task performance* (pp.
1049 3-34). CRC Press.

1050 Wilkie, R., Mole, C., Giles, O., Merat, N., Romano, R., & Makkula, G. (2019, June). Cognitive
1051 load during automation affects gaze behaviours and transitions to manual steering control.
1052 In *Driving Assessment Conference* (Vol. 10, No. 2019). University of Iowa.

1053 Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new
1054 explanation for the effects of mental underload on performance. *Human factors*, *44*(3), 365-
1055 375.

1056 Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? The impact of
1057 visual-cognitive load on driver take-over quality after conditionally automated
1058 driving. *Accident analysis & prevention*, *92*, 230-239.

1059

1060

1061

1062

1063

1064

1065

1066