Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2023 Annual Conference

de Waard, Dick ; Hagemann, Vera; Onnasch, Linda; Toffetti, Antonella; Coelho, Denis; Botzer, Assaf; De Angelis, Marco; Brookhuis, Karel; Fairclough, Stephen

Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2023 Annual Conference

Edited by

Dick de Waard, Vera Hagemann, Linda Onnasch, Antonella Toffetti, Denis Coelho, Assaf Botzer, Marco de Angelis, Karel Brookhuis, and Stephen Fairclough
ISSN 2333-4959 (online)

Contents

AUTOMATION & UX

HUMAN FACTORS IN HEALTHCARE

WORKLOAD

Assessment of a hybrid cursor control device for highly dynamic cockpit environments

Engelbert Hartmann
Airbus Defence and Space GmbH – Human Factors Engineering and Flight Deck
Germany

Abstract

In today's fifth generation fighter aircraft, the human-machine interaction is largely based on cursor control devices. The challenging complexity of future missions suggests the need to further increase the efficiency of interaction modalities. This work proposed a hybrid cursor control device, where head-control is used for long distance coarse aiming, followed by fine control of the cursor with a trackball in the final phase of the approach. A research prototype was evaluated to answer the principal question, whether the hybrid concept could provide performance advantages over the respective single modalities in a dynamic cockpit environment. The experimental study exploited a within-subjects design with 12 fighter pilots in a laboratory setup, using a motion platform to simulate various flight dynamic conditions. Throughput was used as a single measure for performance, while the variation of pointing distance refined the results. Furthermore, user behaviour, workload and subjective attitudes towards the use of the modalities were captured to supplement the comparative evaluation. Overall it was concluded, that the hybrid cursor control modality can provide significant advantages for long distance point and select tasks under light atmospheric turbulences.

Introduction

Human-machine interfaces of modern combat aircraft rely largely on display interaction, which is physically enabled by cursor control devices. Kellerer et al. (2018) proposed a control concept for fighter cockpits comprising a touch-sensitive display combined with a trackball. On this basis, a functional research prototype was produced, implementing a miniature trackball in a throttle control lever. Trackballs are capable of high movement velocities (Stanton et al., 2013), partially achieved by spinning the trackball for large distance movements of the cursor. This advantage was traded off by the miniaturization of the trackball.

Helmet mounted displays are implemented in many modern cockpits and will remain relevant in future (Bibb, 2021). Although head-tracking technology is incorporated in helmet display implementations, head control is not used as a cursor control device for head-down display interaction. Smith et al. (2008) described impairing effects of turbelences on cognitive performance, head tracking, and manual control. Thomas et al. (2015) concluded that vibration and turbulence stabilization systems and algorithms are needed to make head cursor control feasible.

This study addresses the limitations of the prototyped device and proposes a hybrid cursor control concept for long distance pointing tasks. The concept is based on a two-phase control model (Lin et al., 1992), where head control is used for large distance coarse aiming in the ballistic phase, followed by fine control of the cursor with the trackball in the terminal phase (see figure 1). The proposition is made that the hybrid concept can provide performance advantages for large distance pointing tasks, when compared to the single modalities of head control and trackball. Consequently, the work is relevant for current and future helmet mounted displays and their potential use as cursor control device.
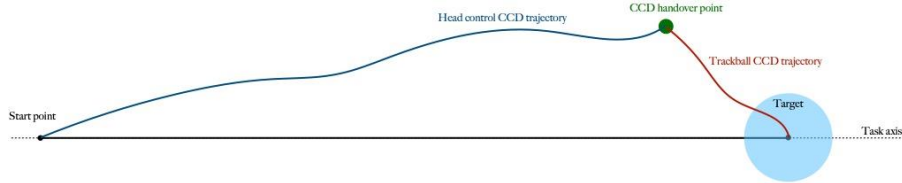
Figure 1. Illustration of a pointing task with the proposed hybrid cursor control device. The cursor control modality is changed at the handover point from head control to trackball.

Zhai et al. (1999) found a non-logarithmic increase in task completion times with increasing target distances, when using the hybrid concept of manual and gaze input cascaded (MAGIC) pointing. In contrast, Drewes and Schmidt (2009) and Stellmach and Dachselt (2012) found no dependency between target distance and task completion times, despite using different manual devices. Furthermore, the advantages of MAGIC control appear to be more evident for large distances. This suggests that pointing task completion times for the proposed hybrid cursor control device might be unaffected by target distance and could outperform the trackball in long distance point and select tasks. Biswas and Langdon (2015) found lower workload for a multimodal cursor control device in contrast to manual cursor control with the joystick only.

The objective of this work is to evaluate the proposed hybrid cursor control concept, comparing it to the single modalities under varying flight dynamic conditions regarding pointing performance, user acceptance, user behaviour, and workload. The variation of pointing distance shall refine the results. In summary, this leads to the following research questions:

> Is there a difference in performance (RQ1) and/or workload (RQ2) between the hybrid and single modalities and how is it affected by various flight dynamic conditions and pointing distances?

> What is the users' attitude towards the hybrid concept compared to the single modalities (RQ3)?

Methods and tools

Participants
Participants have been recruited from a pool of experienced fighter pilots, as their expertise, subjective attitudes and familiarity with flight dynamics were considered

highly valuable. Six active and six retired aircrew (N = 12; 11 males, 1 female) participated in the experiment, with an average flight experience of 3190 hours (SD = 2140) in the aircraft types Eurofighter Typhoon, F-4 Phantom II, F/A-18 Hornet, Mirage 2000 D/N, and Tornado IDS/ECR. Five volunteers had flight test experience. All participants attended voluntarily without compensation payments, reported no health issues or impaired vision.

The sudy was conducted according to the guidelines of the Declaration of Helsinki and approved by the ethics committee of the University of Nottingham. Informed consent was obtained from all participants before experiments.

Equipment

The experiment setup was installed on a motion platform, with pointing tasks displayed on a 32" screen as depicted in figure 2.
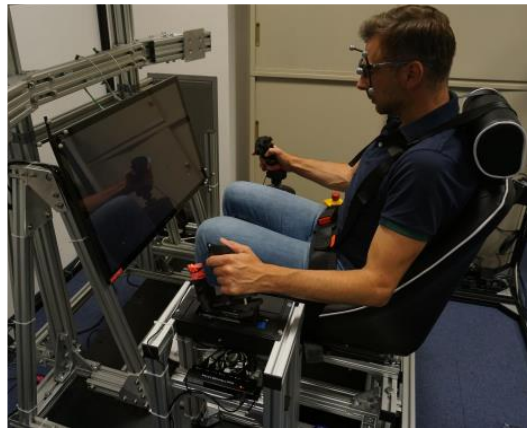
Figure 2. Overview of the experiment setup (Hartmann, 2022).

The head cursor was controlled by a head wearable target and an optical tracking system. To simulate a helmet mounted display, the head cursor gain was unaccelerated and set to one. Trackball gain enabled an accelerated linear cursor movement of 120 mm (slow) to 180 mm (fast). Workload ratings were collected with the NASA TLX application (NASA, n.d.). The acceptance questionnaire was conducted as a pen and paper version.

Independent variables

Cursor control modalitiy alternatives
The proposed hybrid cursor control concept offers different modes, which allows the user to choose the modality. Figures 3 and 4 illustrate the associated control elements and symbology.

(a) Throttle - TB   (b) Throttle - WB   (c) Flight Stick - ATB/IB
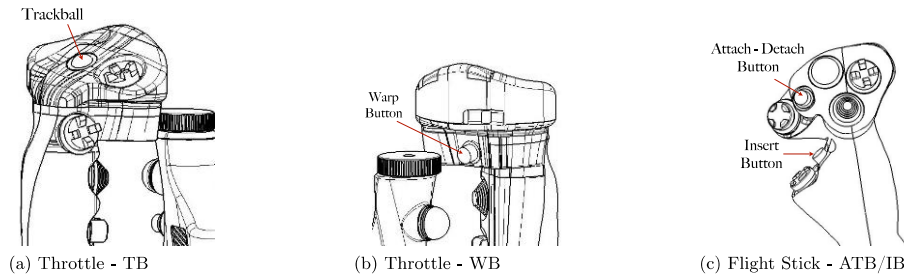
Figure 3. Overview of the control elements (Hartmann, 2022).

- The concept is based on two cursors, a trackball cursor, controlled by the trackball (TB - see figure 3a), and a head cursor, controlled by head tracking.
- A discrete insert button (IB - see figure 3c) is tied to the trackball cursor, which means inserts (selections) can only be made with the trackball cursor.
- A discrete warp button (WB - see figure 3b) is used to warp the trackball cursor to the current position of the head cursor with an instant motion, upon press of the warp button.
- A discrete attach-detatch button (ATB - see figure 3c), which is used to attach and detach the trackball cursorand cursor control device. The attached cursors are moved by head control only.
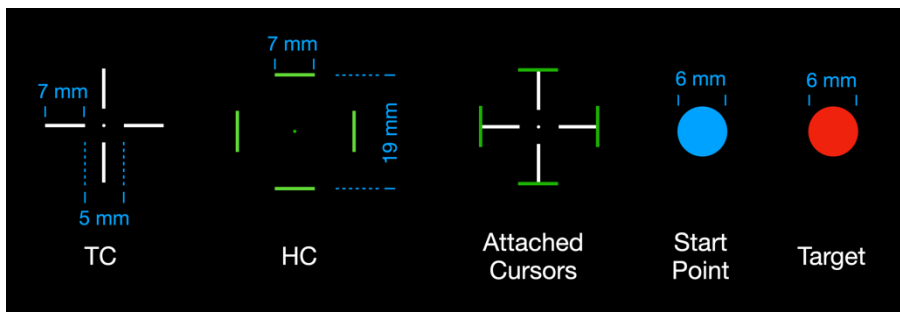
Figure 4. Illustration of symbology used for testing (Hartmann, 2022), including trackball cursor (depicted as TC) and head cursor (depicted as HC).

To test the single modalities in the experiment, the use of the cursor control concept could be constrained to the following factor levels:

- Trackball - the cursor control was constrained to trackball use, with only the trackball cursor visible.
- Head control - the cursor control was constrained to head control use. Only the attached cursors were visible, without the possibility to detach the cursors.
- Hybrid - the cursor control was constrained to hybrid use. Both cursors were visible, with mandatory use of the warp function and without the possibility to attach the cursors.

Index of difficulty alternatives
Due to the overarching fighter cockpit context, a fixed symbol target symbol size of 6 mm was defined for the experimental setup based on MIL-STD-1472H-2020-09. With constant symbol sizes, the factor levels for the index of difficulty (ID) (MacKenzie, 2013) were defined by variations in pointing distance (D). Seven factor levels were specified for this study, as depicted in table 1.

Table 1. Defined factor levels for ID and associated distances (Hartmann, 2022).

ID	4	4.5	5	5.5	6	6.5	7
$D(mm)$	90	138	186	282	378	570	762

Flight dynamic alternatives
Two factor levels for the flight dynamic condition were defined:

- Static flight condition - no mechanical disturbances were introduced
- Light atmospheric turbulences

The motion profile was derived from the Eurofighter Typhoon flight control model and simulated a flight in 3000 feet with Mach 0.4 ($\approx$300 knots) under light atmospheric turbulences. Simulated turbelences were validated with two experienced test pilots.

Dependent variables

Throughput and related objective measures
Based on the CEN ISO/TS 9241-411:20014-08, throughout (TP) was used as a single measure of performance, including effective values for width (W_e) and distance (D_e) (MacKenzie, 2015):

$$TP = \frac{Log_2\left(\frac{D_e}{W_e} + 1\right)}{MT} = \frac{\log_2\left(\frac{D_e}{4.133 \cdot SD_x} + 1\right)}{MT}$$

Further dependent variables of interest in the context of pointing performance are:

- Movement time (MT) – the elapsed time for a seletion task
- Dispersion - the spread of hits around the center of the target in mm
- Warp time - the elapsed time in seconds from task start to the activation of the warp function
- Warp dispersion - the two-dimensional cursor dispersion in mm of the head cursor at the activation of the warp function when using the hybrid cursor control modality

Workload
Workload refers to the aggregated unweighted subjective workload (Hart, 2006), measured with the NASA TLX (NASA, n.d.).

Acceptance measures
This refers to the median of the acceptance measures outlined by the technology acceptance model (Venkatesh & Davis, 2000; Venkatesh et al., 2003), limited to predictors related to perceived usefulness and perceived ease of use.

User behaviour regarding preference of modality
With the unconstrained hybrid cursor control, users could freely choose the cursor control modality. This variable refers to the preference of the users per ID in relation (%) to the number of pointing tasks.

Design

The testing for RQ1 and RQ2, related to the dependent variables throughput and workload under controlled conditions, was comparative in the light of the independent variables with the granularity of ID restricted to 4, 5, 6, and 7. A within-subjects design with three factors and repeated measures was used. The evaluation of modality use characteristics based on RQ3 was descriptive in nature and defined by the dependent variables modality choice, perceived usefulness and perceived ease of use. With turbulent flight conditions and the unconstrained cursor control modality kept constant, ID was the only independent variable and the granularity increased to all seven levels.

Procedure

Prior testing, the participants were briefed about data protection, ethics, and safety procedures. Tests were arranged in three blocks:

- Block 1: All levels of cursor control modality with static flight dynamics
- Block 2: All levels of cursor control modality with light turbulences
- Block 3: Free choice modality with light turbulences

Participants were instructed to select the target as fast and as precise as possible. Before the first use of each modality, a training phase was conducted, which was terminated by the participants based on subjective confidence and understanding. Familiarization with the motion platform was conducted with a dry run.

A series of 20 discrete pointing tasks for each ID = {4; 5; 6; 7} was performed for each modality (240 total), with task axis' along the diagonals of the display, as illustrated in figure 5. For block 3, the series contained 20 tasks for all seven levels of ID (140 total). To prevent order effects, task sequences were simply randomized regarding task ID and task axis for each uneven participant number, and counterbalanced for every even number with the reversed task sequence of the preceding participant.

Figure 5. Example of a pointing task setup, with x_{tc}/y_{tc} and x_{hc}/y_{hc} showing the coordinates for the trackball and head cursor (Hartmann, 2022).

To balance potential effects from learning and fatigue, the sequence of modalities was randomized within each block using a latin square for every subsequent uneven participant number, while the blocks 1 and 2 were alternated between even and uneven participant numbers. In every case, block 3 was conducted as the last one. The workload questionnaire was administered after testing each cursor control modality in each flight dynamic condition and the acceptance was filled in after block 3.

Data Analysis

Statistical analysis
Statistical analysis was performed with SPSS Statistics, v27 (IBM, n.d.). Error bars in all graphs show the 95% confidence interval. An alpha level of .05 is assumed for all tests, unless mentioned otherwise. The statistical assumptions for inferential statistics were tested and are only reported for irregularities or violations. For ease of read, results are reported without unit signs.

Results

Thoughput and related objective measures

The aggregated average throughput was 1.24 ($SD = .26$) for the head control, 1.81 ($SD = .26$) for the hybrid cursor control device, and 1.95 ($SD = .21$) for the trackball, as shown in figure 6.

Figure 6. Throughput per modality (Hartmann, 2022).

Broken down to flight dynamics, the average throughput for each cursor control modality is shown in table 2. Mean throughput for each condition is illustrated in figure 7.

Table 2. Average throughput (TP) for each modality broken down to flight dynamics (Hartmann, 2022).

Flight dynamics	Average TP head control	Average TP hybrid	Average TP trackball
Static	1.44 $(SD = .30)$	1.85 $(SD = .33)$	1.97 $(SD = .27)$
Turbulences	1.04 $(SD = .27)$	1.77 $(SD = .27)$	1.92 $(SD = .19)$

A Shapiro-Wilk test indicated a deviation from normality for the head control in turbulent conditions and ID7 ($W(12) = .73, p = 0.05$). Removal of the causal extreme value of participant 4 resulted in a non-significant normality-test, but had no obvious impact on the q-q plot or the parametric test results. Consequently, this datapoint was retained.
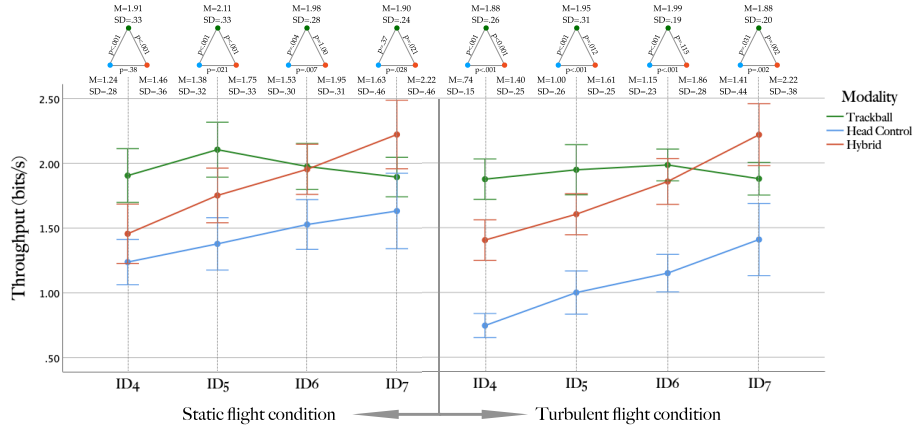
Figure 7. Throughput for all conditions. The triangles illustrate the pairwise comparisons between modalities per ID, including means and p-values (Hartmann, 2022).

A 3-way repeated-measures ANOVA was conducted to investigate the impact of modality, flight dynamics and ID on throughput. Bonferroni corrected Tukey HSD pairwise comparisons were used post-hoc, to determine specific differences in throughput. Maulchy's test indicated a violation of sphericity for modality ($\chi^2(2) = 12.16, p = 0.002$) and ID ($\chi^2(5) = 17.48, p = 0.004$). A Greenhouse-Geisser correction, with $\varepsilon = .59$ and $\varepsilon = .49$ respectively, was applied. The ANOVA indicated statistically significant main effects for modality, $F(1.17,12.91) = 33.25, p < 0.001, \eta_n^2 = .75$, flight dynamics, $F(1.00,11.00) = 7.46, p = .02, \eta_n^2 = .40$, and ID, $F(1.45,16.00) = 56.56, p < .001, \eta_n^2 = .84$. Furthermore, statistically significant interaction between modality and dynamics, $F(1.53, 16.83) = 13.83, p < .001, \eta_n^2 = .56$, modality and ID, $F(3.12,36.50) = 21.75, p < .001, \eta_n^2 = .66$, and flight dynamics and ID, $F(2.36, 26.00) = 5.14, p = .010, \eta_n^2 = .32$, were found. All other interactions were not indicated as statistically significant.

Post-hoc comparisons showed the overall throughput for the trackball modality being significantly higher than the hybrid modality ($.69, 95\% CI[.036,.24], p = .009$), while throughput for the hybrid was significantly higher than for the head control modality ($.55.95\% CI[.25,.85], p = .001$).

Throughput over all modalities and IDs was significantly higher for static flight dynamics, compared to turbulences ($.16, 95\% CI[.032,.30], p = .020$).

Furthermore, throughput over all modalities and flight dynamics was significantly higher for ID7 than ID6 ($.13, 95\% CI[.009,.26], p = .033$), higher for ID6 than ID5 ($.11, 95\% CI[.056,.17], p < .001$), and higher for ID5 than ID4 ($.19, 95\% CI[.13,.26], p < .001$).

In static flight conditions, the trackball ($.53, 95\% CI[.20,.85], p = .002$) and hybrid cursor control device ($.40, 95\% CI[.051,.75], p = .024$) showed significantly higher throughput than head control, while the trackball and hybrid cursor control did not differ significantly. In turbulences though, significantly higher throughput was

achieved with the trackball compared to the hybrid cursor control device ($.15, 95\% \, CI[.013, .288], p = .031$), while throughput with the hybrid cursor control was higher compared to head control ($.70, 95\% \, CI[.40, .99], p = .001$).

The trackball and hybrid cursor control devices showed no statistically significant difference in throughput regarding flight dynamics. The head control was significantly better in the static condition ($.37, 95\% \, CI[.23, .51], p < .001$), compared to turbulences.

Despite no significant three-way within-subjects interaction effects indicated in the ANOVA omnibus test ($F(3.16, 34.78) = 1.55, p = .056$), the contrast analysis provided hints for significant pairwise contrasts in the three-way interaction. The contrasts of modalities for every ID and flight condition appear important for RQ1 and RQ2 and are shown in figure 7, with the respective p-value, mean and 95%CI.

The average task completion time for the head control was 2.79 ($SD = .93$), 3.09 ($SD = .64$) for the hybrid cursor control, and 2.88 ($SD = .52$) for the trackball. Table 3 shows the task completion times broken down to flight dynamics.

Table 3. Average movement time (MT) for each modality broken down to flight dynamics (Hartmann, 2022).

Flight dynamics	Average MT head control	Average MT hybrid	Average MT trackball
Static	2.75 ($SD = 1.00$)	2.98 ($SD = .68$)	2.82 ($SD = .53$)
Turbulences	2.84 ($SD = .92$)	3.19 ($SD = .74$)	2.95 ($SD = .61$)

The average two-dimensional dispersion for the head control was 11.47 ($SD = 4.25$), 2.28 ($SD = 1.59$) for the hybrid cursor control device, and 2.30 ($SD = 2.22$) for the trackball. Table 4 shows the dispersion broken down to flight dynamics. Figure 8 illustrates the task completion times (8a) and dispersion at target selection (8b) for all conditions.

Table 4. Average dispersion (D_x) for each modality broken down to flight dynamics (Hartmann, 2022).

Flight dynamics	Average D_x head control	Average D_x hybrid	Average D_x trackball
Static	7.68 ($SD = 3.74$)	2.41 ($SD = 1.77$)	2.37 ($SD = 2.26$)
Turbulences	15.27 ($SD = 5.51$)	2.15 ($SD = 1.44$)	2.24 ($SD = 2.19$)
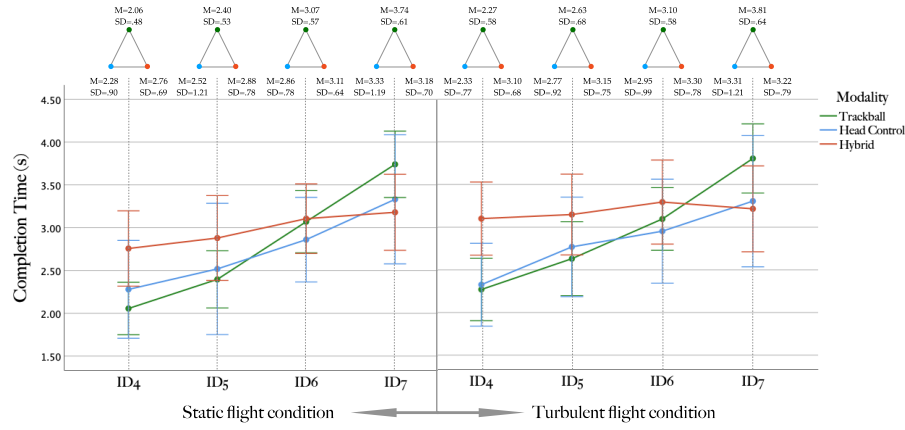
(a) Breakdown of task completion times for all conditions
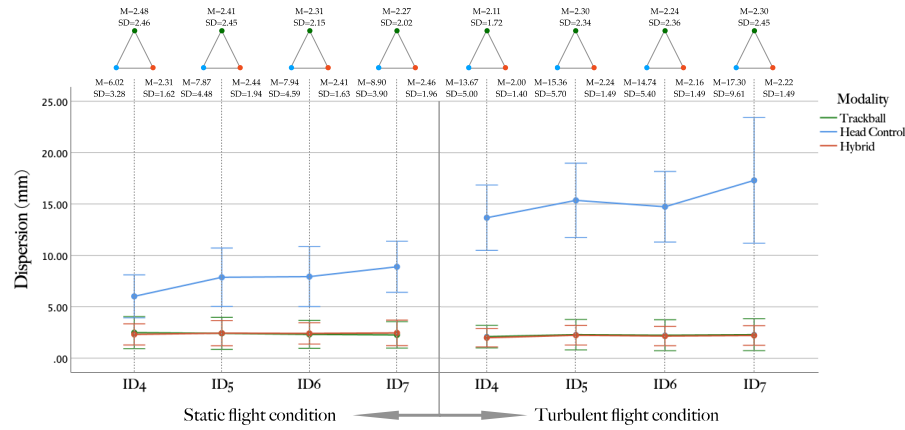
(b) Breakdown of two-dimensional dispersion for all conditions

Figure 8. Task completion times and dispersion in two dimensions for all conditions. The triangles illustrate the means per ID (Hartmann, 2022).

For the hybrid cursor control, the overall average warp dispersion was 18.79 ($SD = 4.85$) with an average warp time of 1.29 ($SD = .36$). Table 5 shows the warp time and distances broken down to flight dynamic conditions, while figure 9 provides an overview of all conditions, including ID.

Table 5. Average warp time and warp dispersion broken down to flight dynamics (Hartmann, 2022).

Flight dynamics	Average Warp dispersion (mm)	Average warp time (s)
Static	15.80 ($SD = 5.51$)	1.39 ($SD = .49$)
Turbulences	21.68 ($SD = 4.49$)	1.19 ($SD = .29$)
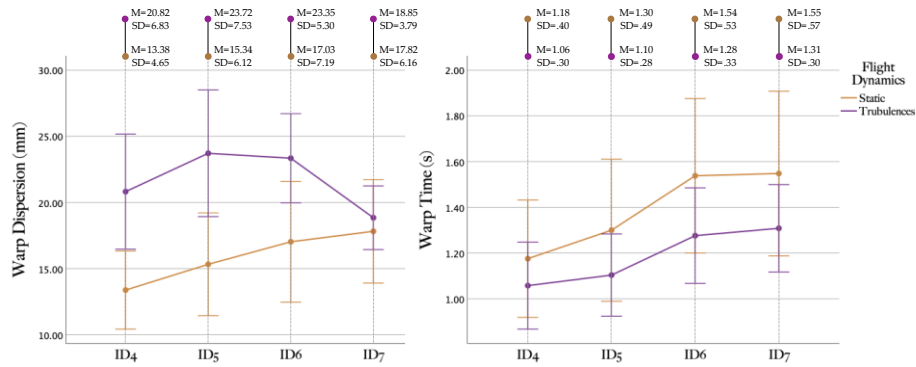
Figure 9. Warp times and warp dispersion for all conditions (Hartmann, 2022).

Workload

Figure 10 illustrates the accumulated workload ratings. Separate Friedman tests were used to determine differences in accumulated workload between modalities in the respective dynamic flight conditions.
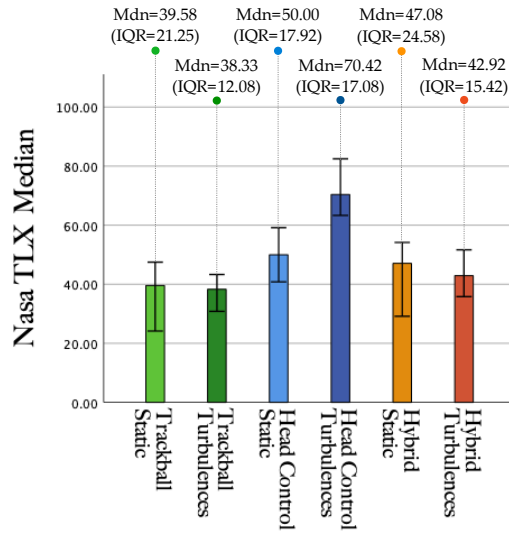
Figure 10. Workload ratings per condition (Hartmann, 2022).

In static flight conditions, no statistically significant difference was found. Under turbulent flight conditions, workload was significantly different between the modalities, $\chi^2(2) = 17.17, p < .001, W = .71$. Post-hoc pairwise comparisons were performed with Bonferroni corrected Wilcoxon tests, with statistical significance accepted at p < .0167. Under turbulences, head control showed significantly higher workload ratings than the trackball ($z = -3.06, p = .002, r = .88$) and hybrid modality ($z = -2.90, p = .004, r = .84$).

Separate Wilcoxon tests at the p < .05 level were conducted to compare workload ratings for each modality between flight conditions. The workload for the head control was rated significantly higher under turbulent conditions, compared to static flight conditions ($z = -3.06, p = .002, r = .88$).

Preference of modality

Head control was not used by any participant. Table 6 shows the average number of warps per ID in relation to the average task numbers. The values represent the frequency of use for the hybrid modality, as the warp function is exclusive to the hybrid cursor control.

Table 6. Average number of warps and warp ratio in block 3 (Hartmann, 2022).

ID	Mean Nr of Tasks	Mean Nr of Warps	Warp Ratio
ID4	19.83 ($SD = .58$)	11.92 ($SD = 8.08$)	61%
ID4.5	20.00 ($SD = NA$)	15.92 ($SD = 6.76$)	80%
ID5	19.92 ($SD = .29$)	16.67 ($SD = 6.73$)	84%
ID5.5	19.92 ($SD = .29$)	17.75 ($SD = 5.4$)	89%
ID6	19.75 ($SD = .62$)	17.83 ($SD = 5.04$)	90%
ID6.5	20.00 ($SD = NA$)	19.08 ($SD = 3.18$)	95%
ID7	19.67 ($SD = .78$)	18.83 ($SD = 2.66$)	96%

Acceptance

A Shapiro-Wilk test indicated non-normality for the perceived usefulness of head control ($W(12) = .31, p = .001$), the perceived usefulness of the hybrid modality ($W(12) = .31, p < .001$), and the perceived ease of use of the hybrid modality ($W(12) = .26, p = .003$). Weighing the advantages and disadvantages of parametric and non-parametric tests as discussed by Zimmerman (1998), the decision was made to analyze the differences between the modalities regarding perceived usefulness and perceived ease of use with Friedman tests. Pairwise comparisons for the three modalities were performed separately for perceived usefulness and perceived ease of use with Bonferroni corrected Wilcoxon tests, where statistical significance was accepted at the p < .0167 level.

Figure 11 shows the median perceived usefulness and perceived ease of use scores for each modality. Perceived usefulness was statistically significant different between the modlities ($\chi^2(2) = 14.85, p < .001, W = .62$). Perceived usefulness for head control was rated significantly lower than for the trackball ($z = -2.87, p = .004, r = .83$) and the hybrid modality ($z = -2.98, p = .003, r = .86$). No other significant differences in perceived usefulness ratings were found.
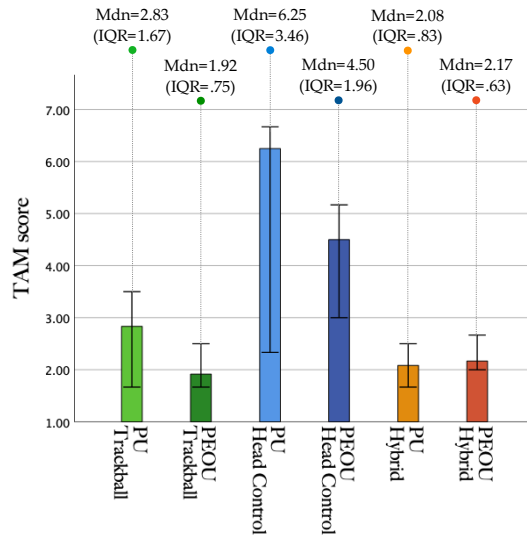
Figure 11. Acceptance ratings for perceived usefulness (PU) and perceived eae of use (PEOU). Lower scores reflect higher acceptance (Hartmann, 2022).

Perceived ease of use was statistically significant different between the modlities ($\chi^2(2) = 19.60, p < .001, W = .82$). Perceived ease of use for head control was rated significantly lower than for the trackball ($z = -3.06, p = .002, r = .89$) and the hybrid modality ($z = -3.07, p = .002, r = .89$). No other significant differences in perceived ease of use ratings were found.

Discussion

Interpretation of findings

This study evaluated and compared the proposed hybrid cursor control concept to the single modalities under varying flight dynamics. A medium effect for throughput confirms a general difference between modalities on pointing performance. While the hybrid modality showed higher performance than the head control, it was outperformed overall by the trackball cursor control device. A significant interaction effect between modality and flight dynamic condition on performance was found. While the results of the main effect for modality are reflected in the comparisons under turbulences, in static condition, no significant difference was found between trackball and hybrid modalities. At a closer look, significant differences were found between modalities under various combinations of ID and flight dynamic conditions. Superior throughput with the trackball could only be achieved for distances of 90 mm (ID4) and 186 mm (ID5). A cross-over point in performance appears to happen around ID6 (378 mm), where equal performance was achieved with the hybrid and trackball cursor control devices. In contrast to the static flight condition, head control is outperformed for all IDs under turbulences.

Based on the descriptive data for accuracy and pointing speed, the relation in throughput performance between trackball and hybrid modality appears to be purely

based on task completion times, as dispersion is fairly equal for all pointing distances. This is plausible, as for both modalities the cursor is controlled with the trackball in the terminal phase. In contrast, the difference between trackball and head control in throughput appears to be based on different accuracies, as the completion times in static and dynamic flight conditions suggest similar behaviour over all IDs. In this light, the stable mean task completion time over all IDs of the hybrid cursor control appears paradoxical, as it combines the other two modalities, each with increasing completion times for increasing IDs. The warp dispersion was overall at $\approx$19 mm ($\pm$ $\approx$2 mm) within $\approx$1.3 s ($\pm$ $\approx$0.1 s) and appears to be fairly stable. Hence, the remaining distance to the target for the trackball movement was also fairly constant over all IDs, which is in line with the findings of Drewes and Schmidt (2009). Hence, for the head control modality, the increase in completion time with increasing ID is apparently related to achieving the required accuracy in the final phase, which is more challenging under turbulences and results in lower output.

With regard to RQ1, it can be concluded that the hybrid modality may provide advantages in pointing performance for long distances.

Under static flight conditions, subjective workload is not significantly different for all modalities. In contrast, workload ratings for head control were significantly higher under turbulent conditions. This appears to be related to physical effort in the terminal phase of pointing, as the hybrid concept did not differ significantly from the trackball. The difference in workload between static and turbulent flight conditions was also significant for the head control, while no significant difference was found for the trackball and hybrid modality.

The analysis of subjective workload ratings for the hybrid warp operation does not suggest higher workload demands, confirming the findings of Biswas and Langdon (2015). This is in line with the observed participant behaviour, as the hybrid was used with a high frequency even for short distance tasks. Although, the high rate of hybrid use could eventually be related to the repetitive character of the pointing tasks and associated difficulties to switch between modalities in the experiment scenario. Nevertheless, the results of the acceptance ratings back up the preference for the hybrid modality and the avoidance of head control.

In light of RQ2, it can be concluded that the hybrid cursor control device may provide an objective performance advantage at distances larger than 378mm.

With regards to RQ3, high user acceptance and perceived benefits might have triggered hybrid modality usage for shorter distances, where actually a performance disadvantage could have been expected.

Limitations

While atmospheric turbulences have been simulated, acoustics and other sources of vibrations in the cockpit environment might cause bio-mechanical effects interfering with cursor control modalities, as described by Smith et al. (2008). Future flight test is therefore suggested, which also addresses the additional weight and shift in head center of gravity by a flight helmet. Furthermore, Statistical power for future studies

can be increased by recruiting participants beyond the population of combat pilots, provided the professional expertise is not of relevance.

Conclusion

In summary, best performance overall was achieved with the trackball. Nevertheless, for pointing tasks with an index of difficulty of 6 or higher, the use of the hybrid cursor control device is promising. This is based on the potential of head control to be effectively used for coarse pointing, even under turbulent flight conditions. In contrast, head control as a single modality should not be considered for precise pointing, especially under turbulent conditions. High user acceptance and usage of the hybrid modality demonstrates acceptance even for short pointing distances. Overall, the hybrid modality provides significant advantages for long distance point and select tasks under light atmospheric turbulences, without the burden of an increased workload compared to head control and the trackball. Based on the acceptance ratings of participating combat aircrew, the potential for implementation in current cockpit environments should be considered.

Acknowledgement

References

Bibb, C. (2021). *Human-Machine Interaction Methods within an Augmented Reality Cockpit: A Human Centred Review* (Technical Report Version 1.2). Birmingham, UK: University of Birmingham.

Biswas, P., & Langdon, P. (2015). Multimodal Intelligent Eye-Gaze Tracking System. *International journal of human-computer interaction, 31*, 277–294.

CEN ISO/TS 9241-411:2014-08. (2014). *Ergonomics of human-system interaction - Part 411: Evaluation methods for the design of physical input devices* (Standard). London, UK: British Standards Institution (BSI).

Drewes, H., & Schmidt, A. (2009). The MAGIC Touch: Combining MAGIC-Pointing with a Touch-Sensitive Mouse. In *Human-Computer Interaction – INTERACT 2009*. Lecture Notes in Computer Science, vol 5727 (pp. 415-428). Berlin, Germany.

Hart, S.G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50 (pp. 904-908).

Hartmann, E.F. (2022). *Evaluation of a multimodal cursor control device (CCD) for combat aircraft cockpit use* (Unpublished master thesis). Nottingham, UK: University of Nottingham.

IBM. (n.d.). *XPSS Software*. Accessed August 2022. Retrieved from https://www.ibm.com/analytics/spss-statistics-software.

Kellerer, J., Rabl, A., & Eichinger, A. (2018). Touch vs. Track-Evaluation of Cockpit Controls for Future Fighter Aircraft in divers Workload Context Conditions. In H. Becherstorfer, J. Lasry, and M. Schwarz (Eds.), *Proceedings of the 33rd Conference of the European Association for Aviation Psychology*. Netherlands.

Lin, M.L., Radwin, R.G., & Vanderheiden, G.C. (1992). Gain effects on performance using a head-controlled computer input device. *Ergonomics, 35*, 159–175.

MacKenzie, S.I. (2013). A Note on the Validity of the Shannon Formulation for Fitts' Index of Difficulty. *Open Journal of Applied Sciences, 03*, 360-368.

MacKenzie, S.I. (2015). Fitts' throughput and the remarkable case of touch-based target selection. In *Proceedings of the 17th International Conference on Human-Computer Interaction, HCI 2015* (pp. 238-249). Cham, Switzerland: Springer.

