



**HAL**  
open science

## Estimating the out-of-the-loop phenomenon from visual strategies during highly automated driving

Damien Schnebelen, Camilo Charron, Franck Mars

### ► To cite this version:

Damien Schnebelen, Camilo Charron, Franck Mars. Estimating the out-of-the-loop phenomenon from visual strategies during highly automated driving. *Accident Analysis & Prevention*, 2020, 148, pp.105776. <10.1016/j.aap.2020.105776>. <hal-02962111>

**HAL Id: hal-02962111**

**<https://hal.science/hal-02962111v1>**

Submitted on 9 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

1 Estimating the out-of-the-loop  
2 phenomenon from visual strategies during  
3 highly automated driving

4

5

6 Authors : Damien SCHNEBELEN<sup>1</sup>, Camilo CHARRON<sup>1,2</sup>, Franck MARS<sup>1</sup>

7 <sup>1</sup> Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France

8 <sup>2</sup> Université de Rennes 2, F-35000 Rennes, France

9

10 Corresponding author: Franck MARS

11 E-mail address: [franck.mars@ls2n.fr](mailto:franck.mars@ls2n.fr),

12 Postal address:

13 LS2N - Centrale Nantes

14 1 rue de la Noë

15 B.P. 92101

16 F- 44321 Nantes Cedex 03

17 FRANCE

18

## 19 Abstract

20 During highly automated driving, drivers no longer physically control the vehicle but they might need  
21 to monitor the driving scene. This is true for SAE level 2, where monitoring the external environment  
22 is required; it is also true for level 3, where drivers must react quickly and safely to a take-over request.  
23 Without such monitoring, even if only partial, drivers are considered out-of-the-loop (OOTL) and safety  
24 may be compromised. The OOTL phenomenon may be particularly important for long automated  
25 driving periods during which mind wandering can occur. This study scrutinized drivers' visual  
26 behaviour for 18 min of highly automated driving. Intersections between gaze and 13 areas of interest  
27 (AOIs) were analysed, considering both static and dynamic indicators. An estimation of self-reported  
28 mind wandering based on gaze behaviour was performed using partial least squares (PLS) regression  
29 models. The outputs of the PLS regressions allowed defining visual strategies associated with good  
30 monitoring of the driving scene. This information may enable online estimation of the OOTL  
31 phenomenon based on a driver's spontaneous visual behaviour.

32

33 Keywords: mind wandering; gaze behaviour; autonomous vehicles; PLS regression; driver monitoring

## 34 1. Introduction

35

36 The deployment of highly automated vehicles on the roads is imminent; it could occur anywhere  
37 between 2020 and 2030 (Chan, 2017). Among the expected benefits of autonomous vehicles  
38 (environmental, societal, etc), road safety is expected to improve (Fitch et al., 2014). The number of  
39 crashes caused by human error could be reduced by 90%, according to Stanton and Salmon (2009).  
40 However, to meet that target, drivers must be clearly aware of their role in the vehicle. That role partly  
41 depends on the level of automation.

42

43 There are five levels of automation (SAE International, 2016), corresponding to a different balance of  
44 tasks between the automation and the driver. At levels 0 and 1, the driver is in full or partial control of  
45 the vehicle commands. Automated driving starts at level 2, when longitudinal and lateral control of the  
46 vehicle is performed by the automation, but the driver must continuously monitor the driving scene and  
47 intervene when needed, even without a request from the system. At level 3 (conditional automation), the  
48 driver may engage in secondary tasks, but must be able to regain vehicle control when required by the  
49 system. This implies that monitoring the driving scene is only required starting with the take-over  
50 request. At level 4 (high automation), under certain conditions, the automation is able to perform all  
51 driving functions and can handle critical situations without requesting a take-over, although driver

52 override may be possible. In level 5 (full automation), the vehicle is autonomous in all conditions and  
53 the driver's action is no longer required.

54

55 In all cases, as soon as the operational part of the driving task is automated, drivers become supervisors  
56 of the automation system and of the driving scene. The level of expected supervision decreases as the  
57 level of automation increases. In manual driving, drivers must gather information from the driving scene  
58 and the vehicle (perceptual process), and must interpret that information (cognitive process) and act  
59 appropriately (motor process). Their actions in turn generate information. By contrast, starting with level  
60 2 automation, the perceptual-motor loop is neutralized, which has consequences for perception and  
61 cognition (Mole et al., 2019). At level 3, driver engagement in secondary tasks may intensify those  
62 consequences, with long periods of distraction from the driving scene. These consequences are referred  
63 to in the literature as the out-of-the-loop (OOTL) phenomenon.

64

65 The OOTL phenomenon was first observed in the aviation field (Endsley and Kiris, 1995), where  
66 automated piloting has long existed. Parasuraman and Riley (1997) showed that human pilots may be  
67 poor supervisors of the system. They may enter into a passive state when interacting with highly  
68 automated systems, causing a lack of situation awareness (Endsley, 1995). Recently, Merat et al. (2019)  
69 proposed an operational definition of OOTL in the context of automated driving. To be OOTL, drivers  
70 must lack physical control of the vehicle (motor process) and must not be monitoring the driving scene  
71 (perceptual or cognitive process). By contrast, when the driver is in manual control, they are considered  
72 to be in-the-loop. An intermediate state, namely on-the-loop (OTL), was introduced to designate cases  
73 where the driver correctly monitors the situation during autonomous driving.

74

75 In cars, the OOTL phenomenon has mainly been investigated through comparisons of driver behaviour  
76 between automated and manual driving. As most of the information processed during driving is visual  
77 (Sivak, 1996), the analysis of gaze behaviour has received much interest. Compared to manual driving,  
78 simulated automated driving involves more horizontal dispersion of gaze (Mackenzie and Harris, 2015;  
79 Louw and Merat, 2017) and a lower percentage of glances towards the road centre (Louw et al., 2015a;  
80 Mackenzie and Harris, 2015). Similarly, in curve driving, automated driving was shown to enhance  
81 long-term anticipation through look-ahead fixations, to the detriment of short-term anticipation used to  
82 guide the vehicle (Mars and Navarro, 2012; Schnebelen et al., 2019). When a secondary task was  
83 performed by drivers, automated driving was associated with relatively frequent fixations on those tasks  
84 (Merat et al., 2012).

85

86 When drivers are required to regain control of the vehicle, their behaviour after the take-over request is  
87 also considered to be an indicator of the OOTL phenomenon. Differences between automated and  
88 manual driving, as observed in critical scenarios, indicate that automated driving leads to impaired

89 visuomotor coordination during take-over (Mole et al., 2019). Navarro et al. (2016) showed that gaze  
90 distribution was widely dispersed, resulting in difficulties in steering around unexpected obstacles.  
91 Furthermore, drivers had longer reaction times to critical events, and vehicular control was impaired  
92 (Neubauer et al., 2012; Gold et al., 2013; Saxby et al., 2013; Louw et al., 2015b; Zeeb et al., 2015, 2017;  
93 Eriksson and Stanton, 2017). Such changes in driver behaviour during take-over were attributed to  
94 drivers being OOTL during automated driving.

95

96 In addition, the OOTL phenomenon seems to increase with a prolonged period of automation. In the  
97 aviation field, Molloy and Parasuraman (1996) showed that monitoring performance decreased with  
98 time. In the automated driving context, some studies have indicated that prolonged periods of automated  
99 driving rendered drivers further OOTL (Körber et al., 2015; Feldhütter et al., 2017; Bourelly et al.,  
100 2019). For example, Bourelly et al. (2019) observed longer reaction times (+0.5 s) to a critical event  
101 after 1 h of automated driving, compared to reactions to the same event after 10 min.

102

103 According to the definition of OOTL in automated driving (Merat et al. 2019), OOTL drivers do not  
104 correctly monitor the driving situation. The OOTL state may be experimentally induced by modifying  
105 perceptions of the driving environment or through instructions given to the drivers (i.e. using a secondary  
106 task). Louw et al. (2015b, 2016, 2017) reduced the visual information available for drivers using  
107 simulated fog to examine whether the driver was further OOTL when the fog was dense. In that sense,  
108 the OOTL state mostly occurred through impairment of perception. In other studies (Carsten et al., 2012;  
109 Merat et al., 2012), monitoring of the driving environment was altered by a secondary task. In these  
110 cases, the degradation of both perceptual (eyes off-road) and cognitive processes (mind off-road) yielded  
111 the OOTL phenomenon.

112

113 However, OOTL may also spontaneously and progressively occur without any modification of the  
114 driving environment or the presence of a secondary task. Due to lack of activity, drivers can experience  
115 mind wandering (MW), progressively disengaging from the supervision task even in the absence of an  
116 external source of distraction (Körber et al., 2015; Feldhütter et al., 2017; Burdett et al., 2019). Gouraud  
117 et al. (2017) proposed a detailed review of the links between the OOTL phenomenon and MW. Both  
118 phenomena are characterized by a decoupling of the immediate task and leads to similar safety issues.  
119 The decoupling can be the result of a reduction in the perception of the environment relevant to the task  
120 (sensory attenuation), which can have a negative impact on the construction of an accurate situation  
121 model. Both MW and the OOTL phenomenon have been associated with a slower response or detection  
122 failure when a critical event occurs. Gouraud et al. (2017) conclude that MW markers could help study  
123 OOTL situations. MW can be assessed by various means, including physiological and behavioural  
124 indicators, although self-report measures remain widely used because of their robustness. Specific to the  
125 case of autonomous driving, MW may occur at level 2, when the driver is supposed to monitor the

126 driving scene. It may also happen at level 3 if the driver chooses not to engage in secondary tasks. In  
127 both cases, MW can be more difficult to assess than visual distraction, since gaze is not diverted from  
128 the driving scene.

129  
130 The OOTL state was defined in previous studies in a relative way: either drivers are more OOTL when  
131 automation is on, compared to manual driving; or they are more OOTL during prolonged automation,  
132 compared to a shorter duration. In all cases, the OOTL phenomenon impacts driver safety during take-  
133 over, especially during long automated drives. Hence, determining whether a driver is OOTL or OTL  
134 must be performed before any situation requiring a take-over.

135  
136 This study estimated the driver's state based on the observation of spontaneous gaze behaviour. The  
137 participants experienced an 18-min drive of automated driving (similar to Feldhuetter et al. 2017), with  
138 the OOTL phenomenon occurring spontaneously. Quantitative assessment of the OOTL state was based  
139 on the self-reported time of MW. The drivers' gaze behaviour was analysed by considering 13 areas of  
140 interest, using static and dynamic indicators. Static indicators refer to the percentage of time the gaze is  
141 directed to one AOI and dynamic indicators refer to transitions from one AOI to another. An original  
142 method was used that involved multiple partial least squares (PLS) regression analyses. The goals were  
143 1) to identify and select the most important gaze indicators, and 2) to generate models for the relationship  
144 between gaze behaviour and MW score. The principle of this methodological choice was to identify the  
145 essential indicators from the overall spontaneous gaze behaviour, without relying on theoretical  
146 preconceptions.

147  
148 The research was guided by two research questions:

- 149
- 150 1. Is it possible to identify gaze behaviour that is characteristic of OOTL drivers? (i.e. what  
151 constitutes inadequate monitoring of a driving situation?)
  - 152 2. Is it possible to estimate the driver's OOTL state from the observation of spontaneous gaze  
153 strategies?

## 154 2. Materials and Methods

### 155 2.1. Participants

156 The study involved 12 participants (N = 12; 3 female, 9 male), with a mean age of 21.4 years (SD = 5.34  
157 y). To facilitate the recording of accurate gaze data, volunteers were required to have either normal

158 vision or vision corrected with contact lenses. They all held a valid driver's license, with average driving  
159 experience of 9950 km/year (SD = 5500). They signed written informed consent to participate.

## 160 2.2. Experimental setup

161 Figure 1 presents the driving simulator setup. This fixed-base simulator consisted of an adjustable seat,  
162 a steering wheel with force feedback, a gear lever, clutch, accelerator and brake pedals. The driving  
163 scene was generated with SCANeR Studio (v1.6) and displayed on three large screens in front of the  
164 driver (field of view  $\approx 120^\circ$ ). A dashboard screen indicated the speed of the vehicle. An HMI screen  
165 was added to the right-hand side of the driver, approximately where a vehicle's centre console is located.  
166 The description of the HMI can be found in the Procedure section.