MIL-STD-1472H-2020-09. (2020). *Department of Defense Design Criteria Standard - Human Engineering* (Standard). USA: Department of Defense.

NASA. (n.d.). *NASA Task Load Index*. Accessed August 2022. Retrieved from https://humansystems.arc.nasa.gov/groups/TLX/tlxapp.php.

Smith, S.D., Goodman, J.R., & Grosveld, F.W. (2008). Vibration and Acoustics. In J.R. Davis, R. Johnson, and J. Stepanek (Eds.), *Fundamentals of Aerospace Medicine, Fourth edition* (pp. 110-141). Philadelphia, USA: Lippincott Williams & Wilkins.

Stanton, N.A., Harvey, C., Plant, K.L., & Bolton, L. (2013). To twist, roll, stroke or poke? A study of input devices for menu navigation in the cockpit. *Ergonomics, 56*, 590-611.

Stellmach, S., & Dachselt, R. (2012). Look & touch: Gaze-supported target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (pp. 2981-2990). New York, USA: Association for Computing Machinery.

Thomas, P.R., Biswas, P., & Langdon, P. (2015). State-of-the-Art and Future Concepts for Interaction in Aircraft Cockpits. In M. Antona, and C. Stephanidis (Eds.), *Universal Access in Human-Computer Interaction. Access to Interaction.* UAHCI 2015. Lecture Notes in Computer Science, vol 9176 (pp. 538-549). Cham, Switzerland: Springer.

Venkatesh, V., & Davis, F. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science, 46*, 186-204.

Venkatesh, V., Morris, M.G., Davis, G.B., & Davis, F.D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27(3)*, 425–478

Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99)* (pp. 246–253). New York, USA: Association for Computing Machinery.

Zimmerman, D.W. (1998). Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education, 67*, 55-68.

A new experimental method to investigate multitasking strategies in flight environments via the use of gamification

Sophie-Marie Stasch & Wolfgang Mack
University der Bundeswehr München
Germany

Abstract

Aviate, navigate, communicate: this approved axiom for pilots demonstrates that the use of an appropriate multitasking strategy is essential to operate an aircraft safely. Different multitasking strategies have been identified recently in basic research studies. However, they have not been experimentally manipulated in a flight environment. Therefore, a suitable experimental method is required to establish a precise connection between a multitasking strategy and the related performance outcomes. The present approach manipulates the stability-flexibility-dilemma of cognitive control as a multitasking strategy using the Multi-Attribute Task Battery (MATB). The MATB simulates four flight tasks that must be completed simultaneously. By the use of a novel gamification method, participants had to either complete the flight tasks in a more stable or more flexible control state. Results indicate the successful manipulation of the multitasking strategy by differences in the task performance and the distribution of visual attention. Future studies may adopt the described method to investigate other multitasking strategies in the cockpit. The presented method is not limited to the field of aviation, but can also be used to examine multitasking strategies in other domains, such as air traffic control or driving.

Multitasking strategies in the cockpit

The working environment in the cockpit is highly demanding: Pilots need to constantly monitor and adjust different flight instruments, communicate with air traffic control, navigate the aircraft or respond to emergencies. Finding an efficient multitasking strategy is the key to mastering these high workload scenarios without significant performance decrements. The A-N-C axiom (Aviate – Navigate - Communicate) is a well-known example of a multitasking strategy, which dictates to prioritise the most safety-critical aviate task. In recent years, different multitasking strategies have emerged in the field of aviation. For instance, Schulte and Donath (2011) identified two multitasking strategies that pilots apply in highly dynamic environments to control mental workload. While load-sharing strategies describe the transfer of tasks to other teammates or automation, load-shedding strategies alter the way tasks are worked on.

Working memory capacity has traditionally been considered a strong predictor of multitasking ability. However, recent studies highlight the importance of attentional

control in the context of multitasking and complex task performance. For instance, Pak et al. (2023) point out that attention control is a better predictor of task performance, compared to measures of working memory, in dynamic task environments.

Cognitive control and multitasking

Closely related to the concept of attention control is the concept of cognitive control. Cognitive control is viewed as a collection of mechanisms that are accountable for the flexible adaption of information processing in the pursuit of task goals (Musslick & Cohen, 2021). Meanwhile, cognitive control is subject to the stability-flexibility-dilemma. This dilemma describes the antagonistic demands of cognitive control in regulating human behaviour (Goschke & Bolte, 2014; Hommel, 2015). On the one hand, cognitive stability is required to shield information processing and task execution efficiently from distractors. On the other hand, cognitive flexibility is crucial to quickly adapting information processing to constantly varying environmental requirements. The challenge is to balance cognitive stability and cognitive flexibility with the current demands of the task. An inadequate degree of cognitive stability can lead to perseveration and reduced flexibility, which aggravates task switching in multitasking situations. In contrast, improperly employed flexibility leads to distraction from irrelevant stimuli in multitasking situations. The stability-flexibility-dilemma of cognitive control can also be regarded as a multitasking strategy, which has been studied in basic research of multitasking and voluntary task switching (Brüning & Manzey, 2018). A distinct feature of each cognitive control state is the number of task switches: a stable control state is associated with a lower number of task switches, whereas a flexible control state is linked to a higher number of task switches (Dreisbach & Fröber, 2019).

These findings are derived from basic cognitive research with simple stimulus-reaction experiments. This paper presents a novel method to experimentally investigate the stability-flexibility-dilemma as a multitasking strategy in a human factors related task. In particular, two multitasking strategies are induced on the basis of gamification.

Gamification in Experiments

Gamification is the use of game mechanics and game design elements to non-game environments to enhance motivation and performance (Deterding et al., 2011). Sailer et al. (2013) describe how different game elements relate to motivation outcomes. Points, as a simple game mechanic, function as immediate positive reinforcement and are viewed as virtual rewards for executed actions. Leaderboards foster competition and address achievement and power motives. Landers et al. (2017) demonstrate that leaderboards successfully motivate participants in a sense that they implicitly set their goals close to the top of the leaderboard.

In order to actually induce the multitasking strategies in regard to the stability-flexibility-dilemma, a reinforcement method based on points and a leaderboard was chosen in this study. The experiment was framed as a type of game to appeal to the subjects' motivation. Additionally, the points awarded provide direct feedback on the

demonstrated multitasking strategy. By presenting a leaderboard, subjects were supposed to implicitly set the goal of continuously improving their score.

General Method

The openMATB (Cegarra et al., 2020) served as the simulated cockpit with the advantage of not requiring prior knowledge of flying. Four flight-like tasks needed to be operated concurrently: the tracking task required the participant to keep the controller within a square in the middle. In the system monitoring task, participants had to react to moving scales or blinking buttons. The communication task required participants to respond to radio messages. In the resource management task, participants had to keep the level of fuel in tank A and B on an indicated level (see Figure 1).
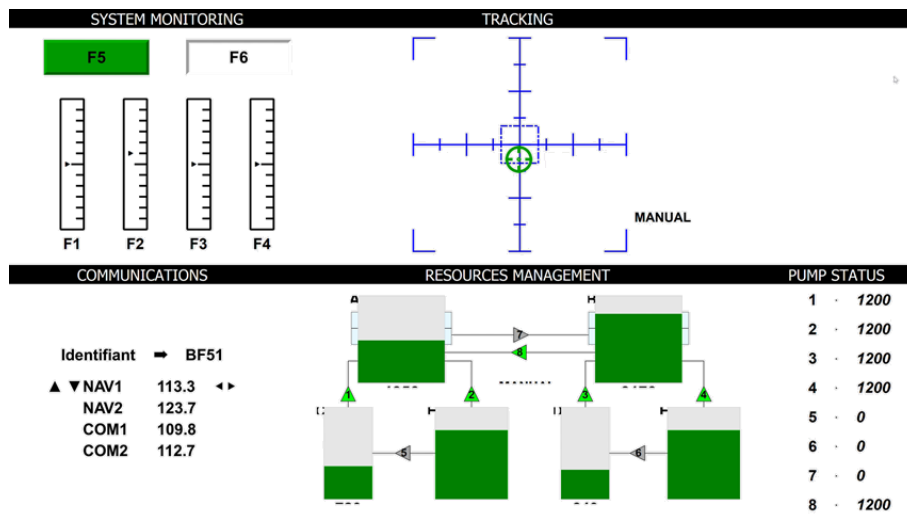
Figure 1. Image of the experimental environment openMATB (Cegarra et al., 2020).

During the experiments, participants operated both strategies in two experimental scenarios to exclude scenario-related effects. Scenarios included the same number of events. However, events differed in the sequence in which they occurred in a scenario. This schema resulted in four experimental conditions. To prevent sequence effects, the sample was divided into two groups receiving a counter-balanced experimental design (see Figure 2).
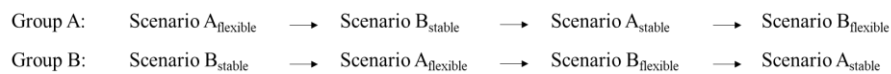
Figure 2. Test sequence of conditions in the pilot study and the experiment.

Based on a g*power analysis (Faul et al., 2007) a sample size of n=34 was determined. The reference study of van Steenbergen et al. (2009) found a medium effect size in the manipulation of cognitive control.

Experiment 1

36 participants were tested in the first experiment. Eye-tracking data was measured using the Eyelink 1000 Plus. The multitasking strategy (levels: flexible, stabil) was die independent variable. The dependent variables comprise the eye movement data, the performance of the individual MATB tasks as well as the feedback score.

Procedure

At the beginning of each condition, participants were told that they will operate the virtual aircraft in two weather conditions, requiring a different multitasking strategy. The good weather condition aimed to induce a flexible multitasking strategy with high task switching. The stormy weather condition primed the participants to prioritise the tracking task with low task switching to avoid a plane crash. One condition consisted of five trials. After each trial, participants received feedback in form of a score varying between 0 and 100. The feedback score was explained as a measure of how well the subjects operated the virtual aircraft according to the mentioned weather conditions. The way the score was calculated differed between the pilot experiment and the second experiment and is described in further detail in the following section. Before each condition, a leaderboard with feedback scores from a pilot experiment was presented to the participants to enhance motivation. The feedback scores displayed were the same for all subjects in all conditions in order to create equal prerequisites for all subjects. One trial was 90 seconds long. With five trials per condition and four conditions in total, the experimental time was half an hour.

Calculation Feedback Score

Step 1. Calculation of deviations (in %):

Tracking task: 35-25 = 10
System monitoring task: 20-25 = (-) 5
Communications task: 10-25 = (-) 15
Resource management task: 35-25 = (-) 10

Step 2. Summation of deviations (in %):
10 + 5 + 15 + 10 = 40

Step 3. Divide the sum the factor:
40 / 1.2 ≈ 33

Step 4. Subtracting result from max. score (100):
100 - 33 = 67 P.

Optimal fixation distribution
Actual fixation distribution
Worst fixation distribution

Figure 3. Calculation example of the flexible feedback score in the pilot study.

The feedback score was calculated based on the relative distribution of the number of fixations on each MATB subtask during one trial. Hereby, each fixation on the respective task was counted and compared to the total number of fixations per trial.

That resulted in a relative percentage of fixations on each task. For each multitasking strategy, an optimal percentage distribution per task was defined in advance. In the flexible condition, each subtask had to be equally prioritised. This resulted in the optimal distribution of 25% fixations on each of the four MATB tasks. Participants demonstrating that behaviour received the maximum feedback score of 100. The worst performance for the flexible condition was defined as follows: If 85% of fixations were on the tracking task and 5% of fixations respectively on the system monitoring task, the communications task and the resource management task, the subject received a feedback score of 0. The opposite calculation scheme applied to the stable condition. Participants showing no task prioritisation received a feedback score of 0. If 85% of fixations were on the tracking task and 5% of fixations respectively on the three other tasks, the participant received a feedback score of 100.

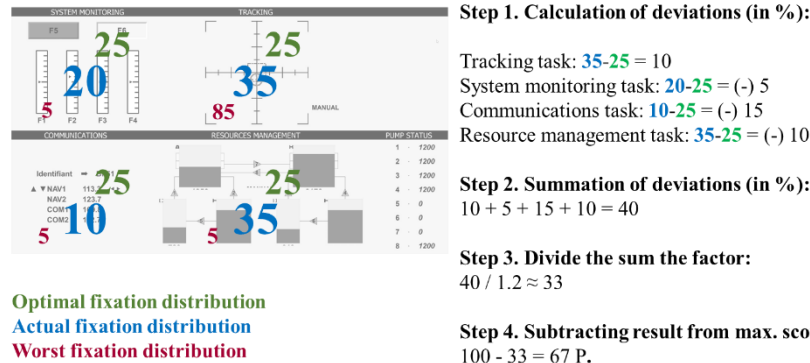In practice, the score was calculated as follows (see Figure 3): The first step was to determine the difference between the actual fixation distribution (in %) and the optimal fixation distribution (in %) for each subtask. The deviations were summed up and divided by a factor in the second and third step. In the current case, it was 1.2. This factor can be attributed to the fact that the absolute amount of the deviation was taken in the previous step. Furthermore, it is defined by the definition of the worst performance. Lastly, the deviation score was subtracted from the maximal possible score (of 100). This result was displayed to the subject as the feedback score.

Statistical analysis experiment 1

Statistical analysis was based on fitting linear mixed models using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022). Subject was included as a random factor in every model. Model selection was based on bayesian hypothesis testing using the Bayesian information criterion (BIC) approximation (Wagenmakers, 2007). Models were built up and compared in the following order: random intercept model, model including strategy as a fixed effect, model including strategy and the number of trials as fixed effects, model including strategy and the number of trials as fixed effects as well as the interaction between strategy and number of trials. $BF_{10} > 3$ was regarded as model improvement to prevent overfitting.

Results experiment 1

To predict the feedback score, the best model included strategy as a fixed factor. The model's intercept, corresponding to the flexible strategy, is at 65.65 (95% CI [62.95, 68.35], $t(676) = 47.76$, $p < .001$). The effect of the stable strategy is statistically significant and negative (beta = -29.95, 95% CI [-32.47, -27.44], $t(676) = -23.37$, $p < .001$; Std. beta = -1.29, 95% CI [-1.40, -1.18]). On the one hand, this result indicates that the participants received much lower values in the stable condition compared to the flexible condition (cf. Figure 4). On the other hand, no influence of the number of trials on the feedback score could be proven in this experiment (cf. Table 1). However, the score difference could be interpreted as a confounding factor, since a lower feedback score in the stable condition could be associated with emotions of negative valuence.

Experiment 2

The first experiment demonstrated that participants received much lower values in the stable condition compared to the flexible condition. To account for this difference, a different calculation procedure was employed in the second experiment (n=42). The instruction on task prioritisation remained the same. The feedback score consisted of two components, which were equally weighted: the number of task switches and the variance in performance. Task switches were operationalised by change of fixation from one task to another and the variance of performance between the tasks. The variance in performance was determined by normalising the performance of each subtask to values between 0 and 100. The results from the first experiment were used as a reference for the minimum and maximum values to capture the approximate performance level of the task. In case if participants scored lower than the defined minimum on a score component, the score was rounded up to 0. Similarly, if participants scored above the defined maximum, the score was rounded off to 100.

An optimal flexible strategy was defined by a high number of task switches and a low number of variances between tasks, since no task prioritisation should be done. Participants showing that behaviour received a feedback score of 100. The opposite was defined as the optimal stable strategy. Participants received a feedback score of 0 if they demonstrated a low number of task switches and a high variance between the main task (tracking task) and the three subtasks. The calculation scheme is similar to that of the first experiment. Additionally, a manipulation check at the end of the experiment controlled how participants were performing the task in each condition. Participants indicating that they did not prioritise tasks as instructed were excluded from the statistical analysis (n=6), resulting in a sample of n=36.

Results Experiment

Feedbackscore
Similar to the first experiment, a linear mixed model was fitted to test the effect of strategy and number of trials on predicting the feedback score. Subject was again included as a random factor. Bayesian hypothesis testing identified the model including strategy as a fixed factor as the best fitting model. The model's intercept, corresponding to the flexible strategy, is at 65.65 (95% CI [62.95, 68.35], $t(676) = 47.76$, $p < .001$). The effect of stable strategy is statistically significant and negative (beta = -29.95, 95% CI [-32.47, -27.44], $t(676) = -23.37$, $p < .001$; Std. beta = -1.29, 95% CI [-1.40, -1.18]).

Table 1. Descriptive statistics of the feedback score per condition.

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5	
	M	*SD*	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experiment 1										
flexible	64.37	11.51	66.72	9.76	66.72	9.64	65.00	10.76	65.68	10.45
stable	30.31	16.14	35.99	19.98	34.56	18.47	38.93	20.61	38.74	20.24
Experiment 2										
flexible	80.29	17.55	83.89	18.22	81.75	18.35	79.56	19.21	81.28	20.37
stable	51.78	21.82	60.44	18.14	59.61	20.22	60.22	19.55	61.84	17.52
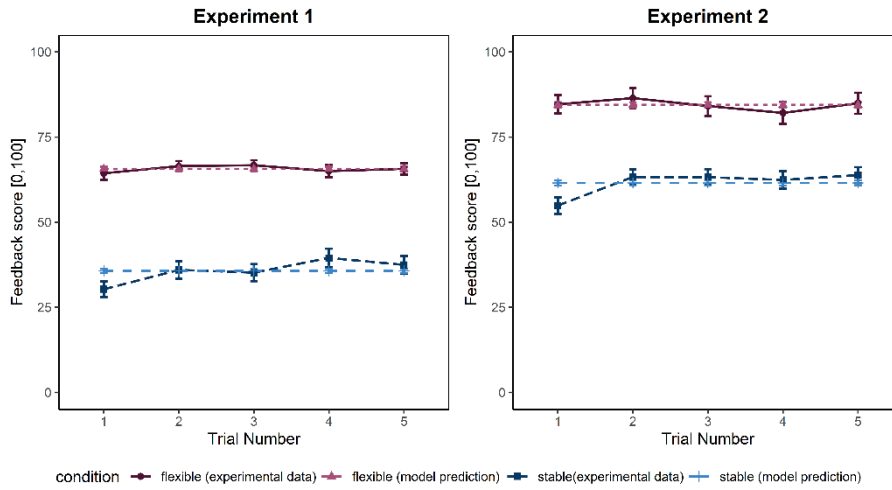
Figure 4. Feedback score of the flexible and stable condition including values of the selected model prediction.

Performance

To investigate how the strategies influenced task performance, performance metrics were operationalised in the following way: For the tracking task, the Root Mean Square Error (RMSE) was calculated. For the system monitoring task and the communications task, the reaction time (RT) of the hits was calculated respectively. The absolute deviation of the optimal tank level as a performance metric for the resource management task. All metrics were z-standardised.

For the tracking task, the best fitting model included strategy as a fixed effect only. The model's intercept, corresponding to the flexible strategy, is at -0.14 (95% CI [-0.34, 0.06], t(560) = -1.38, p = 0.168). The effect of the stable strategy is statistically significant and positive (beta = 0.40, 95% CI [0.26, 0.54], t(560) = 5.68, p < .001; Std. beta = 0.40, 95% CI [0.26, 0.54]).

For the system monitoring task, the best fitting model included strategy as a fixed effect only. The model's intercept, corresponding to the flexible strategy, is at -0.15 (95% CI [-0.30, -5.40e-03], t(697) = -2.04, p = 0.042). The effect of the stable strategy is statistically significant and positive (beta = 0.40, 95% CI [0.26, 0.54], t(560) = 5.68, p < .001; Std. beta = 0.40, 95% CI [0.26, 0.54])

For the communications task, the best fitting model included strategy as a fixed effect only. The model's intercept, corresponding to the flexible strategy, is at -0.14 (95% CI [-0.34, 0.06], t(560) = -1.38, p = 0.168). The effect of the stable strategy is statistically significant and positive (beta = 0.31, 95% CI [0.17, 0.45], t(697) = 4.36, p < .001; Std. beta = 0.31, 95% CI [0.17, 0.45]).

For the resource management task, the best fitting model included the strategy as a fixed effect only. The model's intercept, corresponding to the flexible strategy, is at -0.32 (95% CI [-0.54, -0.10], t(716) = -2.84, p = 0.005). The effect of the stable strategy

is statistically significant and positive (beta = 0.64, 95% CI [0.54, 0.75], t(716) = 12.29, p < .001; Std. beta = 0.64, 95% CI [0.54, 0.75]).
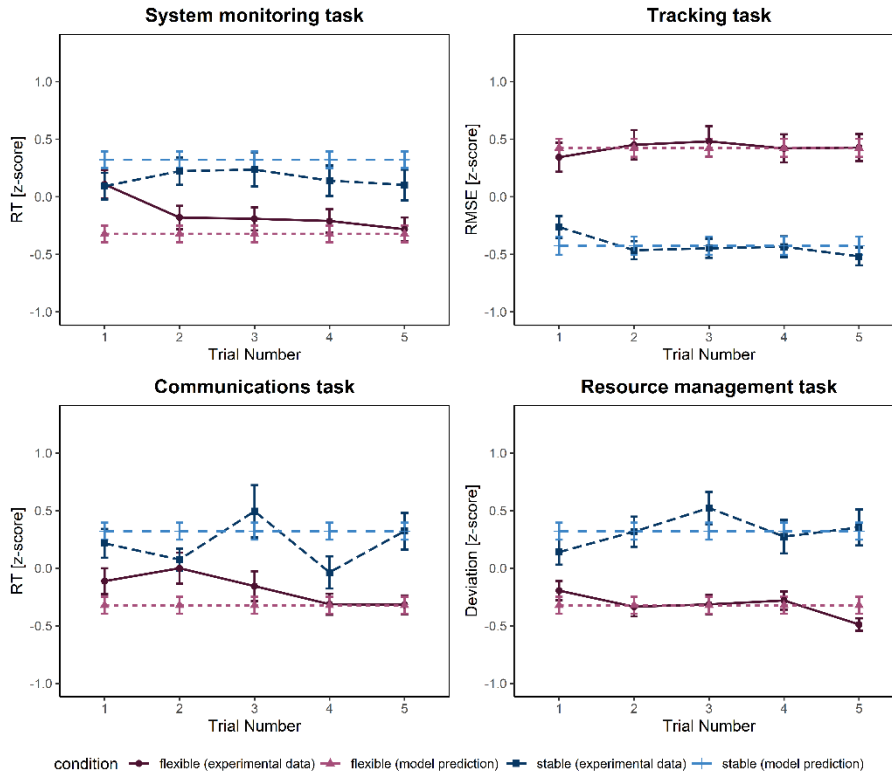
Figure 5. Performance of the four MATB subtasks of the flexible and stable condition including values of the selected model prediction.

Task switches

Task switches were best predicted by the interaction model including strategy and the number of trials. The model's intercept, corresponding to the flexible condition and Trial = 0, is at 116.54 (95% CI [107.53, 125.56], t(677) = 25.38, p < .001). Within this model, the effect of the stable strategy is statistically significant and negative (beta = -15.37, 95% CI [-23.02, -7.72], t(677) = -3.94, p < .001; Std. beta = -0.79, 95% CI [-0.89, -0.69]). The effect of trial is statistically non-significant and negative (beta = -0.12, 95% CI [-1.74, 1.50], t(677) = -0.14, p = 0.886; Std. beta = -5.07e-03, 95% CI [-0.07, 0.06]). However, the interaction effect of number and trial on the stable strategy is statistically significant and negative (beta = -3.57, 95% CI [-5.88, -1.27], t(677) = -3.05, p = 0.002; Std. beta = -0.15, 95% CI [-0.25, -0.05]).
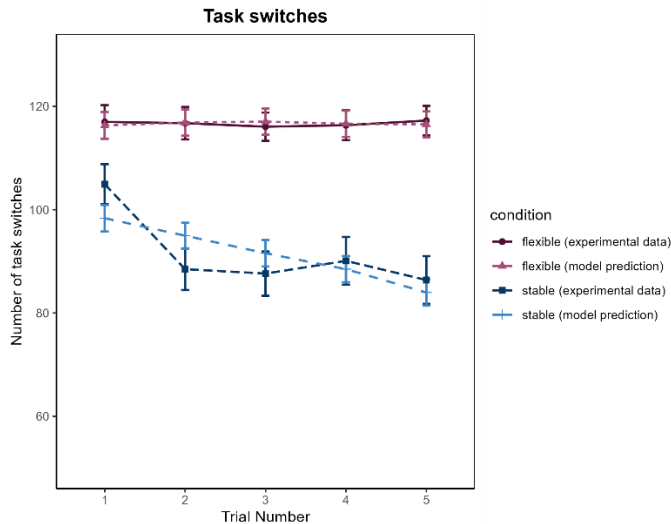
Figure 6. Number of task switches of the flexible and stable condition including values of the selected model prediction.

Discussion

The current experiment demonstrates how a multitasking strategy in a low-fidelty flight simulator can successfully be manipulated by means of the presented gamification method (feedback score + leaderboard). The induced multitasking strategies influenced the performance of every single task of the MATB. The reinforcing effect of the feedback score can be discerned in the interaction effect between the number of trials and the stable equation in the prediction of task changes (see Figure 6). In the flexible condition, the number of task switches remained the same for every trial. However, in the stable condition, the number of task switches decreased from trial to trial. Task switches are a striking correlate of the cognitive control mode (Dreisbach & Fröber, 2019), which served as the underlying basis for the manipulated multitasking strategy in this study. This finding illustrates the usefulness of awarding points to induce a stable multitasking strategy, which was the key factor of the presented gamification method. Furthermore, the leaderboard acted as an additional motivator for the respective multitasking strategy, as subjects could compare their feedback score of a trial with the target implicitly set by the leaderboard.

The lack of interaction between trial and strategy for predicting task switches in the flexible condition could indicate that participants are already using a flexible task strategy. Thus, no flexible multitasking strategy could be further induced. This argument is supported by the fact that the number of trials did not influence the single task performance of the MATB. Furthermore, the observation that subjects received considerably higher feedback scores in the flexible condition compared to the stable condition underlines this notion. The scenarios are designed in a way that constant task switches are required to successfully complete all tasks. Tasks like the system monitoring task and the communication task operate via cues that are very salient. Consequently, it is more difficult to suppress reaction tendencies towards them. In this

respect, it is worth highlighting that forced task switches shift the cognitive control mode in the direction of cognitive flexibility (Fröber & Dreisbach, 2017), which could likewise be the case in this study.

A direct influence of the number of trials on performance and on the feedback score could not be proven (see Figure 4 and Figure 5). However, if the values of the first trial are compared with the last trial of a block, a clear tendency of an improved feedback score can be discerned for the stable condition (see *Table 1*). This result could indicate that subjects were initially in an exploratory phase to discover the optimal multitasking strategy for improving their feedback score. Future studies could increase the number of trials per block to further investigate this supposition.

Application to other domains

Multitasking strategies and the stability-flexibility-dilemma of cognitive control describe multitasking behaviour on a broader level compared to the classical, resource-limited view. Not only for this reason is an investigation of multitasking strategies in other domains important. It is also crucial because multitasking behaviour of an operator is an essential safety-critical factor (Broccia et al., 2019). An application in more realistic flight scenarios is also feasible. Here, a score could be displayed directly in the cockpit for continuous feedback additional to a leaderboard at the beginning of the experiment. This allows for the currently achieved score to be compared to the target score implicitly set by the leaderboard. The multitasking strategies described above can thus be examined in relation to the pilot's flight performance in different flight phases. Applications in the field of air traffic control or driving of the presented gamification method to manipulate multitasking strategies are conceivable, too. For example, multitasking strategies, which occurred without deliberate manipulation in field studies, could be reproduced in the laboratory. Brumby et al. (2009) investigated task interleaving strategies while participants had to do a simple phone dialling task and a steering control task. Also in this study, subjects received feedback to reinforce task prioritisation. Seven strategies could be successfully identified that differed in interleaving at chunk boundaries. The method presented in this paper can therefore be linked to studies with a similar research question. The targeted replication of those strategies in the laboratory can contribute to the development of human-centred automation that takes multitasking strategies into account. This includes, but is not limited to, the study of performance metrics or physiological correlates of mental states. Further studies could also examine the interaction of multitasking strategies with different degrees of automation to investigate their effect on the aforementioned metrics.

Conclusion

The presented study illustrates a method to experimentally manipulate multitasking strategies in the low-fidelity flight simulator MATB. The stability-flexibility dilemma of cognitive control was used as the basis for manipulating the multitasking strategy. The results show a clear effect of the induced strategy on the performance metrics of the MATB as well as the number of task switches. Further studies in the area of human-machine interaction are needed to establish the generalisability of the presented method.

Acknowledgement

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67,* 1–48.

Broccia, G., Milazzo, P., & Ölveczky, P.C. (2019). Formal modeling and analysis of safety-critical human multitasking. *Innovations in Systems and Software Engineering*, 15, 169–190.

Brüning, J., & Manzey, D. (2018). Flexibility of individual multitasking strategies in task-switching with preview: are preferences for serial versus overlapping task processing dependent on between-task conflict? *Psychological research, 82,* 92–108.

Cegarra, J., Valéry, B., Avril, E., Calmettes, C. & Navarro, J. (2020). OpenMATB: A Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability. *Behavior research methods, 52,* 1980–1990.

Deterding, S., Khaled, R., Nacke, L.E., & Dixon, D. (2011). Gamification: Toward a definition. In *CHI 2011 gamification workshop proceedings* (p. 15). Vancouver, BC, Canada: ACM.

Dreisbach, G., & Fröber, K. (2019). On How to Be Flexible (or Not): Modulation of the Stability-Flexibility Balance. *Current Directions in Psychological Science, 28,* 3–9.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39,* 175–191.

Fröber, K., & Dreisbach, G. (2017). Keep flexible – Keep switching! The influence of forced task switching on voluntary task switching. *Cognition, 162,* 48–53.

Goschke, T., & Bolte, A., (2014). Emotional modulation of control dilemmas: The role of positive affect, reward, and dopamine in cognitive stability and flexibility. *Neuropsychologia, 62,* 403–423.

Hommel, B. (2015). Chapter Two - Between Persistence and Flexibility: The Yin and Yang of Action Control. Andrew J. Elliot (Eds.), *Advances in Motivation Science*, 33–67.

Landers, R.N., Bauer, K.N., & Callan, R.C. (2017). Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior, 71,* 508–515.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in cognitive sciences, 25,* 757–775.

Pak, R., McLaughlin, A.C., & Engle, R. (2023). The Relevance of Attention Control, Not Working Memory, in Human Factors. *Human factors, 0,* 1-12.

R Core Team (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Sailer, M., Hense, J., Mandl, H., & Klevers, M. (2013). Psychological perspectives on motivation through gamification. *Interaction Design and Architecture(s) Journal, 19,* pp. 28–37.

Schulte, A., & Donath, D. (2011). Measuring Self-adaptive UAV Operators' Load-Shedding Strategies under High Workload. In Don Harris (Eds.), *Engineering Psychology and Cognitive Ergonomics: 9th International Conference, EPCE 2011.* (pp. 342–351). Berlin, Heidelberg.

Van Steenbergen, H., Band, G.P.H., & Hommel, B. (2009). Reward counteracts conflict adaptation. Evidence for a role of affect in executive control. *Psychological science, 12,* 1473–1477.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804.

What would you expect? Insights from a customer journey analysis of passenger drone flights

Patricia B. Appel[1,2], Henrike Böck[1,2], & Andreas Riener[1,2]
[1]Technische Hochschule Ingolstadt, Germany
[2]Johannes Kepler Universität Linz, Austria

Abstract

Urban Air Mobility (UAM) includes, among others, passenger drones that transport passengers from A to B along fixed routes and predefined stops (vertiports). As with any future technology, there are various issues to be resolved before widespread use by potential customers can be expected. Among the main factors, high user acceptance and customer experience are responsible for the success of this technology. To better understand what customers expect from a passenger drone flight along the entire journey (including booking, check-in, security control, boarding, actual flight, etc.), several user workshops with a total of $N = 21$ participants were conducted. The results are presented in an attempt to help other researchers in the field of UAM to get started. The findings include the need for safety briefings, the display of information about flight status and weather, as well as flexible luggage options. In addition, future passengers expressed a desire for a clean environment, clear security measures, time-saving processes, and good accessibility. The greatest emphasis was placed on qualified personnel at every single stage of the customer journey. In summary, for a successful start of UAM, user requirements must be considered to create an inclusive and accepted future mobility solution.

Introduction

Urban Air Mobility (UAM) is a current research topic encompassing different transportation options, e.g., delivery drones and passenger drones. The latter are the topic of this research paper and further explained. Passenger drones are designed to fly passengers from A to B. They are usually equipped with electric or hybrid-electric propulsion (Doo et al., 2021) to avoid air pollution. Passenger drones are typically constructed as eVTOLs – electric vertical take-off and landing vehicles (Doo et al., 2021). Most of the recent developments of passenger drones envision the capacity of one to four people. These aerial vehicles are not fiction anymore, several companies are already working on their own passenger drones, e.g., Volocopter or Lilium. For 2024, it is planned that the first commercial passenger drones will fly around Paris, where the Olympic summer games will also be taking place (Volocopter, 2023). Landing spots (vertiports) are being designed and the passenger drones by Volocopter will be the first aerial vehicles of this kind in Paris (Volocopter, 2022).

However, plans only made by companies or governments will not automatically lead to a high acceptance among citizens. Especially for a relatively new mode of

transportation, it is important to involve the public in the entire process of integrating passenger drones into their daily lives (Edwards & Price, 2020). Because passenger drones are currently not operating commercially in cities, therefore it may not be clear how passenger drones will affect daily lives. A study found that about 67 % of people who have never seen a passenger drone flying would be willing to fly a Volocopter passenger drone (Planing & Pinar, 2019). Thus, according to this study in 2019, two-thirds of the people surveyed were already interested in flying with this new mode of transportation. On the other hand, that still leaves one-third who were not willing to take a flight with a passenger drone. Although there are also non-users e.g., residents affected by the noise of the drones, this paper only focuses on potential users.

For a successful launch, it is crucial to work with potential users to develop solutions that suit everyone. First, it is necessary to understand what the public (users, affected non-users) expect from passenger drone flights. By understanding expectations, acceptance by the public can be positively influenced. There are several factors that contribute to public acceptance of passenger drone flights. The most important issues of passenger drones according to the expectations of participants are accessibility, environmental awareness, movement of the vehicle, passenger wellbeing as well as perception of safety and soundscape (Edwards & Price, 2020). Those factors are generic and can be applied to specific steps along the customer journey – e.g., (potential) customers interacting with either products, institutions or brands through various mediums or touchpoints (Harris et al., 2021) – of a passenger drone flight. There are already customer journey maps for air travel (e.g., Inversini, 2017) which can be used as a basis for eVTOL flights, since a passenger drone flight basically involves the same steps as a conventional flight. Some stages are also similar to train rides or bus rides, e.g., the purchase of a ticket and the first mile to the station. Considering that every mode of transportation has its individual requirements and passengers expect different specifics, passenger drone flights need to be further investigated (e.g., through an adapted customer journey) to identify specific requirements and expectations of potential users and to develop solutions.

Method

The goal of this study is not to compare different customer journeys or to find out if customer journeys of air travels are similar, but to achieve specific insights with regard to participants needs in every single stage throughout a twelve-step customer journey of a passenger drone flight. The ultimate goal is to provide the reader with an idea of what people expect from a drone flight.

Participants

In total, $N = 21$ participants were invited to three workshops in August 2022 in the *Wissenschaftsgalerie* in the pedestrian street of Ingolstadt (Germany). The single workshops included $n = 6$, $n = 7$ and $n = 8$ participants. During the recruitment process, particular attention was given to considering different age groups and balanced gender distribution. Detailed demographic data about age and gender was not collected, participants were only observed. Students, employees, families, but also people of retirement age were invited. All participants received a monetary compensation of €30. German as a mother tongue was required for participation in the workshops.

Materials

In previous work by Janotta and Hogreve (2022), twelve customer journey stages were predefined. A questionnaire was prepared to collect further insights and specific information at certain stages during the workshops. Furthermore, three personas of potential target groups (business, leisure traveller, and families), were created and handed to the participants to better empathise with the target groups.

Procedure

All participants first received an introduction to the topic of urban air mobility, including insights into current concepts of passenger drones, vertiports and potential operational scenarios. Also, a short explanation of the concept of customer journeys was given, followed by a short introduction round among the participants. Afterwards, the main part of the workshop, evaluating people's expectations of the customer journey of passenger drones began. The process was inspired by a brainwriting method, called 6-3-5 method (Litcanu et al., 2015). All participants received the printed customer journey stages, questionnaire, and personas that they were asked to empathise with during the workshop. In each group a study leader kept track of time and answered questions from the participants. The customer journey stages were worked through sequentially. For each step, participants had five minutes to discuss their expectations and potential pain points, before completing the relevant section of the questionnaire. They wrote down their thoughts on sticky notes and pasted them directly onto the printed paper. Each participant was given additional sticky dots to highlight particularly important expectations or pain points. All completed journey stages were then clearly visible pinned to a pin board. After working on the twelve steps, all participants had an open discussion to find a solution to the stated pain points. The brainwriting and solution finding sessions were audio recorded and later transcribed.

Statistical Evaluation

Designing the study as participative workshop, it contained a qualitative survey. Statistical evaluation of the questionnaire was performed with the statistic programmes JASP (version 0.16.4) and RStudio (version 1.1.456). The collected data and results within the workshops were clustered, categorised and evaluated by frequency.

Results

As mentioned before, the results of the study are based on the workshop and questionnaire findings. However, this paper will primarily focus on the expectations, supported by insights gained through the additional questionnaire.

Workshops

During the workshops, many ideas and expectations of the participants could be collected that were clustered and evaluated by frequency. For each step of the

customer journey, the most frequently mentioned expectations will be presented. Results were also processed into a chart (*Figure 1*).
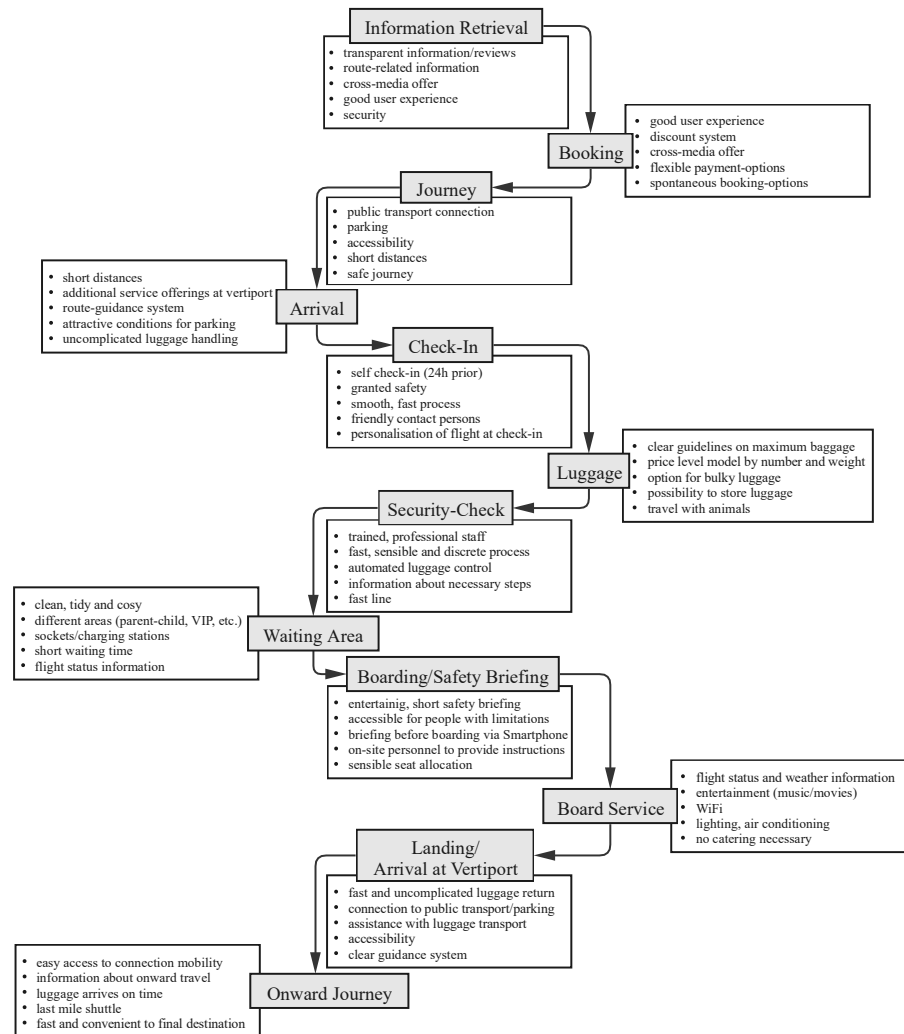
Figure 1. Participants expectations along the customer journey of passenger drones. The five most important expectations in descending order are listed for every step of the customer journey.

Information retrieval and booking
Regarding information retrieval before the actual booking, participants placed a high value on transparent information (e.g., reviews and experience reports) as well as route-related information. Information retrieval should be possible on multiple platforms such as apps and websites of airlines, blogs etc. In the booking process, participants mainly wished for a good user experience (easy to use (booking) websites

and apps, but also the option to have personal consultation during the booking process).

Journey to vertiport and inside vertiport

In the next step of the customer journey, participants had to think about traveling to the vertiport. People expected a very good accessibility and connection to the vertiport by public transport or by car, offering enough parking spaces. At the vertiport, people wished for a clear route guidance system to always know which way to go to their flight. Check-in should be possible independently around 24 hours before the flight to save time. The contact to (friendly) staff was among the most important expectations participants had from this point on through the whole journey inside the vertiport up to the security check. During the check-in process, people wished for the possibility to have some additional booking options (e.g., snacks, entertainment) to their flight and thereby personalise it. Since passenger drones are quite small, luggage is a big topic. Clear guidelines on maximum baggage weight, allowed objects were the most important required expectation of the participants. Also, they wished for different price models depending on the size and weight of the luggage and a place to store luggage during the flight. For security check, the participants asked for clear information about the process and necessary steps, e.g., taking off belts or shoes. Regarding the waiting area and the building of the vertiport the greatest expectation was that the building is clean, tidy, cosy, equipped with Wi-Fi and has enough sockets to charge phones, laptops, etc. A separation into different areas e.g., children-, VIP-, silent-areas was also often mentioned.

Boarding and flight

In the next step, the actual flight experience begins with boarding as well as a short and entertaining safety briefing. Also, an option to have the security briefing already before boarding via smartphone was an important option for many participants. Boarding the drone should be easily accessible, also for people with wheelchairs, strollers, or walkers. As with the previous steps, participants wished for on-site personnel to provide instructions and help. During the flight, permanent displays of flight status and weather were mentioned as important requirements.

Arrival at destination and onward journey

After landing and arriving at the vertiport, luggage had the highest priority. Participants mentioned that baggage claim should work out quick and easy. Just like at the arrival at the departure vertiport, a clear guidance system and short distances are needed to get to connections like public transport or parking spaces. Accessibility is highly recommended by the participants also in this step. A live status information about the onward journey in the drone and at the vertiport is desired. To have a fast and convenient transport to the final destination, participants wished for shuttles that can be booked additionally to drive passengers to their destination.

The criterion of guaranteed safety was never rated as a top expectation but was applied sporadically throughout the entire customer journey.

Questionnaire

The aim of the questionnaire was to get additional insights into specific topics and customer actions along the customer journey. Specific questions on e.g., the preferred luggage transport options and willingness to pay were asked. In total, all $N = 21$ participants filled out the questionnaire.

The participants ($n = 20$) would pay on average $M = €60.10$ ($SD = 45.47$), further results depicted in Figure 2. The majority ($n = 17$; 80.95 %) of participants ($n = 21$) would introduce reduced prices or even tickets free of charge ($n = 1$; 4.76 %) for small children. A reduction should be in $MD = 50.00$ % ($M = 42.00$ %, $SD = 18.59$). Detailed results of the questionnaire can be found in Tables 1 and 2.
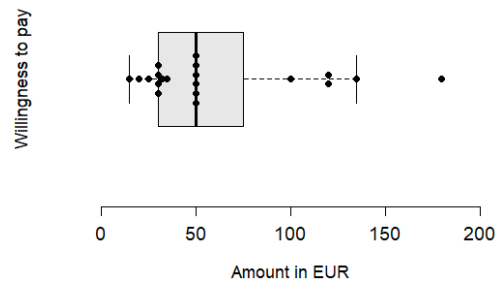
Figure 2. Boxplot of the amount participants were willing to pay for a flight from Munich Airport to Ingolstadt Nordbahnhof (ca. 53 km). For the same route, prices were given for Taxi (€ 135), Airport Express Bus (€ 25), Car (€ 28) and Train ride (€ 26).

Table 1. Part 1 of the results of the questionnaire. *Multiple answers were possible.

Item	Response option	n	%
Where would you book a drone flight?*		21	
	Apps	13	61.90
	Booking website	11	52.38
	Travel agency	4	19.05
	Vending machine	10	47.62
	Counter at airport/railway station	8	38.10
When would you book the ticket?*		21	
	On the day of flight	10	47.62
	A day before	10	47.62
	A week before	13	61.90
	A month before	5	23.81
	Several months before	1	4.76
Would you wish a motorised transport to the vertiport?		21	
	Yes	8	38.96
	Rather yes	9	42.86
	Neither yes nor no	4	19.05
	Rather no	0	0.00
	No	0	0.00

*Table 2. Part 2 of the results of the questionnaire. *Multiple answers were possible.*

Item	Response option	n	%
Is the vertiport airside or landside of the airport?*		21	
	Airside	6	28.57
	Landside	15	71.43
How strict should the security control be?*		21	
	As known from the airport	4	19.05
	Easier security control than at airport	15	71.43
	No security control	3	14.26
How long should the security control incl. waiting time max. be?		21	
	0-5 min	8	38.10
	6-10 min	8	38.10
	11-15 min	3	14.29
	16-20 min	2	9.52
	Longer than 20 min	0	0.00
When would you check your baggage?*		21	
	Check through after flight	3	14.29
	At check-in at the vertiport	9	42.86
	Hand over to staff at drone	8	38.10
	No checked baggage necessary	4	19.05
How should the transportation of your luggage be?*		20	
	Ground transportation	1	5.00
	Baggage compartment in drone	19	95.00
Where would you place your pets (dogs and cats) in the drone?*		21	
	Separate compartment	8	38.10
	Box underneath the seat	7	33.33
	Next to you, instead of a neighbour	7	33.33
Where would you place stroller and walker?*		21	
	Baggage compartment in drone	17	80.05
	Passenger compartment	5	23.81
	Ground transportation	0	0.00
What is most important to you regarding perceived safety? (Ranking)		20	
	Appealing design of the drone	5	25.00
	Smooth operation	11	55.00
	Neat and friendly staff	2	10.00
	Comprehensive safety briefing	2	10.00

Discussion

The aim of the workshops was to investigate people's expectations on the concept of passenger drones. Therefore, three workshops in the city of Ingolstadt, Germany, were conducted. Ultimately, the goal was to compile a collection of the expectations that were most cited during the workshops. Summing up all mentioned expectations, main areas can be named that should be particularly focused in the process of implementation of passenger drones as a mode of transport.

Pricing

The first barrier that determines if passengers are flying with a passenger drone is the price of a flight. Participants would pay ~ €61 for a ~53 km linear distance flight from Munich airport to Ingolstadt Nordbahnhof (~ €1.15 per km). This desired price might be lower than the actual costs of a flight. Ploetner et al. (2020) assume among others €4.94/km for a 40 km long flight. The workshop participants want a reduced ticket fee for small children. This is common in airliners. If it is also applicable in a small drone must be investigated by future research. It could be the case that a drone flight is not economically viable if only two or three passengers in a four-seat drone are paying the full price and a child is only paying half of the adult price, because the drone still needs to take off and head to the destination. On the other hand, the energy consumption calculated by the weight of a child is probably less than the energy consumption of a grown adult. So, maybe a discount in the scope of the difference of the energy consumption of child and adult might be a compromise.

Staff

The high need for (competent) staff in nearly all steps of the customer journey could be derived from a severe lack of employees at airports (check-in/security control and many more departments) in 2022 (e.g., Dissemond, 2022). The workshops took place in August 2022 where news about a staff shortage was all over the media. Maybe the absence of staff led to a realisation of the importance of employees along the customer journey of a flight. Whether workshop participants were affected by long waiting times at the airport was not evaluated. As a controversy, it is noticeable that the participants do not link neat and friendly staff to their perceived safety of a drone flight.

Accessibility and connectivity

If passenger drones should be perceived as an acceptable means of transport, it is important to make it accessible for almost anybody. Therefore, people with limitations as well as families with children must be considered during the design of every single stage of the customer journey (e.g., Soltani et al., 2012). Guidance systems are needed to show people the fastest way to their destination. Vertiports must be very well-connected to public transport and have a good infrastructure for individual mobility to keep distances short and accessible. Most participants in the study by Shaheen, Cohen, and Farrar (2018), conducted in the US, would maximally travel between 0 and 30 minutes to the next vertiport and in total 25 % would go to the vertiport by

public transport and 39 % by car. This underlines the importance of vertiports close to traffic nodes.

Baggage options

As mentioned before, baggage is a big topic when it comes to passenger drones. During the workshops, participants expressed various expectations regarding baggage transport, e.g., pricing models depending on the amount and weight of the baggage. From the perspective of power consumption, leaving out baggage would save energy, but as shown in the workshops, most of the participants want their luggage to be also flown together with them to their destination. To increase acceptance among people, this wish should be considered. As there are no studies (to the authors knowledge), the topic of luggage and acceptable transportation options from the perspective of both, companies and passengers, needs to be further investigated.

Comfort

Regarding vertiport and drone design in general, different needs of all people should be met, e.g., with a separate space for kids or VIPs. Furthermore, air conditioning and adequate lighting are main factors that contribute to passengers' well-being together with a low ambient noise level. It must be mentioned that the workshops were conducted on very warm days in August, so this could have influenced participants. However, it underlines the importance of air conditioning but also heating in winter to give passengers comfort which in fact also contributes to a higher acceptance of passenger drone flights.

Organisation

When looking at the booking section in the customer journey, it is noticeable that many participants want a multimodal offer for information seeking and booking process. There should be apps, websites, ticket counters or venue machines available to the people, and they also would like to have testimonials or influencers telling them more about a drone flight – especially in the beginning, when the technology is new and not many people have flown with it before. Transparency along all steps of the customer journey is a big factor to strengthen the acceptance of this new technology and the faith in the safety of a drone (e.g., Eißfeldt & Biella, 2022).

General

Going back to the introduction of this article, where the most important issues of passenger drones according to a study were mentioned (accessibility, environmental idea, vehicle movement, passenger wellbeing, perception of safety, and soundscape (Edwards & Price, 2020)). The workshop results show similar concerns as in Edwards and Price's study. However, consideration of environmental impact and motion were not addressed by the participants. They expect among others a comfortable journey, a vertiport that is easy to access by car or public transportation as well as accessibility for all people and a high safety standard. The latter was not mentioned often, what could be an effect of a strong requirement. The participants could demand safety of the system so strongly, they did not name it in every stage. One could speculate that

participants assume safety as granted. The workshops show that in every customer journey stage, participants have various expectations on a passenger drone flight that need to be met by future companies to get a high acceptance of potential passengers.

Limitations

Since there are currently no passenger drones commercially flying in Europe, the whole study setup was very hypothetical. There are still many open questions that need to be answered before UAM and especially passenger drones prove to be accessible and available means of transport. For that reason, participants had many questions, e.g., about safety checks or luggage transport in their minds that could not be answered adequately because there are no precise rules yet. This made the workshops even more abstract for the participants and they often had to rely on their experience with ordinary planes and helicopters.

Although the participants were every now and then reminded about the use of the personas, they did not fully see every customer journey step from the perspective of their assigned persona. The participants were also influenced by their own interests and expectations.

Even though very good German skills were required for participating in the workshops, there were difficulties regarding communication within the groups. Participants with a limited knowledge of German language could barely interact with other participants and therefore could not contribute to the workshop on a level that would have made a valuable impact on the outcome.

Conclusion

The goal was to identify what people expect from a passenger drone flight along a twelve-step customer journey. The results from the three conducted workshops suggest that it is highly important to give passengers the opportunity to always contact somebody for assistance with any issues coming up along the entire journey. Moreover, a high need for transparency in flight and weather information, booking processes, safety instructions, and safety measures was also a result of the study. People expect passenger drones to be time-saving and a comfortable mode of transportation. Unexpectedly, accessibility was named very often, making it a very important topic to the participants. Surprisingly, safety of passenger drones themselves was not among the most common expectations, which could be a hint that safety is taken for granted. The implementation of the factors price, accessibility and connectivity, baggage options, comfort, and organisation need to be investigated in further research.

Acknowledgements

Katholische Universität Eichstätt, who provided us with the predefined and elaborated customer journey.

References

Doo, J.T., Pavel, M.D. Didey, A., Hange, C., Diller, N.P., Tsairides, M.A., Smith, M., Bennet, E., Bromfield, M., & Mooberry, J. (2021). *NASA Electric Vertical Takeoff and Landing (eVTOL) Aircraft Technology for Public Services - A White Paper* (Report NASA Document Number 20205000636). National Aeronautics and Space Administration Transformative Vertical Flight Working Group 4 (TVF4). https://ntrs.nasa.gov/citations/20205000636.

Dissemond, S. (2022). Summer holiday chaos for understaffed German tourism sector. *DW.* https://www.dw.com/en/summer-holiday-chaos-as-german-tourism-hospitality-sector-plagued-by-staff-shortages/a-62368973.

Edwards, T. & Price, G., (2020). *eVTOL Passenger Acceptance* (Report NASA/CR—2020–220460). National Aeronautics and Space Administration (NASA). https://ntrs.nasa.gov/citations/20200000532.

Eißfeldt, H., & Biella, M. (2022). The public acceptance of drones – Challenges for advanced aerial mobility (AAM). In *Transportation Research Procedia,* 66, 80–88.

Harris, P., Van der Veen, G., & Pol, H. (2021). Customer Journey: From Practice to Theory. In B.B. Schlegelmilch and R.S. Winer (Eds.), *The Routledge Companion to Strategic Marketing* (pp. 67–90), New York: Routledge.

Inversini, A. (2017). Managing passengers' experience through mobile moments. In: *Journal of Air Transport Management,* 62, 78–81.

Janotta, F., & Hogreve, J. (2022). *Personas in Urban Air Mobility – Final Presentation*. Internal Katholische Universität Eichstätt report: unpublished.

Litcanu, M., Prostean, O., Oros, C., & Mnerie, A.V. (2015). Brain-Writing Vs. Brainstorming Case Study For Power Engineering Education. In *Procedia - Social and Behavioral Sciences,* 191, 387–390.

Planing, P., & Pinar, Y. (2019). *Acceptance of air taxis. A field study during the first flight of an air taxi in a European city*. Stuttgart: Hochschule für Technik Stuttgart.

Ploetner, K.O., Al Haddad, C., Antoniou, C., Frank, F., Fu, M., Kabel, S., Llorca, C., Moeckel, R., Moreno, A.T., Pukhova, A., Rothfeld, R., Shamiyeh, M., Straubinger, A., Wagner, H., & Zhang, Q. (2020). Long-term application potential of urban air mobility complementing public transport: an upper Bavaria example. In: *Council of European Aerospace Societies Aeronautical Journal*, 11(4), 991–1007.

Shaheen, S., Cohen, A., & Farrar, E. (2018). The Potential Societal Barriers of Urban Air Mobility (UAM). National Aeronautics and Space Administration (NASA).

Soltani, S.H.K., Sham, M., Awang, M., Yaman, R. (2012). Accessibility for Disabled in Public Transportation Terminal. In: *Procedia - Social and Behavioral Sciences,* 35, 89–96.

Volocopter (2022). Vertiport Testbed for European Urban Air Mobility Testing Inaugurated in Paris. *Volocopter.* https://www.volocopter.com/newsroom/vertiport-testbed-for-eu-uam-paris/

Volocopter (2023): Volocopter brings urban air mobility to life. *Volocopter.* https://www.volocopter.com/urban-air-mobility/

Towards user friendly autonomous on-demand mobility: insights from a Wizard of Oz pretest

Fabian Schlichtherle[1], Sven Hornburg[2], Mario Rieker[2], & Wolfram Remlinger[1]
[1]University of Stuttgart, [2]University of Applied Sciences Schwäbisch Gmünd
Germany

Abstract

Wizard of Oz (WoOz) experiments are a good and well established means of testing technologies whose practical applicability has not yet progressed far enough. This also applies to the field of highly automated vehicles where driver vehicle interaction in passenger cars has been especially the focus of extensive attention. One aspect of WoOz research in the automobile context that has received much less awareness to date is shared automated vehicles (SAVs) as part of ride-pooling services. Findings of previous research on human vehicle interaction in WoOz environments have provided valuable insights. However, fundamental differences in the interior layouts as well as use cases and user groups of both passenger cars and SAVs, make it necessary to extend the application of this research method. We report on a preliminary test in which we realized the entire usage process of a ride with an autonomous ride-pooling service including multi-passenger interaction as a WoOz experiment. The main emphasis was on questions regarding technical feasibility as well as interface design aspects. Results indicated that WoOz is a suitable method to investigate user needs in SAVs and that this technique is a useful addition to the user-centred design approach.

Introduction

Necessity and current status of SAVs

High volumes of traffic, especially from private vehicles, are increasingly pushing cities to the limits of their infrastructure. In this context, automated driving can function as a key technology to make transportation of people as well as goods more efficient. However, vehicle automation alone will not solve the capacity problems of urban transport. On the contrary, it could even lead to an increase of car usage in the modal split, hence emphasizing the growing importance of sharing automated mobility concepts in the future (May et al., 2020). In this framework, SAVs and their associated business models offer promising solutions. In the use case of ride-pooling, SAVs could be a helpful complement in the middle of the two poles of public and private transport by combining similar routes of multiple individuals into one shared journey. The sharing of the trips and the absence of driving personnel suggest a low cost of travel. Although this is primarily a clear advantage of SAVs, at the same time it poses a latent danger to urban transportation systems, as low-priced alternatives to

public transport could lead to a cannibalization of the latter (Krueger et al., 2016). It is therefore highly important to design SAVs carefully and to choose attributes, like interior features or overall comfort level, in such a way that primarily private car users are persuaded to use the service. This necessitates very precise knowledge of the needs and requirements of various user groups. According to Dorynek et al. (2022), it cannot yet be observed that the vehicles currently used to provide manually operated ride-pooling services meet these requirements and makeshift modifications by operators are the consequence. This suggests that as of now, insufficient attention is being paid to the vehicle interior and the aspect of user-oriented design is being neglected by manufacturers.

Limitations of SAV pilot experiments

In recent years, SAVs have been tested under real traffic conditions in numerous pilot projects. The experimental operation of these novel vehicles makes an important contribution to their faster implementation as potential users are able to experience automated rides first-hand and provide researchers with valuable insights for instance regarding their acceptance of the technology (Riener et al., 2019), experience on board (Dreßler & Höfer, 2022) or issues regarding safety and security (Salonen, 2018). However, this approach to user opinion acquisition also has its drawbacks. For example, Heikoop et al. (2020) report that the limited speed of the vehicles is perceived by the users as problematic. In addition, the prototypical setting of the trips with usually short route lengths as well as the mandatory safety operator on board only allow limited conclusions to be drawn about the later fully autonomous use of the vehicles over a longer trip duration. Kolb et al. (2020) even argue that the shuttle buses under consideration should be assigned to level 2 according to SAE (2018) due to the permanent monitoring required. Given these limitations, combined with the need for upscale comfort levels in automated mobility services, there is a strong demand for research that addresses SAVs of varying comfort levels in highly realistic scenarios. Even though experimental settings in virtual reality already provide an ever-increasing visualisation quality and thus represent a good experimental solution, the use of the WoOz technique is still the one alternative in this issue that is able to provide the most realistic experience for users. In the automotive context, this means that a participant interacts with what appears to be an automated vehicle. The seemingly automated system is however operated by one or more human agents who covertly operate the vehicle and its interaction with the passenger.

Previous research and current shortcomings of WoOz research in SAVs

The practice of WoOz experiments already has a long tradition in the field of automated vehicle research. From a methodological point of view, the considerations of Baltodano et al. (2015) provide an important groundwork for this paper. With the Real Road Autonomous Driving Simulator (RRADS), the authors introduce a framework for conducting inexpensive WoOz studies, using regular, commercially available passenger cars. They manage to convincingly hide the driving wizard behind a partition separating the driver and passenger seats, which largely eliminates the need for structural modifications to the vehicle and significantly reduces the cost of the experiment. In addition, they present a protocol procedure that serves as a guideline for the chronological sequence of a WoOz study.

In another significant study, Meurer et al. (2020) report about providing subjects with an on-demand robotaxi service for one week, operated as a WoOz trial. This is particularly noteworthy as the experimental setting is one of the very few WoOz studies conducted specifically on SAVs and because it is largely integrated into the test subjects' everyday life hence the rides they order are individual as well as actually necessary and useful. One important finding in their work is that participants wish to receive real time information about the ride in terms of upcoming stops and delays but also about the current traffic situation and potentially hazardous incidents. A similar approach is described in Alpers et al. (2020), who refrain from integrating the test drives into the daily routine of their participants, but instead apply a more elaborate interior design to the WoOz vehicle. They also report a clear advantage of a higher level of information – here in the form of a digital assistant.

Overall, only very few literature can be found that reports the application of the WoOz method in the context of SAVs. Previous research in this area focuses mostly on user experience than the methodological aspects that such experimental settings entail. At the same time, methodologically oriented papers regarding WoOz in automated driving exclusively involve passenger cars (Baltodano et al., 2015; Wang et al., 2017), which limits transferability of their findings to the area of shared mobility. Though some basic considerations certainly also apply to SAVs, there is a strong need for an alignment and an extension of the methodological framework specifically considering shared mobility in WoOz studies. The aim of this paper is to point out major methodological challenges in said experiments by conducting a WoOz pretest.

Method

Test vehicle and modifications

The seven-seater version of a Hyundai Staria was chosen as test vehicle. Its exterior appearance was modified in three ways. The word mark of a fictitious ride-pooling provider ("muvit") was attached as a sticker to the centre of each side of the vehicle. Two more stickers with the words "self driving shuttle" were placed on the sides of the last seat row to increase the credibility of the vehicle's autonomy. On the right sliding door, there was a sticker representing a camera lens, which served as a mock-up entry control. Test subjects had to hold a QR code in front of it to gain access to the shuttle. Tinted side windows at the front of the vehicle to prevent views of the driving wizard's workplace as in Alpers et al. (2020) were not feasible due to legal limitations in Germany.

The largest modification to the vehicle's interior was a partition made of lightweight foam boards between the driver's and the second seat row which ensured that the researchers could not be seen operating the vehicle or the interaction prototypes. In addition to two word mark stickers on the left and right of the partition, there were also two stickers representing emergency stop buttons. These were placed in such a way that they were easily accessible from the left and right seat respectively. For video recordings of the participants, an action camera was mounted at mid-height in the left corner between the partition and the wall of the vehicle. The test vehicle was equipped with a built-in camera that allowed observation but not recording of participants throughout the experiment. In the middle, right-hand centre of the partition, two

tablets were mounted next to each other at approximately head height, which served as monitors for a wireless webcam broadcast of the current traffic situation (left) as well as to display information about the journey (right). To reinforce the impression of a ride-pooling service, magazines were placed in the map pockets for the participants to read during the ride. Figure 1 provides an overview of the vehicle's seat layout and the respective modifications made. Images of the vehicle's exterior as well as its interior can be taken from figure 2.
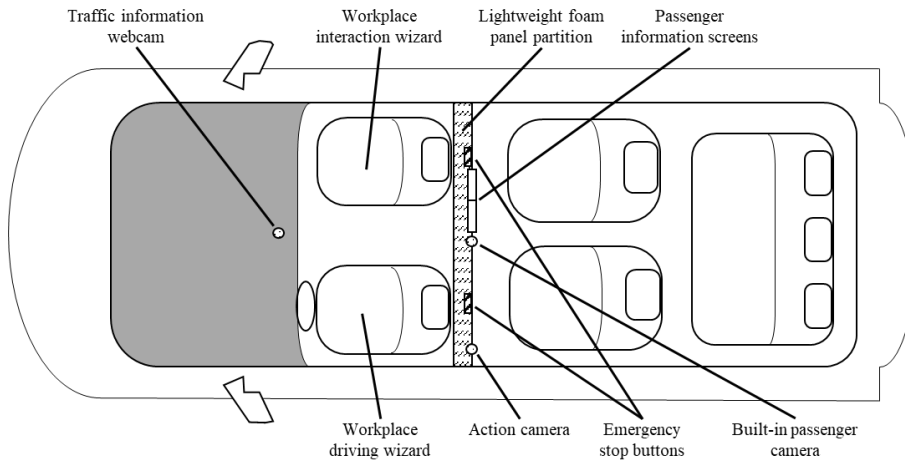
Figure 1. Seat layout and modifications of the test vehicle's interior.

Figure 2. The exterior (left) and the interior (right) of the WoOz test vehicle.

Interaction prototype

Participants' interaction with the mobility service in general and the vehicle in particular mainly involved two phases. These were (1) the booking process before entering the vehicle, when participants ordered a ride by using a prototypical mobile phone application and (2) during the ride, where two screens and a voice assistant provided information about the current traffic situation as well as the journey. The prototype for the mobile phone application was created with the UX prototyping software Figma and represented a mobility-as-a-service application that offered a choice of transport options including the "muvit" ride-pooling service. Participants were able to go through the entire booking process on a smartphone and finally book a ticket, whereupon they were presented with a QR code that served as identification

when boarding the vehicle. During the ride, participants could observe a webcam-broadcast of the current traffic situation on the left of the two screens. In addition, as suggested by Fröhlich et al. (2019), a streaming software was used to show the vehicle's current speed and highlight relevant events or road users by placing overlays in real time. The right of the two screens had the function of welcoming passengers, displaying the route and current position of the vehicle, and providing general information about the journey. Figure 3 gives an impression of the different interfaces used in the experiment.
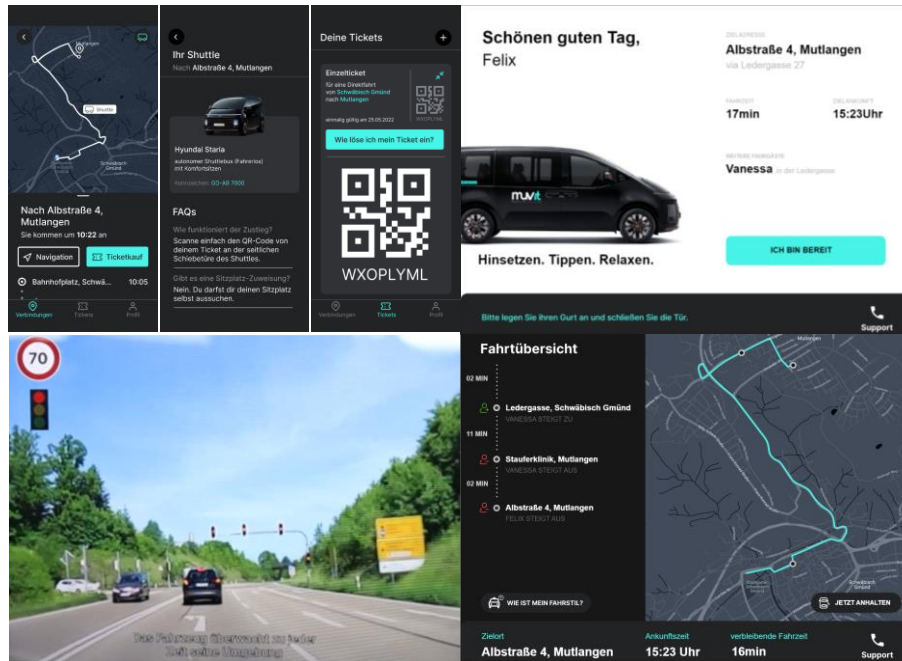
Figure 3. Mobile booking app (top left), passenger welcome screen (top right), traffic information screen (bottom left), and route information screen (bottom right).

Test route and study protocol

The experiment took place in May 2022 in the German city of Schwäbisch Gmünd with n = 8 participants. A driving route of approximately 15 minutes was chosen between the train station of Schwäbisch Gmünd and a suburban residential area in the neighbouring municipality of Mutlangen. As one of the aims of the trial was to integrate a multi-user component into WoOz testing, two additional stops, one shortly after the station and the other at a clinic in Mutlangen, were set along the route, at which a second participant got on and off. The route and relevant events during the journey are shown in figure 4.
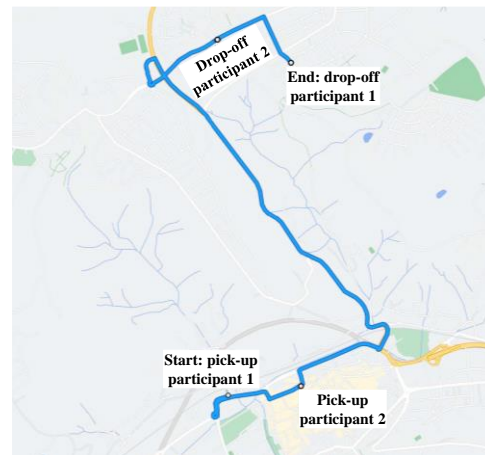
Figure 4. Route map of the experimental drive and relevant events.

The course of the trial can be divided into three phases: before, during and after the ride. Each run required two participants, who waited at separate locations for the trial to begin (participant 1: starting point; participant 2: pick-up point). Participant 1 was introduced to the testing procedure by the interaction wizard and was subsequently told to imagine they had just arrived at the station by train but missed their connecting bus. As an alternative to waiting for the next regular bus, participant 1 was instructed to book a ticket for a ride with an autonomous ride-pooling service on a smartphone they were given for this task. After the participant had booked the ticket, they approached the test vehicle from behind and entered it by presenting the QR-code to the camera lens on the sliding door. The interaction wizard took the front passenger seat, which was justified with legal requirements. A female voice assistant played over the vehicle's speakers welcomed the participant, instructed them to take their seat and put on a seatbelt. After safely stowing possible luggage, the participant tapped the "Start now" button on the welcome screen, whereupon the driving wizard began the journey. Throughout the entire ride, the interaction wizard triggered different voice outputs and integrated digital overlays into the traffic video stream. After a short ride of two minutes, the voice assistant announced that another passenger is going to join and after the vehicle stopped, participant 2 entered by presenting their QR code that had been sent to them earlier in order to minimize effort in the scope of this pretest. Once ready to start their ride, they also tapped the start button and the trip continued for approximately eight minutes until participant 2 reached their destination and exited the vehicle. After another minute of riding alone, participant 1 also reached their final stop and left the shuttle. In the post ride phase of the experiment, participant 1 was debriefed and offered a ride back to the starting point during which participant 2 was also picked up. Subsequently, both participants were asked about their experience with the ride. For this pretest setting, the opinions of the subjects were collected solely in the context of a short qualitative interview in which they were asked what should be continued, improved, stopped or added to the ride-pooling service.

Results

Qualitative post-ride interviews

Participants were positive about the comfort level of the service and the possibility to determine the departure of the vehicle by pressing the start button. The ingress process by presenting the QR-code was also mentioned as user friendly. Suggestions for improvement addressed the information screen display. A wish was expressed for the route progress information to be displayed more clearly during the journey. One point of criticism was that the live broadcast of the traffic ahead could be annoying or suggest that the passenger's attention is necessary. In addition, it was not always clear to the participants where exactly they should get off. A clearer marking of the exit location in the app was mentioned as a solution to this problem. Additional features that participants wished for were more space between the two seats and a table for longer journeys. Stronger individualisation of the digital displays was also suggested. Lastly, it was noted that a physical emergency stop switch instead of a sticker could provide more realism. With regard to the credibility of the vehicle's autonomy, there were mixed results. While some of the participants were quite convinced that they were sitting in an actually automated vehicle, others were at least sceptical due to their background knowledge of the current state of the art of automated driving.

Lessons learned

Conducting a feasibility study regarding a WoOz trial in a multi-passenger setting provided important insights and highlighted critical requirements for such testings. It can be assumed that there are distinct differences between WoOz testing for private passenger cars and SAVs. Based on the experiences made, the following basic guidelines for WoOz testing of SAVs are suggested:

1. Vehicle choice: Ideally, a larger van or minibus is used to convey the feeling of a purpose built vehicle that does not create the impression of a private passenger car.
2. Vehicle interior and positioning of participants: A generous amount of space should be offered which entails the avoidance of bench seats unless the examination of lower comfort interiors is deliberately aimed at. A back seat setup as is already common in representing level 5 automation (Bengler et al., 2020) is advisable. A partition that separates the front row from the rear of the vehicle provides a comfortable workplace for the interaction as well as the driving wizard.
3. Consideration of the holistic travel process: In order to map the entire process of a journey with an autonomous ride-pooling service, the associated booking process should also be integrated into the experiment. This includes both the prototypical interface of a booking app and a frame story that gives participants a comprehensible reason for their journey. Similarly, the experiment should not end inside the vehicle, but the drop-off process including complete egress should also be part of the investigation.
4. Consideration of multi-user settings: SAVs will only be able to contribute to reducing traffic volume if they are in fact shared by multiple passengers.

Therefore, WoOz studies in the context of SAVs should be designed in such a way that realistic interaction between multiple passengers is possible.

Discussion

Based on a review of current literature, this pretest is the first of its kind explicitly aimed to explore difficulties in realizing a WoOz study in the context of SAVs and to identify the prerequisites for it as well as establishing guidelines by focussing on methodological aspects like vehicle interior modification, interface design, or interaction among passengers. Essentially, it can be stated that this experimental method is well suited for evaluating the needs of future users and it should find a larger place in the repertoire of user-centred shared mobility research. However, it must be taken into account that established procedures from the field of WoOz research on passenger cars should not just be blindly adopted. While methodological problems such as the consistency of the driving wizard's driving style are certainly also a challenge, other areas specific to SAVs must be taken closely into account.

First, the selection of the test vehicle plays a greater role. The vehicle interior needs to receive more attention in the development of purpose built vehicles (Dorynek et al., 2022). Therefore, by using spacious cars, research in future WoOz studies for SAVs could focus particularly on ingress-egress processes with multiple passengers, as well as on privacy or in-vehicle security issues. Also imaginable are studies on alternatives to the classic seating layout, such as campfire or vis-à-vis arrangements. Second, interaction concepts like passenger information screens or voice assistants as well as a frame story that justifies the trip are supposed to immerse participants as much as possible. As was shown by Baltodano et al. (2015) and Alpers et al. (2020) the use of a welcoming digital assistant improved participants' attitude towards the vehicle. In this context, the recommendations of Large et al. (2019) regarding conversational user interfaces should be followed. Finally, the ideal solution in terms of a naturalistic usage of a ride-pooling service would be the approach by Meurer et al. (2020) to integrate the experiment into participants' daily life. However, this is hardly feasible especially if more than one participant is supposed to take part in the experiment at the same time. Nevertheless, a multi-user setting is indispensable in the development and testing of SAVs. As long as this type of vehicle is not tested in the shared state under real road conditions, it is difficult to draw reliable conclusions about the behaviour and experience of future users in shared interiors. It is debatable whether the additional passenger should be another naïve participant or a confederate researcher. While it is true that the simultaneous study of two participants in the same vehicle presents a methodological challenge and greater susceptibility to the influence of confounding variables, this way of conducting WoOz studies would represent the most realistic way of conducting shared rides.

Bengler et al. (2020) rightfully call for strictly monitored consistency in WoOz studies, especially in the characteristics of the driving wizard. This certainly also applies to studies of this kind on SAVs. Additionally, to counteract the methodological issues of WoOz experiments, results from these studies should always be compared to and put in relation with findings derived from other research methods like virtual reality studies, interviews, questionnaires or workshops. Still, the

outcomes of this pretest can be described as positive and increased usage of this method will certainly contribute to a more user-centred design of SAVs.