167



168

169

170

171

Figure 1: Driving simulator setup

172 Gaze data were recorded using a Smart Eye Pro (V5.9) eye-tracker with four cameras; two were below  
173 the central screen and one below each peripheral screen. The calibration was performed in two steps.  
174 First, a 3D model of the driver's head was computed using an 11-point head calibration procedure with  
175 the head and gaze oriented toward the points. Then, the gaze was calibrated using 15 points: nine on the  
176 central screen, two on each peripheral screen, one on the dashboard and one on the HMI screen; with  
177 the head oriented to the central screen and gaze directed to the points. Gaze data were synchronized and  
178 recorded with vehicle data at 20Hz by the driving simulator software.

179

180 Most of the road was a 40-km two-lane dual carriageway, with a speed limit of 130 km/h in accordance  
181 with French regulations. Occasional changes in road geometry and speed limits were included to make  
182 the driving less monotonous. This included temporary 3-lane traffic flow, highway exits, slope variation

183 and variation of the speed limit (130 km/h to 110 km/h). In both directions on the highway, traffic was  
184 fluid, with eight overtaking situations.

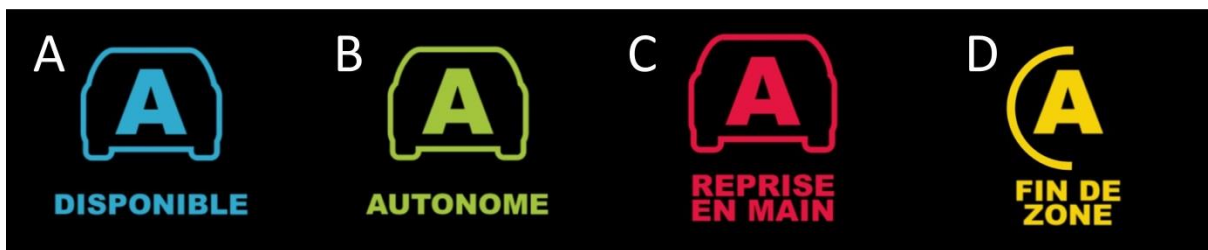
### 185 2.3. Procedure

186 The participants first adjusted the seat position, and the gaze calibration procedure was performed. Then,  
187 they drove manually along a training track to become accustomed to the driving environment and the  
188 vehicle's reactions. Once this training session was completed, instructions for automated driving were  
189 given orally.

190  
191 Drivers were told that the automated function would only be available for a portion of the road. The  
192 distance and time remaining in the autonomous mode were displayed on the left of the HMI. When  
193 activated by pressing a button, the automation controlled the lateral position and speed of the vehicle  
194 appropriately, accounting for traffic, speed limits and other conditions, and overtaking other cars if  
195 necessary. Participants were instructed to take control of the vehicle when requested by the system.

196  
197 Two possible use cases were presented to the participants. In the first case, the vehicle was approaching  
198 the end of the automated road section. The drivers would receive mild auditory and visual warning  
199 signals and would have 45 s to regain control. The second use case was an unexpected event, such as  
200 the loss of sensors. In this case, an intense auditory alarm would sound and a new pictogram would be  
201 displayed, and drivers would have only 8 s to resume control. All the pictograms and sounds used by  
202 the HMI were presented to participants before they began a second training session. The pictograms are  
203 shown in Figure 2.

204



205

206

207 Figure 2: Pictograms displayed on the HMI. A, autonomous driving available;

208 B, autonomous driving activated; C, critical take-over request (8 s); D, planned take-over request (45 s).

209

210 The second training session allowed participants to experience semi-automated driving (SAE level 1):  
211 cruise control with the driver in charge of the steering wheel; and level 3: conditional automation. At  
212 level 3, there were four transitions to manual control, two in each use case presented in the instructions.  
213 All take-overs were properly performed during the training session.

214

215 The experiment itself then commenced. The study followed a within-participant design, with all  
216 participants driving under both automated and semi-automated conditions. These conditions were  
217 similar, but no critical case happened in the semi-automated condition. However, because this paper  
218 concerns the analysis and modelling of gaze behaviour during automated driving, only the results  
219 obtained during the automated condition are presented here.

220

221 In the automated driving condition, participants activated the automated driving mode just before  
222 entering a highway. Gaze data were recorded as soon as the vehicle was correctly inserted in the lane  
223 and had reached 130 km/h. No major driving events or take-over request occurred in the first 15 min on  
224 the road, to allow enough time for the driver to become OOTL. The driver did not perform any secondary  
225 task during that time.

226

227 Between minutes 16 and 17, the two vehicles that would be involved in the critical case appeared in the  
228 driving scene. One overtook the participant's vehicle on the left and positioned itself in the right lane,  
229 300 m ahead. The other vehicle remained in the left lane, slowly approaching the participant's vehicle.  
230 During minute 18 of automated driving, participants experienced a critical take-over request occurring  
231 in response to unexpected braking from the lead vehicle. The warning signals were delivered as soon as  
232 the lead vehicle started to brake, with a time-to-collision of 8 s. At that moment, the lead vehicle was in  
233 the adjacent lane, in the blind spot of the participant. Changing lane would lead to a collision. To  
234 successfully handle the critical situation, drivers had to brake, remain in the right lane until the  
235 overtaking vehicle had passed, and then change lanes to avoid the lead vehicle. The scenario ended 30  
236 s after the critical case.

237

238 Participants were then asked to report on a continuous Likert scale the proportion of time they had spent  
239 thinking about something other than the driving task, throughout the trial. This simple method of MW  
240 self-assessment has been shown to be sensitive to driver disengagement during prolonged driving  
241 sessions (Mars et al., 2014).

## 242 2.4. Data structure and annotations

### 243 2.4.1 Definition of the MW score Y

244

245 In the absence of a secondary task, the evaluation of the percentage of time spent thinking about  
246 something other than the driving task was regarded as a self-assessment of the MW. The higher the  
247 percentage, the more the driver had estimated being OOTL. Percentages for all participants were stored

248 in a vector with 12 elements. After standardization of the data (conversion to a z score), the vector was  
249 denoted as Y and named MW score.

## 250 2.4.2 Definition of the matrix of gaze behaviour X

251

252 The driving scene was divided into 13 areas of interest (AOI), as shown in Figure 3. These are described  
253 below.

254 a) The central screen contained six AOIs:

- 255 • central mirror (CM)
- 256 • road centre (RC), defined as a circular area of 8° radius in front of the driver
- 257 • four additional areas, defined relative to the road centre (Up, Left, Down, Right).

258

259 The percentage road centre (PRC) is defined as the proportion of time spent in RC, as introduced by  
260 Victor (2005). A decrease in PRC was found to be a reliable indicator of distraction during driving;  
261 drivers reduced their PRC when visually or auditorily distracted (Victor et al., 2005).

262

263 b) Each peripheral screen contained two areas, with two items in each:

- 264 • lateral mirror (LM, RM)
- 265 • the remaining peripheral scene (LS, RS)

266

267 c) The dashboard (D)

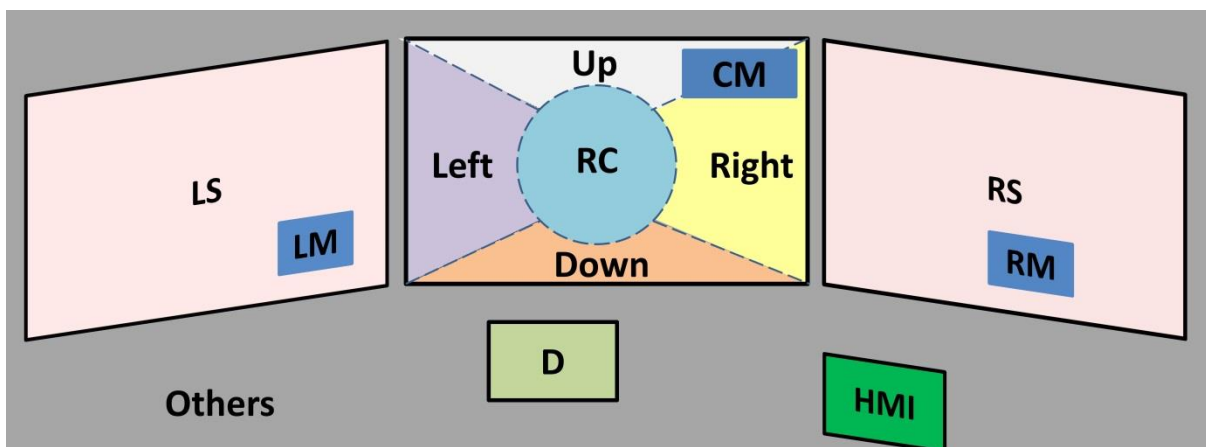
268

269 d) The HMI (HMI)

270

271 e) All data for gazes directed outside all the above areas were grouped as “other areas” (Others).

272



273

274

Figure 3: Division of the driving environment into 13 areas of interest

275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308

The percentage of time spent gazing at each AOI was computed, as was a matrix of transitions between AOIs. Using Markov logic, the matrix of transitions between AOIs corresponded to the probability of shifting from one AOI to another. The probability that gaze remained in the same AOI was also computed. This constituted the diagonal of the transition matrix. Probabilities were estimated by the observations of the participants (see Gonçalves et al., 2019, for an application of Markov chains in case of a lane change manoeuvre). Before a given AOI was intersected for the first time, a prior probability was associated with it. A uniform law was used for prior probabilities.

The 13 AOIs defined the entire world. Thus, the transition matrix was a 13\*13 matrix. If rows contained the current AOI and columns the probabilities, the sum of each row was equal to 1. In all cases, next gaze intersection appears in one of the 13 areas of interest.

In this study, the driver's gaze behaviour for each participant was considered to be the combination of static and dynamic indicators of gaze behaviour. Static refers to percentage of time in one AOI; dynamic refers to the transition matrix. Thus, the gaze strategy of a participant was represented by a vector of 182 numerical indicators (= 13\*13 transitions + 13 percentage of time on each AOI). When considering all participants, the matrix of gaze strategies was named X, and its size was 12 (participants) \* 182 (visual indicators).

Supplementary indicators that were sensitive to drivers' drowsiness were also computed. These included percentage of eye closure (PERCLOS; see Wierwille et al., 1994) and the blink-rate (Stern et al., 1994).

### 2.4.3 Computation of training and validation datasets

The objective was to predict the MW score as a function of gaze behaviour. Therefore, a training dataset (i.e., a gaze-behaviour matrix to create the model) and a validation dataset (a gaze-behaviour matrix to evaluate the model) were required.

The gaze-behaviour matrix obtained during the last two minutes (16 and 17) of automated driving was chosen as the validation dataset. The rationale was that because OOTL increases with automation duration, the final two minutes might reflect the visual consequences of the OOTL phenomenon most accurately. The validation dataset did not include the gaze data recorded once the take-over request was initiated.

309 Nevertheless, the question of the speed of appearance of OOTL and the observation time required to  
310 model it was of interest. To answer this question, 15 training datasets (i.e. 15 gaze-behaviour matrixes),  
311 labelled  $X_t$ , were calculated. The difference between the sets was the integration time – that is, the  
312 duration of automated driving while the matrix of gaze behaviour was computed. The integration time  
313 varied from 1 min to 15 min. The reference for the time window was the 15<sup>th</sup> minute of simulation (i.e.  
314 immediately before the validation dataset). Thus,  $X_1$  considered gaze behaviour during one minute (the  
315 15<sup>th</sup> minute), whereas  $X_7$  was computed from seven minutes of automated driving (between the 9<sup>th</sup> and  
316 15<sup>th</sup> minutes). The rationale was to evaluate whether using a short time window was enough to capture  
317 the consequences of OOTL to create a satisfactory model, or whether aggregating more data by  
318 enlarging the time window would make the model more robust.

319

320 In summary, 15 matrixes of gaze behaviours (labelled  $X_t$  for  $t$  between 1 and 15 min) were calculated,  
321 and these constituted the training datasets for the model. Once the best model had been selected, we  
322 validated it by comparing the MW score to the model's prediction based on the two final minutes of  
323 driving (16<sup>th</sup> and 17<sup>th</sup> minutes).

324

#### 2.4.4 Choice of model for prediction: PLS regression

325 Regarding the data structuring described above, the aim was to predict  $Y$  (MW score) from  $X_t$  (gaze-  
326 behaviour matrix), given the following conditions:

327

- 328 - Visual indicators ( $X_t$ ) are correlated. The driving environment was divided into 13 AOIs, hence,  
329 the percentage of time spent on 12 AOIs enabled calculating the percentage of time spent on the  
330 13<sup>th</sup> AOI. In mathematical terms,  $X_t$  might not be full rank.
- 331 - The number of visual indicators (182) that could explain  $Y$  is higher than the number of  
332 observations made on the participants (12).