Acknowledgements

References

Alpers, B.S., Cornn, K., Feitzinger, L.E., Usman, K., So Yeon, P., Bardia, B., Hills-Bunnell, D.J., Trevor, H., Kaustubh, D., Rieko, Y., Larry, L., & Aquino Shluzas, L. (2020). Capturing Passenger Experience in a Ride-Sharing Autonomous Vehicle. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 83–93). New York. ACM.

Baltodano, S., Sibi, S., Martelaro, N., Gowda, N., & Ju, W. (2015). The RRADS platform. In G. Burnett, J. Gabbard, P. Green, S. Osswald, A. Eren, and V. Antrobus (Eds.), *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 281–288). New York. ACM.

Bengler, K., Omozik, K., & Scheiter, A.I. (2020). The Renaissance of Wizard of Oz (WoOz) - Using the WoOz methodology to prototype automated vehicles. In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (Eds.) (2020). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference* (pp. 63-72). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Dorynek, M., Derle, L.-T., Fleischer, M., Thanos, A., Weinmann, P., Schreiber, M., Schumann, S., Tunc, T., & Bengler, K. (2022). Potential Analysis for a New Vehicle Class in the Use Case of Ride-Pooling: How New Model Developments Could Satisfy Customers and Mobility Makers. *Vehicles*, *4*, 199–218.

Dreßler, A., & Höfer, E. (2022). User experience of a self-driving minibus - reflecting vision, state and development needs of automated driving in public transport. In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.) (2022). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference* (pp. 51-61). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Fröhlich, P., Schatz, R., Buchta, M., Schrammel, J., Suette, S., & Tscheligi, M. (2019). "What's the Robo-Driver up to?" Requirements for Screen-based Awareness and Intent Communication in Autonomous Buses. *I-Com Journal of Interactive Media*, *18*, 151–165.

Heikoop, D.D., Nuñez Velasco, J.P., Boersma, R., Bjørnskau, T., & Hagenzieker, M.P. (2020). Automated bus systems in Europe: A systematic review of passenger experience and road user interaction. In *Advances in Transport Policy and Planning. Policy Implications of Autonomous Vehicles,* 5, pp. 51–71.

Kolb, J.C., Wech, L., Schwabe, M., Ruzok, C., & Trost, C. (2020). Technische Aspekte des automatisierten Fahrens am Projekt des autonomen Shuttlebusses in Bad Birnbach. In A. Riener, A. Appel, W. Dorner, T. Huber, J. C. Kolb, and H.

Wagner (Eds.), *Autonome Shuttlebusse im ÖPNV,* pp. 57–91. Springer Berlin Heidelberg.

Krueger, R., Rashidi, T.H., & Rose, J.M. (2016). Preferences for shared autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, *69*, 343–355.

Large, D.R., Burnett, G., & Clark, L. (2019). Lessons from Oz. In C.P. Janssen, S. F. Donker, L.L. Chuang, and W. Ju (Eds.), *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (pp. 335–340). New York. ACM.

May, A.D., Shepherd, S., Pfaffenbichler, P., & Emberger, G. (2020). The potential impacts of automated cars on urban transport: An exploratory analysis. *Transport Policy*, *98*, 127–138.

Meurer, J., Pakusch, C., Stevens, G., Randall, D., & Wulf, V. (2020). A Wizard of Oz Study on Passengers' Experiences of a Robo-Taxi Service in Real-Life Settings. In R. Wakkary, K. Andersen, W. Odom, A. Desjardins, & M.G. Petersen (Eds.), *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (pp. 1365–1377). New York. ACM.

Riener, A., Appel, A., Dorner, W., Huber, T., Kolb, J.C., & Wagner, H. (2019). *Autonome Shuttlebusse im ÖPNV: Analysen und Bewertungen zum Fallbeispiel Bad Birnbach aus technischer, gesellschaftlicher und planerischer Sicht*. Springer Berlin Heidelberg.

SAE (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (SAE J3016). 400 Commonwealth Drive, Warrendale, PA, United States. SAE International.

Salonen, A.O. (2018). Passenger's subjective traffic safety, in-vehicle security and emergency management in the driverless shuttle bus in Finland. *Transport Policy*, *61*, 106–110.

Wang, P., Sibi, S., Mok, B., & Ju, W. (2017). Marionette: Enabling On-Road Wizard-of-Oz Autonomous Driving Studies. In B. Mutlu, M. Tscheligi, A. Weiss, and J.E. Young (Eds.), *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 234–243). New York. ACM.

Sound and vision: how can auditory displays support supervision of automated driving?

Canmanie T. Ponnambalam, Rins de Zwart, & Reinier J. Jansen
SWOV Institute for Road Safety Research
The Netherlands

Abstract

Higher levels of vehicle automation allow for less constant attention on the driving task, enabling drivers to engage in non-driving related activities (NDRAs). Shifts between levels of automation require self-regulation of NDRA engagement to remain sufficiently attentive for future changes in automation level. A visual HMI can provide information about current and upcoming changes in automation level but requires visual attention, compromising the convenience of automation. Auditory displays have been used to provide continuous information during monitoring tasks, but research into how sound can aid automation supervision in the driving context is limited. This online video study explored how auditory displays can augment a visual HMI of a self-driving vehicle while participants were engaged in an NDRA, comparing sounds that convey information about current and (time to) future system states through changes in volume, inter-pulse interval, harmonic series, and pitch. Adding sound improved perceived direction of change and remaining time until changes in automation level. Despite best efforts in sound design, sounds in the vehicle were initially perceived as negative even when they indicated something positive such as an upcoming increase in automation level. This and other implications for HMI research and design are discussed further.

Introduction

The mass adoption of highly automated vehicles promises many gains such as safety benefits, improved productivity, and lower adverse impacts on the environment (Greenblatt & Shaheen, 2015; Tafidis et al., 2022). However, it is largely accepted that the transition period will include the introduction of partially automated vehicles that will require a human driver to be present and available to take control of the vehicle under certain circumstances. This transfer of control between the automated vehicle and human driver is a safety critical task requiring a timely and effective takeover, during which the human must remain attentive and have suitable situational awareness.

Existing literature into auditory displays for supervision of automated driving often centres around the moment of transfer, with less attention given to how the driver can

be supported in the time leading up to this safety-critical point (de Winter et al., 2021). Transitions between different automation levels can occur not only when takeover is required, but also when moving between levels of automation that have different requirements of the driver (Lu et. al., 2016). With increasing data and improved technology, it will be possible to give advance notice on when some transitions are expected. For example, in the case where a vehicle is following a given route and will shift to a lower automation level when the vehicle leaves the highway and enters an urban area. This lower automation level will require the driver to have their eyes on the road. In such scenarios, information can be provided to the driver to help them self-regulate their engagement in non-driving related activities. Solís-Marcos et al. (2018) found that performance on a visual non-driving related activity (NDRA) decreased during automated driving compared to manual driving. The authors suggest that the supervisory task required in the automated driving condition (i.e., frequently looking at a system state symbol) was more disruptive to NDRA performance than controlling the vehicle in the manual driving condition, which supports research into how to make monitoring as non-demanding as possible. Previous work has looked at the use of visual displays to enhance situational awareness of drivers during automated driving by providing cues about automation level and time-to-takeover request (Tinga et al., 2022).

The present study investigates how auditory displays can be used in addition to visual displays to assist drivers (who may be occupied with NDRAs) in maintaining mode awareness and conveying time budget. Research has shown that sound displays can be highly beneficial during monitoring tasks, particularly when visual attention is compromised (Hermann et al., 2011, Nees & Walker, 2011). In increasingly automated vehicles, sound is becoming an important signal for crucial time-imminent information (e.g., blind spot monitoring, collision warnings), however it remains underutilized in delivering non-crucial but important information.

An online video study was conducted to investigate the benefits a sound display can provide to a visual information system that aids the driver in supervising autonomous driving. Several studies have proposed different properties (e.g., pitch, volume) of auditory signals that can convey concepts such as time and uncertainty. Research has shown that higher urgency and lower time budget can be communicated by increasing pitch (Baldwin & Lewis, 2014; Ko et al., 2021) or decreasing the inter-pulse interval between signals (Baldwin et al., 2012; Baldwin & Lewis, 2014; Nadri et al., 2021; Song et al., 2022). Pitch has also been used to indicate different severity levels in a medical patient monitoring system (Andrade et al., 2021), indicating the potential to communicate different levels of functioning. In terms of spectral composition, Edworthy et al. (1991) found that sounds with a regular harmonic series (e.g., all harmonics are integer multiple of the base frequency) were experienced as less urgent than sounds with an irregular harmonic series (e.g., non-integer multiples). Furthermore, Gray (2011) has shown that volume increase during playback (e.g., looming) can be used to communicate time-to-collision, resulting in a faster brake response time. Given their promise in past studies, the present study will investigate parameter manipulation in pitch change, volume, harmonic series, and inter-pulse interval in communicating direction of change of automation level and time budget. No studies were found that investigate the use of sound in the non-takeover

autonomous driving-monitoring context considered here. This study aims to answer the following two questions: 1) Does an auditory display provide additional support to a visual display in conveying upcoming changes in automation level? 2) Is there an added benefit to manipulating pitch, harmonics, and inter-pulse interval in conveying these upcoming changes?

Method

In a previous study, different visual displays were evaluated based on how well they communicated automation level, upcoming changes to automation level, reasons for the change, and remaining time until said change (Tinga et al., 2022). The present study builds on these results by incorporating an auditory display into their recommended visual display and by investigating different types of sound signals to see which best support a visually distracted supervising driver. An online user study was designed consisting of a self-guided survey with embedded videos. In the survey, participants watched videos that placed them in the driver's seat of a partially automated vehicle while it is driving itself (see Figure 1). The visual information system uses emoji-type icons to display information on the current and upcoming automation level when a change is pending. A bar that lights up across the dashboard indicates when a change is upcoming; the depletion of the bar indicates the time remaining until said change. Sound was presented when the vehicle is 60 seconds away from an upcoming change (the moment the bar appears), 30 seconds away, and 10 seconds away. The study was approved by the ethical board of SWOV – Institute for Road Safety Research.

Figure 1. A screenshot from one of the videos included in the study (adapted from (Tinga et al., 2022)). The smiley face icons indicate the current automation level and the upcoming change, and the green progress bar indicates time remaining until the change. Participants monitored the driving while counting the number of times a "4" appeared on the mobile phone.

Participants

Participants were recruited via Prolific (https://prolific.co), a commercial web-based tool for participant recruitment considered to produce high data quality in comparison

to other platforms for online behavioural research (Eyal et al., 2021). As the survey is in English, English fluency was a requirement, in addition to having a valid driver's license and owning a car. Participants were recruited only from the European Union or United Kingdom, where the road signs are consistent with those visible in the interface. They were required to continue with the study only if they were using a laptop or desktop computer, and not a mobile or tablet device.

In total, 200 people completed the study in full. Of these, four were rejected due to concerns about data quality related to the speed at which they finished the study and/or the low-effort responses received. Participants were paid at a rate of £9.90/hour. On average, the study took 18 minutes to complete. The participants were gender balanced, with 97 identifying as male, 96 identifying as female, 2 identifying as non-binary and 1 preferring not to identify. The mean age was 31 years (range: 19-73, *SD* = 10.3).

Sound Sets

This study investigates whether sounds encoded with information can provide intuitive understanding in support of a visual HMI, particularly when drivers are visually distracted. The body of research on auditory display design for supervision tasks is dwarfed by that focused on alarms and warnings. In the latter, three main types of signals have been investigated: auditory icons, earcons and speech or speech-based sounds. In attempt to combine the intuitiveness of auditory icons with the flexibility of earcons, an *echo* metaphor was designed to communicate the time until a change in automation level. In an echo, the time it takes for the sound to return, and the volume of the returned sound are indications of distance to the reflecting surface. This is an intuitive concept that may help to communicate the time until a change in automation reliability. To communicate the direction of change in automation level, an increase in pitch is used to indicate an increase in level (i.e., higher level of automation). Additionally, inharmonic tones may be experienced as a machine that is not running smoothly. This study examines whether this metaphor can be used to communicate a transition from or to a lower level of automation reliability. To isolate the effects of manipulations of these parameters, three sound sets were designed in Max (v8.3.1) that share properties as much as possible and build up with increasing complexity in terms of the manipulated parameters: 'basic', 'simple', and 'complex'. The 'basic' set is a single tone with an exponential decay. The 'simple' set contains two tones where the pitch and inter-pulse interval are manipulated to indicate whether automation level is going up (more automated) or down and in how long. In the 'complex' sound set, the pitch and inter-pulse interval as well as the harmonics and volume are manipulated. The sounds and their detailed descriptions can be found at: https://github.com/canmanie-swov/HMI_sound_study. A fourth 'silent' sound set did not include sound. Figure 2 provides a graphical depiction.

Experimental Conditions

The direction of an upcoming change in automation level (e.g., 'up' or 'down') was manipulated in addition to manipulating the sound set. Thus, a 4 (Sound set) × 2 (Direction) within-subjects experimental design was conceived, corresponding with a total of 8 experimental conditions. To restrict the total duration of the experiment,

each participant was presented with 4 conditions, with the constraint that all sound sets were experienced once and that each direction of change in automation level was experienced twice. Participants experienced the sound sets in a random order. A 'down' drive was always followed by an 'up' drive and vice versa.
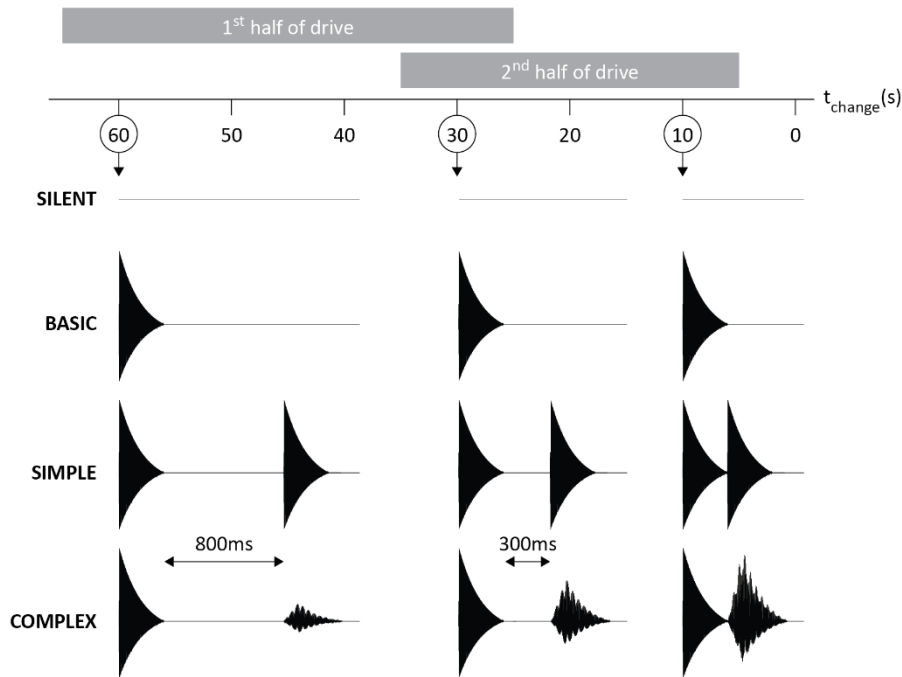
Figure 2: Timeline of each drive, divided into two halves presented separately. Four sound sets were available. Sounds (visualized here as waveforms) were presented at 60s, 30s and 10s before the upcoming change in automation level. The drive halves ended 25s and 5s prior to the change.

The drives corresponding to each experimental condition were split into two halves, the first half (duration: 40s) containing sounds at 60s and 30s until a change of automation level, and the second half (duration: 30s) containing sounds at 30s and 10s until a change (Figure 2). Experimental conditions featuring the 'silent' sound set did not include sounds throughout the drive. Each participant watched 8 videos in throughout the experiment: two drive halves by four different sound sets.

Non-Driving Related Activity (NDRA)

In a rapid serial visual presentation task, participants keep track of the number of times a target appears on the screen (Lee et al., 2006). This activity has no manual or physical component and does not require interaction with the system or an experimenter during the experiment, fitting with the online self-guided nature of this study. For this study, a single digit was visible for 500ms at 1000ms intervals, and the participants were asked to keep track of the number of times a "4" appeared in a

random sequence. The NDRA was present in every drive and condition and participants were instructed to complete it to the best of their ability.

Questionnaires

At the beginning of the study, participants answered demographic questions and questions about their driving experience and experience with partially automated vehicles. After each sound condition, participants scored their subjective demand according to three dimensions of the NASA Task Load Index (TLX) (Hart & Staveland, 1988): temporal, mental and frustration. They also answered questions related to the sounds heard in the videos, particularly about how noticeable, annoying, complex, understandable, and appropriate they found the sounds.

Procedure

After providing informed consent and filling out the pre-drive questionnaires, participants were shown the same 45-second familiarization video consisting of the car driving in the environment with a pending decrease in automation level due to upcoming road work. The instructions for the NDRA were given and participants were shown another 15-second drive with the NDRA present, then asked how many times they counted the target appearing. This was followed by an audio check with a test tone at 3 dB below the volume of the stimuli to ensure that participants had functional speakers that were set at a comfortable volume. They were instructed not to adjust their audio settings beyond this point for the remainder of the experiment.

Initially, no information was given about the visual or auditory information systems present in the vehicle. The video of the first half of the first drive was presented, and participants were asked *"What do you think the information system is indicating? Choose your best guess."* With four multiple-choice options relating to a vehicle error, automation error, takeover request, or automation level change (the correct answer). It was then revealed that the vehicle contains an information system that communicates the automation level, and the time remaining until the change. After this information was given, participants were asked how soon (in seconds) they felt the upcoming change would happen. The video of the second half of the drive was then shown, and participants were asked whether they thought the automation level was going to go up or down and, again, in how many seconds, see Figure 3.
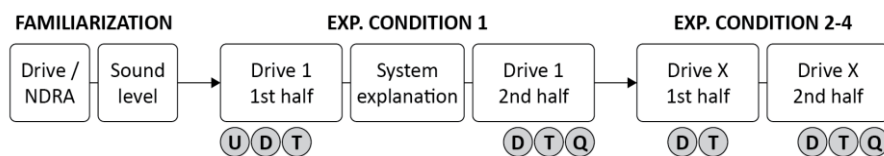
Figure 3: Schematic representation of the experimental procedure. Measures are depicted in gray circles. U = Intuitive understanding, D = Direction of upcoming change in automation level, T = estimated time until change, Q = Questionnaires on task load, usability and user experience.

In the remaining experimental conditions (2-4) the videos (6 in total) were accompanied by the sentence explaining the information system. After each video

participants were asked if they thought the automation level was going up or down and in how many seconds. At the end of each experimental condition (including the first one), participants completed the adjusted TLX, questions on usability and user experience.

Results

Analysis was performed using R. All reports of significance are compared against an alpha of .05.

Understanding of direction of change

A measure of intuitive understanding was compared between-subjects, investigating the effect of sound set on whether participants answered the question "What do you think the information system is indicating?" correctly. This question was asked only once (before participants were given an explanation about the system). The results of a chi-square test are significant ($\chi^2(1) = 7.80$, $p < .05$, Cramer's $V = .52$). However, pairwise comparisons with Bonferroni correction do not reveal significance between any pairs of sound sets. Only in the 'silent' (53%) and 'simple' (60%) conditions were most participants correct. Participants were 35% and 41% correct in the 'complex' and 'basic' conditions, respectively.

Binomial logistic regression was used to investigate the relationship between sound condition and whether participants gave the correct answer when asked in which direction automation level was going to change (either 'up' or 'down'). An additional First predictor (here, and in analyses below) was a binary flag indicating whether it was the first experimental condition (where information about the system was only given halfway, see Figure 3) to distinguish between intuitive understanding and understanding after having been given a system explanation. A Direction predictor indicated whether the actual transition was up or down. The model included participant as a random effect. Initially a full factorial model was tested. In the revised model, shown in Equation (1), only significant effects were kept.

$$Direction_{Correct} = Direction + First + Direction{:}First +$$
$$Direction{:}Sound\_set + (1|Participant\_ID) \tag{1}$$

The corresponding ANOVA yielded significant main effects of Direction ($\chi^2(1) = 13.32$, $p < .001$) and First ($X2(1) = 50.71$, $p < .001$). When Direction was 'down' the proportion of correct answers was significantly higher (76%) than when Direction was 'up' (66%). In the first experimental condition fewer correct answers (58%) were given than in subsequent experimental conditions (76%). The interaction between Direction and First was significant ($\chi^2(1) = 34.90$, $p < .001$), showing fewer ($\beta = -1.91$) correct responses when Direction was 'up' and First was 'true' compared to 'up' and 'false', respectively, $z = 9.20$, $p < .001$. Finally, the interaction between Direction and Sound set was significant ($\chi^2(6) = 15.31$, $p < .05$). Pair-wise comparisons showed that participants showed more ($\beta = 1.69$) correct responses when Sound set and Direction were {'complex','down'} than {'complex','up'}, $z = 5.64$, $p < .001$. Likewise, participants showed more ($\beta = 1.14$) correct responses when Sound set and Direction were {'simple','down'} than {'simple','up'}, $z = 4.01$, $p < .01$. Thus,

participants were more likely to answer correctly when Direction was 'down', except with the Silent and Basic sound sets.

Time Estimates

Participants estimated how much time was remaining before the automation level would change when the true answer was 25 seconds and 5 seconds. A poisson regression model investigated the relationship of Sound set, Actual (time remaining, a categorical value of either 25 or 5 seconds) and First (experimental condition) on how far time estimates deviated from the actual time remaining. Next to these main effects, the model included an interaction between Sound set and First, as well as participant as a random effect. Data were shifted so there were no negative values. Again, the revised model, shown in Equation (2), included only significant effects. In this model, only Actual was a significant predictor (χ^2 (1) = 729.29, $p < .001$).

$$Deviation \ = Actual + (1|Participant\_ID) \qquad (2)$$

The percentage of videos in which participants correctly guessed that time budget was decreasing (i.e., the second time estimation was lower than the first time estimation) was 56% in the 'silent' sound set, 59% in the 'basic' sound set, 62% in the 'simple' sound set and 68% in the 'complex' sound set. The effect of Sound set on whether participants estimated that the time budget was decreasing was initially investigated with a full factorial binary logistic regression model, with First (experimental condition) and Sound set as predictors, and with participant added as a random effect. Equation (3) describes the revised model, including only significant effects.

$$Decreased \ = First + Sound\_set + (1|Participant\_ID) \qquad (3)$$

The ANOVA of the model showed that First (χ^2 (1) = 17.07, $p < .001$) and Sound set (χ^2 (3) = 20.62, $p < .001$) were significant predictors. When First was 'true', significantly fewer ($\beta = -.62$) participants indicated a decreasing time budget than when First was 'false', $z = 4.12$, $p < .001$. Pairwise comparisons showed that the 'complex' sound set was associated with a significantly higher likelihood of a decreasing time budget than the 'silent' sound set ($\beta = .82$, $z = 4.40$, $p < .001$) as well as the 'basic' sound set ($\beta = .58$, $z = 3.11$, $p < .05$).

Within the instances where participants correctly estimated that time was decreasing, the gap between their time-estimates was closest to the correct time of 20s (i.e., the elapsed time between 25s and 5s prior to the automation level change) in the 'basic' sound set ($M = 18$s), in which the single tone notably does not include timing information. The gap in the 'complex' sound set deviated most from the target gap ($M = 11$s), with means of the 'silent' and 'simple' sound sets being both 14s.

Questionnaires

ANOVAs were performed on the NASA TLX scales, using sound set and order as independent variables. Sound set did not yield significant results. However, significant effects of First showed that workload was experienced as higher in the first experimental condition than in subsequent conditions on mental load ($F(1,776) =$

11.69, $p < .001$), frustration ($F(1,776) = 4.77$, $p < .05$), as well as on temporal load ($F(1,776) = 12.05$, $p < .001$). A significant interaction effect between First and Sound set ($F(3,776) = 4.65$, $p < .01$) showed that the effect of First on temporal load was mainly driven by the 'complex' and 'silent' sound sets. These effects of First can be expected as participants were given information about what the HMI communicated after the first half of the first video.

The usability and user experience questions were answered on a 5-point Likert scale and analysed for influence of Sound set with a Kruskal-Wallis test, in which no significant impact was found. Median responses are reported only if the result of a Wilcoxon signed rank test for significant deviance from the neutral position was significant. For all sounds, participants generally noticed them (median response = "Strongly agree") and felt it was obvious why they occurred (median response = "Agree"). The median response to the question "I found the sound notifications unnecessarily complex" was "Disagree" in only the 'basic' and 'simple' sound sets. And only in the 'basic' sound set were participants more likely to be in the direction of disagreeance for the question "I would shut off the system in my own car".

Discussion

The findings show that on first exposure to the information system, sound did not improve intuitive understanding regardless of the manipulated sound parameters. In fact, all sound sets evoked an association toward negative change (automation level going down). The most common usage of auditory signals within a vehicle is to indicate an issue or alert the driver some action is needed. Participants might have assumed this was the case and expected a system failure or takeover request to be the cause of the sound, instead of the intended message of upcoming automation level change. Further investigation is needed to determine if more complex sound designs are effective as well as which natural responses may be impacting the intuitive understanding of different sound manipulations.

Participants were most likely to understand that time budget was decreasing in the 'complex' sound set, particularly when compared to the 'basic' and 'silent' sound sets. Contrary to the latter sound sounds, time budget was explicitly encoded in the 'complex' sound set (through manipulation of inter-pulse interval). The 'simple' sound set also included this information, but performance increase compared to the 'silent' and 'basic' sound sets was non-significant. These results suggest sound could be an effective way to convey decreasing time budget, but more complex sound design may be needed than mere manipulation of inter-pulse interval. Interestingly, participants were closest in their estimate of a decreasing time gap in the 'basic' sound set. In one previous study, it was found that a single master alarm (such as in the 'basic' condition) was no worse than information-rich auditory icons in reaction time and accuracy (Cummings et. al., 2007). The present study hints that there may be some factors which can be communicated effectively through sound (time budget is decreasing), and others (how much time is available) that are better communicated by a visual HMI, supporting the idea that both should be used harmoniously.

Some studies have warned against the use of auditory feedback in vehicles that may increase distraction or load (Donmez et al., 2006, 2007). While no main effect of

Sound set on the NASA TLX was found, participants indicated that they were more likely to shut off the system in their own car in both the simple and complex sound sets, suggesting that there may be room in sound design to find an effective trade-off between increasing load with providing information.

References

Andrade, E., Quinlan, L., Harte, R., Byrne, D., Fallon, E., Kelly, M., Casey, S., Kirrane, F., O'Connor, P., O'Hora, D., Scully, M., Laffey, J., Pladys, P., Beuchée, A., & ÓLaighin, G. (2021). Augmenting critical care patient monitoring using wearable technology: Review of usability and human factors. *JMIR Human Factors, 8*(2). https://doi.org/10.2196/16491

Baldwin, C.L., Eisert, J.L., Garcia, A., Lewis, B., Pratt, S.M., & Gonzalez, C. (2012). Multimodal urgency coding: Auditory, visual, and tactile parameters and their impact on perceived urgency. *Work, 41*(SUPPL.1), 3586–3591. https://doi.org/10.3233/WOR-2012-0669-3586

Baldwin, C.L., & Lewis, B. A. (2014). Perceived urgency mapping across modalities within a driving context. *Applied Ergonomics, 45*, 1270–1277. https://doi.org/10.1016/j.apergo.2013.05.002

Cummings, M. L., Kilgore, R. M., Wang, E., Tijerina, L., & Kochhar, D. S. (2007). Effects of single versus multiple warnings on driver performance. *Human Factors, 49*, 1097–1106. https://doi.org/10.1518/001872007X249956

De Winter, J., Stanton, N., & Eisma, Y.B. (2021). Is the take-over paradigm a mere convenience? *Transportation Research Interdisciplinary Perspectives, 10*. https://doi.org/10.1016/j.trip.2021.100370

Donmez, B., Boyle, L.N., & Lee, J.D. (2006). The Impact of Distraction Mitigation Strategies on Driving. *Human Factors, 48*, 785–804.

Donmez, B., Boyle, L.N., & Lee, J.D. (2007). Safety implications of providing real-time feedback to distracted drivers. *Accident Analysis and Prevention, 39*, 581–590. https://doi.org/10.1016/j.aap.2006.10.003

Edworthy, J., Loxley, S., & Dennis, I. (1991). Improving auditory warning design: relationship between warning sound parameters and perceived urgency. *Human Factors, 33*, 205–231.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods.* https://doi.org/10.3758/s13428-021-01694-3

Godthelp, J. (1984). *Studies on human vehicle control.* PhD thesis, University of Delft. Soesterberg, The Netherlands: Institute for Perception TNO.

Gray, R. (2011). Looming Auditory Collision Warnings for Driving. *Human Factors, 53*, 63–74. https://doi.org/10.1177/0018720810397833

Greenblatt, J. B., & Shaheen, S. (2015). Automated Vehicles, On-Demand Mobility, and Environmental Impacts. Current Sustainable/Renewable Energy Reports, 2(3), 74–81. https://doi.org/10.1007/s40518-015-0038-5

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology, 52,* 139–183.

Hermann, T., Hunt, A., & Neuhoff, J. G. (2011). *The sonification handbook -* Sonification for Process Monitoring. Logos Verlag.

Ko, S., Zhang, Y., & Jeon, M. (2019). Modeling the effects of auditory display takeover requests on drivers' behavior in autonomous vehicles. Adjunct Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019, 392–398. https://doi.org/10.1145/3349263.3351508

Lee, J.D., McGehee, D.V., Brown, T.L. & Marshall, D. (2006). Effects of Adaptive Cruise Control and Alert Modality on Driver Performance. *Transportation Research Record: Journal of the Transportation Research Board, 1980*, 49–56.

Lu, Z.i & Happee, R. & Cabrall, C., Kyriakidis, M., & de Winter, J. (2016). Human Factors of Transitions in Automated Driving: A General Framework and Literature Survey. *Transportation Research Part , 43*. 183-196. https://doi.org/10.1016/j.trf.2016.10.007

Nadri, C., Lee, S. C., Kekal, S., Li, Y., Li, X., Lautala, P., Nelson, D., & Jeon, M. (2021). Effects of auditory display types and acoustic variables on subjective driver assessment in a rail crossing context. *Transportation Research Record, 2675*(9), 1457–1468. https://doi.org/10.1177/03611981211007838

Nees, M. A., & Walker, B. N. (2011). Auditory Displays for In-Vehicle Technologies. *Reviews of Human Factors and Ergonomics, 7*(1), 58–99. https://doi.org/10.1177/1557234X11410396

Solís-Marcos, I., Ahlström, C., & Kircher, K. (2018). Performance of an Additional Task During Level 2 Automated Driving: An On-Road Study Comparing Drivers With and Without Experience With Partial Automation. *Human Factors, 60*, 778–792. https://doi.org/10.1177/0018720818773636

Song, J., Wang, Y., An, X., Ma, S., Wang, D., Gan, T., Shi, H., Yang, Z., & Liu, H. (2022). Novel sonification designs: Compressed, iconic, and pitch-dynamic auditory icons boost driving behavior. *Applied Ergonomics, 103*, 103797. https://doi.org/10.1016/j.apergo.2022.103797

Tafidis, P., Farah, H., Brijs, T., & Pirdavani, A. (2022). Safety implications of higher levels of automated vehicles: a scoping review. *Transport Reviews, 42*, 245–267. https://doi.org/10.1080/01441647.2021.1971794

Tinga, A. M., Cleij, D., Jansen, R. J., van der Kint, S., & van Nes, N. (2022). Human machine interface design for continuous support of mode awareness during automated driving: An online simulation. *Transportation Research Part F, 87*, 102–119. https://doi.org/10.1016/j.trf.2022.03.020

Conceptualisation and evaluation of adaptive driver tutoring for conditional driving automation

Nikolai Ebinger, Sandra Trösterer, Norah Neuhuber, & Peter Mörtl
Virtual Vehicle Research GmbH
Austria

Abstract

Drivers are often overwhelmed by the complexity of automated driving. Recent research shows that driver tutoring is a promising approach to improve drivers´ understanding and use of driving automation. Presenting information in a stepwise, adaptive, and context-sensitive manner could help drivers to better understand and safely use conditional driving automation (SAE Level 3). The focus of the developed tutoring is to improve the transition of control from the automation to the driver. In a simulator study, drivers watched a pre-drive tutoring video and received auditive explanations and adaptive feedback based on their take-over performance while driving. To evaluate the tutoring, a group with tutoring was compared to a baseline group that only received written information. The results show that the tutored drivers had a more correct mental model and an improved attention allocation while taking back manual control. However, the take-over times did not differ significantly. Half of the tutored drivers reported that they changed their behaviour in response to the adaptive feedback. The findings presented in this paper confirm that adaptive and context-sensitive tutoring helps drivers to better understand the driving automation and to perform safer transitions of control.

Theoretical background and previous work

Need for driver tutoring

Automated driving aims to improve road safety and driver comfort, but also brings challenges for drivers. Conditional driving automation (SAE International, 2021, Level 3) can take over control in certain situations and allows drivers to perform non-driving related activities (NDRAs). This likely improves safety and comfort but challenges drivers to safely re-engage and regain control of manual driving after automated driving. Drivers should get a sufficient overview and take control in a suitable time frame which is not too fast and hectic but also not too slow. Indeed, research shows that getting the driver back in the loop can be challenging and error-prone (see Morales-Alvarez et al., 2020).

Current methods of how drivers learn to understand and use driving automation result in insufficient understanding and interaction skills. For learning how to use an automated vehicle, drivers usually rely on the owner´s manual as the primary source of information (Boelhouwer, van den Beukel, van der Voort, Hottentot, et al., 2020).

This is problematic as owner´s manuals do not sufficiently improve drivers´ general understanding and do not teach them to decide when to rely on automation (Boelhouwer et al., 2019b). As a result, drivers learn to use automation by trial and error as they try out different interaction strategies (Strand et al., 2011). For this and other reasons, drivers may have an insufficient understanding. For example, less than 60% of drivers who own a vehicle with adaptive cruise control are aware of safety-critical limitations (DeGuzman & Donmez, 2021).

Current approaches to driver tutoring

One approach to better prepare drivers for automated driving is tutoring. Building on previous studies that have used the term tutoring, we define driver tutoring for automated vehicles as technology-enhanced learning support focused on the safe use of automated vehicles. Existing training and tutoring approaches can be categorised into pre-drive tutoring, during-the-drive tutoring, and tutoring that combines pre- and during-the-drive tutoring.

Pre-drive tutoring approaches use different methods to introduce drivers to automated driving, such as simulator-based driver training (Hergeth et al., 2017; Payre et al., 2016), interactive quizzes (Forster et al., 2019; Forster et al., 2020), gamified information (Feinauer et al., 2022), video (Noble et al., 2019), or a virtual reality tour (Ebnali et al., 2021; Sportillo et al., 2018). Different pre-drive tutoring methods were also already applied combined (Payre et al., 2017).

Compared to an owner´s manual, pre-drive tutoring improves aspects such as drivers´ understanding (e.g., Forster et al., 2020) and interaction (e.g., Payre et al., 2017). In most cases, pre-drive tutoring did not improve transitions between automation levels (Forster et al., 2019; Forster et al., 2020) and mode awareness (Feinauer et al., 2022). Results on whether pre-drive tutoring impacts trust are mixed (e.g., Feinauer et al., 2022; Hergeth et al., 2017).

Boelhouwer, van den Beukel, van der Voort, Verwey, and Martens (2020) investigated tutoring while driving via visual overlay on the windscreen with verbal explanations. The explanations were adapted to the driving context and drivers´ behaviour. Tutored drivers had fewer collisions, better reliance calibration for the first of three drives, and a partially improved take-over performance. Rukonić et al. (2022) tutored drivers while driving using a voice assistant that provided verbal explanations.

Neuhuber et al. (2022) combined pre-drive and during-the-drive tutoring to present complex information before driving and to repeat important aspects in related driving situations. In addition to repeating important information, the tutoring while driving adapted to drivers´ reliance calibration. The tutoring increased drivers´ overall monitoring and taking manual control in high-risk situations. However, trust calibration did not significantly change.

Design of tutoring

Teaching literature suggests that multimodality, adaptivity, and stepwise information presentation are important for effective learning. Learning is more effective when

information is provided through multiple modalities, e.g., auditive and visual (Mayer et al., 1995; Wild & Möller, 2020). The need for adaptivity becomes apparent in an analysis of driving school instructors. They adapt their explanations to the situation and the drivers´ behaviour, e.g., by providing corrective feedback as soon as a critical situation is over (Boelhouwer et al., 2019a). A stepwise information presentation is part of most common training processes. Training is typically structured as a coordinated learning of content that builds on each other (e.g., visiting beginner lectures, then intermediate, …). The theory of knowledge spaces (Heller et al., 2006) provides a framework for understanding the interdependence of knowledge aspects.

Existing tutoring approaches show that teaching methods find application in driver tutoring. Most tutoring approaches use different modalities to explain information by combining auditive and visual information (e.g., Payre et al., 2017). Adaptivity was implemented via situation-sensitive information presentation and feedback on trust and reliance calibration (Boelhouwer, van den Beukel, van der Voort, Verwey, & Martens, 2020; Neuhuber et al., 2022). Neuhuber et al. (2022) applied a stepwise presentation of information by combining a pre-drive tutoring video with auditive during-the-drive tutoring.

Present study

The present study combines methods from previous training and tutoring approaches to improve the challenging transition of control from conditionally automated to manual driving. In addition to the transition of control, the developed tutoring supports the drivers in all aspects relevant for conditionally automated driving. Our hypotheses are: that the tutoring group acquires a better mental model of the driving automation used (H1), has an improved attention allocation during take-overs (H2), and takes over manual control less hectically, i.e. slower (H3) than the baseline group without tutoring.