333

334 Considering these constraints, the PLS regression model was selected. PLS regression yields the best  
335 estimation of  $Y$  available with a linear model given the matrix  $X_t$  (Abdi, 2010). All PLS regressions  
336 performed in this study used the PLS regression package (Wehrens and Mevik, 2007) from R (Core  
337 Team et al., 2013).

338

## 2.5. Data analysis

339 Four sequential stages composed the analysis of the training datasets ( $X_t$ ) (see Figure 4):

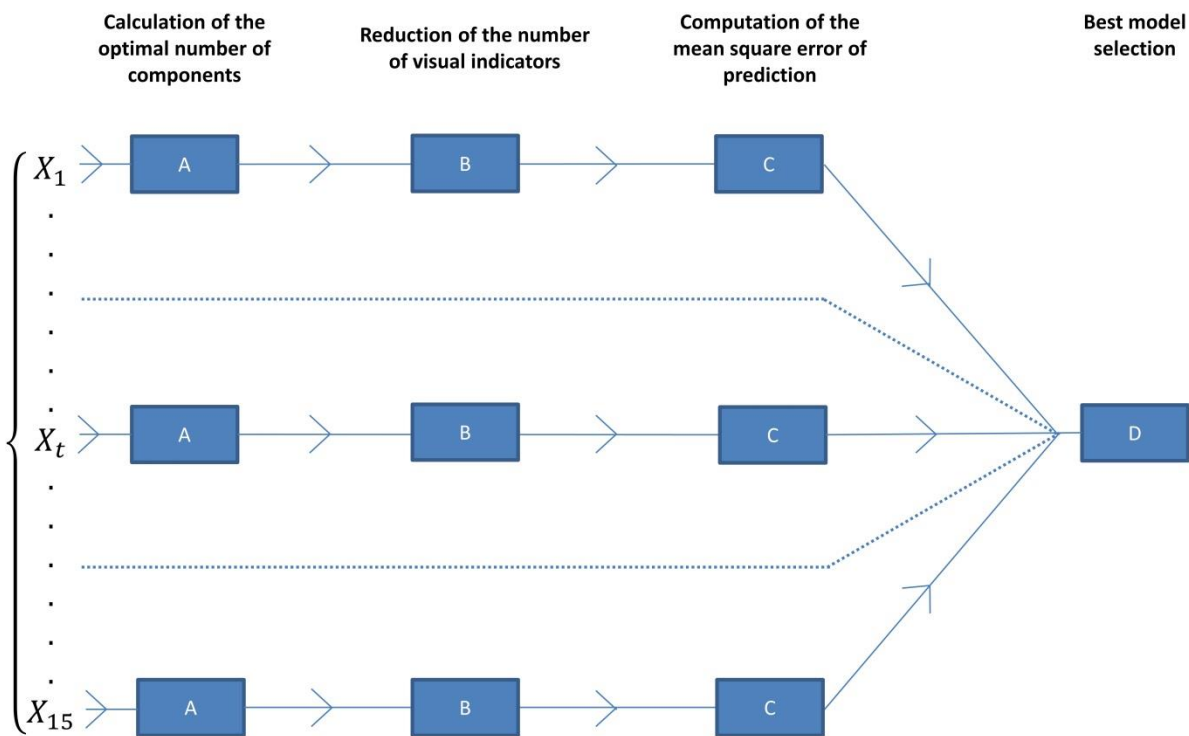
340

341

1. Step A entailed finding the optimal structure (i.e. the optimal number of components) of the  
PLS regression model for predicting the MW score.

- 342 2. Step B entailed selecting the visual indicators that significantly contributed to MW score  
 343 prediction.  
 344 3. In Step C, we considered the optimal parameters (components and visual indicators) of the  
 345 prediction model, and evaluated the model's accuracy using the mean square error of prediction  
 346 (MSEP) for both the training and validation datasets.  
 347 4. In Step D, we selected the model with the least validation error.  
 348

349 The step-by-step procedure of the data analysis is presented in the appendix. Only the final results, which  
 350 lead to model selection, are presented in the next section (Results).



351  
 352  
 353 Figure 4: Data analysis was performed in four sequential steps. First, the best parameters of the PLS regression  
 354 models were identified. This entailed finding the optimal number of components and reducing the number of  
 355 visual indicators (Steps A and B). Then, in this optimal configuration, the accuracy of the models was considered  
 356 by computing the mean square error of prediction, for both training and validation datasets (Step C). In the last  
 357 step (Step D), the model with the lowest validation error was selected and interpreted.

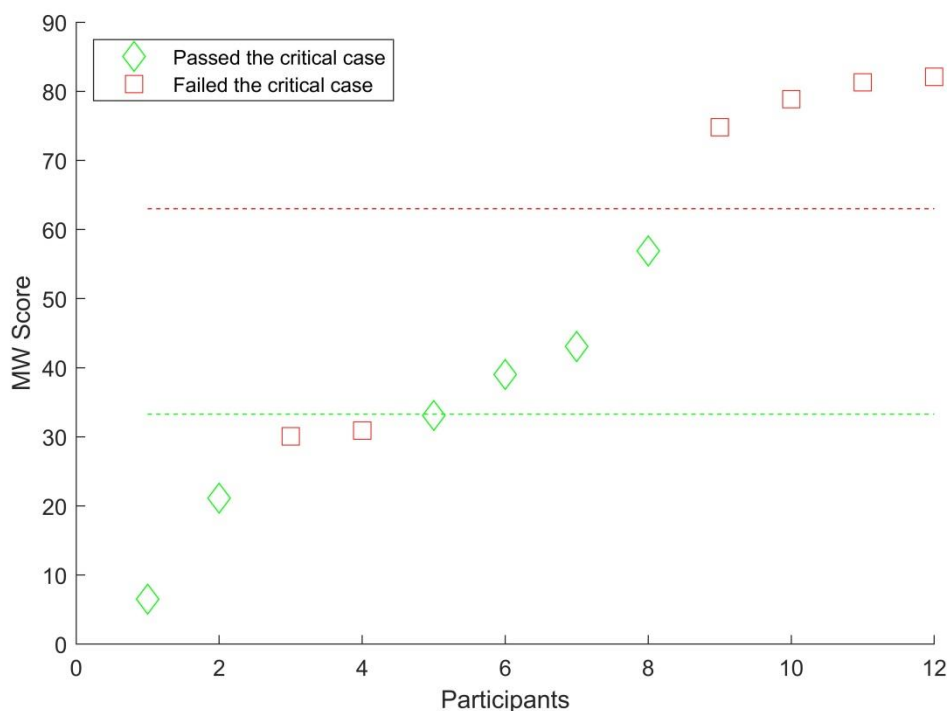
### 358 3. Results

#### 359 3.1. MW scores and drowsiness indicators

360 Figure 5 shows that the self-reported MW scores varied widely among participants. The median score  
361 was 43%, but four participants (participants 9 to 12) estimated that they had spent more than 70% of the  
362 time thinking about something other than the driving task. Those four participants failed in the critical  
363 case. Two other participants (3 and 4) also failed, with moderate MW scores of around 30%. All other  
364 participants managed to avoid a collision.

365

366



367

368

369 Figure 5: MW scores reported by the participants according to the outcome in the critical case. The colour of the  
370 marker indicates whether participants passed (green) or failed (red) the critical case. The dashed lines represent  
371 the average of the successful or failed groups. Participants are sorted according to their MW score, not their  
372 order of passing the experiment.

373

374 Mean PERCLOS and blink-rate per participant showed no significant correlation with MW score ( $r =$   
375  $0.25$  and  $r = -0.49$  respectively). The highest PERCLOS score was  $4.73\%$ , obtained by participant 2.  
376 Furthermore, a paired t-test revealed no significant differences for either PERCLOS or blink-rate  
377 between the first and last five minutes of automated driving ( $p = 0.28$  and  $p = 0.11$  respectively).

378

### 3.2. Selection of best model for MW score prediction

379

Results obtained for all different training datasets at the end of the PLS regression procedure (see

380

appendixes for details) are presented in Table 1.

381

382

383

384

385

386

387

Table 1: Optimal number of components, number of selected visual indicators and mean square error of

388

prediction (for both training and validation datasets) as a function of the time window. A minimum of validation

389

error was found for gaze data computed over 10 minutes of automated driving (0.419).

| Time Window (min) | Optimal number of components (step A) | Number of selected visual indicators (step B) | Mean square error of prediction with training dataset (step C) | Mean square error of prediction with validation dataset (step C) |
|-------------------|---------------------------------------|---|--|--|
| 1                 | 3                                     | 56  | <0.001   | 0.478  |
| 2                 | 1                                     | 26  | 0.200  | 0.625  |
| 3                 | 1                                     | 26  | 0.122  | 0.697  |
| 4                 | 1                                     | 16  | 0.054  | 0.915  |
| 5                 | 3                                     | 52  | 0.002  | 0.721  |
| 6                 | 1                                     | 19  | 0.233  | 0.542  |
| 7                 | 4                                     | 88  | <0.001   | 0.563  |
| 8                 | 5                                     | 80  | < 0.001  | 0.522  |
| 9                 | 1                                     | 11  | 0.224  | 0.443  |
| <b>10</b>         | <b>1</b>                              | <b>12</b>                                     | <b>0.211</b>   | <b>0.419</b>   |
| 11                | 1                                     | 8   | 0.130  | 0.446  |
| 12                | 1                                     | 7   | 0.131  | 0.496  |
| 13                | 1                                     | 57  | 0.248  | 0.533  |
| 14                | 1                                     | 3   | 0.141  | 0.598  |
| 15                | 1                                     | 3   | 0.147  | 0.595  |

390

391

Table 1 shows that the training error was small (<0.25) in all cases and it depended on the structure of

392

the model (i.e. number of components). The models having the most components (time windows of 1,

393 5, 7 and 8 minutes) predicted the learning dataset almost perfectly (MSEP <0.001). However, MSEP for  
 394 the validation dataset was comparatively high. As the time window increased, the validation error  
 395 decreased, attaining the minimal value (0.419) for 10 min of data. Thereafter, adding more data by  
 396 expanding the time window increased the MSEP. Thus, ultimately the best model for predicting MW  
 397 score from the last two minutes of driving was obtained by aggregating the gaze data for the 10 min  
 398 that preceded those final minutes.

### 399 3.3. Final prediction of MW score

400 After identifying the best model, it remains to be determined what visual indicators were most  
 401 important for predicting the MW score and how well the predicted scores were correlated to the actual  
 402 scores.

#### 403 3.3.1 Visual indicators

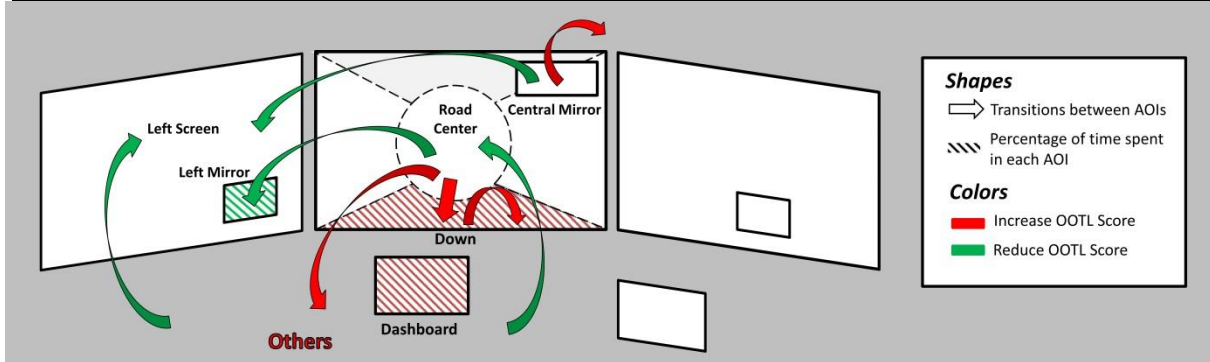
404 The best PLS model retained only 12 visual indicators to predict the MW score. The PLS regression  
 405 coefficients and the correlation coefficients between MW score and the 12 selected visual indicators  
 406 are presented in Table 2. Figure 6 illustrates the visual indicators retained by the model.