Driver tutoring

The driver tutoring developed for this study is based on data (the same as used in Neuhuber et al., 2022) about typical challenges with driving automation and an analysis of drivers´ knowledge requirements. The resulting knowledge and skill requirements were structured following the theory of knowledge spaces (Heller et al., 2006). Figure 1 shows the resulting hierarchically dependent knowledge requirements from drivers of conditionally automated vehicles.

The tutoring presented information and explanations in a multimodal way. The pre-drive tutoring consisted of a video with audio. Drivers could replay individual video chapters while driving conditionally automated. Tutoring reminders and feedback during the drive were presented via audio. In this way, the tutoring explained the same information in a stepwise way and using different modalities.
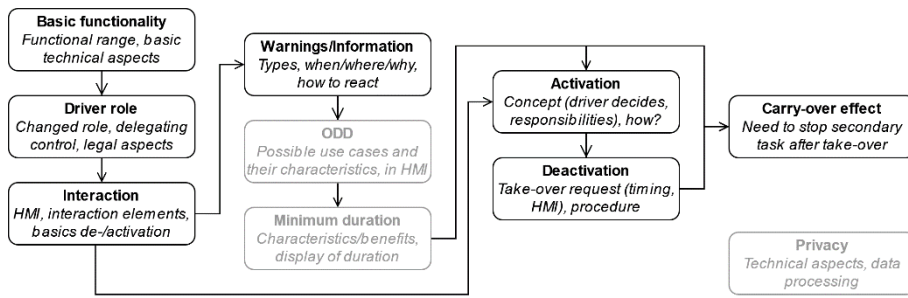
Figure 1. Hierarchical dependent knowledge requirements for conditionally automated driving. Grayed-out requirements were not applicable in the study setup and thus not included.

Tutoring while driving was context-sensitive and adaptive. Pre-recorded tutoring explanations and reminders were automatically triggered based on events in the driving situation, such as the availability of conditionally automated driving. Feedback on drivers´ take-overs were adapted to the take-over behaviour. As shown in Figure 2, the tutoring intensity decreased as the same driving events occur repeatedly.
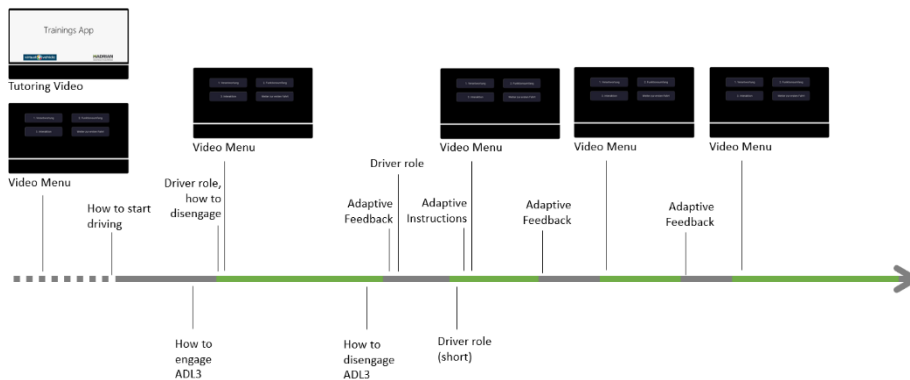
Figure 2. Overview of tutoring elements before and during the drive (dotted line = before driving, grey line = manual driving, green line = conditionally automated driving).

The tutoring video contained all the relevant information required to use conditional driving automation safely and was structured according to the hierarchical dependent knowledge requirements. Participants started the video by tapping on a mounted tablet. When the video ended, the tablet displayed an overview of the individual video chapters. Within the menu, the "continue to drive" button triggered an auditive explanation of how to start driving.

Auditive explanations during the drive were presented depending on the driving context. The auditive explanations included reminders on important aspects (e.g., driver role) and a detailed explanation of the interaction (e.g., how to activate the automation). Figure 2 gives an overview of the tutoring elements during the drive.

Adaptive feedback and take-over instructions were triggered based on the take-over time and an experimenter rating on the drivers´ monitoring behaviour. The experimenter observed the eye-tracking video and rated the visual monitoring based on pre-defined rules. Sufficient monitoring required that drivers gazed at the speedometer and the road before taking manual control. The take-over time was automatically categorized into too early (within 5 sec), in time (between 5 to 10 sec), too late (more than 10 sec). For an overview of the adaptive feedback, see Table 1.

Table 1. Adaptive feedback and instructions based on the driver behaviour.

Driver behaviour	Adaptive feedback	Adaptive instructions
Monitoring 👎 Take-over time 👍	"Good, you took control. Next time, get an overview of the situation and then take control. That means check your surroundings, traffic, and speed before you take over."	"Always remember to check your surroundings, traffic and speed before taking over."
Monitoring 👍 Take-over time 👍	"Great! You got an overview of the situation and then took control."	-
Monitoring 👍 Take-over time 👎	"Good! You got an overview of the situation. After that, you need to take control quickly. This means braking, steering, or pressing the red button."	"Always remember to take control after getting an overview."
Monitoring 👎 Take-over time 👎	"Next time, remember to get an overview of the situation and take control. This means checking your surroundings, traffic, and speed before taking over. Then brake, steer, or push the red button."	"Always remember to check your surroundings, the traffic, and your speed before taking over. To take over, brake, steer, or press the red button."

Baseline information

Participants in the baseline group received printed (one double-sided DinA4 page) written information that was inspired by the owner´s manuals that traditionally come with vehicles. The written information contained the same information as presented in the tutoring and was available to the drivers throughout driving. The written information included pictures of the user interface.

Method

Participants

The sample consisted of 20 participants with a gender distribution of 10 women and 10 men. Hereby two tested participants were replaced with new participants due to technical problems with the study setup during testing. The age of the participants ranged from 19 to 37 years ($M = 24.90$, $SD = 4.48$). On average, participants had held their driving licence for 2.60 years ($SD = 0.75$). Criteria used to select participants

included having a driving license and no or very limited experience with advanced driver assistance systems.

Experimental design and procedure

The study was realised as a mixed design with the between-subjects factor group (tutoring vs. baseline) and the within-subjects factor number of take-over (1 to 4). The hypotheses focused on the between-subjects factor group. Participants were randomly assigned taking into account gender balance.

Participants were welcomed and informed about the study. After signing the informed consent and answering a pre-questionnaire, participants sat in the driving simulator for eye-tracker calibration and a short familiarisation drive. Participants of the baseline group received written information and the tutoring group watched the tutoring video. Each participant drove on a countryside road for about 20 minutes and had to take back manual control four times. As NDRA, the participants played a video game ("Cut the Rope") on a hand-held tablet. After the drive, participants answered a set of questionnaires, including a mental model questionnaire and were interviewed by the experimenter. Each participant received a compensation of €20.

Setup and Materials

The study was conducted in a fixed-based driving simulator with automated driving functionality equal to SAE Level 3 (conditional driving automation; SAE International, 2021). A mounted tablet with external speakers displayed the tutoring and the status of the conditional driving automation, including take-over requests (see Figure 3).
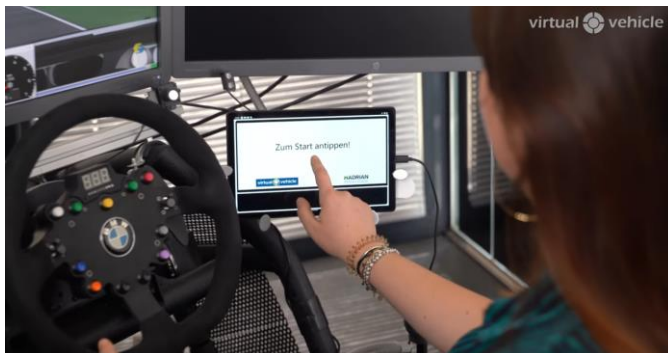
Figure 3. The fixed based simulator with a mounted tablet for displaying the tutoring and automation status.

Mental model

A set of 12 open-ended knowledge questions was used to assess the mental model. The items covered all aspects explained in the written information and tutoring and were answered in written form on a laptop. For example, the functionality item was "What tasks can the driving system perform?" and the transition of control item was

"What should be done after notification that the driving system will soon be unavailable?". A maximum of one point was awarded for each answer and points were scored proportionally. A mean score was calculated for each participant and used for the statistical analyses.

Eye-tracking
The "SmartEye Pro" eye-tracker was used to track the drivers´ gaze in relation to the "road and speedometer" area (see Figure 4). The data was stored as a dependent variable.

Figure 4. Picture of the road and speedometer area.

Take-over time
The time participants needed to take over manual control was processed live to trigger the adaptive feedback and stored as a dependent variable.

Results

The analyses had two aims: first, to assess whether the tutoring was more beneficial than written information, and second, to learn how the tutoring could be improved. To evaluate the tutoring, the tutoring group was compared with the baseline group regarding the dependent variables. The interview provides insights into what the drivers liked and disliked about the tutoring and how it could be improved.

Mental model

The open-ended mental model questions were scored (range 0 to 1), and a mean value was used for the statistical analysis.

The results support H1 that the tutoring leads to a more correct mental model than the baseline information. The tutoring group had a significantly more correct mental model ($M = 0.79$) than the baseline group ($M = 0.66$; $t(18) = -3.20$, $p = .005$, $d = -1.43$).

Gaze behaviour

Participants´ gaze during take-overs was analysed using the percentage of gaze on the "road and speedometer" area in the time between the take-over request and the deactivation of the automation. One participant is not included in this analysis due to missing eye-tracking data.

The results support H2 that the tutoring improves the attention allocation during take-overs. On average, the tutored participants gazed significantly more at the road and speedometer while taking manual control than the baseline group did.

A mixed ANOVA showed that there was a significant main effect for the group ($F(1, 17) = 5.49$, $p = .032$, $Eta^2 = 0.144$). The tutoring group gazed significantly more at the road and speedometer during take-overs than the baseline group (see Figure 5). There was no significant main effect for the take-over number ($F(3, 51) = 0.46$, $p = .712$, $Eta^2 = 0.013$) and no significant interaction between both factors ($F(3, 51) = .131$, $p = .942$, $Eta^2 = 0.004$).
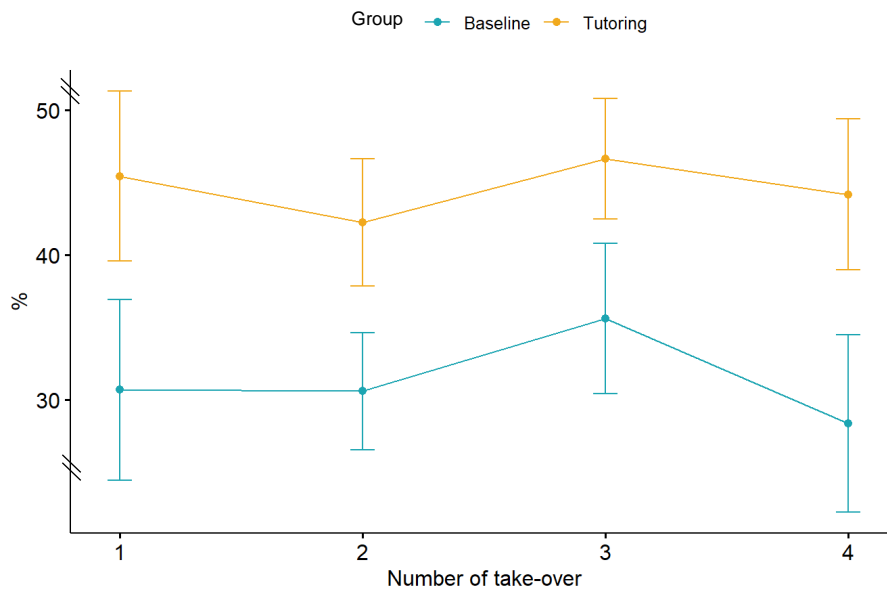
Figure 5. Percentage of gazing at the road and speedometer by group and number of take-over.

Take-over times

Participants´ take-over times were analysed using the time from the start of take-over request until the driver deactivated the automation. A robust version of the mixed ANOVA was used due to a violation of test assumptions (Mair & Wilcox, 2020). One participant is not included in this analysis due to missing take-over time data.

The results do not support H3 that tutored drivers take over manual control less hectically.

A robust mixed ANOVA showed that there is no significant main effect for the group ($F(1, 16.15) = 1.272$, $p = .276$), no significant main effect for the number of take-over ($F(3, 10.89) = 1.04$, $p = .412$), and no significant interaction between the group and the number of take-over ($F(3, 10.86) = 2.58$, $p = .107$). For descriptive statistics, see Figure 6.
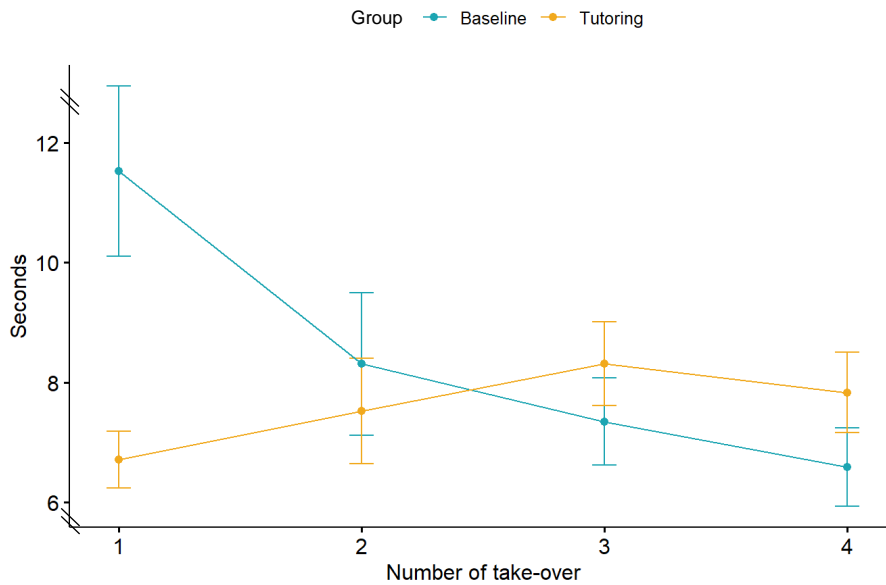
Figure 6. Take-over time (seconds) by group and number of take-over.

Interview analysis

In the interview, eight participants stated that the pre-drive video was the most important part of the tutoring. Six participants commented that the information was short and precise, that it was structured (2 participants), and that the level of repetition was good (2 participants). Half of the participants liked the context-sensitive explanations while driving (5 participants), three participants were neutral, and three participants found them too repetitive (one of them still liked it). Seven participants explained that they changed their behaviour after perceiving feedback from the tutoring. More specifically, four participants reported that they had increased their monitoring. However, two participants stated that they did not change their behaviour and one said "probably yes". All participants stated that they would like to have the support of tutoring when driving in a real automated vehicle. The fact that all

participants who tried the tutoring wanted to use it in a real car indicates a high level of acceptance.

Discussion

The study compared driver tutoring with a written owner´s manual. The tutoring consisted of a pre-drive video and context-sensitive and adaptive explanations while driving. Drivers who experienced the tutoring had a more accurate mental model and better attention allocation during take-overs than those who read an owner´s manual. However, the take-over times did not differ between the groups. Yet, in the interviews, drivers reported that they had changed their behaviour after receiving corrective feedback, suggesting that adaptive feedback had influenced elements of their behaviour other than their response times.

Combining pre- and during-the-drive tutoring improves drivers´ mental model of the driving automation. Tutoring drivers via video alone does not improve all aspects of the mental model (Noble et al., 2019). Combining it with during-the-drive tutoring, within a hierarchical knowledge needs framework (Heller et al., 2006), resulted in an overall improved mental model.

The tutoring used in this study is an effective method to improve drivers´ attention allocation during take-overs. Drivers with tutoring had a better attention allocation than drivers using an owner´s manual. By gazing more at the street while taking manual control, drivers had likely gained a better overview and thus took control more safely. That the attention allocation was improved from the first take-over on suggests that it was caused by both, the pre-drive tutoring, and the context sensitive tutoring. Accordingly, four drivers stated that they had increased their gazing on the street during take-overs after receiving adaptive feedback.

Tutoring does not lead to longer take-over times, which we would interpret as less hectic take-overs. Descriptive inspection of the data shows that the driver´s take-over times are generally within an acceptable range. However, applying our scoring system for adaptive feedback on take-over behaviour, the average time for the first take-over of the baseline group would be considered to be too slow. In comparison, the take-over times of the tutoring group were more consistent from the first take-over onward. Nevertheless, there was no significant main-effect for the take-over times between tutoring- and baseline group.

Developing tutoring with the right intensity is critical and could be made easier by including personalisation options. The interview showed that half of the participants liked the context-sensitive explanations. Although the tutoring intensity decreased over time, three drivers found it too repetitive. As different drivers are likely to have different needs, allowing drivers to change the intensity of the tutoring on an individual basis could be beneficial.

Outlook

Analysing the benefits of each tutoring element individually could allow a more systematic development of tutoring systems. Recent studies, including this one, often

analyse overall benefits of a set of tutoring elements. An analysis of each component (e.g., adaptive feedback) would be needed to learn which elements should be used for effective tutoring. This would contribute to the potential of tutoring to improve the user experience with automated driving, but also with other technologies.

Acknowledgments

References

Boelhouwer, A., van den Beukel, A.P., van der Voort, M.C., Hottentot, C., de Wit, R. Q., & Martens, M.H. (2020). How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands. *Transportation Research Interdisciplinary Perspectives, 4*. https://doi.org/10.1016/j.trip.2020.100103

Boelhouwer, A., van den Beukel, A.P., van der Voort, M., & Martens, M. (2019a). Designing a naturalistic in-car tutor system for the initial use of partially automated cars: Taking inspiration from driving instructors. *In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (pp. 410–414). ACM. https://doi.org/10.1145/3349263.3351504

Boelhouwer, A., van den Beukel, A.P., van der Voort, M.C., & Martens, M.H. (2019b). Should I take over? Does system knowledge help drivers in making take-over decisions while driving a partially automated car? *Transportation Research Part F, 60*, 669–684.
https://doi.org/10.1016/j.trf.2018.11.016

Boelhouwer, A., van den Beukel, A. P., van der Voort, M. C., Verwey, W. B., & Martens, M. H. (2020). Supporting Drivers of Partially Automated Cars through an Adaptive Digital In-Car Tutor. *Information, 11*(4), 185.
https://doi.org/10.3390/info11040185

DeGuzman, C.A., & Donmez, B. (2021). Drivers Still Have Limited Knowledge About Adaptive Cruise Control Even When They Own the System. *Transportation Research Record: Journal of the Transportation Research Board, 2675*(10), 328–339. https://doi.org/10.1177/03611981211011482

Ebnali, M., Lamb, R., Fathi, R., & Hulme, K. (2021). Virtual reality tour for first-time users of highly automated cars: Comparing the effects of virtual environments with different levels of interaction fidelity. *Applied Ergonomics, 90*, 103226. https://doi.org/10.1016/j.apergo.2020.103226

Feinauer, S., Schuller, L., Groh, I., Huestegge, L., & Petzoldt, T. (2022). The potential of gamification for user education in partial and conditional driving automation: A driving simulator study. *Transportation Research Part F, 90,* 252–268. https://doi.org/10.1016/j.trf.2022.08.009

Forster, Y., Hergeth, S., Naujoks, F., Krems, J., & Keinath, A. (2019). User Education in Automated Driving: Owner's Manual and Interactive Tutorial Support Mental Model Formation and Human-Automation Interaction. *Information, 10*(4), 143. https://doi.org/10.3390/info10040143

Forster, Y., Hergeth, S., Naujoks, F., Krems, J.F., & Keinath, A. (2020). What and how to tell beforehand: The effect of user education on understanding, interaction and satisfaction with driving automation. *Transportation Research Part F, 68*, 316–335.
https://doi.org/10.1016/j.trf.2019.11.017

Heller, J., Steiner, C., Hockemeyer, C., & Albert, D. (2006). Competence-Based Knowledge Structures for Personalised Learning. In (pp. 75–88). Association for the Advancement of Computing in Education (AACE).
https://www.learntechlib.org/p/21759/

Hergeth, S., Lorenz, L., & Krems, J.F. (2017). Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. *Human Factors, 59*, 457–470. https://doi.org/10.1177/0018720816678714

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, 52, 464–488.
https://doi.org/10.3758/s13428-019-01246-w

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review, 20*, 709–734.

Morales-Alvarez, W., Sipele, O., Léberon, R., Tadjine, H. H., & Olaverri-Monreal, C. (2020). Automated Driving: A Literature Review of the Take over Request in Conditional Automation. *Electronics, 9*(12), 2087.
https://doi.org/10.3390/electronics9122087

Neuhuber, N., Ebinger, N., Pretto, P., & Kubicek, B. (2022). How am I supposed to know? Conceptualization and first Evaluation of a Driver Tutoring System for Automated Driving. In D. de Waard, S.H. Fairclough, K.A. Brookhuis, D. Manzey, L. Onnasch, A. Naumann, R. Wiczorek, F. Di Nocera, S. Röttger, and A. Toffetti (Eds.). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2022 Annual Conference* (pp. 17-28). ISSN 2333-4959 (online). Available from http://hfes-europe.org

Noble, A. M., Klauer, S. G., Doerzaph, Z. R., & Manser, M. P. (2019). Driver Training for Automated Vehicle Technology – Knowledge, Behaviors, and Perceived Familiarity. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 2110–2114. https://doi.org/10.1177/1071181319631249

Payre, W., Cestac, J., Dang, N.-T., Vienne, F., & Delhomme, P. (2017). Impact of training and in-vehicle task performance on manual control recovery in an automated car. *Transportation Research Part F*, *46*, 216–227.
https://doi.org/10.1016/j.trf.2017.02.001

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully Automated Driving: Impact of Trust and Practice on Manual Control Recovery. *Human Factors, 58*, 229–241.
https://doi.org/10.1177/0018720815612319

Rukonić, L., Mwange, M.-A., & Kieffer, S. (2022). Teaching Drivers about ADAS using Spoken Dialogue: A Wizard of Oz Study. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (pp. 88–98). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010913900003124

SAE International (2021). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles: J3016.

Sportillo, D., Paljic, A., & Ojeda, L. (2018). Get ready for automated driving using Virtual Reality. *Accident, Analysis and Prevention, 118*, 102–113. https://doi.org/10.1016/j.aap.2018.06.003

Strand, N., Nilsson, J., Karlsson, I., & Nilsson, L. (2011). Exploring end-user experiences: self-perceived notions on use of adaptive cruise control systems. *IET Intelligent Transport Systems, 5*(2), 134. https://doi.org/10.1049/iet-its.2010.0116

Wild, E., & Möller, J. (Eds.). (2020). *Pädagogische Psychologie*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61403-7

Comparing accepted gaps in traffic flow for conducting left turn actions in different intersection scenarios – A driving simulator study

Ann-Christin Hensch, Matthias Beggiato, & Josef F. Krems
Chemnitz University of Technology, Chemnitz, Germany

Abstract

Automated vehicles (AVs) are required to interact predictably and intuitively with conventional road users to support traffic safety and efficiency. This might be achieved by implementing established interaction capabilities of manual drivers in AVs. When coordinating interactions, drivers anticipate the development of driving scenes and adapt own manoeuvres by selecting specific gaps in traffic flow (gap acceptance, GA) to provide adequate safety margins to surrounding road users. Therefore, the current driving simulator study aimed at deriving manual drivers' GA for conducting left turn manoeuvres at intersections as a basis for AVs' driving functions. Thus, two intersection scenarios including interaction vehicles (approach trajectories from the left vs. head-on trajectories from the ego-vehicle's perspective) encountering with two different speed levels (30 km/h vs. 50 km/h) were investigated by a within-subject design including $N = 28$ participants. The results revealed that smaller time gaps were selected at higher approach speeds and when the interaction vehicles encountered the intersection from the left of the ego-vehicle's perspective. The findings imply that there is no one unique time gap value for conducting turning manoeuvres. Rather, drivers seem to consider the speed of approaching vehicles and thus anticipate resulting safety margins to surrounding traffic participants as well as specific required temporal durations for conducting turning manoeuvres in different intersection scenarios. The revealed gaps could provide a basis for intuitive driving functions of AVs in such scenarios.

Introduction

Road traffic represents a social system requiring interactions between different traffic participants (Domeyer et al., 2020). Since there has been a steady increase of automated driving functions in recent years, facilitating for enhanced road safety, traffic efficiency, driving comfort and accessibility (ERTRAC, 2017), vehicles of different levels of automation will share the road in the near future (Litman, 2017). For this, transparent and efficient coordination of actions are required in mixed traffic, comprising conventional traffic participants and automated vehicles (AVs, i.e., SAE Level 3 or higher; SAE, 2018).

To prevent from breakdowns of interaction processes ensuring safe and efficient encounters, drivers are required to obtain adequate Situation Awareness (SA) of the driving environment (Endsley, 1995; Krems & Baumann, 2009). According to

Endsley (1995), the concept is described as "knowing what is going on" (Endsley, 1995; p. 36) and comprises three levels: the perception of relevant elements of a situation, the integration and comprehension of their meaning as well as the prediction and anticipation of futures states of the respective situation (Endsley, 1995). The capabilities of predicting and anticipating future states of driving scenes and prospective movements of surrounding road users enable drivers for a proactive behaviour by adapting own manoeuvres and thus ensure smooth encounters (Endsley, 1995; Stahl et al., 2014). Due to potentially conflicting trajectories of the involved road users, coordinating and anticipating driving actions is highly important in intersection scenarios (Hancock et al., 1991; Wilbrink et al., 2018). Since these manoeuvres cause a higher mental load of drivers compared to straight driving, potentially leading to an increased accident rate (Hancock et al., 1989), automated turning actions might foster traffic safety and efficiency.

To support the involved traffic participants in achieving adequate SA, AVs' driving behaviour needs to be transparent and predictable (Habibovic et al., 2018). Therefore, established and intuitive interaction capabilities should be implemented in AVs' driving functions (Schieben et al., 2019) to provide a common ground when coordinating actions, comprising mutual knowledge and assumptions about interactions (Clark & Brennan, 1991). Specific parameters of manual driving might be derived and applied as a basic orientation for AVs' driving functions. However, to provide intuitive and expectable interactions in mixed traffic, AVs are required to consider diverse influencing factors when conducting driving manoeuvres comparable to human drivers (Beggiato et al., 2018; Stange et al., 2022). Thus, potentially influencing factors and their particular impact on driving parameters need to be further investigated.

When coordinating actions, drivers mainly apply temporal and spatial thresholds to surrounding traffic participants (Gibson & Crooks, 1938; Hall, 1969; Summala, 2007). These thresholds describe drivers' individual gap acceptance behaviour (GA; Summala, 2007). Previous research showed that drivers' GA was influenced by various factors, such as the approach speed of encountering traffic (e.g., Hancock et al., 1991). A stable effect was reported for approach speed on GA, particularly resulting in smaller accepted time gaps when initiating driving manoeuvres during higher speeds of encountering traffic participants (e.g., Hancock et al., 1991; Petzoldt et al., 2017). The effect was shown for different approach speeds (e.g., Schleinitz et al., 2019), diverse encountering interaction partners (Hensch et al., 2021) and from various perspectives (e.g., for a pedestrians' perspective see Beggiato et al., 2018; for a drivers' perspective see Hensch et al., 2023). Besides situational factors, drivers might also consider infrastructural conditions and anticipate required manoeuvres and resulting temporal durations to provide sufficient thresholds of space and time to surrounding road users. Thus, different intersection scenarios and anticipated driving manoeuvres might have an impact on GA, since these scenarios are considered as highly relevant for coordinating actions (Wilbrink et al., 2018). For instance, Avsar et al. (2021) investigated accepted gaps in traffic flow in a driving simulator study. The authors reported more accepted gaps for left turn manoeuvres at intersections when the interaction vehicles approached the ego-vehicle's perspective from the left compared to head-on approach trajectories of the interaction vehicles. In contrast,

Petzoldt et al. (2017) found no effect of intersection scenarios on drivers' GA for left turn manoeuvres in their test track study. Thus, it seems reasonable to investigate the influence of intersection scenarios and specifically anticipated manoeuvres on drivers' GA for left turn actions in closer detail.

Objectives of the study

The current study aimed at investigating manual drivers' accepted time gaps in traffic flow for conducting left turn manoeuvres at intersections. The quantified GA parameters could provide information about influencing factors on manoeuvring decisions for intuitive interactions. Therefore, the derived time gap values of manual drivers might be used as a basic orientation for AVs' driving functions in these scenarios, particularly as accepted time ranges for initiating automated turning manoeuvres, to support transparent and predictable interactions in mixed traffic. The influence of the approach speed of encountering interaction vehicles was manipulated (i.e., 30 km/h vs. 50 km/h) within the current study. Two different intersections, requiring left turn manoeuvres, were also considered in the study, since the coordination of actions between different traffic participants is particularly important in these scenarios.

Method

Research Design

The present driving simulator study investigated participants' GA in traffic flow for conducting left turn manoeuvres in two different intersection scenarios. A 2x2x2 within-subject design was applied. In particular, two approach speeds of the interaction vehicles (30 km/h vs. 50 km/h) and two intersection scenarios (Figure 1) were investigated. Each condition was repeated twice to stabilize the results. Participants' accepted gaps served as dependent variable.

Apparatus and Material

The study was conducted in a fixed-based driving simulator comprising a mock-up including adjustable seats, steering wheel, speedometer, brake, accelerator pedal for vehicle control, rear and side mirrors. The simulation of the urban environment including the investigated intersection scenarios was programmed in SILAB 7.0 (WIVW, 2021). The driving scenery was projected with a resolution of 7680x1600 pixels by three projectors at 180° field view. Participants' driving behaviour was logged with a sampling frequency of 60 Hz.

The simulation included eight urban T-intersections comprising two different scenarios (Figure 1), which were presented in a balanced order. Both implemented scenarios required the participants in the ego-vehicle to conduct left turn manoeuvres and thus crossing the path of the approaching interaction vehicle stream (i.e., exclusively passenger cars due to controlling for influencing factors, such as vehicles size on GA; Hancock et al., 1991). In scenario 1, the ego-vehicle was required to turn from a side road into a main road and the interaction vehicles approached from the left of the ego-vehicle's perspective (Figure 1a). Whereas, in scenario 2, the ego-vehicle was located on a main road and the interaction vehicles encountered the intersection with head-on

approach trajectories from the ego-vehicle's perspective (Figure 1b). Thus, different routes of the ego-vehicle were required despite conducting left turn manoeuvres in both scenarios. When approaching the intersections, the ego-vehicle always needed to give priority to the interaction vehicles demanding a full stop and a subsequent start from standstill. The interaction vehicles encountered the intersections with an approach speed of either 30 km/h or 50 km/h respectively, which was varied per intersection. The gaps within the approaching interaction vehicle stream increased ascendingly (i.e., one additional second per gap). When the oncoming vehicles approached with 30 km/h, gaps in traffic flow up to 10 seconds were presented and gaps up to 12 seconds were presented at 50 km/h approach speed. The maximum gaps were deduced from previous online studies of the authors that examined comparable scenarios (Hensch et al., 2021; Hensch et al., 2022). Road signs were placed at the intersections to remind participants of the instructed left turn manoeuvres. To provide a conventional driving experience and reduce the risk of simulator sickness, straight route parts were included in the simulation between the investigated intersection scenarios.
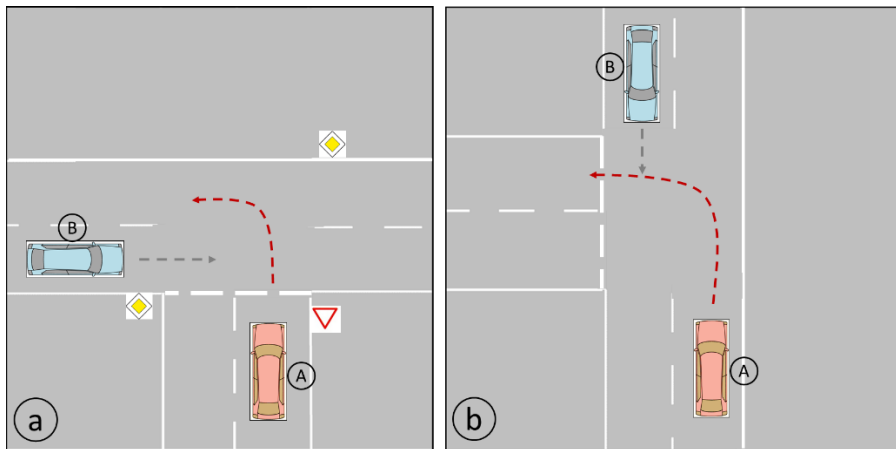
Figure 1. Intersection scenarios of the simulator drive requiring the ego-vehicle (A) for accepting gaps within traffic flow (B) and thus conducting left turn manoeuvres. a) Scenario 1: The ego-vehicle was required to turn from a side road into a main road while the interaction vehicles approached the intersection from the left of the ego-vehicle's perspective. b) Scenario 2: The ego-vehicle was located on a main road while the interaction vehicles encountered the intersection with head-on approach trajectories from the ego-vehicle's perspective.

Procedure

Since drivers' GA was investigated within the current study, holding a valid driver's licence was a precondition for participating to the study. At first, participants were welcomed, general information was provided and informed consent was obtained. Then, sociodemographic data were collected. In the subsequent simulator drive, participants' accepted gaps in traffic flow for conducting left turn manoeuvres at intersections were collected. The participants were instructed to conduct left turn manoeuvres at the smallest gaps they still perceived as comfortable. To familiarize the participants with the simulation environment and the vehicle guidance particularly

when conducting turning manoeuvres at intersection, a test drive was conducted. Afterwards, drivers' GA was collected in a drive that included eight urban intersection scenarios, presented in a balanced order. At the end, participants received a monetary compensation of €25 for contributing to the study.

Participants

Overall, $N = 41$ participants contributed to the study. However, $n = 12$ participants had to be excluded from data analysis due to simulator sickness. Moreover, $n = 1$ participant did not accept any of the presented gaps, which did not allow a calculation of mean accepted time gaps for conducting turning manoeuvres in traffic flow. All participants of the remaining sample of $n = 28$ participants ($M = 34$ years, $SD = 14.54$) held a valid driving licence ($M = 15.88$ years, $SD = 13.33$) and reported a mean annual mileage of $M = 12588.75$ km ($SD = 13654.77$).

Data preparation and analysis

Participants' mean accepted gaps in traffic flow when conducting left turn manoeuvres in the examined intersection scenarios were calculated according to Lobjois et al. (2013) and Petzoldt (2014), who applied comparable approaches when investigating pedestrians' gap selection during road crossing actions. Logistic regressions with binary outcomes (i.e., yes/no response for accepting a specific gap in traffic flow) were calculated to consider for varying decision patterns to accept or reject gaps, since participants not always showed a consistent decision behaviour (Petzoldt, 2014). Four individual regression models were calculated per participant (one per approach speed condition of 30 km/h vs. 50 km/h; one per intersection scenario namely scenario 1 vs. scenario 2, respectively). The individual regression models predicted the probability that a gap in traffic flow was accepted for conducting a turning manoeuvre (i.e., yes response for accepting a specific gap) for each participant. The respective transition points of the individual regression models (i.e., 50% probability of accepting a specific gap) were defined as the mean GA for each participant and per condition and were used for further analysis (Lobjois et al., 2013; Petzoldt, 2014). As an indicator of model fit, mean Nagelkerke's R^2 are reported for the calculated regression models (Nagelkerke, 1991).

Results

The calculated individual regression models for each participant per speed level of the approaching traffic (mean Nagelkerke's $R^2_{30\ km/h} = .92$; mean Nagelkerke's $R^2_{50\ km/h} = .96$) and per scenario (mean Nagelkerke's $R^2_{scenario\ 1} = .93$; mean Nagelkerke's $R^2_{scenario\ 2} = .95$) are plotted in Figure 2. In particular, participants accepted smaller gaps in traffic flow when conducting turning manoeuvres at higher speeds of approaching vehicles. This is also depicted in Figure 2 were the individual regression models are distributed closer towards the y-axis when the interaction vehicles approached the intersection with 50 km/h compared to 30 km/h. This difference can be also found when comparing the mean accepted time gaps ($M_{30km/h} = 6.13$s, $SD_{30km/h} = 1.56$; $M_{50km/h} = 5.60$s, $SD_{50km/h} = 1.14$; Figure 3). Moreover, smaller time gaps were accepted for conducting turning manoeuvres in scenario 1 compared to scenario 2, which is also displayed in Figure 2. In detail, the individual regression models of scenario 1 are somewhat distributed closer towards

the y-axis compared to the regression models of scenario 2. This difference is also shown when comparing the mean accepted time gap values for the scenarios ($M_{\text{scenario 1}} = 5.60$s, $SD_{\text{scenario 1}} = 1.40$; $M_{\text{scenario 2}} = 6.13$s, $SD_{\text{scenario 2}} = 1.43$; Figure 3). According to the 2 (approach speed) x 2 (scenario) repeated measures ANOVA, the differences in mean accepted time gaps were revealed to be statistically significant for both factors, approach speed ($F(1,27) = 17.80$, $p < .001$, $\eta^2_p = .397$) and scenario ($F(1,27) = 7.67$, $p = .009$, $\eta^2_p = .228$). Additionally, a significant interaction effect for approach speed x scenario was found ($F(1,27) = 11.80$, $p = .002$, $\eta^2_p = .304$).
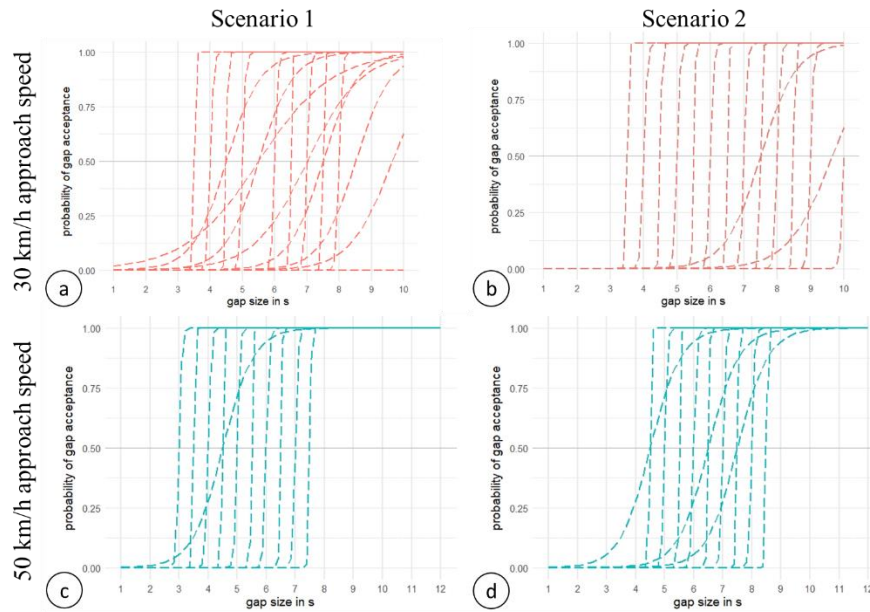
Figure 2. Individual regression models for participants' GA for a) scenario 1 and b) scenario 2 with an approach speed of 30 km/h of the interaction vehicles; and c) scenario 1 and d) scenario 2 with an approach speed of 50 km/h of the interaction vehicles. The individual regression models might overlap due to similar decision patterns of different participants (i.e., accepted and rejected gaps in traffic flow). The solid horizontal lines in the respective graphs highlight the transition points of the individual regression models (i.e., 50% probability of accepting a gap).

Discussion

The present driving simulator study aimed at investigating and quantifying manual drivers' GA when conducting left turn manoeuvres with approaching traffic in different intersection scenarios. The derived gaps could be used as a basic orientation for transparent and intuitive driving functions in AVs, specifically as accepted time ranges for initiating automated turning manoeuvres at intersections (for comparable approaches see Beggiato et al., 2018; Stange et al., 2022). The results showed that there is no one unique time gap value that is always considered as appropriate when conducting left turn manoeuvres at intersections. Rather, drivers' decisions for selecting a gap in traffic flow was influenced by the approach speed of the
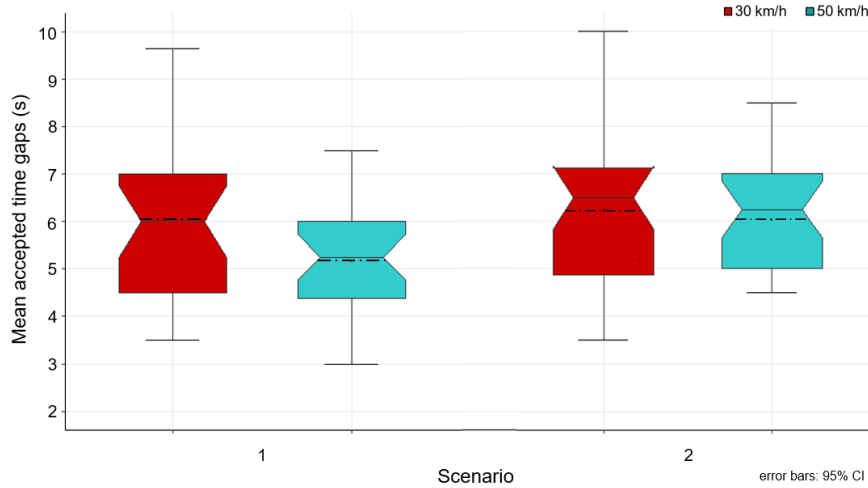
Figure 3. Boxplots of mean GA (in seconds) by investigated scenarios (1 = ego-vehicle required to turn from a side road into a main road while the interaction vehicles approached from the left of the ego-vehicle's perspective; 2 = ego-vehicle located on a main road while the interaction vehicles encountered with head-on approach trajectories from the ego-vehicle's perspective; see Figure 1) and approach speed of encountering vehicles. The boxes represent the mean 50% of GA values, the lower whiskers represent the minimum GA values, the upper whiskers the maximum GA values that were selected respectively. The solid line within the boxes displays the median value, the dashed line the mean value for the respective condition.

encountering vehicles. Moreover, GA differed between the investigated intersection scenarios.

The approach speed of the encountering interaction vehicles was shown to influence drivers' GA in traffic flow when conducting left turn manoeuvres at intersections. In particular, smaller time gaps were accepted at higher speeds of the approaching vehicles. The results corroborate previous studies that also reported smaller GA at higher speeds of encountering interaction partners (Hancock et al., 1991; Petzoldt et al., 2017). The findings might be explained by the approach that drivers rather rely on spatial distances when selecting gaps in traffic flow than on time-to-arrival estimates when initiating driving manoeuvres (Law et al., 1993). To provide an intuitive interaction behaviour in mixed traffic, the results imply that AVs need to consider the speed of the approaching traffic when selecting gaps in traffic flow for left turn manoeuvres (Hensch et al., 2023).

In addition, the intersection scenario itself also affected drivers' GA. In detail, smaller GA for conducting left turn manoeuvres was revealed when the interaction vehicles approached the ego-vehicle from the left (i.e., scenario 1) compared with head-on trajectories of the interaction vehicles (i.e., scenario 2). The findings are in line with Avsar et al. (2021). The authors reported more accepted gaps in intersection scenarios were the interaction vehicles encountered from the left of the ego-vehicle's perspective (comparable with scenario 1 of the current study) in contrast to intersection scenarios were the interaction traffic approached from the opposite

direction of the ego-vehicle's perspective (comparable with scenario 2 of the current study; Avsar et al., 2021). The results of the current study might be explained by larger distances for passing the encountering interaction vehicles in scenario 2 compared to scenario 1 (i.e., passing the potential point of collision to prevent accidents). Therefore, the findings indicate that drivers anticipate specific route distances and thus temporal durations that are required to execute prospective manoeuvres. Specifically, drivers seem to anticipate varieties in route distances and thus resulting differences considering temporal durations for passing the potential point of collision during turning manoeuvres at intersections with encountering interaction traffic. This could have influenced GA parameters to provide sufficient time and space to surrounding traffic participants (Gibson & Crooks, 1938; Summala, 2007). Thus, AVs should also be enabled to consider specific route distances and thus required manoeuvre durations for conducting driving actions when selecting gaps in traffic flow and therefore support smooth interactions with other road users. However, the results might be also explained by the observers' perspective on the approaching traffic (Petzoldt et al., 2017). This might have led to more accurate time-to-arrival estimates due to the viewing angle on the approaching interaction vehicles (i.e., from the left of the observers' perspective) resulting in smaller GA for scenario 1 and more conservative decisions in GA due to the frontal view in scenario 2 (Schiff & Oldak, 1990).

It should be mentioned that the current driving simulator study merely investigated left turn scenarios including interaction vehicles from one specific direction to achieve standardized conditions. To receive a comprehensive understanding of human drivers' interaction behaviour in road traffic, future studies should investigate further interaction partners and driving parameters besides GA (e.g., acceleration behaviour; Bellem et al., 2018). Additional factors, such as different age groups and their impact on the specific driving parameter of GA, should be also quantified in future driving simulator studies (as investigated for deceleration behaviour e.g., by de Haan et al., 2022). Moreover, perspectives of other road users, such as pedestrians during crossing actions, need to be considered to receive an extensive understanding of human users' interaction behaviour in road traffic (e.g., Beggiato et al., 2018).

Conclusion

For smooth interactions, drivers anticipate required temporal durations for their prospective manoeuvres and the development of driving scenes (Stahl et al., 2014). Coordinating actions is particularly relevant at intersections due to potentially conflicting trajectories of involved road users (Hancock et al., 1991; Wilbrink, 2018). Therefore, drivers adapt their own manoeuvres accordingly to provide sufficient safety margins to surrounding road users (Gibson & Crooks, 1938; Summala, 2007). To support intuitive encounters in mixed traffic, established interaction capabilities from manual drivers could be used as a basic orientation for AVs' driving functions (Beggiato et al., 2018; Stange et al., 2022). With regard to GA as one specific driving parameter, the results of the current driving simulator study imply that there is no one unique time gap value for conducting left turn manoeuvres at intersections in traffic flow. In fact, the speed of approaching interaction vehicles and the particular traffic scenario influenced drivers' accepted gaps. Drivers seem to consider infrastructural

conditions, contextual factors, anticipate particular manoeuvres and thus resulting routes as well as temporal durations when conducting driving actions. To support traffic safety and efficiency as well as the acceptance of AVs, this differentiated knowledge of experienced human drivers (Krems & Baumann, 2009) might be also implemented in AVs' driving functions to support intuitive interactions in mixed traffic.

Acknowledgements

References

Avsar, H., Utesch, F., Wilbrink, M., Oehl, M., & Schießl, C. (2021). Efficient communication of automated vehicles and manually driven vehicles through an external human-machine interface (eHMI): Evaluation at T-junctions. In C. Stephanidis, M. Antona, & S. Ntoa, (Eds.), *HCI International 2021 - Posters. HCII 2021. Communications in Computer and Information Science, vol 1421* (pp. 224-232). Springer: Cham. https://doi.org/10.1007/978-3-030-78645-8_28.

Beggiato, M., Witzlack, C., Springer, S., & Krems, J.F. (2018). The right moment for braking as informal communication signal between automated vehicles and pedestrians in crossing situations. In N. Stanton (Ed.), *Advances in Human Aspects of Transportation. AHFE 2017. Advances in Intelligent Systems and Computing* (pp. 1072-1081). Springer: Cham. https://doi.org/10.1007/978-3-319-93885-1.

Bellem, H., Thiel, B., Schrauf, M., & Krems, J.F. (2018). Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transportation Research Part F, 55*, 90-100.
https://doi.org/10.1016/j.trf.2018.02.036.

Clark, H.E., Perrone, J.A., & Isler, R.B. (2013). An illusory size-speed bias and railway crossing collisions. *Accident Analysis and Prevention, 55*, 226-231. https://doi. org/10.1016/j.aap.2013.02.037.

De Haan, T., Stuiver, A., Lorist, M.M., & de Waard, D. (2022). Other road users' adaptations to increase safety in response to older drivers' behaviour. *Transportation Research Part F, 84,* 277-286.
https://doi.org/10.1016/j.trf.2021.12.009.

Domeyer, J.E., Lee, J.D., & Toyoda, H. (2020). Vehicle automation-other road user communication and coordination: Theory and mechanisms. *IEEE Access, 8*, 19860-19872. https://doi.org/10.1109/ACCESS.2020.2969233.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.
https://doi.org/10.1518/001872095779049543.

European Road Transport Research Advisory Council (ERTRAC). (2017). *Automated driving roadmap.* Report of the ERTRAC working group "Connectivity and Automated Driving". https://www.ertrac.org/wp-content/uploads/2022/07/ERTRAC_Automated_Driving_2017.pdf

Gibson, J.J., & Crooks, L. E. (1938). A Theoretical Field-Analysis of Automobile-Driving. *The American Journal of Psychology, 51,* pp. 453-471. https://doi.org/10.2307/1416145.

Habibovic, A., Lundgren, V. M., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., Fagerlönn, J., Edgren, C., Fredriksson, R., Krupenia, S., Saluäär, D., & Larsson, P. (2018). Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology, 9*, 1336. https://doi.org/10.3389/fpsyg.2018.01336.

Hall, E.T. (1969). *The Hidden Dimension*. Doubleday: Anchor Books.

Hancock, P.A., Caird, J.K., & Shekhar, S. (1991). Factors influencing drivers' left turn decisions. *Proceedings of the Human Factors 35th Annual Meeting, 35*(15), 1139-1143. https://doi.org/https://doi.org/10.1177/154193129103501525.

Hancock, P.A., Wulf, G., Thom, D.R. & Fassnacht, P. (1989). Contrasting driver behavior during turns and straight driving. *Proceedings of the Human Factors 33rd Annual Meeting, 33(15),* 918-922. https://doi.org/10.1177/154193128903301504.

Hensch, A.-C., Beggiato, M., & Krems, J.F. (2022). Should I wait or should I go? - Deciphering implicit communication cues for cooperative interactions in left-turn scenarios. *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (IEEE CIVEMSA 2022)*, 1-5. doi: 10.1109/CIVEMSA53371.2022.9853676.

Hensch, A.-C., Beggiato, M., & Krems, J.F. (2023). Drivers' gap acceptance during parking maneuvers as a basis for initiating driving actions in automated vehicles. *Transportation Research Part F, 92*, 133-142. https://doi.org/10.1016/j.trf.2022.11.008.

Hensch, A.-C., Beggiato, M., Schömann, M.X., & Krems, J. F. (2021). Different types, different speeds – The effect of interaction partners and encountering speeds at intersections on drivers ' gap acceptance as an implicit communication signal in automated driving. In H. Krömker (Ed.), *HCI in Mobility, Transport, and Automotive Systems. HCII 2021. Lecture Notes in Computer Science* (pp. 517-528). Springer: Cham. https://doi.org/10.1007/978-3-030-78358-7_36.

Krems, J.F., & Baumann, M.R. (2009). Driving and situation awareness: A cognitive model of memory-update processes. In M. Kurosu (Ed.), *Lecture Notes in Computer Science, vol. 5619 LNCS* (pp. 986-994). Springer: Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02806-9_113.

Litman, T. (2017). *Autonomous Vehicle Implementation Predictions*. Victoria, British Columbia, Canada: Victoria Transport Policy Institute. https://www.vtpi.org/avip.pdf.

Lobjois, R., Benguigui, N., & Cavallo, V. (2013). The effects of age and traffic density on street-crossing behavior. *Accident Analysis & Prevention, 53*, 166-175, 2013. http://dx.doi.org/10.1016/j.aap.2012.12.028.

Law, D.J., Pellegrino, J.W., Mitchell, S.R., Fischer, S.C., McDonald, T.P., & Hunt, E.B. (1993). Perceptual and cognitive factors governing performance in comparative arrival-time judgments. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 1183-1199. https://doi.org/10.1037/0096-1523.19.6.1183.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692. https://doi.org/10.1093/biomet/78.3.691.

Petzoldt, T. (2014). On the relationship between pedestrian gap acceptance and time to arrival estimates. *Accident Analysis & Prevention, 72*, pp. 127-133. https://doi.org/10.1016/j.aap.2014.06.019.

Petzoldt, T., Schleinitz, K., Krems, J.F., & Gehlert, T. (2017). Drivers' gap acceptance in front of approaching bicycles – Effects of bicycle speed and bicycle type. *Safety Science, 92*, 283-289. https://doi.org/10.1016/j.ssci.2015.07.021.

SAE International's new standard J3016. (2018). *Levels of driving automation.* Warrendale, Pennsylvania, US: SAE International. https://doi.org/P141661.

Schieben, A., Wilbrink, M., Kettwich, C., Madigan, R., Louw, T., & Merat, N. (2019). Designing the interaction of automated vehicles with other traffic participants: Design considerations based on human needs and expectations. *Cognition, Technology and Work, 21*, 69-85. https://doi.org/10.1007/s10111-018-0521-z.

Schiff, W. & Oldak, R. (1990). Accuracy of judging time to arrival: effects of modality, trajectory, and gender. *Journal of Experimental Psychology. Human Perception and Performance, 16,* 303-316. doi: 10.1037//0096-1523.16.2.303.

Schleinitz, K., Petzoldt, T., & Gehlert, T. (2019). Drivers' gap acceptance and time to arrival judgements when confronted with approaching bicycles, e-bikes, and scooters. *Journal of Transportation Safety and Security, 12*, 3-16. https://doi.org/10.1080/19439962.2019.1591551.

Stahl, P., Donmez, B., & Jamieson, G. A. (2014). Anticipation in driving: The role of experience in the efficacy of pre-event conflict cues. *IEEE Transactions on Human- Machine Systems, 44*, 603-613. https://doi.org/10.1109/THMS.2014.2325558.

Stange, V., Goralzik, A., Ernst, S., Steimle, M., Maurer, M., & Vollrath, M. (2022). Please stop now, automated vehicle! – Passengers aim to avoid risk experiences in interactions with a crossing vulnerable road user at an urban junction. *Transportation Research Part F, 87*, 164-188. https://doi. org/10.1016/j.trf.2022.04.001.

Summala, H. (2007). Towards understanding motivational and emotional factors in driver behaviour: Comfort through satisficing. In P. C. Cacciabue (Ed.), *Modelling driver behaviour in automotive environments: Critical issues in driver interactions with intelligent transport systems* (pp. 189-207). Springer: London. https://doi.org/10.1007/978-1-84628-618-6.

Wilbrink, M., Schieben, A., Markowski, R., Weber, F., Gehb, T., Ruenz, J., Tango, F., Kaup, M., Willrodt, J.-H., Portouli, V., Merat, N., Madigan, R., Markkula, G., Romano, R., Fox, C., Althoff, M., Söntges, S., & Dietrich, A. (2018). *Designing cooperative interaction of automated vehicles with other road users in mixed traffic environments. InterACT D1.1 Definition of interACT use cases and scenarios.*

https://www.interact-roadautomation.eu/wp-content/uploads/
interACT_WP1_D1.1_UseCases_Scenarios_1.1_approved_UploadWebsite.pdf
Würzburger Institut für Verkehrswissenschaften GmbH, WIVW (2021). *SILAB 7.0.*
https://www.wivw.de.

Implicit communication on the motorway slip road: a driving simulator study

Sofie Ehrhardt, Raphael Roß, & Barbara Deml
Karlsruhe Institute of Technology, Germany

Abstract

The mutual use of road infrastructure by (partially) automated and manually driven vehicles requires an understanding of human communication in traffic. This paper investigates the situation of the motorway slip road. In a driving simulator study (N=32), participants evaluate the behaviour of an interaction partner regarding their cooperativity and criticality of the situation. The interaction partner was either marked as automated by an external Human-Machine-Interface (eHMI) or an unmarked vehicle. Three implicit communication means were tested: Accelerating, decelerating, and maintaining speed. Participants experienced this both from the perspective of the slip road and the right lane while the interaction partner merged (2x3x2 design). Participants on the slip road evaluated the behaviour of partners with eHMI as significantly less critical. If they are already on the motorway themselves, an eHMI on an oncoming vehicle had no influence on the evaluation. However, participants maintain a significantly greater distance from merging vehicles with eHMI. When entering the motorway, deceleration or speed maintaining of the cooperation partner in the right lane is perceived as less critical and more cooperative than acceleration. If the cooperating vehicle is on the slip road, its deceleration is assessed as less critical and more cooperative than acceleration.

Introduction

The more vehicles feature functions that enable semi-automated driving, the more frequent interactions between human-controlled and (temporarily) automated vehicles will become. Smooth, safe road traffic requires constant communication between all road users. There is now a moderately good understanding of what means of communication drivers use among themselves (for a list, see Schaarschmidt et al., 2021). However, how each is used in specific traffic situations is not yet fully understood. Understanding the importance of the communication tools and how they are evaluated is, however, important to ensure smooth traffic. Furthermore, it needs to be explored whether these communication tools are assessed in the same way in visibly automated vehicles and thus whether there is a transferability from findings regarding human-to-human communication on automated driving. In this paper, the assessment of implicit communication through driving dynamics (e.g., braking and accelerating) is considered in the situation of slip roads. Two research questions ar stuedied: 1. How are different means of communication via driving dynamics evaluated at motorway slip roads? 2. Is this evaluation dependent on whether a vehicle is marked as automated?

Theoretical background

Especially at high speeds, where the time windows for mutual understanding are shorter, effective communication is crucial for safe and efficient driving. As a result, communication at motorways often occurs implicitly, e.g., via driving dynamics, to communicate one's intentions (Moore et al., 2019; Risto et al., 2017; Schaarschmidt et al., 2021). Driving dynamics include lateral movements as well as deceleration and acceleration. Human drivers rely on their experience to understand the message in the behaviour of other drivers. However, automated vehicles lack this experience, and it is necessary to implement implicit communication effectively for safe integration into highway traffic (Kauffmann et al., 2018).

Research on the role of driving dynamics for communication purposes has been conducted for different traffic scenarios: In urban traffic, people tend to view slowing down as defensive behaviour, while accelerating or maintaining speed is seen as aggressive (Imbsweiler, 2019). In their naturalistic driving study, accelerating was found to be an unambiguous way to communicate the intention to make a manoeuvre ahead of another vehicle. However, participants rated this as rather uncooperative. Björklund and Åberg (2005) found that when approaching an intersection, maintaining speed and accelerating are associated with driving intention, while slowing down is perceived as yielding. Similar results have been found from a driving simulator study observing a bottleneck scenario: maintaining was understood as intention to drive while reducing speed was understood as yielding (Rettenmaier et al., 2021). To identify implicit forms of cooperative behaviour on motorways, Kauffmann et al. (2018) investigated the evaluation of different behaviours conducted by the vehicle behind participants that changed lanes in dense traffic. They found that the harder the rear vehicle braked, the more cooperative and less critical the behaviour was perceived by the participants. The study, however, refers to lane changes on motorways which fall into the category of unforced lane changes. However, at motorway slip roads, forced lane changes take place as the acceleration lane ends. Further research is needed to investigate implicit cooperative behaviour during these forced lane changes.

In addition to the evaluation of longitudinal driving behaviour as a means of communication, this work will explore the impact of an external human-machine interface (eHMI). An eHMI which labels a car as automated (status eHMI) can increase acceptance of unexpected behaviour, similar to the markings of driving school cars (Fuest et al., 2020). Studies on the influence of eHMI in different traffic situations show no significant effects on the behaviour and evaluation of participants (Fuest et al., 2020; GATEway Project, 2017; Joisten et al., 2019). In a driving simulator study, no differences in gap size acceptance were found when participants encountered automated vs. manual vehicles (GATEway Project, 2017). And pedestrians in a controlled field experiment did not differ in their crossing behaviour when interacting with a vehicle equipped with a status eHMI. Also, their self reported perceived safety did not differ significantly (Joisten et al., 2019). However, only the study by Fuest et al. (2020) refers to the interaction of two vehicles on the highway; the other two studies took place in urban traffic. In three different scenarios (road works, traffic jam, lane change) participants followed a car on the highway, which

was either labelled as automated by an eHMI or not. The vehicle obeyed the traffic rules but doing so was a deviation from the rest of the traffic. The subjective opinion of the participants on the behaviour of the vehicle and the distance behaviour (Time Headway, THW) was surveyed. No significant influence of the eHMI was found. Yet, in Fuest et al. (2020), no cooperation between the participants and the vehicles with eHMI took place. This paper aims to fill this research gap.

Material and methods

Four hypotheses were investigated in the present study:

H1a. Drivers on the slip road perceive a deceleration of the vehicle in the right lane as more cooperative and less critical than acceleration and maintaining speed.

H1b. Drivers in the right lane perceive deceleration of the vehicle on the slip road as more cooperative and less critical than acceleration and maintaining speed.

H2a. There is no significant difference in the participants perceptions of cooperation and criticality between manual and automated vehicles with identical behaviour.

H2b. There is no significant difference in the distance behaviour toward manual and automated vehicles with identical behaviour.

Independent and dependent variables

Three independent variables were investigated in the study: First, the level of automation of the cooperation partner which was manipulated using an eHMI. Participants were instructed that vehicles without eHMI were manually controlled by humans and vehicles with eHMI drove automatically (for information on the eHMI see subchapter *"Setting"*). Secondly, the behaviour of the cooperation partner was varied: After being at the same level as the participants at the beginning of the acceleration lane, they either accelerated, maintained their speed, or braked (for detailed information on cooperation partners' behaviour see subchapter *"Setting"*). Thirdly, the situation was examined from both perspectives; the participants drove onto the motorway themselves and passed another slip road on which a cooperation partner was driving. This resulted in a total of twelve situations (2x3x2), two of which were driven through in each lap of the test circuit. The six resulting test runs were randomised. Each test participant drives through all situations according to a within-subject design.

The dependent variables are perceived cooperativeness and criticality on the one hand and the participants' distance behaviour on the other. The perceived cooperativeness of the other road user was assessed on a seven-point Likert scale according to Imbsweiler (2019). It ranged from "very uncooperative" (1) to "very cooperative" (7). Perceived criticality was represented on an eleven-point scale from (0) to (10), which is derived from Neukum and Krüger (2003). The rating is divided into five categories. A rating of (0) corresponds to the category of nothing noticed, (1-3) falls into the category of noticeability, (4-6) describes an interference with driving, (7-9) classifies the situation as dangerous, and a rating of (10) classifies the vehicle as no longer

controllable. The distance behaviour was determined using the "time headway" (THW). It is defined as the distance of the following vehicle to the vehicle in front, divided by the speed of the vehicle behind. It allows a statement about the distance behaviour in relation to the speed driven (Maurer et al., 2015).

Setting

The experiment was conducted in a static driving simulator based on a VW Golf 6 automatic. Unlike the original vehicle, the simulator has a speedometer display and centre console display, and the mirrors have been replaced by displays. A curved screen and three projectors provide a 180-degree field of vision (see Fig. 1). The test track was realised with the simulation software SILAB 7.0®.

Figure 1. Driving simulator at Karlsruhe Institute for Technology.

The track was designed as a circuit. The starting point is a country road that leads to a junction and then via an entry to a three-lane motorway. Overgrowth prevents an early view of the flowing traffic before the first cooperative on-ramp situation takes place on the slip road. The route then leads about two kilometres at a speed limit of 120 km/h past an obscured motorway car park, where the second cooperative on-ramp situation occures. The route then leads along the motorway for a further stretch until an exit leads back onto the country road, the starting point of the route. On both slip roads only one vehicle was present (the participant's vehicle or the cooperating vehicle). There was also only one vehicle in the right-hand lane in these situations.
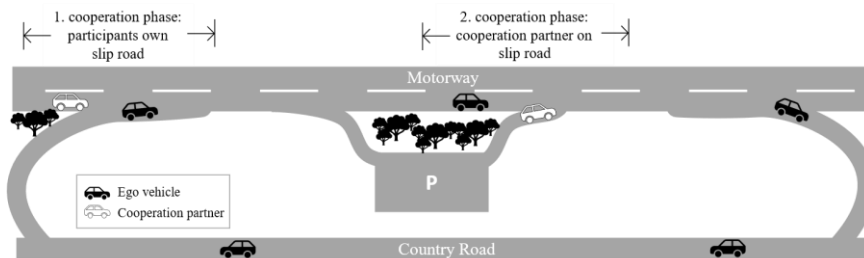
Figure 2. Overview over the track.

Initially, the participants drove a lap on the circuit without interacting with other road users for familiarisation purposes. For demonstration they drive past a parked vehicle that is marked as automated by eHMI. The eHMI consisted of a surrounding turquoise stripe (see Fig. 3). The design was adapted from Willbrink et al. (2021). Participants were instructed that this eHMI indicates that the vehicle was driving autonomously.

Figure 3. Screenshot from the simulated cooperation vehicle with eHMI.

Relevant for the study were the two slip road situations during the test drive: The first situation is the slip road of the participants onto the motorway. Here, the participants drive through a long right-hand curve restricted to 70km/h onto the acceleration lane of the motorway, where a speed limit of 120km/h applied. A cooperative vehicle is placed in the right lane of the motorway so that it is at the same level and travels at the same speed as the subject. After the two vehicles have been at the same level for about one second and the lane marking between the two lanes allows a lane change, the cooperative vehicle performs one of the following three actions:

- It accelerates from the subject's speed to about 130 km/h.
- It maintains exactly the last speed the subject was travelling at.
- It decelerates to 50km/h from the initial speed of approx. 80 km/h.

The second relevant situation is when another road user enters the motorway. To guarantee that the participants are in the right lane in this situation, they drive through a stretch of road with little traffic before the car park. At the beginning of the test drive, they are asked to respect the obligation to drive on the right. To make it difficult for the participants to evade into the middle lane, a convoy of three vehicles overtakes shortly before the car park with a significant speed difference and small distance between the individual vehicles. The vehicle on the slip road is positioned to be on the same level as the subject's vehicle and travelling at the same speed. It sets the indicator at the beginning of the lane and then performs one of the following three actions:

- It accelerates to 145 km/h and changes to the right lane at a predetermined point on the acceleration lane in front of the subject.

- It maintains the subject's speed for a further 50 metres. After 75% of the acceleration lane, it brakes sharply and then merges onto the motorway behind the subject.
- It decelerates until it is driving about 15% slower than the participant. Then it changes to the right lane of the motorway.

Procedure

To be included, participants had to hold a valid driving licence to drive a passenger car. The acquisition of participants took place in associations and interest groups close to the university. The participants were first informed about the purpose of the study and informed consent was given. This was followed by the collection of demographic data using the LimeSurvey questionnaire software. During this, the eHMI was introduced as well as the scales for the later rating of cooperativeness and criticality. This was followed by the familiarisation phase in the driving simulator. During the entire drive, the experimenter sat in the passenger seat. At first, the experimenter instructed the participants to adhere to the traffic rules, especially the obligation to drive on the right. This was followed by the test drive, during which the experimenter asked the participants to rate their cooperativeness and criticality after each cooperative situation. Finally, the subjects completed a short post-survey.

Results

Thirty-six participants participated in the study, four of whom had to discontinue the test due to simulator sickness. Of the remaining 32 participants, ten were female (31%), the mean age of the participants was $M = 25.44$ years ($SD = 3.72$) with an average driving experience of $M = 7.15$ years ($SD = 3.58$).

First-person perspective of the slip road

First, the perspective of the own slip road is analysed: The two-factor MANOVA with repeated measures showed a statistically significant influence of the cooperation vehicles' behaviour on the combined dependent variables ($F(4,28) = 7.97$, $p < .001$, $\eta_p^2 = .533$). The behavioural pattern 'accelerating' was rated as significantly more critical compared to 'decelerating' ($F(1,31) = 7.60$, $p = .010$, $\eta_p^2 = .197$). 'Maintaining speed' was not rated differently from the other two behaviours in terms of criticality ($F(1,31) < 1$, $p = .740$, $\eta_p^2 = .004$; see Fig. 4a). Accordingly, 'accelerating' was rated less cooperative ($F(1,31) = 13.41$, $p < .001$, $\eta_p^2 = .302$) compared to 'decelerating'. Furthermore, 'mainting speed' was rated significantly less cooperative than 'decelerating' ($F(1,31) = 7.48$, $p = .010$, $\eta_p^2 = .194$; see Fig. 4b).

Figure 4a. Influence of the bahviour of the cooperation partner, who was driving on the motorway, on perceiced criticality. Circles reflect mean ratings; error bars reflect standard deviations. ** p < .001, * p = .010, n.s. = not significant.

Figure 4b. Influence of the bahviour of the cooperation partner, who was driving on the motorway, on perceiced cooperativity. Circles reflect mean ratings; error bars reflect standard deviations. ** p < .001, * p = .010, n.s. = not significant.

Across all behaviours, the behaviour of the vehicle marked as automated via eHMI was rated as less critical (F (1,31) = 8.32, p = .007, η_p^2 = .212). The assessment of the cooperativeness of the partner did not differ significantly between manual and automated vehicles (F (1,31) = 3.13, p = .087, η_p^2 = .092). However, there is no significant interaction effect between the behaviour rating and the presence of the eHMI (F (4,28) = 0.22, p = .924).

Slip Road of the cooperation partner

A significant influence of the cooperation partners' behaviour on the two independent variables is also shown for the cooperation partners' slip road (F (4,27) = 16.97, p < .001, η_p^2 = .715). The behavioural pattern 'decelerating' was rated significantly less critical than 'accelerating' (F (1,30) = 13.62, p < .001., η_p^2 = .312) and 'maintaining speed' (F (1,30) = 20.11, p < .001., η_p^2 = .401; see Fig. 5a). The evaluation of the criticality of 'accelerate' and 'maintaining speed' did not differ significantly (p = .211). The behavioural pattern 'decelerating' was rated the most cooperative, significantly more cooperative than 'accelerating' (F (1,30) = 16.40, p < .001., η_p^2 = .353). The latter, in turn, was rated significantly more cooperative than 'maintaining speed' (F (1,30) = 44.76, p < .001., η_p^2 = .599; see Fig. 5b).

In contrast to the slip road from a first-person perspective, for the third-party driveway the degree of automation of the cooperative vehicle has no significant influence on the evaluation of criticality and cooperation (F (2,29) = 0.73, p = .488, η_p^2 = .048). As with own driveway, there is no significant interaction effect between the degree of automation and behaviour (F (4,27) = 0.39, p = .816, η_p^2 = .054).
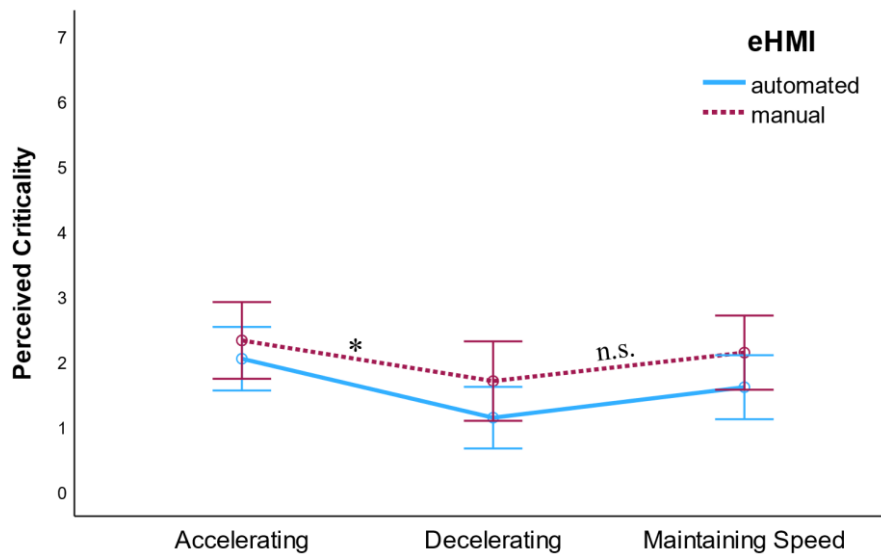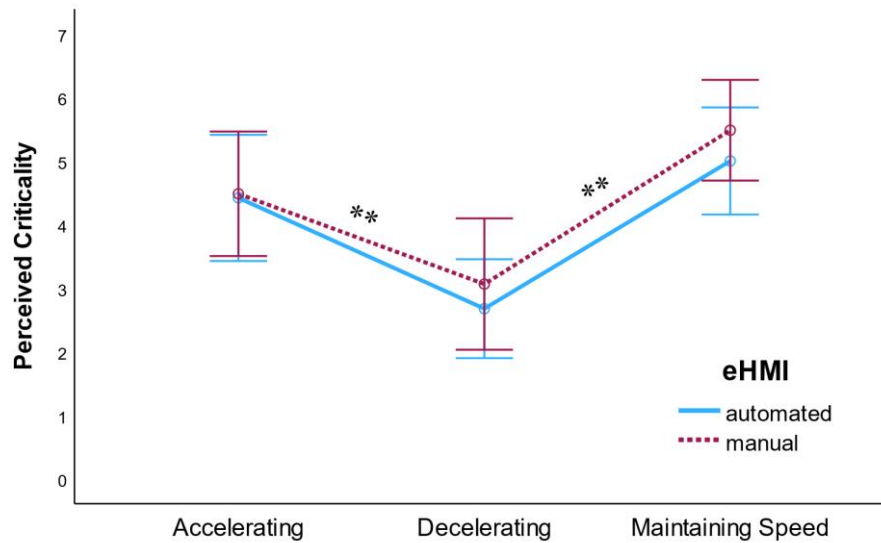

*Figure 5a. Influence of the behaviour of the cooperation partner, who was driving on the slip road, on perceived criticality. Circles reflect mean ratings; error bars reflect standard deviations. ** p < .001.*
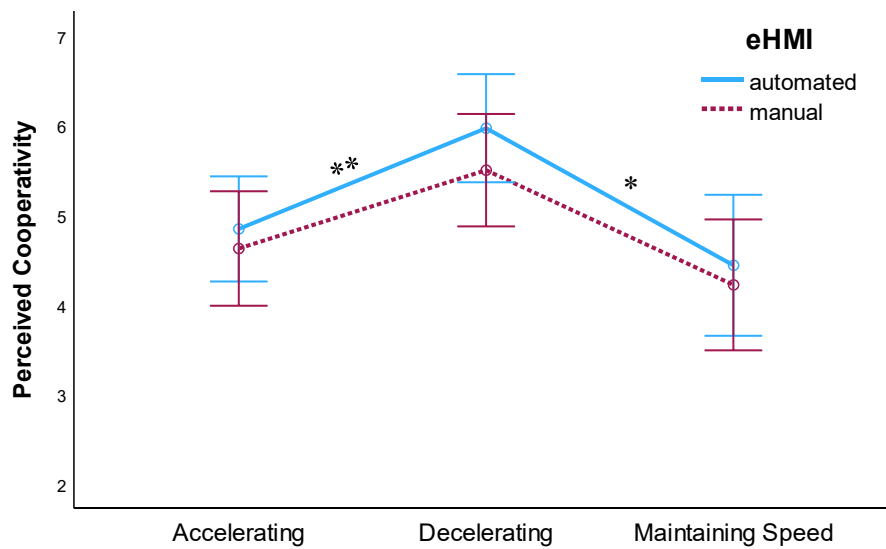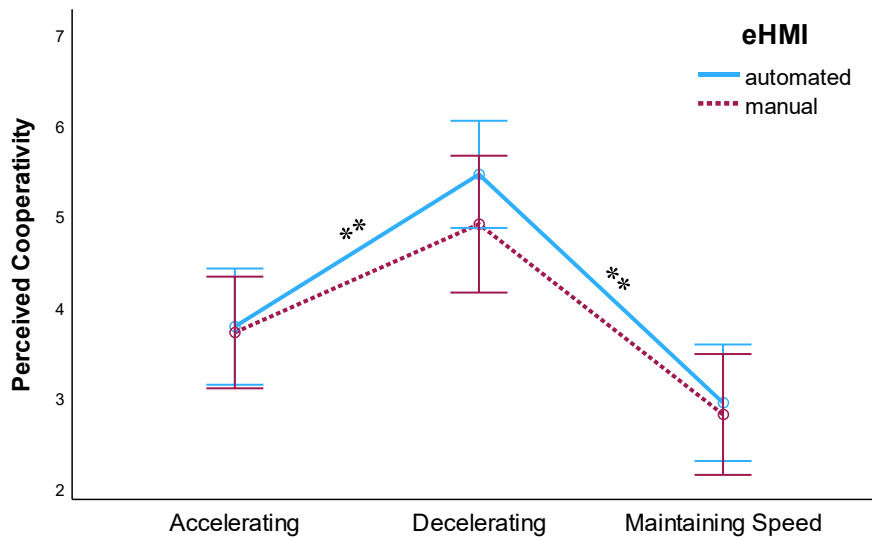
*Figure 5b. Influence of the behaviour of the cooperation partner, who was driving on the slip road, on perceived cooperativity. Circles reflect mean ratings; error bars reflect standard deviations. ** p < .001.*

Influence of the eHMI

The influence of the degree of automation indicated by the eHMI on participants' evaluation of the cooperation was already evaluated in the mutivariate models for the analysis of behaviour (*see earlier subchapters).* The analysis showed an influence of the eHMI on the evaluation of the cooperation partner when the subjects were on the slip road themselves, but not when the cooperation partner was on the slip road. In addition to the influence of the eHMI on the participants' evaluation, the influence on their distance behaviour (time headway, THW) was also investigated.