407

408 Table 2: Coefficients of the PLS regression and correlation coefficients between the MW score and the visual  
 409 indicators (\* p<0.05; \*\* p< 0.01). The 12 first indicators correspond to those shown in Figure 6, which were  
 410 used for the final prediction of MW score. The last four static indicators were not selected by the PLS regression,  
 411 but are reported for discussion.

| Impact on the prediction | Visual indicators                                     | PLS regression coefficients | Correlation coefficients |
|--------------------------|---|-----------------------------|--------------------------|
| Decreases the MW score   | Transition from the Central Mirror to the Left Screen | -0.098                      | -0.65*                   |
|                          | Transition from Others to the Left Screen             | -0.096                      | -0.63*                   |
|                          | Transition from the Others to the Road Centre         | -0.093                      | -0.61*                   |
|                          | Transition from the Road Centre to the Left Mirror    | -0.091                      | -0.63*                   |
|                          | Percentage of time spent in the Left Mirror           | -0.089                      | -0.58*                   |
| Increases the MW score   | Percentage of time spent in the Others Area           | 0.087                       | 0.58*                    |
|                          | Percentage of time spent in the Dashboard             | 0.094                       | 0.62*                    |
|                          | Transition from the Central Mirror to the Others Area | 0.099                       | 0.66*                    |
|                          | Transition from the Road Centre to the Others Area    | 0.101                       | 0.67*                    |
|                          | Transition from the Road Centre to the Down area      | 0.102                       | 0.66*                    |
|                          | Multiple gazes in the Down area                       | 0.108                       | 0.71**                   |

|                                |   |       |        |
|--------------------------------|---|-------|--------|
|                                | Percentage of time spent in the Down area         | 0.115 | 0.76** |
| Non-selected static indicators | Percentage of time spent in the Road Centre (PRC) |       | -0.16  |
|                                | Percentage of time spent in the Central Mirror    |       | -0.14  |
|                                | Percentage of time spent in the Right Mirror      |       | -0.09  |
|                                | Percentage of time spent in the HMI               |       | 0.47   |



412  
413  
414

Figure 6: Visual indicators relevant for MW score prediction.

415 The influence of a visual indicator on MW score is displayed in colour: red shows an increase in MW score, and  
416 green shows reduced MW score. Arrows represent transitions between AOIs. Filled areas, or name written in red  
417 (in the case of Others) means that percentage of time spent gazing in the AOI was selected by PLS regression.

418

419 Of the 12 indicators, eight were dynamic (transitions between AOI) and four were static (percentage of  
420 time spent in the area). The signs of the coefficients (see Table 2) indicate that seven of them contributed  
421 to an increase of the MW score estimation. These are shown in red in Figure 6, and can be summarized  
422 as follows: 1) taking the gaze off the central mirror to look away from the driving scene, 2) taking the  
423 gaze off the road centre area, to look down or away from the driving scene, and 3) spending too much  
424 time in the down area or on the dashboard.

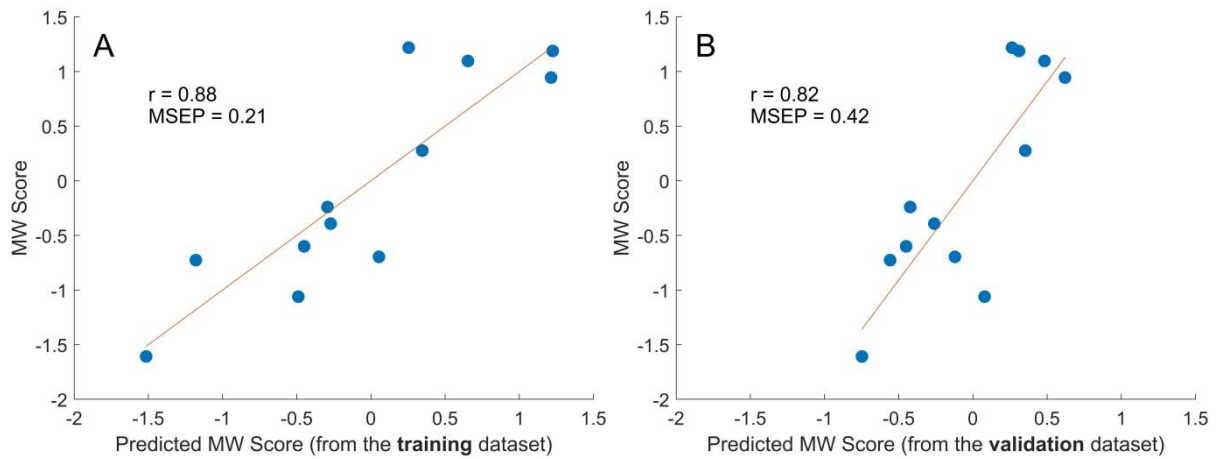
425

426 By contrast, five indicators contributed to a reduction in MW score estimation, shown in green in  
427 Figure 6. These consisted of 1) redirecting the gaze to the road from any area outside the driving scene  
428 (Others), 2) regularly checking the surroundings by looking at the left view mirror or the left side screen.

### 429 3.3.2 Prediction of MW score

430 Figure 7 presents the results of estimating the MW scores using the best model (time window = 10 min;  
431 12 visual indicators; 1 component) for the training dataset (A) and the validation dataset (B).

432  
433



434

435

436

437

438

439

440

441

442

443

444

445

446

447

Figure 7: Correlation plots between the (real) MW score and the predictions of the MW score by PLS regressions on the training dataset (A) and the validation dataset (B). All the values presented here are standardized.

## 448 4. Discussion

449

450

451

452

453

454

455

456

457

458

459

The OOTL phenomenon results from the combination of a driver not having physical control of their vehicle and incorrectly monitoring the driving situation (Merat et al, 2019). The alternative state, OTL, indicates a driver who satisfactorily monitors the driving environment during automated driving. Monitoring of the driving situation may be impaired when actively engaged in a secondary task. It can also take the form of MW, i.e cognitive disengagement from the driving task due to lack of activity. This study specifically investigated the second form of the OOTL phenomenon. Whatever the condition under which the OOTL phenomenon occurs, the question of how to model and quantify what constitutes proper monitoring of the driving scene remains debatable. The study addressed this question by distinguishing OTL and OOTL drivers in a highway driving context.

460 The method consisted of using multiple PLS regressions to identify characteristic elements of the gaze  
461 behaviour of OTL and OOTL drivers. The multi-step approach began with 182 indicators as an input  
462 matrix, accounting for both static elements (cumulative time spent in 13 AOIs) and dynamic elements  
463 (transitions between AOIs). These visual indicators were computed for different time windows to assess  
464 the evolution of the OOTL phenomenon over time. Once the optimal parameters of each model were  
465 calculated, the MW scores were predicted. Finally, a linear combination of the most important indicators  
466 enabled estimating the driver's MW score accurately.

#### 467 4.1. What constitutes good monitoring of a driving situation?

468  
469 The results revealed that drivers with relatively low MW scores made many transitions from the road  
470 centre to the left rearview mirror. In total, they spent more time looking at the mirror. They also looked  
471 at the left screen, where they could monitor traffic, immediately after gathering information from the  
472 central mirror. After spending time looking at areas unrelated to driving ("Others"), they frequently  
473 returned their gaze to the road (road centre area) or to the left screen. By contrast, drivers with relatively  
474 high MW scores made many transitions from the road centre to areas irrelevant to driving, where they  
475 gazed for a considerable time. They also gazed often at the lower part of the front screen and the  
476 dashboard.

477  
478 These findings can be interpreted in terms of the adequacy of the driver's gaze strategy to maintain  
479 sufficient situation awareness (Endsley, 1995) in autonomous mode. According to Merat (2019),  
480 situation awareness during automated driving involves three dimensions: perception, comprehension  
481 and projection. In the current study, OTL drivers remained dynamically aware of their surroundings by  
482 regularly checking the road ahead, the left lane and the left rearview mirror. Doing so helped them to  
483 anticipate future hazards and avoid difficulties when reacting in the final critical case. These gaze  
484 strategies allowed them to perceive, comprehend and project the future state of the driving situation  
485 appropriately. In other words, they had adequate situation awareness.

486  
487 By contrast, the OOTL drivers' gaze was more strongly attracted by irrelevant information outside the  
488 simulator. Even when looking at the driving environment, the driver favoured gazing at the road  
489 immediately in front (Down Area) and monitored their vehicle speed on the dashboard. Although OOTL  
490 drivers continued to correctly perceive the state of their vehicle and its position in the lane, they showed  
491 relatively few signs of visual anticipation by looking far ahead and they paid relatively little attention to  
492 other vehicles on the road.

493

494 Previous studies (Carsten et al., 2012; Louw et al., 2016) showed that manual drivers displayed  
495 a higher percentage of looking at the road centre (PRC) than drivers with automation. In these studies,  
496 secondary tasks were allowed or required by the experimental procedure. By contrast, in the current  
497 study, PRC did not emerge as a critical indicator (see Table 2). This was the case for most static  
498 indicators. This finding indicates that that a good monitoring of the situation is not so much a matter of  
499 spending considerable time looking at the road. It looks more important to regularly check relevant  
500 objects in the scene. Thus, considering the transitions between areas is more important than the amount  
501 of time spent on a particular area. For instance, moving the gaze away from the road centre might  
502 indicate a relatively high level of disengagement if the driver repeatedly looks at irrelevant areas.  
503 However, it could also contribute to situation awareness if the gaze moves to the left rearview mirror.

504  
505 The multiple PLS regression method allowed us to identify markers of the OOTL phenomenon  
506 from a rich set of data on drivers' gazes. This identification might have been achieved with simpler  
507 approaches, for example by examining the mean difference in visual indicators between two groups of  
508 participants, namely those having high or low MW scores. It might also have been possible to examine  
509 only the correlations between MW score and individual indicators. In this case, however, an a-priori  
510 selection should have been made, given the number of possibilities (182 indicators, with 13 AOI).  
511 Furthermore, individual correlations would not account for relationships between indicators. The  
512 advantage of PLS regression analysis is that all visual indicators are considered together to optimize the  
513 prediction of MW scores.

## 514 4.2. Is it possible to estimate drivers' OOTL state from the 515 observation of spontaneous gaze strategies?

516  
517 The results of the modelling work show that the best estimation of the MW score was obtained by  
518 considering 10 minutes of gaze data. The least prediction error occurred when the training dataset was  
519 used; however, performance was also good for the validation dataset (i.e. only two minutes of driving  
520 data, which were not used for the model determination).

521  
522 It can be concluded that the influence of the driver's state on their gaze strategy was qualitatively similar  
523 during the last two minutes of automated driving as it had been during the previous 10 minutes. This  
524 point suggests that detecting the OOTL state could perhaps be performed using a shorter time window  
525 than 10 minutes. This is interesting from the perspective of defining an algorithm for real automated  
526 vehicles to monitor the driver's state. However, further tests are necessary, notably with other  
527 participants and in other driving contexts.

528

529 The results also suggest that the OOTL phenomenon took some time to appear and that it increased with  
530 the duration of automated driving (Körber et al., 2015; Feldhütter et al., 2017; Bourrelly et al., 2019).  
531 The prediction error decreased when more gaze data were considered, but the minimal value occurred  
532 for 10 minutes. Beyond this point, the error gradually increased. This finding can probably be explained  
533 by the fact that considering more than 10 minutes of gaze data meant aggregating data from the  
534 beginning of the scenario. During the initial minutes, drivers had not had enough time to become OOTL.  
535 In order to evaluate the actual evolution of MW over time, an online measurement would have been  
536 preferable. Probe techniques that interrupt the participant could hardly have been considered under our  
537 driving conditions, as they could have prevented the OOTL phenomenon from developing.  
538 Physiological indicators such as pupil diameter, skin conductance or cardiac measurements could be  
539 considered in the future, although their robustness remains to be demonstrated.

540

541 The influence of the OOTL phenomenon has often been assessed in terms of its consequences in a take-  
542 over situation. Drivers who are OOTL usually react relatively late and inefficiently when a critical event  
543 occurs, especially after a long time spent in automation (Neubauer et al., 2012; Gold et al., 2013; Saxby  
544 et al., 2013; Louw et al., 2015b; Zeeb et al., 2015, 2017; Eriksson and Stanton, 2017, Bourrelly et al.,  
545 2019). More direct assessment of the OOTL phenomenon may also be performed using post-trial  
546 questionnaires (see for instance Lu et al., 2017, for an interesting approach in evaluating situation  
547 awareness). In the present study, the drivers' state was assessed by asking the participants to report their  
548 level of MW. The results showed that drivers with the highest level of MW all failed to take over  
549 adequately, which implies that they were indeed OOTL. Nevertheless, failure to correctly manage the  
550 critical situation also occurred for two participants, who reported only a moderate level of MW. This  
551 finding suggests that the quality of a take-over is not entirely determined by the degree of OOTL before  
552 the take-over request. The skills of the driver to quickly recover situation awareness and to handle the  
553 vehicle in the time allocated may also be essential aspects.

554

555 A last point worth mentioning is that the MW scores were not correlated to PERCLOS and blink-rate  
556 indicators, which have been shown to be highly predictive of drowsiness (Jacobé de Naurois et al., 2019)  
557 . The two indicators remained lower than values typically associated with drowsy states during driving.  
558 (PERCLOS of >12.5% was used to categorize drowsy drivers in Hanowski et al. 2008). Thus, we  
559 concluded that the MW score did not reflect driver fatigue in our experiments. This result might be  
560 different for an extended period of automated driving, during which participants might fall asleep  
561 (Vogelpohl et al., 2019). This was observed by Bourrelly et al. (2018) in a 1-h test run of automated  
562 driving. It remains to be determined whether the multiple PLS approach we propose can discriminate  
563 between the two phenomena. The value of the PLS approach to realize an integrated diagnostic must  
564 also be determined.

## 565 5. Conclusion

566

567 We investigated whether drivers' gaze behaviour could be used to detect the OOTL phenomenon during  
568 automated driving. The results indicate that the gaze dynamics appear to be a crucial point: being OTL  
569 required frequent gaze shifts to the road while also obtaining tactical information about the oncoming  
570 situation.

571

572 It remains to be determined whether this conclusion can be generalized to other conditions under which  
573 the OOTL phenomenon may occur. It was induced in this study by a relatively monotonous driving task  
574 in the absence of external distraction. However, being out of the loop may also be due to the engagement  
575 of attention in a secondary task, an expected consequence of Level 3 automation. In this case, changes  
576 in driver behaviour will most likely be characterised by a massive redirection of attention to external  
577 displays. It would be interesting to assess whether dynamic indicators will be essential to improve the  
578 estimation of the driver's state or whether static indicators, such as the time spent without looking at the  
579 road, are sufficient. The method used in this study could help to answer this question.

580

581 To more accurately detect the OOTL phenomenon during automated driving, the analysis of gaze  
582 behaviour could perhaps be coupled with other approaches. For example, physiological measurements  
583 or the analysis of the driver's posture could be incorporated in the diagnosis.

## 584 6. Declaration of Competing Interest

585 The authors declare that they have no known competing financial interests or personal relationships that  
586 could have appeared to influence the work reported in this paper.

## 587 7. Acknowledgments

588 The authors would like to thank Denis Creusot for his invaluable help on the driving simulator, as well  
589 as the participants involved in the experiment. This study was supported by the French National  
590 Research Agency (Agence Nationale de la Recherche, AUTOCONDUCT project, grant n°ANR-16-  
591 CE22-0007-05).

## 592 8. References

593 Abdi, H. (2010). Partial least squares regression and projection on latent structure regression

594 (PLS Regression): PLS REGRESSION. *WIREs Comp Stat* 2, 97–106. doi:10.1002/wics.51.

595 Bourrelly, A., Jacobé de Naurois, C., Zran, A., and others (2018). Gaze behavior during take-

596 over after a long period of autonomous driving: A pilot study'. in *Proc. Int. Conf. Driving*

597 *Simulation Conference Europe VR, Antibes, France*.

598 Bourrelly, de Naurois, C. J., Zran, A., Rampillon, F., Vercher, J.-L., and Bourdin, C. (2019).

599 Long automated driving phase affects take-over performance. *IET Intelligent Transport*

600 *Systems*. doi:10.1049/iet-its.2019.0018.

601 Burdett, B. R. D., Charlton, S. G., and Starkey, N. J. (2019). Mind wandering during everyday

602 driving: An on-road study. *Accident Analysis & Prevention* 122, 76–84.

603 doi:10.1016/j.aap.2018.10.001.

604 Carsten, O., Lai, F. C., Barnard, Y., Jamson, A. H., and Merat, N. (2012). Control task

605 substitution in semiautomated driving: Does it matter what aspects are automated? *Human*

606 *Factors* 54, 747–761. doi:10.1177/0018720812460246.

607 Chan, C.-Y. (2017). Advancements, prospects, and impacts of automated driving systems.

608 *International Journal of Transportation Science and Technology* 6, 208–216.

609 doi:10.1016/j.ijst.2017.07.008.

610 Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human*

611 *Factors* 37, 32–64.

612 Endsley, M. R., and Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and

613 Level of Control in Automation. *Human Factors* 37, 381–394.

614 doi:10.1518/001872095779064555.

615 Eriksson, A., and Stanton, N. A. (2017). Driving performance after self-regulated control

616 transitions in highly automated vehicles. *Human Factors* 59, 1233–1248.

617 doi:10.1177/0018720817728774.

618 Feldhütter, A., Gold, C., Schneider, S., and Bengler, K. (2017). How the duration of

619 automated driving influences take-over performance and gaze behavior. *Advances in*

620 *Ergonomic Design of Systems, Products and Processes*, 309–318. doi:10.1007/978-3-662-

621 53305-5\_22.

622 Fitch, G. M., Bowman, D. S., and Llaneras, R. E. (2014). Distracted driver performance to

623 multiple alerts in a multiple-conflict scenario. *Human Factors* 56, 1497–1505.

624 doi:10.1177/0018720814531785.

625 Gold, C., Damböck, D., Bengler, K., and Lorenz, L. (2013). Partially automated driving as a

626 fallback level of high automation. in *6. Tagung Fahrerassistenzsysteme. Der Weg zum*

627 *automatischen Fahren*. (TÜV SÜD Akademie GmbH).

628 Gonçalves, R., Louw, T., Madigan, R., and Merat, N. (2019). Using markov chains to

629 understand the sequence of drivers' gaze transitions during lane-changes in automated

630 driving. in *Proceedings of the International Driving Symposium on Human Factors in Driver*

631 *Assessment, Training, and Vehicle Design* (Leeds).

632 Gouraud, J., Delorme, A., and Berberian, B. (2017). Autopilot, mind wandering, and the out

633 of the loop performance problem. *Frontiers in Neuroscience* 11, 541.

634 doi:10.3389/fnins.2017.00541.

635 Hanowski, R. J., Bowman, D., Alden, A., Wierwille, W. W., and Carroll, R. (2008).

636 PERCLOS+: Development of a robust field measure of driver drowsiness. in *15th World*

637 *Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting, New*

638 York, NY.

639 Jacobé de Naurois, C., Bourdin, C., Stratulat, A., Diaz, E., and Vercher, J.-L. (2019).  
640 Detection and prediction of driver drowsiness using artificial neural network models. *Accident*  
641 *Analysis & Prevention* 126, 95–104. doi:10.1016/j.aap.2017.11.038.

642 Körber, M., Cingel, A., Zimmermann, M., and Bengler, K. (2015). Vigilance decrement and  
643 passive fatigue caused by monotony in automated driving. in *6th International Conference on*  
644 *Applied Human Factors and Ergonomics*, 2403–2409.

645 Louw, Kountouriotis, G., Carsten, O., and Merat, N. (2015a). Driver Inattention During  
646 Vehicle Automation: How Does Driver Engagement Affect Resumption Of Control? in *4th*  
647 *International Conference on Driver Distraction and Inattention*, 1–16.

648 Louw, Madigan, R., Carsten, O., and Merat, N. (2016). Were they in the loop during  
649 automated driving? Links between visual attention and crash potential. *Injury Prevention* 23,  
650 281–286. doi:10.1136/injuryprev-2016-042155.

651 Louw, and Merat, N. (2016). A Methodology for Inducing the Out of the Loop Phenomenon  
652 in Highly Automated Driving. in *International Conference on Traffic and Transport*  
653 *Psychology, Brisbane, Australia*.

654 Louw, Merat, N., and Jamson, H. (2015b). Engaging with Highly Automated Driving: To be  
655 or Not to be in the Loop? in *Proceedings of the Eighth International Driving Symposium on*  
656 *Human Factors in Driver Assessment, Training and Vehicle Design*, 190–196.  
657 doi:[https://doi.org/ 10.17077/drivingassessment.1570](https://doi.org/10.17077/drivingassessment.1570).

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

661 Lu, Z., Coster, X., and de Winter, J. (2017). How much time do drivers need to obtain  
662 situation awareness? A laboratory-based study of automated driving. *Applied Ergonomics* 60,  
663 293–304. doi:10.1016/j.apergo.2016.12.003.

664 Mackenzie, A. K., and Harris, J. M. (2015). Eye movements and hazard perception in active  
665 and passive driving. *Visual cognition* 23, 736–757. doi:10.1080/13506285.2015.1079583.

666 Mars, F., Deroo, M., and Charron, C. (2014). Driver adaptation to haptic shared control of the  
667 steering wheel. in *2014 IEEE International Conference on Systems, Man, and Cybernetics*  
668 *(SMC) (IEEE)*, 1505–1509. doi:10.1109/SMC.2014.6974129.

669 Mars, F., and Navarro, J. (2012). Where we look when we drive with or without active  
670 steering wheel control. *PLoS One* 7, e43858. doi:10.1371/journal.pone.0043858.

671 Merat, N., Jamson, A. H., Lai, F. C. H., and Carsten, O. (2012). Highly Automated Driving,  
672 Secondary Task Performance, and Driver State. *Human Factors* 54, 762–771.  
673 doi:10.1177/0018720812442087.

674 Merat, N., Seppelt, B., Louw, T., Engström, J., Lee, J. D., Johansson, E., et al. (2019). The  
675 “out-of-the-loop” concept in automated driving: Proposed definition, measures and  
676 implications. *Cognition, Technology & Work* 21, 87–98. doi:10.1007/s10111-018-0525-8.

677 Mole, C. D., Lappi, O., Giles, O., Markkula, G., Mars, F., and Wilkie, R. M. (2019). Getting  
678 Back Into the Loop: The Perceptual-Motor Determinants of Successful Transitions out of  
679 Automated Driving. *Human Factors* 61, 1037–1065. doi:10.1177/0018720819829594.

680 Molloy, R., and Parasuraman, R. (1996). Monitoring an automated system for a single failure:  
681 Vigilance and task complexity effects. *Human Factors* 38, 311–322.

682 doi:10.1177/001872089606380211.

683 Navarro, J., François, M., and Mars, F. (2016). Obstacle avoidance under automated steering:  
684 Impact on driving and gaze behaviours. *Transportation Research Part F: Traffic Psychology*  
685 *and Behaviour* 43, 315–324. doi:10.1016/j.trf.2016.09.007.

686 Neubauer, C., Matthews, G., Langheim, L., and Saxby, D. (2012). Fatigue and voluntary  
687 utilization of automation in simulated driving. *Human Factors* 54, 734–746.  
688 doi:10.1177/0018720811423261.

689 Parasuraman, R., and Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse,  
690 Abuse. *Human Factors* 39, 230–253. doi:10.1518/001872097778543886.

691 SAE International (2016). Taxonomy and Definitions for Terms Related to On-Road Motor  
692 Vehicle Automated Driving Systems. Washington, DC: SAE International.

693 Saxby, D. J., Matthews, G., Warm, J. S., Hitchcock, E. M., and Neubauer, C. (2013). Active  
694 and Passive Fatigue in Simulated Driving: Discriminating Styles of Workload Regulation and  
695 Their Safety Impacts. *J Exp Psychol Appl* 19, 287–300. doi:10.1037/a0034386.

696 Schnebelen, D., Lappi, O., Mole, C., Pekkanen, J., and Mars, F. (2019). Looking at the Road  
697 When Driving Around Bends: Influence of Vehicle Automation and Speed. *Front. Psychol.*  
698 10. doi:10.3389/fpsyg.2019.01699.

699 Sivak (1996). The information that drivers use: is it indeed 90% visual? *Perception* 25, 1081–  
700 1089. doi:10.1068/p251081.

701 Stanton, N. A., and Salmon, P. M. (2009). Human error taxonomies applied to driving: A  
702 generic driver error taxonomy and its implications for intelligent transport systems. *Safety*  
703 *Science* 47, 227–237. doi:10.1016/j.ssci.2008.03.006.

704 Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue.  
705 *Human Factors* 36, 285–297. doi:10.1177/001872089403600209.

706 Team, R. C., and others (2013). R: A language and environment for statistical computing.  
707 Victor (2005). Keeping eye and mind on the road. *Digital Comprehensive Summaries of*  
708 *Uppsala Dissertations from the Faculty of Social Sciences*, 83.