In the case of the motorway slip road from a first-person perspective, there is no significant difference in distance behaviour compared to automated vehicles (F (1, 31) <1, p = .894). However, when the cooperation partner is on the slip road, participants maintain a significantly greater distance from vehicles with eHMI (F (1, 19) = 11.99, p = .003, η_p^2 = .387). From both perspectives, the behaviour of the cooperating vehicle has a statistically significant influence on the THW. When driving on the slip road themselves (F (2, 62) = 44.36, p < .001, η^2 = .589), the greatest distance is kept from vehicles that decelerate (M = 2.01 seconds), a slightly lower distance for the 'accelerate' behaviour pattern (M = 1.44 seconds), and maintaining speed resulted in the lowest THW (M = 0.74 seconds). When the other vehicle was merging (F (1.498, 28.458) = 6.039, p = .011, η^2 = 0.441), only the conditions of 'accelerating' and 'maintening speed' differed, 'maintaining speed' resulted in significantly greater THW ($M_{accelerate}$ = 0.47, $M_{maintain}$ = 0.76, p = .002).

Discussion

The results of the driving simulator study show that the preferred behaviour of the cooperating partner is independent of the factor 'deceleration'. This is perceived as both more cooperative and less critical than accelerating or maintaining speed. A constraint exists for the first-person perspective of the slip road in terms of criticality, as there is no significant difference between "decelerating" and "maintaining speed" here. Hypothesis H1b can thus be accepted, Hypothesis H1a can only be accepted with one constraint. The results are consistent with findings from the literature, according to which accelerating or holding speed were judged as uncooperative and aggressive (Imbsweiler, 2019; Kaufmann, 2018) and signaled the intention to drive (Björklund & Åberg, 2005; Rettenmaier et al., 2021). In conclusion, at motorway slip roads, the willingness of drivers on the slip road as well as on the right lane to give way to the other vehicle is perceived as more cooperative and safer than speeding up or holding speed to indicate that one wants to drive or merge in front of the cooperative partner. However, this leads to a dilemma: if participants from both perspectives prefer that the cooperation partner brakes and gives way to the other person, only one of the two cooperation partners can act according to that preference. Otherwise, a conflict would arise. To give automated vehicles a concrete instruction for action in this scenario, the next step would be to investigate which behaviour is appropriate if both cooperation partners brake initially. However, the qualitative data show that beyond the preference towards a deceleration or acceleration behaviour, an evasive lane change may be the preferred behaviour.

Secondly, the influence of the eHMI on the evaluation of the behaviour was examined. The evaluation depends on the perspective: If the participants were on the motorway, an eHMI on the merging vehicle had no influence on the evaluation. However, if the participants were on the motorway slip road, the identical behaviour of a vehicle marked as automated was assessed as more cooperative and less critical. Hypothesis 2a is thus rejected; in one out of two situations, the eHMI has an influence on the evaluation of behaviour. This is inconsistent with previous studies, according to which a status eHMI has no influence on the participants' behaviour evaluation (Fuest et al., 2020; GATEway Project, 2017; Joisten et al., 2019). The reasons for the deviation compared to the literature could be that the previous studies partly took place in urban traffic and the criticality of the situation was possibly assessed differently due to the lower speed. And in contrast to Fuest et al. (2020), the behaviour of the cooperating vehicle potentially influenced the participant's own safety, which may have led to the behaviour being analysed in greater detail.

Hypothesis 2b is also rejected: contrary to previous findings (Fuest et al., 2020; GATEway Project, 2017; Joisten et al., 2019), in some cases the eHMI had a significant effect on participants' behaviour. This effect is also dependent on perspective. Participants kept a significantly greater distance from merging vehicles with eHMI than from those without eHMI. There was no difference in distance behaviour when the participants drove onto the motorway themselves. However, the influence of perspective may also result from the situation: When driving onto the motorway themselves, the possibility of increasing the safety distance to the vehicle in the right lane was limited due to the programming of the cooperation partner.

Nevertheless, it is worth noting that the behaviour of the participants is not consistent with their evaluation of the cooperation partner. Participants kept a significantly greater distance to oncoming vehicles with eHMI, while they did not rate their behaviour differently. When entering the motorway themselves, the behaviour of the cooperating partners with eHMI was perceived as more cooperative and less critical, but the distance did not differ significantly. This may indicate that subjects' explicit and implicit attitudes toward automated driving differ. Future research should look more deeply into how this divergence can be explained.

Limitations

The present study is subject to some limitations: Although the implementation in the driving simulator allows a high standardisation of the situation, it also leads to a limitation in the external validity. Ultimately, the results have to be validated in real traffic. Secondly, the sample was obtained in a university environment and is not representative of the population in terms of age and level of education. Furthermore, all participants were habituated to German traffic. Studies show a cultural influence on the acceptance of certain driving behaviours (Edelmann et al., 2021) and on the acceptance on automated vehicles in general (Schoettle & Sival, 2014). The generalisability of the results beyond Germany will be investigated in a following study.

Conclusion

An important result of the present study is the clear preference for the other vehicle to visibly decelerate, regardless of whether the participants themselves drive onto the motorway or interact with the cooperation partner on the slip road. The influence of the eHMI, on the other hand, is not as clear: it can influence both the participants' evaluation of the behaviour as well as their own behaviour, but not in all situations. The results provide a first basis for the implementation of implicit communication for automated vehicles on motorway slip roads, although further research is certainly needed.

References

Björklund, G.& Åberg, L. (2005). Driver behaviour in intersections: Formal and informal traffic rules. *Transportation Research Part F: Traffic Psychology and Behaviour, 8,* 239-253. https://doi.org/10.1016/j.trf.2005.04.006

Edelmann, A., Stümper, S., & Petzoldt, T. (2021). Cross-cultural differences in the acceptance of decisions of automated vehicles. *Applied Ergonomics*, *92,* 103346. https://doi.org/10.1016/j.apergo.2020.103346

Fuest, T., Feierle, A., Schmidt, E., & Bengler, K. (2020). Effects of Marking Automated Vehicles on Human Drivers on Highways. *Information, 11,* 286. https://doi.org/10.3390/info11060286.

GATEway Project. (2017). *Driver responses to encountering automated vehicles in an urban environment.* Project report PPR807. United Kingdom. Transport Research Foundation.

Imbsweiler, J. (2019). *Kooperation im Straßenverkehr in innerstädtischen Pattsituationen* [Cooperation in road traffic in inner-city standoffs]. Dissertation. Karlsruhe Institute of Technology, Karlsruhe.

Joisten, P., Alexandi, E., Drews, R., Klassen, L., Petersohn, P., Pick, A., Schwindt, S., Abendroth, B. (2019). Displaying vehicle driving mode – Effects on pedestrian behavior and perceived safety. *International Conference on Human Systems Engineering and Design: Future Trends and Applications*, 250-256. Cham, Germany: Springer.

Kauffmann, N., Naujoks, F., Winkler, F., Kunde, W. (2018). Learning the "Language" of Road Users - How Shall a Self-driving Car Convey Its Intention to Cooperate to Other Human Drivers? In. *International Conference on Applied Human Factors and Ergonomics,* 53–63. Cham, Germany: Springer.

Maurer, M., Gerdes, J.C., Lenz, B., & Winner, H. (2015). *Autonomes Fahren* [Autonomous Driving]. Berlin: Springer.

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The Case for Implicit External Human-Machine Interfaces for Autonomous Vehicles. In *ACM Digital Library, Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications,* 295–307. New York, NY, United States: Association for Computing Machinery. https://doi.org/10.1145/3342197.3345320

Neukum, A. & Krüger, H. (2003). Fahrerreaktionen bei Lenksystemstoerungen-Untersuchungsmethodik und Bewertungskriterien [Driver reaction to steering system failures-methodology and criteria for evaluation]. *VDI-Berichte*, 1791.

Rettenmaier, M., Dinkel, S., & Bengler, K. (2021). Communication via motion – Suitability of automated vehicle movements to negotiate the right of way in road bottleneck scenarios. *Applied Ergonomics, 95,* 103438. https://doi.org/10.1016/j.apergo.2021.103438.

Risto, M., Emmenegger, C., Vinkhuyzen, E., Cefkin, M., & Hollan, J. (2017). Human-Vehicle Interfaces: The Power of Vehicle Movement Gestures. In *Human Road User Coordination. Driving Assessment Conference, 9*, 186–192. https://doi.org/10.17077/drivingassessment.1633

Schaarschmidt, E., Yen, R., Bosch, R., Zwicke, L., Schade, J., & Petzoldt, T. (Hrsg.). (2021). Grundlagen zur Kommunikation zwischen automatisierten Kraftfahrzeugen und Verkehrsteilnehmern [Basics of communication between automated vehicles and road users]. *Berichte der Bundesanstalt für Strassenwesen F, Fahrzeugtechnik, 138.* Bremen. Fachverlag NW in Carl Ed. Schünemann KG.

Schoettle, B., & Sivak, M. (2014). A survey of public opinion about autonomous and self-driving vehicles in the U.S., the U.K., and Australia. University of Michigan, Ann Arbor, Transportation Research Institute.

Wilbrink, M., Lau, M., Illgner, J., Schieben, A., & Oehl, M. (2021). Impact of External Human–Machine Inter-face Communication Strategies of Automated Vehicles on Pedestrians' Crossing Decisions and Behaviors in an Urban Environment. *Sustainability, 13,* 8396. https://doi.org/10.3390/su13158396.

Artful journeys: improving the traveller's experience

Yvonne Barnard & Paul Timms
Institute for Transport Studies, University of Leeds
UK

Abstract

Transport planners usually focus on getting travellers from A to B efficiently and safely. However, travellers are not parcels: they have emotions that influence their experience. Art can enhance their experience of travelling. Making journeys by public transport or by walking more enjoyable is also stimulating sustainable travel. Over seven years we supervised 11 international students doing their dissertation on Art and Transport. They investigated different art forms and different transport modes in different countries: visual arts, design, and music, in metro and rail stations, busses, airports, and on streets. User surveys, brainstorm sessions, interviews, and gathering images on social media were methods used. Several students made their own designs, projected on pedestrian routes. A wide range of ways to improve travel experience emerged. Based on these dissertations the paper discusses how art can contribute to the travel experience, the many forms art in transport can take, and research methods to explore this topic.

Introduction

Transport planners usually focus on getting travellers from A to B efficiently and safely. However, travellers are not parcels; during their journey they have experiences and emotions. In this paper we look at how art can be used to enhance the travel experience. This is important not just for the pleasure of travellers, but also for sustainability reasons. With improved travel experience people may be more tempted to use more sustainable forms of transport instead of using their private cars. In the first place this would be public transport. But also walking can be stimulated by providing a more pleasant and interesting walking experience.

Art can enhance the travel experience, and art can evoke different emotions and feelings. Art can be soothing, taking away travel stress, be thought provoking, stimulating the traveller to think differently and be inspired. Art can "transport" the traveller to another place in their mind, such as evoking memories of other places and events. It does not need to be dramatic, art may just bring a smile, or a short disruption from boredom. Art does not even have to please everyone, it may be a point of discussion with fellow travellers or others, live or on social media.

There is another reason to study art in transport. Art is related to creativity, and by stimulating thought about the role of art, transport professionals may be stimulated to be more creative, thinking about doing things differently, whether it is in transport engineering, planning or operations.

This paper is about art; however, we do not take the perspective of art experts, art historians or artists: our perspective is that of transport experts. This means that when we talk about "art" we do not use a strict definition or an aesthetic judgement. We use the term loosely, and may also include art forms that are closer to design and decoration. However, we do not address architecture. Our aim is to look at art that can be added to existing transport vehicles like busses and trains, to buildings such as train stations and bus shelters and to the existing infrastructure such as walkways.

Master dissertations on art in transport

Intrigued by what art could mean for transport professionals, we started in 2016 with offering "artful journeys" as a topic for Master dissertations in transport planning, sustainability in transport, and transport engineering at the Institute for Transport Studies (ITS) of the University of Leeds. The students come from all over the world, some of them already working in the transport industry.

Table 1. Master students' dissertations

Year	Title	Author	Themes
2016	My beautiful busses: the role of art in public transport	Chuqiao Han	Bus and bus shelters, metro, different art forms
2017	Artful journeys: enhancing travellers experience with the high aesthetic quality environment	Kristina Shanidze	Metro stations, visual arts
2018	Artful journeys: enhancing the travel experience in the airport	Ziwei Li	Airports, visual arts
2019	Art and transport: public transit seat aesthetic design and its impacts on travel experience	Minghao Huang	Bus and metro seats, design
2019	Art and transport: Could urban bus shelter aesthetics enhance travellers' experience?	Shiyi Chen	Bus shelters, visual arts
2020	Art and transport: enhance walking experience with visual public art	Zeyu Yang	Walking, footways, visual arts
2021	Art and transport: Do people like live music in public transport?	Nachuan Wang	Train stations, music
2022	Art and transport: Metro background music enhance passenger travel experience	Xuan Fan	Metro, music
2022	Artistic travel: whether the art MTR stations fill people's cultural demand	Ziyu Long	Metro, visual arts
2022	Artful journeys: enhancing urban rail transit experience with public art that represents local culture	Dingye Han	Rail, visual arts
2022	Artful journeys: enhancing the walking experience on pedestrian crossings	Zihua Zhao	Walking, pedestrian crossings, visual arts

At first there was some scepticism amongst our colleagues. However, every year there were students who were interested in the topic, and one of them, Kristina Shanidze,

even won the Frank Lai Award from the Institute in 2017, for the dissertation which is deemed to be the most innovative or to have the most impact. In Taable 1 Master dissertation students at ITS are listed with the topics they investigated.

Students come from different cultural backgrounds and nationalities, with Chinese students showing particular interest in the topic. Students decided their own focus within this general topic, and as is clear from Table 1 there are many different choices: art in waiting areas, in vehicles and in the walking environment. Different art forms were studied: visual art (paintings, photos, statues, decorations) and music (background music, public pianos). These choices were not imposed, but came from the personal interests of the students themselves, and they formulated their own research questions. A few examples are:

- What kind of artistic forms of public arts in urban rail transit does the public prefer and expect?
- What is the role of two-dimensional public art in enhancing walking experience?
- What are both bus users and non-bus users' attitudes towards aesthetic designs of bus shelters?

In this paper we will not describe all these dissertations, but aim to provide a picture of how the topic of art in transport was explored and studied.

Methods and findings

Although the methods used were different, there are commonalities. In the next sections we aim to give an impression of how the studies were conducted and examples of results will be given. The following methods were used:

- Literature search on user experience, travel mode and art forms
- Categorizing attributes of art and travel mode
- Collecting a wide range of examples
- Observing and analysing the target domain
- Brainstorming
- Asking the public, surveys, analysing social media comments and pictures
- Asking experts, in transport research, operations, and art
- Making one's own design (sometimes) and asking for opinions in surveys
- Recommendations for practitioners and research

Literature research

Research, of course, starts with investigating what is already known about a topic. The literature reviewed came from different perspectives. There is literature on art in transport, but not much. For example: Amundsen (1995); Cascetta and Carteni (2014); Enright (2023); Isaacs (2000); Johansson (2016); Kido (2005); Li and Zhang (2017); Liu et al. (2019); Yang et al. (2019). Literature research focussed on several aspects such as attributes of art in transport, the role art can play in transport, impact of art on human behaviour, the role of art in promoting and expressing local culture and identity, and art as a means to provide information, orientation and navigation.

Some of the findings of the impact on human behaviour are related to well-being and stress reduction. Art can also play a role in helping travellers to navigate in complex environments such as underground stations. There are also claims that in public environments art can help to reduce vandalism and anti-social behaviour. Attractiveness of a place is very context sensitive, and from a transport point of view, aesthetics can be addressed not as an attribute of a transport environment, but as a user experience of dealing with this environment (Timms & Tight, 2010).

Categorizing art forms

In order to be able to study the research questions it is necessary to categorise the different art forms. This is important for collecting examples, and for designing questionnaires to explore the public's opinions. Some categorisation was based on literature study; for example Van Leeuwen (2012) uses the category dimension of realistic versus abstract, separated versus integrated, and small or localised versus large. Categories were also formed based on interviews with art experts. For example, categories identified were relation to local culture, and stimulating versus soothing (especially types of music). A categorisation of art in pedestrian spaces, formulated by Yang (2020) and accompanied by some of his own designs, is shown in Table 2.

Collecting a wide range of examples

When transport mode or environment and art categories were defined, examples of art in transport were collected, mostly using internet sources, but also own photos. Examples of metro stations collected from the internet by Long (2022) are shown in Figure 1, whilst an example of public piano playing in London, downloaded from YouTube, is shown in Figure 2.

Table 2. Categorisation of types of art in public space (Yang, 2020)

Type of art	Examples
Type A: Spliced Fragment	
Type B: Figure/Object Portrait	
Type C: Scenery	
Type D: Pattern/Stamp (including letter)	
Type E: Abstract Shapes	
Type F: Functional	

Figure 1. Examples of art in metro stations (Long, 2022)

Figure 2. Example of public piano playing in London from YouTube: www.youtube.com/watch?v=ez1iu-zx7eE. (Wang, 2021)

Observing and analysing the target domain

When studying a specific place, students used their own observations, by visiting the place or because they were acquainted with them: for example, a station in a hometown in China. When visiting a place, students looked at both the art and the behaviour and reaction of people. Also internet sources such as YouTube were used: for example, videos on the way people interact with piano players in railway stations. The categories identified were also used to rate examples: for example, metro stations. Figure 3 shows the public piano at Ningbo Railway Station, China, which was a case study site chosen by Wang (2021) in his investigation into music in railway stations.

Figure 3. Public piano playing in Ningbo Railway Staion, China (Wang, 2021)

Brainstorming

Brainstorming usually played an important role in the exploration of the topic. Brainstorming could take place at the beginning of the study, with friends, students, or staff from the Institute interested in the topic. Brainstorming often took the form of a group of 4-8 people sitting around a table in an informal setting, using (lots of) pictures of art in transport, and discussing, sorting, commenting etc. Also on-line brainstorms took place, partially due to the Covid period, but also with people in the respective home countries. Results were written out, analysed, mindmaps constructed, and fed back into the categorisation, and were used for survey construction. Brainstorming was usually a very pleasant experience for all participants, and people were sometimes surprised by their own strong feelings about art, and by how many creative ideas resulted from these sessions. A picture of a brainstorming session conducted by Chen (2019) is shown in Figure 4.

Asking the public: overview

For finding out what the opinions are of travellers about the role of art two main methods were used: asking directly by interviews and questionnaire surveys, and analysing social media comments and pictures.

Interviews
Interviews were conducted, mostly on-line. Not many interviews took place directly near the art works themselves, although this is a good method for collecting more in-depth information; for practical reasons this was usually not possible. An example of an interview with a rail passenger is shown in Table 3.

Figure 4. Brainstorming session conducted by Shiyi Chen (Chen, 2019)

Table 3. Example of an interview

	Interview quote
Rail passenger	"He was curious and stopped for about ten minutes to watch… he doesn't know much about pianos and rarely listens to piano music. He had passed by Ningbo Station many times on the train and this is the first time that he comes across a public piano improvisation. When asked if he would come to the station earlier next time to watch others play, he said yes, and that compared with his previous trips, this time with a public piano and the playing of others passengers did enhance the experience to a great extent" (Wang, 2021).

Questionnaire surveys

Students designed on-line questionnaires to ask people about their experiences, ideas and preferences. Questionnaires usually contained pictures that people could rate, and a dissertation on music had a questionnaire containing audio files. The respondent size was usually around 100-150. As on-line questionnaires were mostly distributed via snowballing, starting with students' own social networks, the respondents were largely of a younger age group. This is not necessarily a bad thing, as the aim was not to have a representative population but an exploration into travellers' experiences, and young people are an interesting target group for transport. When there was a focus on a specific place, questionnaires were distributed to people living near that place. Some examples of findings from these questionnaire surveys are shown in Table 4.

Table 4. Examples of findings from questionnaires

	Questionnaire findings
Bus users	"Generally, in both UK and China, most people care for travel experience, and quite a few are willing to pay more for a better-experience journey. Bus users fairly feel satisfied with current bus travel, but riders in UK perceive experience waiting at bus stop relatively stressful during a whole bus journey whilst their counterparts in China feel in bus journey the most uncomfortable" (Chen, 2019).
Metro users	"With regard to the subjective assessment performed we may state that in average metro users' sample is satisfied with the delivered aesthetic quality of the stations (4.69 Likert-Type score). The participants assigned the highest scores to the stations with the richest meaning, highest salience and lowest modality. The lowest appreciation was noted in cases of the non-decorated stations" (Shanidze, 2017).
Rail users	"Most participants believe that public art with local culture influence can enhance their sense of belonging and cultural pride and help them to increase their knowledge of history and culture" (Han, 2022).

Analysing social media comments and pictures

A very interesting way of finding out what people think about art in transport is to look at social media. For example, there is a Facebook group dedicated to bus seats. On Instagram there are many pictures of art in transport environments with lots of comments. By doing content analysis, both on the pictures and on the comments, one can gain insight into opinions and ideas. Several students explored this option; a more systematic search and analysis could be a good way to conduct research in the future. An example of the results of analysing social media for metro users is shown in Table 5.

Table 5. Example of results from analysis of social media for metro users

	Conclusion
Social media metro users	"Finally, the visual environment should not be homogeneous. We have an evidence of art objects of different scale being appreciated. Little pieces well integrated into the high salience environment are more likely to be noticed. They catch the attention of one's eye, allowing a sight to stop wondering and rest for a moment" (Shanidze, 2017).

Consulting experts

Experts in transport research, operations, and art were consulted, sometimes in the beginning of the projects, sometimes towards the end to discuss findings.

Making one's own design

Surprisingly for us some students came to ask whether they could make their own designs. Designs were made for a street as part of the walking environment and for pedestrian crossings. The designs were photoshopped into pictures of the real

environment. Some of the designs for public space by Zeyu Yang were shown above in Table 2, whilst designs for pedestrian crossings by Zihua Zhao (Zhao, 2022) are shown in Figure 5. In the questionnaires conducted by these two students the designs were presented to the public, asking for opinions.

Figure 5. Designs for pedestrian crossings by Zihua Zhao (Zhao, 2022)

Table 6. Examples of recommendations

Art form	Recommendation example
Bus seat design	"Seat aesthetic design could create an interaction between operator, designers and users. Special design in seat could inform users about service information, and it could be a tool for art expression" (Huang, 2019).
Airport waiting areas	"Even though people prefer pure art without business advertisements, planners and business stores could use pure art design to attract more people to utilize their waiting time" (Li, 2018).
Public space used by pedestrians	"A high level of aesthetic attribute of walking path can greatly stimulate the walking potential among urban residents" (Yang, 2020).
Pedestrian crossings	"Current reported safety issues with artistic pedestrian crossings: Negative effects: 1. It may have an unfair impact on disabled users. 2. It may distract pedestrians and motorists. 3. The excessive aesthetic appearance may attract too many visitors and disturb traffic. Positive effects: There are reports and ample evidences that artistic pedestrian crossings can be effective in reducing the number of traffic accidents" (Zhao, 2022).
Background music in metro stations	"Most passengers like to have music playing in the background in metro stations. At peak times, when passengers are at rest - on platforms and in carriages - Slow-Paced music is more appropriate. At peak times, when passengers are in motion - in the interchange area and at entry/exit points - Fast-Paced music is more appropriate" (Fan, 2022).
Public pianos in metro stations	"Public transport has been frequently used by an increasing number of people, which shows the potential as well as the responsibility of policymakers to enrich their trips. Various forms of interactions such as public pianos, are recommended to be built at transport stations. It could benefit passengers on trips and the stations in both reputation and revenue if being well operated" (Wang, 2021).

Recommendations for practitioners and research

Finally, on the basis of the findings, students came up with recommendations for both transport practitioners and for further research. Some examples of these recommendations are shown in Table 6.

Conclusions

We consider aesthetics as an aspect of human factors. The Human Factors and Ergonomics Society (https://www.hfes.org/About-HFES/What-is-Human-Factors-and-Ergonomics) uses the following definition:

> *"Human Factors is concerned with the application of what we know about people, their abilities, characteristics, and limitations to the design of equipment they use, environments in which they function, and jobs they perform."*

> *"The goal of human factors is to reduce human error, increase productivity, and enhance safety and comfort with a specific focus on the interaction between a human and the thing of interest".*

Art may not be a traditional topic in human factors research, but as the focus in human factors is on the interaction between humans and their environment and equipment, we think that enhancing the subjective experience of travellers in interaction with public transport and the walking environment is of interest. Aiming to make travel more pleasant, interesting, exciting, and going beyond efficiency and comfort could be an interesting part of human factors research.

Art can contribute to a positive travel experience, but other things need to be taken care of too, such as cleanliness, comfort, efficiency, safety and security. As Han (2016) concludes: "Art is really important to improve people's journey experience from different aspects: inspiring, educational, relaxing, etc. But the practical functions of the public transport is more important than art".

It is not only about the individual experience; art can have a strong social and cultural aspect. According to Long (2022): "Artistic metro stations can indeed generate cultural demand, and different metro stations generate different cultural demands, some obvious and some subtle. This is due to the way artists and designers combine metro stations with different cultures, and the expressive effect of the finished piece". There are many initiatives for art in transport, but there are still far more opportunities to enrich the journey experience. All our students found a large appreciation from travellers for art; art was widely recognized to be beneficial in improving their experience.

When transport students start exploring art they encounter a variety of methods, art forms and user experiences, but there is a lot still to be discovered, and new methods to be developed. We will continue with dissertations, extending to a PhD study, and integrating the topic in the curriculum. A Brazilian university (Federal Centre for Technological Education of Minas Gerais – CEFET-MG) started an art and transport project inspired by this work.

Creativity, emotion, and cultural awareness are not only of interest to travellers, but are also tools for professionals working to enhance the travel experience, because in transport it is not only about efficiency, sustainability, comfort, but also about enjoyment.

References

Amundsen, C. (1995). Public Art and Public Transportation. In *Transportation Research Board Conference Proceedings 1 (8), pp.* 75-86.

Cascetta, E., & Carteni, A. (2014). The hedonic value of railways terminals. A quantitative analysis of the impact of stations quality on travellers behaviour, *Transportation Research Part A: Policy and Practice*, *61*, 41-52.

Chen, S. (2019). *Art and transport: Could urban bus shelter aesthetics enhance travellers' experience?* MSc dissertation. Institute for Transport Studies, University of Leeds.

Enright, T. (2023). Art in transit: Mobility, aesthetics and urban development. *Urban Studies*, *60,* 67–84.

Fan, X. (2022). *Art and transport: Metro background music enhance passenger travel experience*. MSc dissertation. Institute for Transport Studies, University of Leeds.

Han, C. (2016). *My beautiful busses: the role of art in public transport*. MSc dissertation. Institute for Transport Studies, University of Leeds.

Han, D. (2022). *Artful journeys: enhancing urban rail transit experience with public art that represents local culture*. MSc dissertation. Institute for Transport Studies, University of Leeds.

Huang, M. (2019). *Art and transport: public transit seat aesthetic design and its impacts on travel experience*. MSc dissertation. Institute for Transport Studies, University of Leeds.

Isaacs, R. (2000). The urban picturesque: an aesthetic experience of urban pedestrian places. *Journal of Urban Design, 5*, 145-180.

Johansson, M., Sternudd, C., & Kärrholm, M. (2016). Perceived urban design qualities and affective experiences of walking. *Journal of Urban Design, 21,* 256-275.

Li, Y., & Zhang, L. (2017). Research on the Art Design of Public Space of Rail Transit. In: *Proceedings of the 4th International Conference on Education, Management and Computing Technology* (pp. 256-259). Atlantis Press.

Liu, Y., Wang, F., & Ren, B. (2019). Public Art on the London Underground and Wuhan Metro: A Comparative Study on the Art Form of the Metro Public Art Projects. In D. Chen, P.M.A. Desmet, H. Xue, and H. Zhang (Eds.), *Heritage & Vision, Proceedings of the 2019 International Conference on Design for Experience and Wellbeing (*pp. 15-35). Xi'an: Northwestern Polytechnical University.

Li, Z. (2018). *Artful journeys: enhancing the travel experience in the airport*. MSc dissertation. Institute for Transport Studies, University of Leeds.

Kido, E. (2005). Aesthetics Aspects of Railway Stations in Japan and Europe, As a Part of Context Sensitive Design for Railways. *Journal of the Eastern Asia Society for Transportation Studies*, *6,* 4381-4396.

Long, Z. (2022). *Artistic travel: whether the art MTR stations fill people's cultural demand.* MSc dissertation. Institute for Transport Studies, University of Leeds.

Shanidze, K. (2017). *Artful journeys: enhancing travellers experience with the high aesthetic quality environment.* MSc dissertation. Institute for Transport Studies, University of Leeds.

Timms, P.M., & Tight, M.R. (2010). Aesthetic aspects of walking and cycling. *Built Environment, 36,* 487-503 (16).

Van Leeuwen, T. (2012). *Handbook of visual analysis*. London, UK: SAGE Publications.

Wang, N. (2021). *Art and transport: Do people like live music in public transport?* MSc dissertation. Institute for Transport Studies, University of Leeds.

Yang, L., van Dam, K.H., Majumdar, A., Anvari, B., Ochieng, W.Y., & Zhang, L. (2019). Integrated design of transport infrastructure and public spaces considering human behavior: A review of state-of-the-art methods and tools. *Frontiers of Architectural Research, 8*, 429-453.

Yang, Z. (2020). *Art and transport: enhance walking experience with visual public art.* MSc dissertation. Institute for Transport Studies, University of Leeds.

Zhao, Z. (2022). *Artful journeys: enhancing the walking experience on pedestrian crossings.* MSc dissertation. Institute for Transport Studies, University of Leeds.

On the use of bicycle simulators

Bastiaan Sporrel, Arjan Stuiver, & Dick de Waard
University of Groningen, The Netherlands

Abstract

Bicycle simulators (BS) are being used for a wide variety of applications. Their use can be found in the fields of rehabilitation, sports, traffic safety research, and in the study of bicycle dynamics. Each of these fields has different design requirements for the simulators and their associated virtual environments. This paper provides an overview of the different designs of BS and the use of these systems in their respective fields. Additional attention is given to the use of BS in the field of traffic safety research, where BS are being used to examine how people experience different infrastructure layouts or how people react to specific traffic situations. Several literature gaps are discussed, as well as some limitations of current designs. This paper argues that more attention should be given to the behavioural validity of BS, since behavioural responses are the primary outcome measure in most studies, and reports on this subject are scarce. Furthermore, an improvement to current designs is proposed by including a balance component to the control of the BS, as it is necessary for the control of a regular bicycle but lacking in current BS systems.

Introduction

Despite a growing interest in cyclist's safety and improvements in bicycle infrastructure, the number of bicycle crashes has increased over the last 10 years in the Netherlands and several other countries in Europe (Adminaité-Fodor et al., 2020). This in contrast with motorized vehicle safety, where the number of fatalities continues to decline. A significant factor to consider in relation to this contrarian trends is that the current data on bicycle accidents predominantly relies on information sourced from police- and hospital records (de Guerre et al., 2020). This is an important limitation, as these describe the statistics on frequency and severity of bicycle crashes and give insight in the demographics of the parties involved, but they rarely describe the behaviour of road users in the situations leading up to the incident.

In contrast to this, in surveys and self-reports, cyclists often acknowledge the role of their behaviour, or the behaviour of others, in the crash. Most bicycle accidents are single sided crashes, i.e. crashes that are not with another road user (Krul et al., 2022). In these accidents people fall due to a loss of balance, slipping or losing control of the bicycle (Boele-Vos et al., 2017). Adverse environmental conditions (e.g. strong winds, bad road conditions, or obstacles on the road) are not the sole reason for these problems, since the behaviour of cyclists plays a major role in crashes (Utriainen et al., 2023). Knowledge on which behaviours precede loss of bicycle control and balance is necessary to create safer cycling infrastructure and can be helpful for designing programmes for fall prevention of vulnerable cyclists.

On-road studies, where participants interact with traffic, give the most realistic description of a cyclist's behaviour. There are however serious ethical concerns for testing behaviour in field experiments, especially since testing critical situations brings an obvious risk of injury to the participants. Furthermore, finding generalizable results in this type of study is more difficult because of the lack of control over situations on the road. A method where participants can cycle in a safe and controlled environment would accommodate these concerns.

One approach would be the use of a bicycle simulator. A bicycle simulator uses a virtual environment, where the user can navigate through in a similar way as if they were riding a real bicycle. This means that they can change velocity by pedalling and braking, and change direction by steering with a handlebar configuration. A simulator that is sufficiently realistic should lead to similar cycling behaviour as in real-world cycling. This use of virtual environments has been a valuable tool for behavioural studies with motorized vehicles (Bruck et a., 2021) and is becoming increasingly popular in bicycle research as well. Yet, despite this popularity, little is known about what is necessary for a bicycle simulator to provide sufficiently realistic behavioural responses. The aim of this paper is therefore to give an overview of bicycle simulators that are currently in use, and to discuss how bicycle simulators have been used in traffic safety research. Based on this information, directions for future research into (development of) bicycle simulators are provided.

Bicycle simulator types

Four categories of bicycle simulators can be identified in the literature: sport-specific simulators, simulators for traffic safety research, simulators for rehabilitation, and simulators with advanced bicycle dynamics. Regardless of category, bicycle simulators consist of three components: a virtual reality environment (VRE), a bicycle control unit, and peripherals for immersion in the virtual reality. The VRE generates the virtual world that the user moves through. Inputs for this model are generated by the user on the bicycle control unit, which is usually a bicycle-like device with pedals and a handlebar. The peripherals display the simulated environment and provide other forms of feedback for immersion in the VRE. For example, the visual part of the peripherals usually consist either of tv or computer screens, projectors or a VR headset.

Sport

The first bicycle simulators were designed for sport training and entertainment (Carraro et al., 1998). The goal of this simulator type is to stimulate engagement in training and provide the option of competing with other indoor cyclists (Westmattelmann et al., 2021). This is achieved by simulating the physical load of a cycling course while displaying the course from the rider's perspective on a television or computer screen. Physical load is calculated based on the rider's position in the virtual space, and then generated by a stationary home trainer unit that is connected to the user's bicycle. This connection is made either by connecting the back wheel of a regular bicycle to a flywheel or directly with the chain, but specifically designed exercise bikes are also available for this purpose (see for instance the Tacx Neo Bike smart trainer, or the Wahoo KICKR Bike).

These bicycle simulator systems have become increasingly popular over the last decade with the emergence of affordable and commercially available systems (McIlroy et al., 2021), and fit well within the trend of digitally monitoring of sports performance (Thompson, 2021). Online subscription-based platforms such as ZWIFT or Wahoo SYSTM can generate personalized training programmes, where users can track their progress and get real time feedback on their performance during the training. Since the primary goal is training, the main focus of the systems is to have an accurate representation of the physical workload of cycling (Matta et al., 2022; Westmattelmann et al., 2021, 2022), but they are also designed to be a motivating environment for training (Catur Supriyanto & Bing Liu, 2021). While these systems offer a realistic physical workload to the virtual-environment demands, other control inputs such as steering and braking are either not implemented, or have low fidelity. Thus, apart from the physical workload dynamics and the visualisation of the cycling course, these systems offer a limited simulation of cycling.

Figure 1. A bicycle simulator commonly used for traffic safety research. The bicycle is fixed to a static home trainer, the virtual reality environment is displayed on a head mounted display. Picture taken at Royal Dutch Visio rehabilitation centre.

Simulators for traffic safety research

Bicycle simulators for traffic safety research are designed to evaluate cyclists' behaviour. The focus of these systems is to elicit natural behaviour by immersing the user in the VRE and providing stimuli that are realistic enough for the cyclists to react to in a similar manner as they would in real-life. Immersion and the feeling of being in the VRE are important for realistic behavioural responses (Slater, 2009). The VRE therefore needs to be a sufficiently believable rendering of the scenario that is being tested to increase the feeling of being there. Scenarios are custom-built for this purpose using virtual reality editors such as the Unity or Unreal game engine. Immersion is further enhanced by either providing a surround view using head-mounted displays or by a wide field of view with multiple television screens or projector screens placed in front of and around the user. The bicycle control unit of these systems is usually built from bicycles mounted on simple static home trainers.

Sensors on the bicycles are then used for integration with the VRE. Speed, steering angle and, in some systems, braking are measured as input for the model. The main differences between these simulators and simulators that are designed for sport specifically, is that sport-specific simulators focus on realistic physical workload, whereas simulators for traffic safety research focus on a realistic virtual environment and interaction with other road users.

Rehabilitation

Bicycle simulators have also had considerable attention for physical rehabilitation. In this type of application, simulators are equipped with rehabilitation-specific ergometers with additional sensors for integration with custom-written VRE software. Single display setups are used to show the virtual environment. As in sport-oriented simulators, the VRE is used primarily as a means to make the exercises more engaging and help the patient to stay motivated to continue their rehabilitation programme and promote physical activity.