709 Victor, Harbluk, J. L., and Engström, J. A. (2005). Sensitivity of eye-movement measures to  
710 in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*  
711 8, 167–190. doi:10.1016/j.trf.2005.04.014.

712 Vogelpohl, T., Kühn, M., Hummel, T., and Vollrath, M. (2019). Asleep at the automated  
713 wheel—Sleepiness and fatigue during highly automated driving. *Accident Analysis &*  
714 *Prevention* 126, 70–84. doi:10.1016/j.aap.2018.03.013.

715 Wehrens, R., and Mevik, B.-H. (2007). The pls Package: Principal Component and Partial  
716 Least Squares Regression in R. *Journal of Statistical Software* 18, 1–24.

717 Wierwille, W. W., Wreggit, S., Kim, C., Ellsworth, L., and Fairbanks, R. (1994). Research on  
718 vehicle-based driver status/performance monitoring; development, validation, and refinement  
719 of algorithms for detection of driver drowsiness. *National Highway Traffic Safety*  
720 *Administration Final Report*.

721 Zeeb, K., Buchner, A., and Schrauf, M. (2015). What determines the take-over time? An  
722 integrated model approach of driver take-over after automated driving. *Accident Analysis &*  
723 *Prevention* 78, 212–221. doi:10.1016/j.aap.2015.02.023.

724 Zeeb, K., Härtel, M., Buchner, A., and Schrauf, M. (2017). Why is steering not the same as  
725 braking? The impact of non-driving related tasks on lateral and longitudinal driver

726 interventions during conditionally automated driving. *Transportation Research Part F:*  
727 *Traffic Psychology and Behaviour* 50, 65–79. doi:10.1016/j.trf.2017.07.008.  
728

## 729 9. Appendix 1: Step-by-step procedure of PLS 730 regression

731 This appendix develops step-by-step the procedure to predict the MW score (Y) using PLS regressions.  
732 We used a training dataset (matrix of gaze behaviour computed over a given time window t,  $X_t$ ) and a  
733 validation dataset (matrix of gaze behaviour computed over the two final minutes of automated driving,  
734  $X_{val}$ ).  
735 Steps A and B were performed with the training datasets and enabled computing the optimal parameters  
736 (number of components and relevant visual indicators) of the prediction models. With that configuration,  
737 the accuracy of the model was tested for both the training and validation datasets (Step C).

### 738 Step A: Calculating the optimal number of components

739

#### 740 Principle

741

742 PLS models are based on several orthogonal components, which constitute the underlying structure of  
743 the prediction model. With many components, the model will be complex and highly accurate but also  
744 very specific of the data. By contrast, few components will mean a simpler model structure. The model  
745 may lose accuracy but may be more generalizable to other datasets. Thus, an optimal compromise in the  
746 number of components can prevent data overfitting while maintaining high accuracy.

747

748

#### 749 Application

750

751 This compromise was sought by testing several numbers of components (from one to 10 components).  
752 The optimal number of components would minimize the mean square error of prediction (MSEP), with  
753 a leave-one-out procedure.

## 754 Step B: Reducing the number of visual indicators

755 In the previous step, an optimal structure of the prediction model was found, considering all possible  
756 visual indicators (182) to predict the MW score. The aim of this next step was to increase the predictive  
757 power (i.e. the percentage of variance in Y that was explained) by selecting fewer indicators.

758

### 759 Principle

760

761 Because PLS regression is a linear statistical model, the relationship between the training dataset ( $X_t$ )  
762 and the dependent variable to estimate  $\widehat{Y}_{t, \text{train}}$  was linear:

$$763 \quad \widehat{Y}_{t, \text{train}} = X_t * C_t$$

764 where  $C_t$  is the matrix of the regression coefficients.

765

766 Coefficients can be interpreted as follows:

- 767 - Coefficient signs indicate the direction in which a visual indicator (from  $X_t$ ) influenced the  
768 estimation of the MW score ( $\widehat{Y}_{t, \text{train}}$ ). If positive, the MW score increased. By contrast, a  
769 negative coefficient meant that the MW score decreased.
- 770 - A coefficient's magnitude (absolute value) indicates the importance of each indicator relative.  
771 If the magnitude of a coefficient was close to zero, the contribution of this visual indicator to  
772 the prediction would be negligible. By contrast, a large magnitude indicated a crucial indicator  
773 in the prediction.

774

### 775 Application

776

777 To reduce the number of visual indicators of  $X_t$ , coefficient magnitudes were compared with an  
778 increasing threshold value, which ranged from 0.01 to 0.2. A new PLS regression was computed for  
779 each partial matrix (i.e. a matrix comprising only the indicators whose coefficient magnitude exceeded  
780 the threshold value). The threshold was increased by steps of 0.005 until the percentage of variance in  
781 Y that was explained by the partial model no longer increased.

782

783 At the end of this step, a partial matrix of gaze behaviour, which included only the selected visual  
784 indicators, was computed for the training and validation datasets. These partial matrixes are denoted  $X_t^p$   
785 and  $X_{\text{val}}^p$  for the training and validation datasets, respectively.

## 786 Step C: Computing the mean square error of prediction

787 The previous steps found the most appropriate parameters for PLS regression models (number of  
788 components in Step A and relevant visual indicators in Step B) for predicting the MW score. A new  
789 model that considered those parameters was set up, with its coefficients denoted  $C_t^p$ . Then, the  
790 estimations from the training and validation datasets for a given time window  $t$  were calculated as  
791 follows:

$$792 \left\{ \begin{array}{l} \widehat{Y_{t, \text{train}}^p} = X_t^p * C_t^p \\ \widehat{Y_{t, \text{val}}^p} = X_{\text{val}}^p * C_t^p \end{array} \right.$$

794

795 From those estimations, the mean square error of prediction was:

796

$$797 \left\{ \begin{array}{l} \text{MSEP}_t^{\text{train}} = \overline{(Y - \widehat{Y_{t, \text{train}}^p})^2} \\ \text{MSEP}_t^{\text{val}} = \overline{(Y - \widehat{Y_{t, \text{val}}^p})^2} \end{array} \right.$$

798

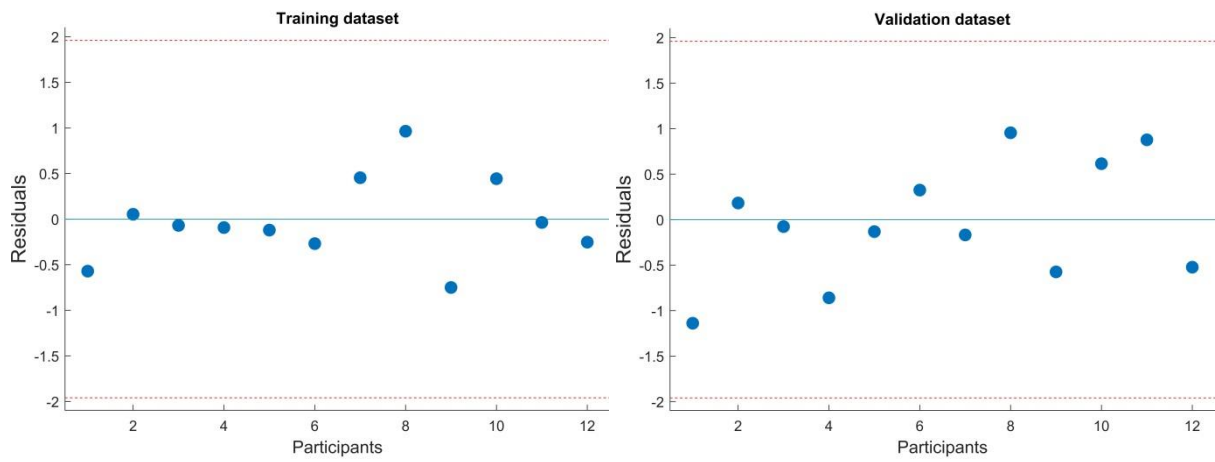
799 where  $C_t^p$  are the coefficients of the PLS regression, computed from the partial matrixes with the optimal  
800 number of components for time window  $t$ ; and the terms  $X_t^p$  and  $X_{\text{val}}^p$  refer to the partial matrixes of gaze  
801 behaviour for the training and validation datasets, respectively.

802

## 803 10. Appendix 2: goodness-of-fit based on the model 804 residuals

805

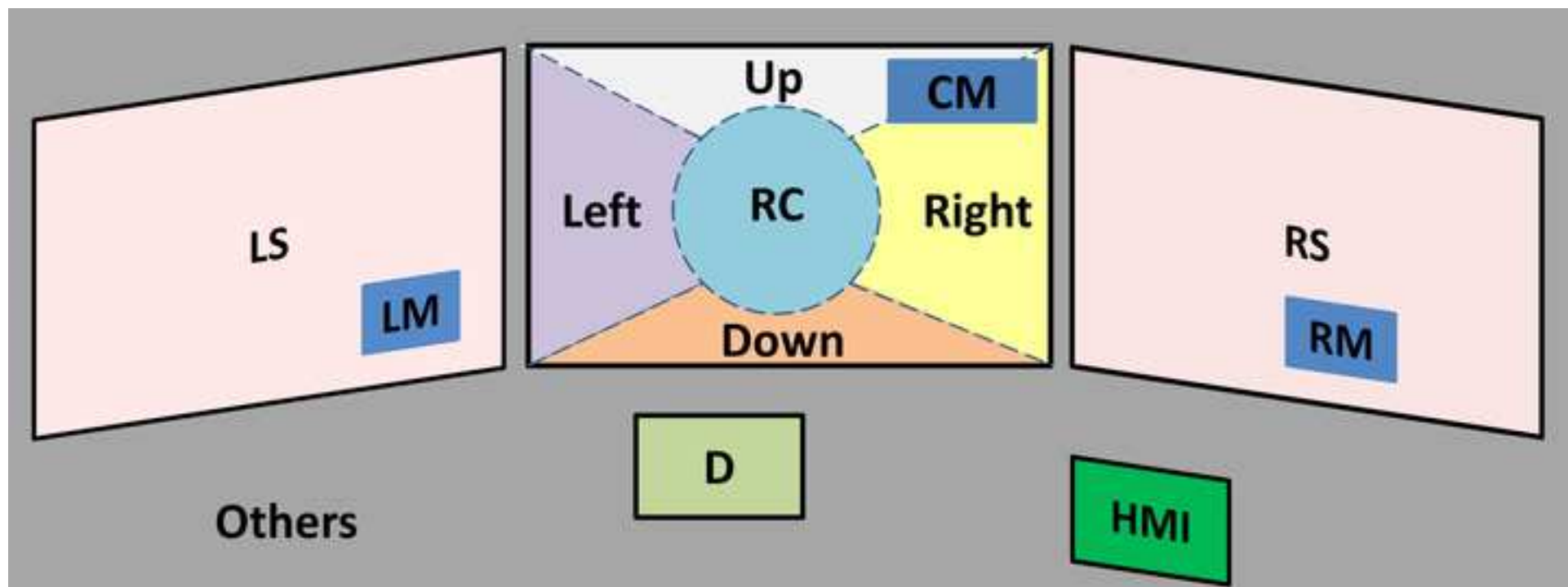
806 Due to the standardization of the data, MW score may be interpreted as Z-scores. Consequently, the  
807 residuals are significantly different from 0 with a 95% confidence level if their value is outside the  
808 range [-1.96, 1.96]. Figure 8 shows that residuals for each different participant are within this range for  
809 both the validation and training datasets. This demonstrates that the model is well-fitted for all  
810 participants in all cases.

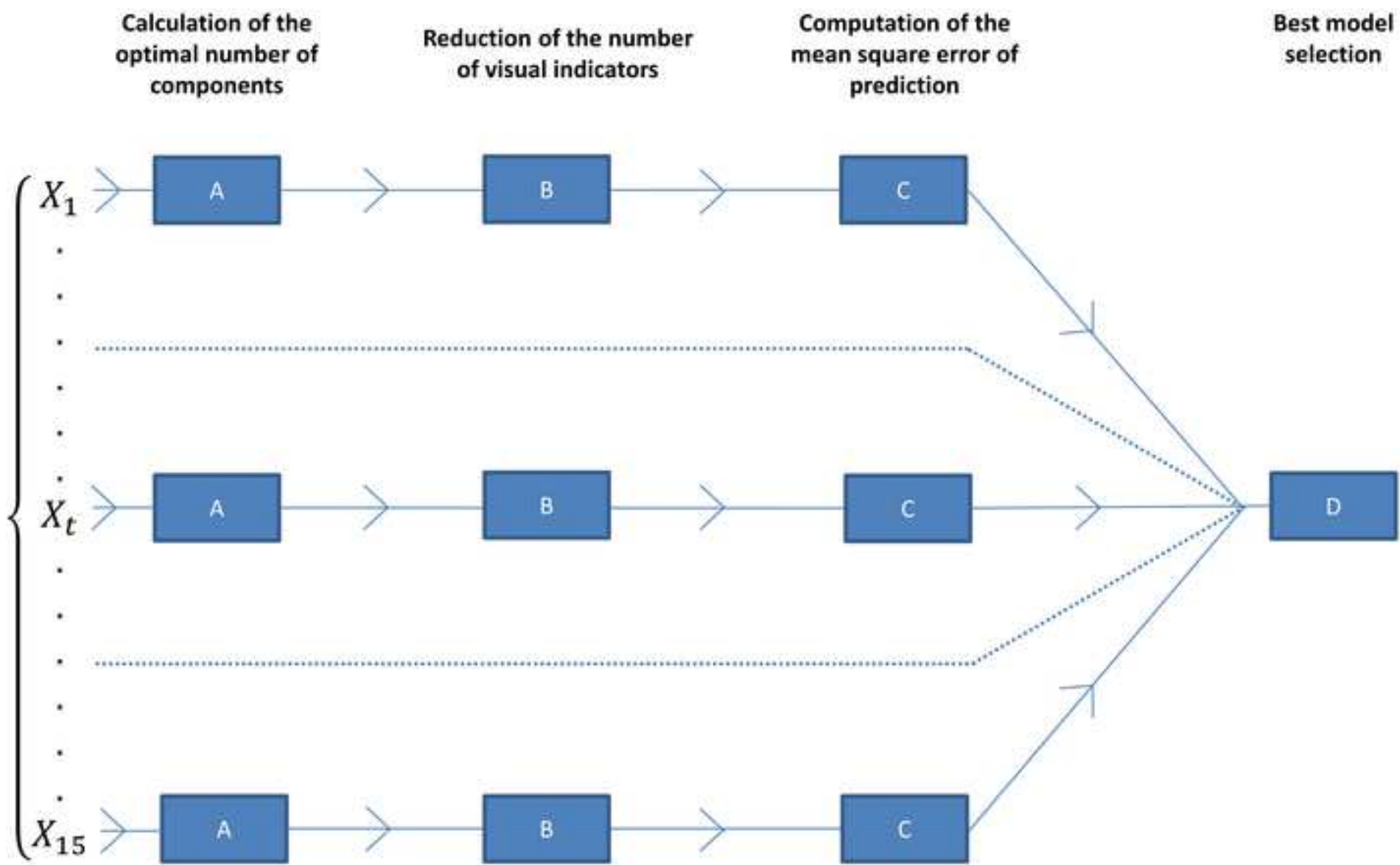


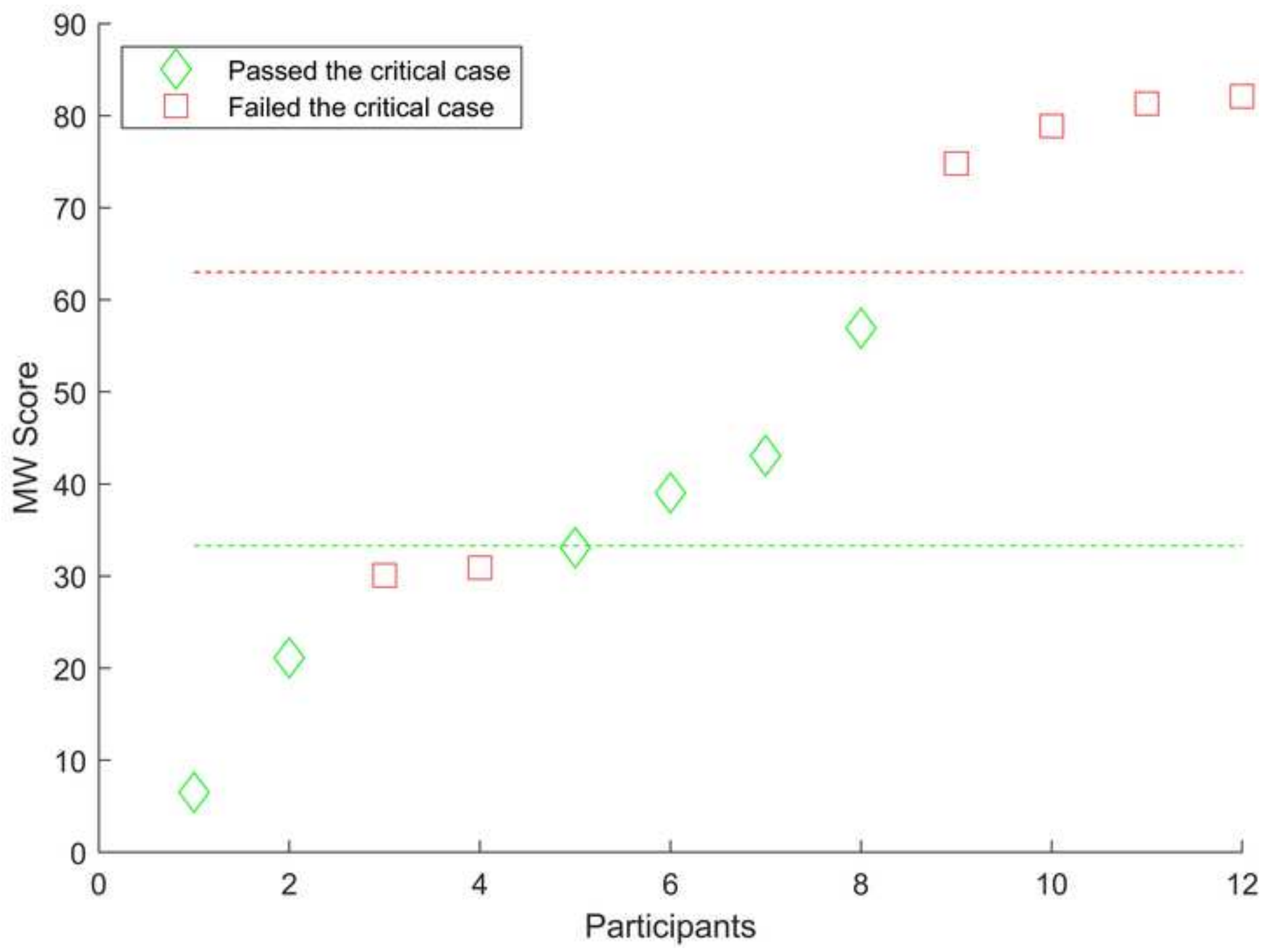
811  
 812 Figure 8: Residuals from the PLS regression models for the training (left) or validation (right) datasets.  
 813 All residuals are within the confidence interval  $[-1.96, 1.96]$  (red dotted lines).  
 814

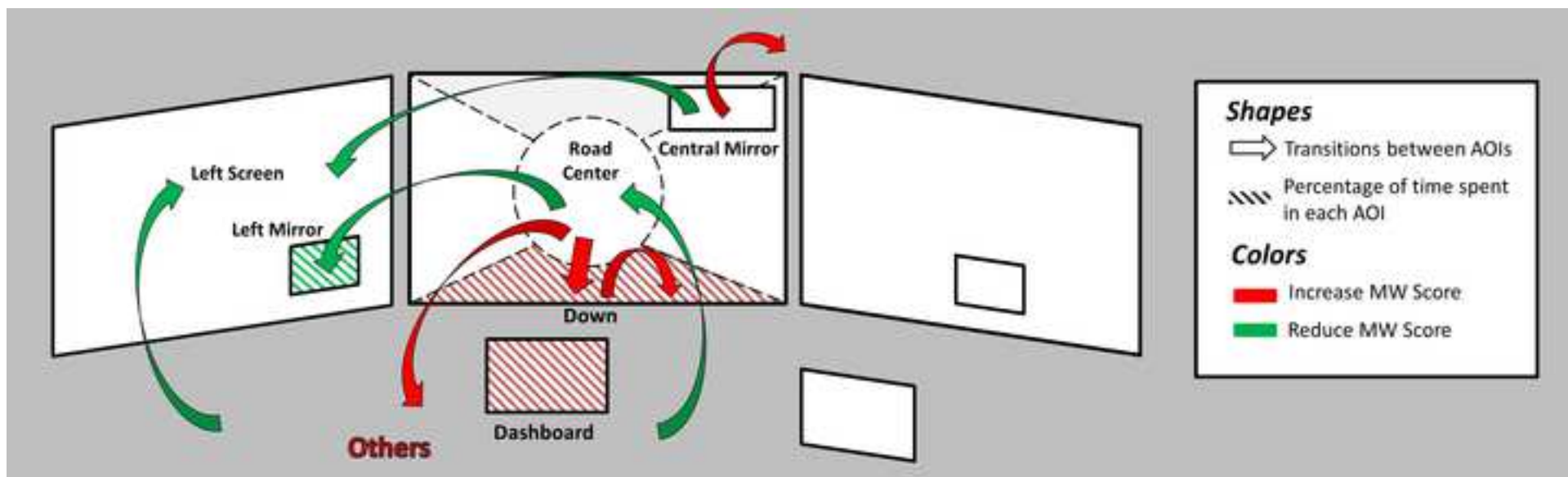


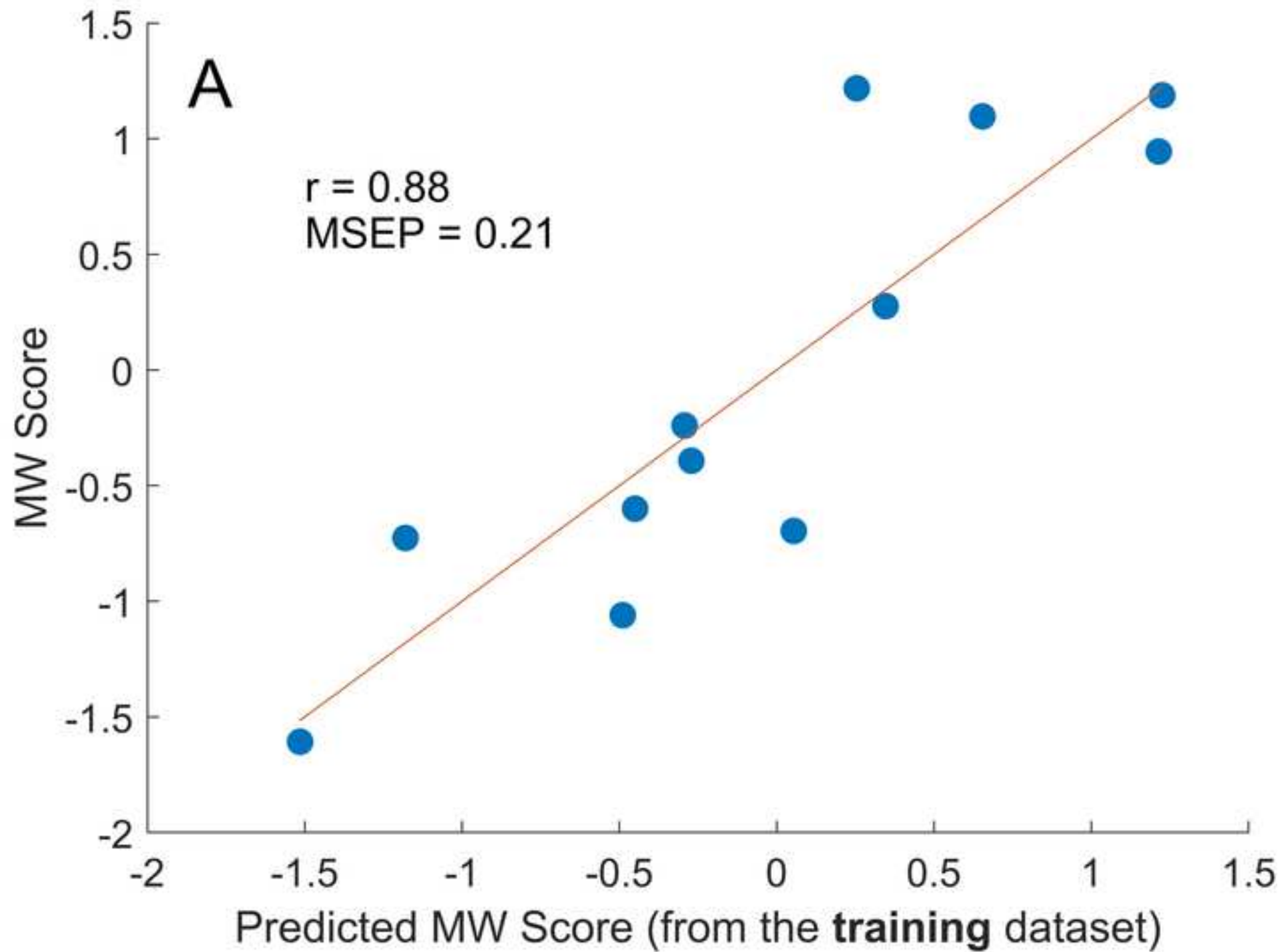


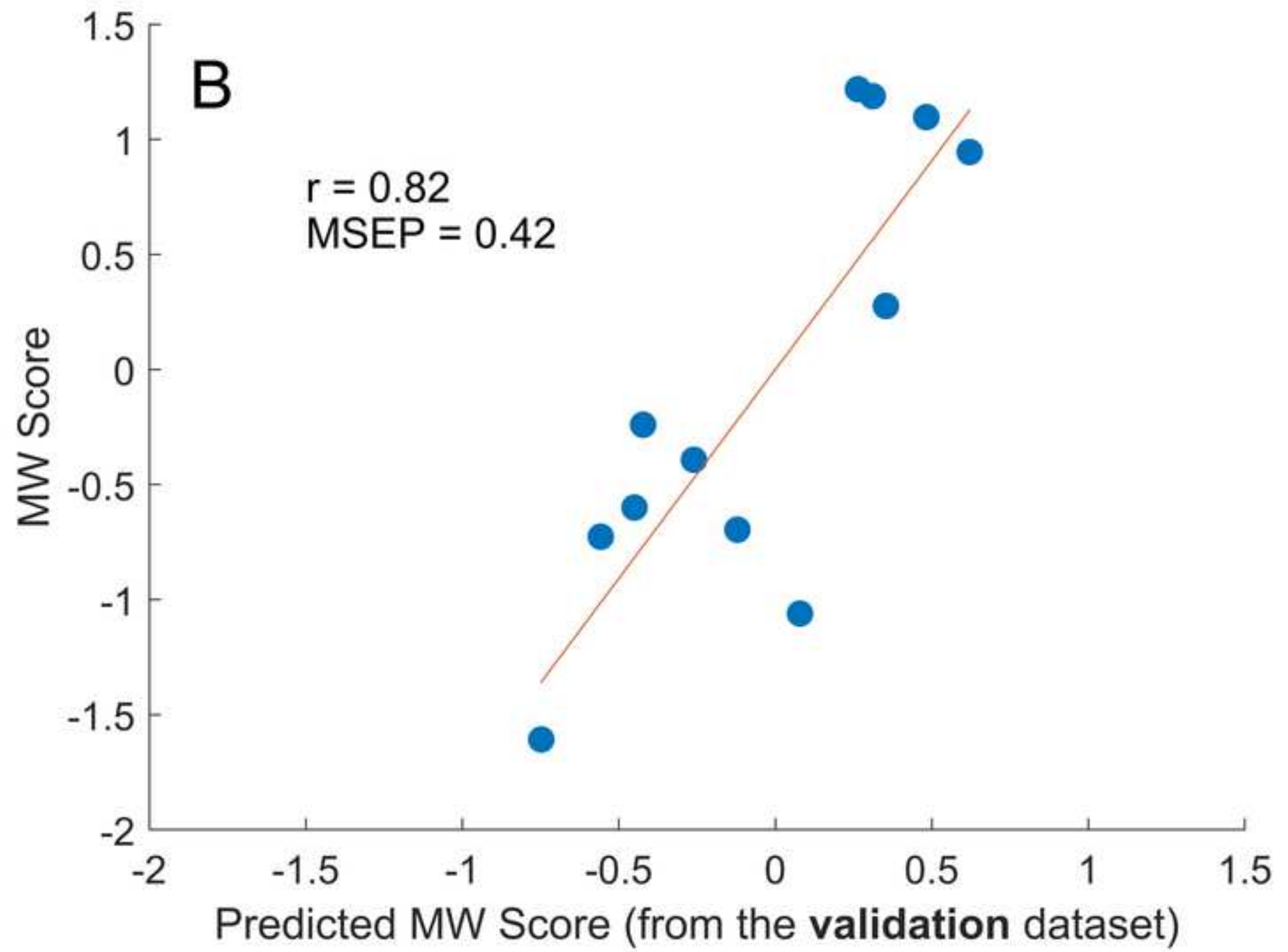




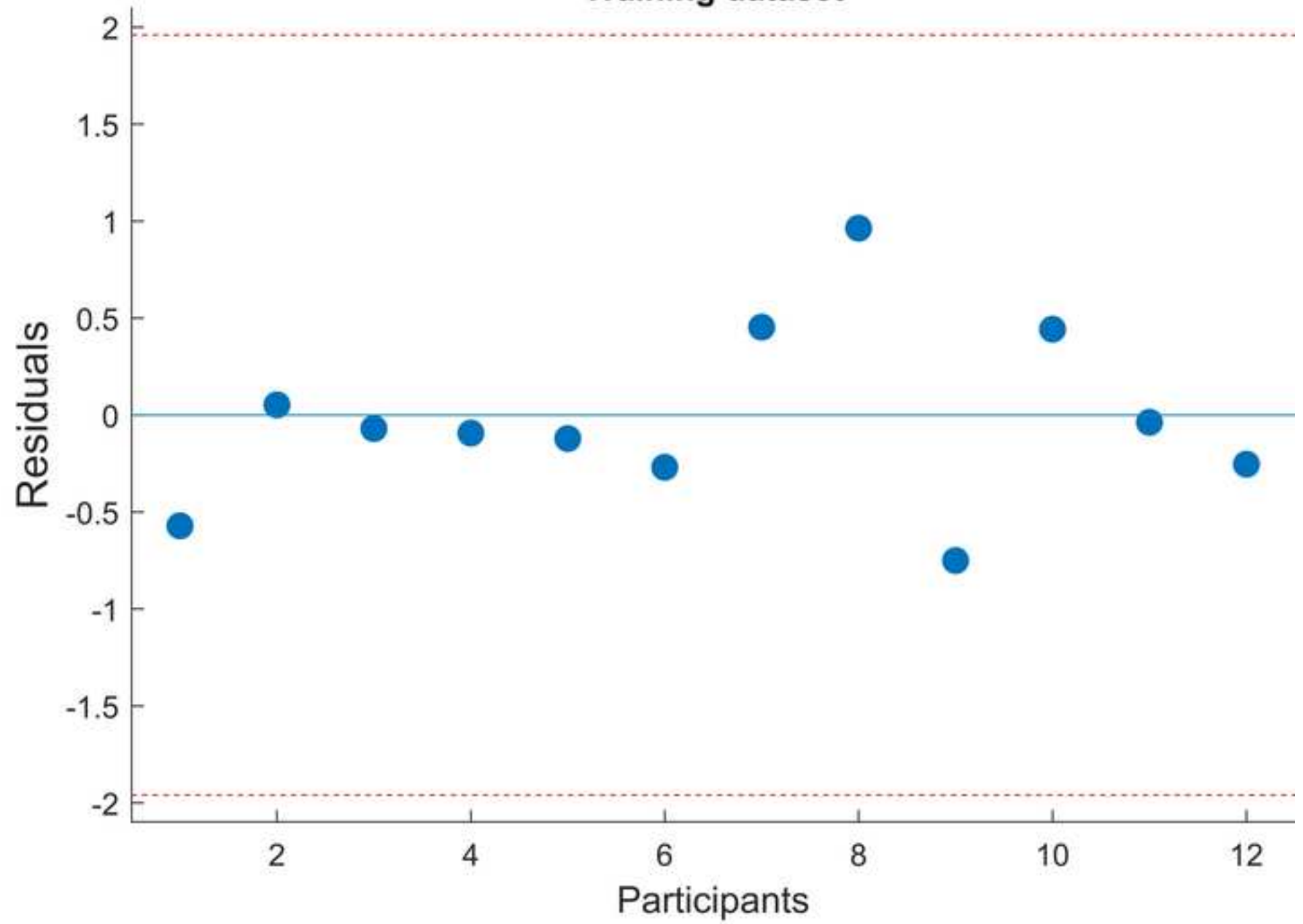




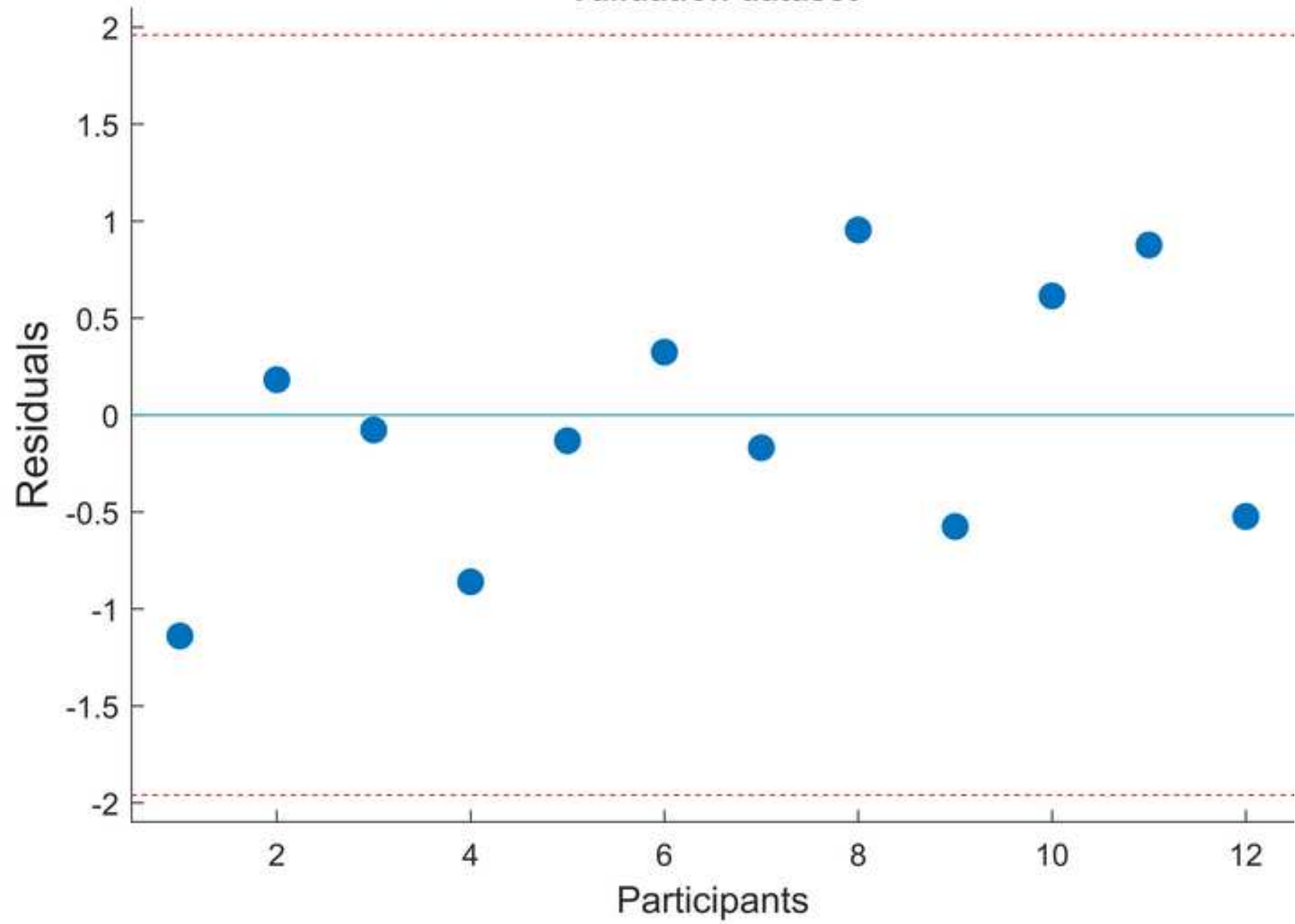




### Training dataset



### Validation dataset



The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Damien Schnebelen:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Visualization **Camilo Charron:** Conceptualization, Methodology, Formal analysis, Writing – review and editing **Franck Mars:** Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Supervision, Project administration, Funding acquisition