While some rehabilitation systems use realistic renderings of environments or recorded bike rides to motivate patients, many use gamification to engage the patient. Realism is thus not the goal in these systems, instead, the VREs are built to facilitate the rehabilitation tasks, which can, in some cases, change the control modes of the bicycle in the VRE, like steering by asymmetric leg forces (Deutsch et al., 2013; Yin et al., 2016), or having the avatar speed up with when the heart rate increases (Deutsch et al., 2013). This shows that in some of these systems, it is not the cycling task that is simulated but that the cycling movement is used to achieve a rehabilitation goal. The VRE in these applications can be used as a means to provide meaningful feedback on the rehabilitation task to help maximise the outcome (Sung Hwan Jeong et al., 2005). Research on these simulators is mostly focussed on the efficacy of the simulators for different patient groups.

Bicycle simulators with advanced bicycle dynamics

The focus of traffic and sport cycling simulators is to achieve a high level of realism by providing a high-fidelity rendering of the environment, or by simulating an accurate physical workload. A realistic vehicle control is, however, not central in these simulators. Steering is simplified by only turning the handlebars, whereas in real life cycling the bicycle steer and lean motion are dynamically coupled, thereby making the bicycle lean motion a major component for realistic balancing and manoeuvring (Moore et al., 2011). The simplification of steering is likely the result of the technical complexity and the associated costs of implementing additional degrees of freedom in the system, which may not have been regarded as necessary for the earlier mentioned two purposes of these simulators. The dynamics of the bicycle have nevertheless been studied extensively (Schwab & Mejaard, 2013), and it has been shown that a necessary condition for balancing bicycle is to steer into the direction of the undesired fall. Several research groups have attempted to build bicycle simulators that incorporate the dynamic behaviour of a real bicycle.
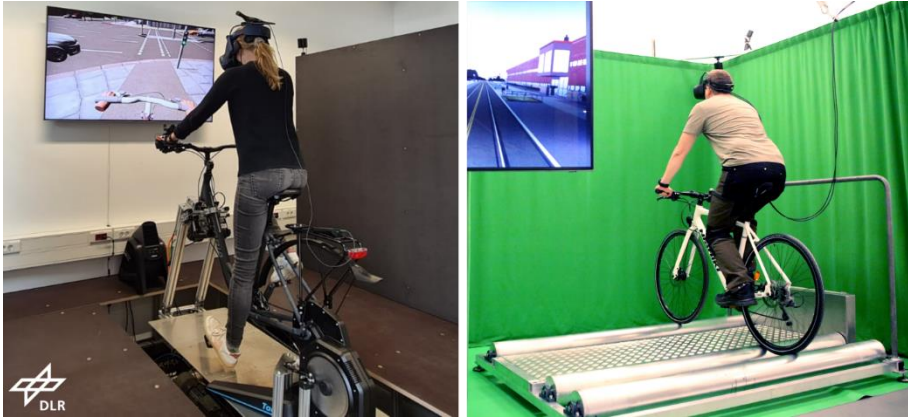
Figure 2. Two simulators with advanced bicycle dynamics. The simulator from DLR (left) uses a motion platform under the floor to add lean to the bicycle. The VTI simulator (right) consists of a bicycle on large rollers, in this setup the bicycle can move freely over the rollers and the user has to balance the bicycle as in real life. At both simulators a head mounted display is used to visualise the virtual reality environment.

The earliest example of a simulator with advanced bicycle dynamics is the KAIST interactive bicycle simulator (Kwon et al., 2001). This simulator consisted of a bicycle mounted on a motion platform with six degrees of freedom (6-DOF; the motion platform can pivot in all directions), and a motor attached to the steering tube was used to simulate the forces on the handlebars. Since then, similar bicycle simulator systems with a bicycle on a motion platform have been built at the German Aerospace Center (DLR), the TU Wien (Wintersberger et al., 2022) and the Hochshule Koblenz (Schulzyk et al., 2009). At the Swedish National Road and Transport Research Institute (VTI) a driving simulator with an 8-DOF motion platform was modified to fit a bicycle simulator (Bruzelius & Augusto, 2018), this system added horizontal movements along with the moving base. A different system of adding bicycle lean to a simulator was developed by Yamaguchi et al. (2018), they used a pully system with cables to change the bicycle lean angle. At the TU Delft a simulator was developed that, without lean motion, provides haptic feedback for steering by generating torque on the handlebars based on bicycle dynamics, (Dialynas et al., 2019). Although these systems can mimic the movements of a bicycle, getting the bicycle dynamics and the general 'feel' of the simulator right remains a challenge (Bruzelius & Augusto, 2018; Wintersberger et al., 2022). Present systems calculate the lean angle of the bicycle kinematically from the virtual speed of the cyclist and the steering angle, while in real life cycling the steer and lean angle are dynamically coupled (Meijaard 2007, Kooijman 2011) and both need to be controlled by the cyclist to maintain balance. Other issues such as computational lag and system inertia also decrease the realism of these systems (Bruzelius & Augusto, 2018). Wintersberger et al. (2022) discussed a more fundamental problem with bicycle tilt in a simulator: In normal cycling, centrifugal forces counteract gravitational forces when negotiating a corner. In a simulator, these centrifugal forces are absent, which results in unrealistic gravitational pull in curves when the same leaning angles are used. A weaker bicycle lean may therefore be beneficial for the feeling of realism when cycling in VR. This example highlights that a one-to-one translation of cycling dynamics to a bicycle simulator on

a motion platform may not be desirable, and that additional research is required to find the best implementation of bicycle dynamics for such simulators, and how this changes the user's bicycle control.

One method to preserve the natural dynamics of a bicycle while keeping the bicycle stationary is by placing the bicycle on rollers (Dressel & Papadopoulos, 2012). This setup has been tested as a bicycle control unit for a simulator (VTI, 2023), but no formal studies have been published using a bicycle simulator on rollers. While this type of simulator offers lifelike bicycle handling, it is not without its challenges. To keep the bicycle balanced it is necessary to continuously cycle, so the task of stopping cannot be evaluated in this simulator. The ability to negotiate a corner is equally challenging, as steering too much would result in the bicycle going off the rollers. The design of the simumulator described in Haasnoot et al. (2023) does not use rollers, but balances the bicycle by laterally displacing the front- and rear wheel contact points during steering, based on a dynamic model. This eliminates the problem of going off the rollers, but is again an approximation of natural cycling dynamics. However, for cycling on a straight road, the realistic bicycle dynamics of these simulators are a major advantage.

Traffic safety research using bicycle simulators

In traffic safety research, bicycle simulators are used for many different types of studies. One major research topic is how people react to different road markings and infrastructure designs (Bialkova et al., 2018; Bialkova & Ettema, 2019; Cobb et al., 2021; Nazemi et al., 2021). The adaptability of virtual environments allows for the researchers to test road factors in a controlled environment without having to modify the physical environment, thereby preventing potentially dangerous situations and saving costs, while also having control over the traffic around the participants. Research topics within this field can be divided into two categories; how people experience different bicycle infrastructure, and the effect of different bicycle infrastructure on cycling behaviour.

How people experience bicycle infrastructures is most often measured with questionnaires after a bicycle ride in a simulator. Perceived level of safety, willingness to cycle and enjoyment are common measures (Bialkova et al., 2018; Bialkova & Ettema, 2019; Nazemi et al., 2021), although galvanic skin response is also sometimes used as a measure of stress (Cobb et al., 2021). The main manipulations in these studies are changing the width as well as the location of the cycling infrastructure (i.e., on the road, separated bicycle path or on pavement), as are the speed and volume of other traffic. Realism and the feeling of being present in the virtual environment are considered to be important for this type of research, and as such, a head mounted display has been the preferred method for displaying the VRE (Bialkova et al., 2018; Bialkova & Ettema, 2019), although Cobb et al. (2021) achieved comparable outcomes while using a single large screen to display the VRE. The effects of display type on the perceived level of safety and stress is a worthwhile topic for future research.

Behavioural effects of different bicycle infrastructures are usually measured as effects on lateral position on the road and cycling speed (Abadi et al., 2022; Abadi & Hurwitz,

2018, 2019; Brown et al., 2017; Huemer et al., 2022; O'Hern et al., 2017). These variables are measured directly in the VRE in these studies. Recommendations for bicycle path markings and other cycling infrastructure are given from the perspective of the study's specific goals, such as guiding cyclists to specific waiting spots at signalised intersections (Brown et al., 2017), improving the safety of cyclists near intersections (Abadi & Hurwitz, 2018) and commercial loading zones (Abadi et al., 2022; Abadi & Hurwitz, 2019), or guiding cyclists to cycle at a safe distance from cars (Huemer et al., 2022; O'Hern et al., 2017). The simulators used in these studies are static simulators, which require no balancing by the cyclist. As such, the lateral space for balancing and the natural swerving behaviour is not present. The extent of these differences is unclear, as few studies have made direct comparisons between virtual bicycling and on-road cycling (see section 'Validation of bicycle simulators').

Interactions with other traffic.

Simulators also offer the possibility to study the interactions with other road users. In a VRE, the positions and behaviours of the other road users can be precisely controlled, making it possible to offer standardised situations and study the behavioural responses of participants. Some examples of this are the work of Liu et al. (2012), who looked at the braking onset of cyclists when confronted with motorcycles cutting in front of the participant, or the work of Thorslund and Lindström (2020), who evaluated cyclist's stopping position at signalised intersections shared with car and truck traffic. Another good example is the work of a research group from the University of Iowa. They used the control of traffic flow to study the crossing behaviour of children and adults. By controlling the distance between cars in the virtual environment, the authors could identify which gap sizes between cars are preferred when crossing a road and how adults and children time their crossing (Plumert et al., 2011). They used a system with three large projector screens for immersion, and since the focus was on street crossing, special attention was given to generating realistic bicycle propulsion on the simulator. The latter is a unique feature of the simulator system, as no other previously mentioned study incorporated realistic bicycle dynamics in any way. This may highlight a limitation of other bicycle simulators; they strive to study realistic human behaviour, but they do not offer realistic physical bicycle behaviour.

Yet, not all bicycle simulators for traffic research are used to test cyclists' behaviour. In the development of highly automated vehicles (HAV), bicycle simulators are used to assess cyclists' preferences for the behaviour and communication of the HAV. For instance, the evaluation of external visual interfaces (Hou et al., 2020; Kaß et al., 2020), or the HAV onset of braking when it has to give way to cyclists (Stange et al., 2021). In this type of studies, the physical properties of bicycle handling may not be as important. As long as the participants can read the intentions of the HAV, the study goals can be achieved. As such, the simulators used in these studies are relatively simple home trainer setups with the ability to steer. Having some agency in the VR is preferred, however, as the ability to swerve and brake in VR changes the participants' preferences to HAV behaviour (Stange et al., 2021), thereby showing the benefits of a bicycle simulator over video presentations in this type of research.

Validation of bicycle simulators

While the use of bicycle simulators in traffic safety research is growing, only a few of these simulators have been validated in experiments. Horne et al. (2018) and Powel et al. (2018) both calibrated their simulators to have the movements in the VE match the speed of the physical bike. Horne et al. (2018) did this by finding the optimal tire pressure of the rear wheel, and Powel et al. (2018) compared the speed output of the simulator to a theoretical model. Both studies took a mechanical approach to the validation to create a realistic visual movement in the VR based on the physical workload generated on the bicycle. While the authors argue that this can be achieved, it remains unclear if the users experience the simulators to be a realistic representation, as no subjective measures were taken, and no direct comparison with real-world cycling was done.

There are more studies that focus primarily on the mechanical and computational parts of the simulators; Shoman and Imine (2021) tested a novel bicycle simulator (the novelty being an asphalt specimen that contacts the real wheel tire to create realistic road vibrations) and asked their participants how realistic the simulator felt, but also focused mostly on the validation of their model of bicycle dynamics. The participants indicated that the simulator felt somewhat realistic with the new asphalt specimen, but since no comparisons were made with real-world cycling behaviour, again, the behavioural validity remains unclear.

Only two studies have compared real-world cycling with cycling in VR. The study by O'Hern et al. (2017) modelled a VR road to an actual road that the participants cycled on and measured lateral position (LP), standard deviation of lane position (SDLP), and speed in both situations. Comparable LP and SDLP were found on the road and in VR, but participants kept a lower speed in VR. Despite the speed difference, the authors concluded that the validity of the simulator is sufficient for their purposes of evaluating infrastructure design. Haasnoot et al. (2023) performed an validation with their simulator where they compared the motions of the bicycle control unit with the motions of a (racing) bicycle during on-road cycling. Since similar movements were found between the simulator and the regular bicycle, the authors conclude that their method of simulating balance on a bicycle is sufficient. Allthough the authors took subjective measures of realism, they interpret these with caution, since the subject's attention was mostly focussed on maintaining the right speed for the experiment task. With the mechanical validity of this simulator tested, a validation of cycling behaviour in a similar study would be a valuable addition.

Discussion

Bicycle simulators are being used for a wide variety of purposes, and for many purposes, a custom bicycle simulator is designed, emphasising specific aspects of the bicycle simulator. Simulators for cycling sport appear to be the most used, with commercial systems being widely available. In this type of simulator, the emphasis lies mainly on creating a realistic physical workload. Cycling simulators for physical rehabilitation purposes are more focussed on creating engaging training environments, where the VRE is used as a tool to provide specific feedback to facilitate the task. Simulators that are used for traffic safety research have the primary

goal of measuring behaviour and experience in the VRE, and therefore the primary focus of these systems is to have a realistic VRE. The bicycle control unit often has a simple design in these systems, consisting of a cycling ergometer or a standard bicycle mounted on a fixed home trainer system. Only a few simulators are more advanced in the sense that they offer additional degrees of freedom next to steering and pedalling. These simulators are hardly ever used in research, as a consequence it remains unclear how these should be tuned for an optimal virtual cycling experience and how cycling differs on these simulators when compared to cycling on the road.

The way VRE is displayed varies greatly between simulators. Systems used include setups with one or more large television screens, multiple projector screens setup, and the use of head-mounted displays (HDM). The latter are becoming more popular due to their increasing affordability, the small space they require, and the high level of immersion that they can achieve. In studies where behavioural validity is important, a high immersion and a realistic VRE is in general preferred (Slater, 2009). The challenge with a highly immersive display of the VRE is that this increases the risk of simulator sickness, which is especially prevalent when using an HMD (Mittelstaedt et al., 2018). While some methods can reduce simulator sickness, such as reducing the time spent in the VRE or simulating airflow (Matviienko et al., 2022), it remains a problem for many individuals. A profound investigation of which display type is best suited for a bicycle simulator is required to find the ideal solution for studying cycling behaviour.

Although cycling behaviour is a primary topic in several studies, little is known about the similarity of the behavioural measures, such as lateral position in VR compared with real-world cycling, since up to now sparse attention is given to testing the validity of bicycle simulators. When validity is being assessed, the focus is primarily on the mechanical aspects of the simulator, behavioural validity is hardly ever addressed. Huemer et al. (2022) compared the results with results from other studies, but a direct comparison between a cyclist's behaviour in a virtual environment and on a real road is limited to the study by O'Hern et al. (2017). Results show that these simulators have relative validity when it comes to position on the road and speed choice (even if cycling speed is often lower in the simulators), but other aspects of behaviour remain untested. Some anecdotal reports indicate that people cycle differently in VR, as exemplified by participants in the study of Hou et al. (2020), who reported that they took more risk in the simulator because they perceived less risk in the simulator than in real traffic. Knowing whether people might modify their behaviour in the bicycle simulators is paramount to know the limitations of the simulators and the research that is done with these devices, thus validation of bicycle simulators deserve a higher priority.

One item that is lacking at the moment in cycling simulators is balance control. Nearly all simulators are static machines with steering and pedalling as a control input. In regular cycling, the dynamic coupling between steer and lean angle is an essential mechanism for balancing and manoeuvring the bicycle (Kooijman & Schwab, 2011; Moore et al., 2011; Shwab & Meijaard, 2013). Also, a model that simulates the relationship between speed and steering is not always implemented (Dialynas et al., 2019). As a result, people get the feeling that they cannot steer properly (Shoman &

Imine, 2021; Stratmann et al., 2019). These two factors are likely to change the cyclist's normal behaviour, and common measures of cycling behaviour, such as swerving could be affected in this type of simulators. This lack of balance control also makes it currently difficult to study cyclists' behaviour leading to a fall, as many accidents are reportedly due to a loss of balance (Boele-Vos et al., 2017). To use a bicycle simulator to safely study cyclists' behaviour in critical situations, balance as a control element for cyclists' movement is a necessary addition.

Conclusion

Bicycle simulators have been developed to achieve different goals; sports and physical rehabilitation simulators focus on providing a motivating environment to enhance exercise, whereas simulators designed for bicycle dynamics and traffic safety research have the goal of creating a realistic cycling experience. In traffic safety research, these simulators have been used to measure cyclists' behaviour in different infrastructure designs and in interactions with other road users. Although bicycle simulators have existed for several decades, their designs could still be improved. We propose that future research should prioritise the behavioural validation of existing bicycle simulators, and the effects of different visualisation options to the behaviour of the users in VR. Newly developed bicycles that are used to study cyclists' behaviour should also strive for more realistic bicycle handling by implementing balance as a control element.

Aknowledgements

References

Abadi, M.G., & Hurwitz, D.S. (2018). Bicyclist's perceived level of comfort in dense urban environments: How do ambient traffic, engineering treatments, and bicyclist characteristics relate? *Sustainable Cities and Society*, *40*, 101–109. https://doi.org/10.1016/j.scs.2018.04.003

Abadi, M.G., Maloney, P., & Hurwitz, D. (2022). Exploring Bicyclists' Visual Attention during Conflicts with Truck Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(11), 137–144. https://doi.org/10.1177/03611981221091709

Adminaité-Fodor, D., Jost, G., Stipdonk, H., Sanz-Villegas, M. T., Thomas, P., Velten, B., Foundation, D., Yannis, G., & Avenoso, A. (2020). How safe is walking and cycling in Europe? In *73*. www.etsc.eu/pin

Bialkova, S., & Ettema, D. (2019). Cycling renaissance: The VR potential in exploring static and moving environment elements. *2019 IEEE 5th Workshop on Everyday Virtual Reality (WEVR)*, 1–6. https://doi.org/10.1109/WEVR.2019.8809586

Bialkova, S., Ettema, D., & Dijst, M. (2018). Urban future: Unlocking Cycling with VR Applications. *2018 IEEE Workshop on Augmented and Virtual Realities for Good (VAR4Good)*, 1–5. https://doi.org/10.1109/VAR4GOOD.2018.8576888

Boele-Vos, M.J., Van Duijvenvoorde, K., Doumen, M.J. A., Duivenvoorden, C.W.A.E., Louwerse, W.J.R., & Davidse, R.J. (2017). Crashes involving cyclists aged 50 and over in the Netherlands: An in-depth study. *Accident Analysis & Prevention*, *105*, 4–10. https://doi.org/10.1016/j.aap.2016.07.016

Brown, H., Sun, C., & Qing, Z. (2017). Investigation of Alternative Bicycle Pavement Markings with the Use of a Bicycle Simulator. *Transportation Research Record: Journal of the Transportation Research Board*, *2662*(1), 143–151. https://doi.org/10.3141/2662-16

Bruck, L., Haycock, B., & Emadi, A. (2021). A Review of Driving Simulation Technology and Applications. *IEEE Open Journal of Vehicular Technology*, *2*, 1–16. https://doi.org/10.1109/OJVT.2020.3036582

Bruzelius, F., & Augusto, B. (2018). *CykelSim Development and demonstration of an advanced bicycle simulator Public report*. www.vinnova.se/ffi

Carraro, G.U., Cortes, M., Edmark, J.T., & Ensor, J.R. (1998). The peloton bicycling simulator. *Proceedings of the Third Symposium on Virtual Reality Modeling Language*, 63–70. https://doi.org/10.1145/271897.274372

Supriyanto, C & Liu, B (2021). Virtual Cycling For Promoting a Healthy Lifestyle. *International Journal of Science, Technology & Management*, *2*(1), 60–71. https://doi.org/10.46729/ijstm.v2i1.114

Cobb, D.P., Jashami, H., & Hurwitz, D.S. (2021). Bicyclists' behavioral and physiological responses to varying roadway conditions and bicycle infrastructure. *Transportation Research Part F: Traffic Psychology and Behaviour*, *80*, 172–188. https://doi.org/10.1016/j.trf.2021.04.004

de Guerre, L.E.V.M., Sadiqi, S., Leenen, L.P.H., Oner, C.F., & van Gaalen, S. M. (2020). Injuries related to bicycle accidents: an epidemiological study in The Netherlands. *European Journal of Trauma and Emergency Surgery*, *46*(2), 413–418. https://doi.org/10.1007/s00068-018-1033-5

Deutsch, J.E., Myslinski, M.J., Kafri, M., Ranky, R., Sivak, M., Mavroidis, C., & Lewis, J.A. (2013). Feasibility of Virtual Reality Augmented Cycling for Health Promotion of People Poststroke. *Journal of Neurologic Physical Therapy*, *37*(3), 118–124. https://doi.org/10.1097/NPT.0b013e3182a0a078

Dialynas, G., Happee, R., & Schwab, A.L. (2019). Design and hardware selection for a bicycle simulator. *Mechanical Sciences*, *10*(1), 1–10. https://doi.org/10.5194/ms-10-1-2019

Dong-Soo Kwon, Gi-Hun Yang, Chong-Won Lee, Jae-Cheol Shin, Youngjin Park, Byungbo Jung, Doo Yong Lee, Kyungno Lee, Soon-Hung Han, Byoung-Hyun Yoo, Kwang-Yun Wohn, & Jung-Hyun Ahn. (2001). KAIST interactive bicycle simulator. *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*, *3*, 2313–2318. https://doi.org/10.1109/ROBOT.2001.932967

Dressel, A., & Papadopoulos, J. M. (2012). Comment on 'on the stability of a bicycle on rollers'. European journal of physics, 33(4), L21. https://doi.org/10.1088/0143-0807/33/4/L21

Ghodrat Abadi, M., & Hurwitz, D.S. (2019). Operational Impacts of Protected-Permitted Right-Turn Phasing and Pavement Markings on Bicyclist Performance during Conflicts with Right-Turning Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(4), 789–799. https://doi.org/10.1177/0361198119837231

Haasnoot, J., Happee, R., van der Wijk, V. and Schwab, A.L. (2023). Validation of a novel bicycle simulator with realistic lateral and roll motion. *Vehicle System Dynamics*, 1-25. https://doi.org/10.1080/00423114.2023.2264418

Hou, M., Mahadevan, K., Somanath, S., Sharlin, E., & Oehlberg, L. (2020). Autonomous Vehicle-Cyclist Interaction: Peril and Promise. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376884

Huemer, A.K., Rosenboom, L.M., Naujoks, M., & Banach, E. (2022). Testing cycling infrastructure layout in virtual environments: An examination from a bicycle rider's perspective in simulation and online. *Transportation Research Interdisciplinary Perspectives*, *14*, 100586.
https://doi.org/10.1016/j.trip.2022.100586

Kaß, C., Schoch, S., Naujoks, F., Hergeth, S., Keinath, A., & Neukum, A. (2020). A Methodological Approach to Determine the Benefits of External HMI During Interactions Between Cyclists and Automated Vehicles: A Bicycle Simulator Study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12213 LNCS* (pp. 211–227). Springer. https://doi.org/10.1007/978-3-030-50537-0_16

Kooijman, J.D.G., & Schwab, A.L. (2011). A Review on Handling Aspects in Bicycle and Motorcycle Control. *Volume 4: 8th International Conference on Multibody Systems, Nonlinear Dynamics, and Control, Parts A and B*, 597–607. https://doi.org/10.1115/DETC2011-47963

Kooijman, J.D., Meijaard, J.P., Papadopoulos, J.M., Ruina, A. and Schwab, A.L., 2011. A bicycle can be self-stable without gyroscopic or caster effects. Science, 332(6027), pp.339-342.

Krul, I., Valkenberg, H., Asscherman, S., Stam, C., & Klein Wolt, K. (2022). *Fietsongevallen en snor-/bromfietsongevallen in Nederland*. Amsterdam: VeiligheidNL.

Matta, G., Edwards, A., Roelands, B., Hettinga, F., & Hurst, P. (2022). Reproducibility of 20-min Time-trial Performance on a Virtual Cycling Platform. *International Journal of Sports Medicine*, *43*, 1190–1195. https://doi.org/10.1055/a-1848-8478

Matviienko, A., Müller, F., Zickler, M., Gasche, L.A., Abels, J., Steinert, T., & Mühlhäuser, M. (2022). Reducing Virtual Reality Sickness for Cyclists in VR Bicycle Simulators. *CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3491102.3501959

McIlroy, B., Passfield, L., Holmberg, H.C., & Sperlich, B. (2021). Virtual Training of Endurance Cycling – A Summary of Strengths, Weaknesses, Opportunities and Threats. *Frontiers in Sports and Active Living*, *3*, 631101. https://doi.org/10.3389/fspor.2021.631101

Meijaard, J.P., Papadopoulos, J.M., Ruina, A. and Schwab, A.L., 2007. Linearized dynamics equations for the balance and steer of a bicycle: a benchmark and review. *Proceedings of the Royal Society* A: mathematical, physical and engineering sciences, 463(2084), pp.1955-1982.

Mittelstaedt, J., Wacker, J., & Stelling, D. (2018). Effects of display type and motion control on cybersickness in a virtual bike simulator. *Displays*, *51*, 43–50. https://doi.org/10.1016/j.displa.2018.01.002

Moore, J.K., Kooijman, J.D.G., Schwab, A.L., & Hubbard, M. (2011). Rider motion identification during normal bicycling by means of principal component analysis. *Multibody System Dynamics*, *25*(2), 225–244. https://doi.org/10.1007/s11044-010-9225-8

Nazemi, M., van Eggermond, M.A.B., Erath, A., Schaffner, D., Joos, M., & Axhausen, K.W. (2021). Studying bicyclists' perceived level of safety using a bicycle simulator combined with immersive virtual reality. *Accident Analysis & Prevention*, *151*, 105943. https://doi.org/10.1016/j.aap.2020.105943

O'Hern, S., Oxley, J., & Stevenson, M. (2017). Validation of a bicycle simulator for road safety research. *Accident Analysis & Prevention*, *100*, 53–58. https://doi.org/10.1016/j.aap.2017.01.002

Plumert, J.M., Kearney, J.K., Cremer, J.F., Recker, K.M., & Strutt, J. (2011). Changes in children's perception-action tuning over short time scales: Bicycling across traffic-filled intersections in a virtual environment. *Journal of Experimental Child Psychology*, *108*(2), 322–337. https://doi.org/10.1016/j.jecp.2010.07.005

Schulzyk, O., Hartmann, U., Bongartz, J., Bildhauer, T., & Herpers, R. (2009). A real bicycle simulator in a virtual reality environment: the FIVIS project. In *IFMBE Proceedings* (Vol. 22, pp. 2628–2631). https://doi.org/10.1007/978-3-540-89208-3_630

Schwab, A.L., & Meijaard, J.P. (2013). A review on bicycle dynamics and rider control. *Vehicle System Dynamics*, *51*(7), 1059–1090. https://doi.org/10.1080/00423114.2013.793365

Shoman, M.M., & Imine, H. (2021). Bicycle Simulator Improvement and Validation. *IEEE Access*, *9*, 55063–55076. https://doi.org/10.1109/ACCESS.2021.3071214

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549–3557. https://doi.org/10.1098/rstb.2009.0138

Stange, V., Goralzik, A., & Vollrath, M. (2021). Keep Your Distance, Automated Vehicle! – Configuration of Automated Driving Behavior at an Urban Junction from a Cyclist's Perspective. In *Lecture Notes in Networks and Systems* (Vol. 270, pp. 393–402). https://doi.org/10.1007/978-3-030-80012-3_45

Stratmann, L., Buse, D. S., Heinovski, J., Klingler, F., Sommer, C., Tunnermann, J., Scharlau, I., & Dressler, F. (2019). Psychological feasibility of a virtual cycling environment for human-in-the-loop experiments. *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft Fur Informatik (GI)*, *295*, 185–194. https://doi.org/10.18420/inf2019_ws21

Sung Hwan Jeong, Yong Jun Piao, Woo Suk Chong, Yong Yook Kim, Sang Min Lee, Tae Kyu Kwon, Chul Un Hong, & Nam Gyun Kim. (2005). The Development of a New Training System for Improving Equilibrium Sense Using a Virtual Bicycle Simulator. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, *7 VOLS*, 2567–2570. https://doi.org/10.1109/IEMBS.2005.1616993

Swedish National Road and Transport Research Institute. (2023, Februari 24). *VTI's new cycling simulator on rollers.* https://www.vti.se/en/archives/news/ archives/ 2023-02-24-vtis-new-cycling-simulator-on-rollers

Thompson, W.R. (2021). Worldwide Survey of Fitness Trends for 2021. *ACSM'S Health & Fitness Journal*, *25*(1), 10–19. https://doi.org/10.1249/FIT.0000000000000631

Utriainen, R., O'Hern, S., & Pöllänen, M. (2023). Review on single-bicycle crashes in the recent scientific literature. Transport Reviews, 43(2), 159–177. https://doi.org/10.1080/01441647.2022.2055674

Westmattelmann, D., Grotenhermen, J.-G., Sprenger, M., & Schewe, G. (2021). The show must go on - virtualisation of sport events during the COVID-19 pandemic. *European Journal of Information Systems*, *30*(2), 119–136. https://doi.org/10.1080/0960085X.2020.1850186

Westmattelmann, D., Stoffers, B., Sprenger, M., Grotenhermen, J.-G., & Schewe, G. (2022). The Performance-Result Gap in Mixed-Reality Cycling – Evidence From the Virtual Tour de France 2020 on Zwift. *Frontiers in Physiology*, *13*. https://doi.org/10.3389/fphys.2022.868902

Wintersberger, P., Matviienko, A., Schweidler, A., & Michahelles, F. (2022). Development and Evaluation of a Motion-based VR Bicycle Simulator. *Proceedings of the ACM on Human-Computer Interaction*, *6*(MHCI), 1–19. https://doi.org/10.1145/3546745

Yamaguchi, T., Yamada, A., Fujiwara, T., Takahashi, R., Shimada, T., Ishida, Y., Ueda, K., & Kaneda, S. (2018). Development of Bicycle Simulator with Tilt Angle Control Tilt Angle. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, *2*, 247–252. https://doi.org/10.1109/COMPSAC.2018.10238

Yin, C., Hsueh, Y.H., Yeh, C.Y., Lo, H.C., & Lan, Y.T. (2016). A Virtual Reality-Cycling Training System for Lower Limb Balance Improvement. *BioMed Research International*, *2016*, 1–10. https://doi.org/10.1155/2016/9276508

Comparing participants' risk perception and sensed presence while driving in CARLA and a self-developed remote driving testbed

Quang Son Le, Avinoam Borowsky, Shai Arogeti, & Yisrael Parmet
Ben Gurion University of the Negev
Israel

Abstract

While automotive research plays a significant role nowadays, most on-road and natural studies activities demand costly platforms and involve safety issues. A large variety of driving simulators proposes an alternative to reduce monetary costs and guarantee safety. However, no matter how sophisticated driving simulators are, they still cannot reflect the actual complexities of the physical world. In this study, we evaluated a self-developed remote driving testbed platform to be used as a hybrid solution that combines the strengths of the simulator and actual vehicle driving in the real world. To increase the external validity of the remote driving station, we implemented a driving torque mechanism. This study introduces an empirical experiment comparing participants' perceptions and behaviour in two driving platforms. The first includes our developed testbed consisting of a remote station and a toy-car driving on a 1:10 scale circular road, and the second includes the CARLA driving simulator. The experimental results demonstrated the sense of risk the drivers in the testbed experienced versus the risk-free experience when driving a conventional simulator. The risk-free experience influenced drivers' confidence and sensed presence. Also, the participants rated the testbed as more realistic than the high-end graphical driving simulator.

Introduction

A driving simulator is a laboratory device that is an alternative solution to outdoor driving experiments. For many years, it was utilized to study the human aspect of automotive driving, and in recent years, scenarios of mixed traffic (i.e., autonomous and human drivers) can be tested in the virtual environment. The driving simulator is a cost-effective and safe solution to study traffic dynamics that involve human drivers (Dosovitskiy et al., 2017).

Although many of the simulation aspects were improved to make the environment more realistic (such as integrating Virtual Reality technology), the virtual world is still a much-simplified version of the actual world and uses high-resolution graphics that drastically increase the processing load. (Taheri et al., 2017). Another prominent

limitation of driving simulators is that the interaction with autonomous cars is only simulated based on computational models, which may not be realistic, especially due to the rapid progress in this field (Punzo & Ciuffo, 2011). To resemble realistic vehicle dynamics, efforts were made to develop vehicle models that integrate into a virtual reality simulation (Shabalina et al, 2019 and Dosovitskiy et al, 2017). However, these models must still be validated in real driving conditions (Shabalina et al, 2019). This is reasonable, because the model, regardless of its complexity, cannot represent the real physical world exactly with its nonlinear and uncertain nature.

Besides the infrastructure and physical model issues, the simulator's ability to provide drivers with real-world driving feeling is still ambiguous (Kemeny & Panerai, 2003). De Winter et al. (2012) point out three disadvantages that participants could face while driving in a simulator but not in a real vehicle: (1) Limited physical, perceptual, and behavioural fidelity, (2) lack of research demonstrating the validity of driving simulators, and (3) discomfort, especially for older people or when demanding conditions are tested. Focusing on the first bullet, a study by De Winter et al. (2012) indicates that despite the safety benefits of using driving simulators, they lack real danger and actual action consequences. This causes participants to have a false sense of safety and responsibility. Similarly, Milleville-Pennel and Charron (2015) compared a simulator with real-world driving. The participants were asked to drive normally on each platform (with similar road conditions). They then answered psychological questions regarding various feelings about their driving experience on each platform. The experiment found that people tended to grade actual driving as more realistic than driving in the simulator and more stressful, possibly due to an actual sense of safety and responsibility in actual driving.

Thus, although on-road studies are eventually unavoidable and important for evaluating (automated) driving performance and drivers' behaviour, they impose safety challenges and risks and are harder to control. Hence, driving simulators are commonly used as a surrogate platform, at least at the early stages of system development. However, the simulator's ability to provide drivers with a sense of real-world driving is still ambiguous (Kemeny & Panerai, 2003; De Winter et al., 2012). This study aimed to answer the following research question: Can we find a new approach to studying human driving that is affordable and safe but still reflects, to some extent, the physical characteristics of actual driving?

To answer this question, we propose a new approach for experimenting with human driving that is affordable and safe but still provides the physical effects of actual driving. The proposed approach is based on a remote driving testbed consisting of a small-scale car-like mobile platform (1/10th model of a real car), a small road, and a remote driving station. In this approach, the remote driver sees a video taken by a camera installed in the moving car, providing drivers with a driving environment that resembles the driving experience of a real car (Le et al., 2021).

This work has two goals: (1) To develop a driving testbed that reflects physical driving but in a safe and controlled lab-like environment, and (2) To compare a driving

simulator platform (CARLA; Dosovitskiy et al., 2017), and a self-developed testbed across various measures of performance and perceptions.

Hypotheses

H1: Drivers would need more training on the testbed than on the Carla simulator to feel ready to begin the main experiment.

As De Winter et al. (2012) specify, drivers in a driving simulator may lack the feeling of responsibility and risk associated with their driving actions. Thus, we speculate that participants, regardless of their driving experience, would need more training on the testbed than the driving simulator.

H2: The driving speed and steering command variations will be higher in the driving simulator than in the testbed.

For the same reasons mentioned above, we speculated that drivers in the testbed will be more risk averse than in the virtual driving simulator and will thus be more cautious when driving it. This cautiousness will be reflected in lower speed and steering command variations in the testbed. However, there is a difference in the scope of speed in the platforms because the remote-driving car is a small-scale version of a real vehicle. Therefore, it has a lower speed range than a simulated real vehicle. To normalize the speed and steering variations on the same scale, we divided the standard deviations of each platform by the average speed. This measure is also known as the coefficient of variation (CV), a relative standard deviation, so the units are percentages in both platforms.

H3: Participants will have a better sense of realism in the driving testbed than in the simulator.

Participants using the testbed experience physical interaction with the actual world and are fed real-time video streams from the car's onboard camera. Then we postulated that the testbed would be perceived as more realistic than the CARLA driving simulator despite running on a high-end graphical engine, Unreal 4.

The participants were asked, "Did you feel that the driving situations were realistic compared to the sort of situations you usually encounter on the road?" when he/she finished a driving session (testbed or CARLA simulator) with a 5-point Likert scale answer.

H4: Participants will report a higher level of mental workload when experiencing the testbed environment.

Free-risk aversion could ease the mental workload of the participants while driving in the simulator. The mean of the NASA-TLX questionnaire serves as a measure of the mental workload.

The experiment

Participants

Fifteen participants, 10 males, and 5 females, participated as unpaid volunteers. The participants' range of age was 23-35 (*M*=27.6, *SD*=3.5). All participants reported having normal or corrected-to-normal visual acuity (6/9 or better). The participants reported their level of driving experience (LOE) on a 5-point Likert scale from no experience (1) to highly experienced (5). The Industrial Engineering and Management Dept. IRB at BGU ethically approved the study.

Experimental Apparatuses

The testbed consists of a driving station (Logitech Gaming G29 steering wheel and a curved 27" LCD; Figure 1 right image) and a small-scale vehicle, remotely driven based on visual information from an onboard camera presented to the driver in real-time. A non-linear algorithm calculates the expected resisting torque in the steering wheel and applies torque feedback to the driver to enhance the realistic driving experience (see Le et al., 2021 for technical details). Figure 1 (left) presents the view from the vehicle's onboard camera. A short video clip demonstrating the video that the participant sees at the remote station can be seen at:
https://www.youtube.com/watch?v=mQXl7B_W9cU&t=4s.
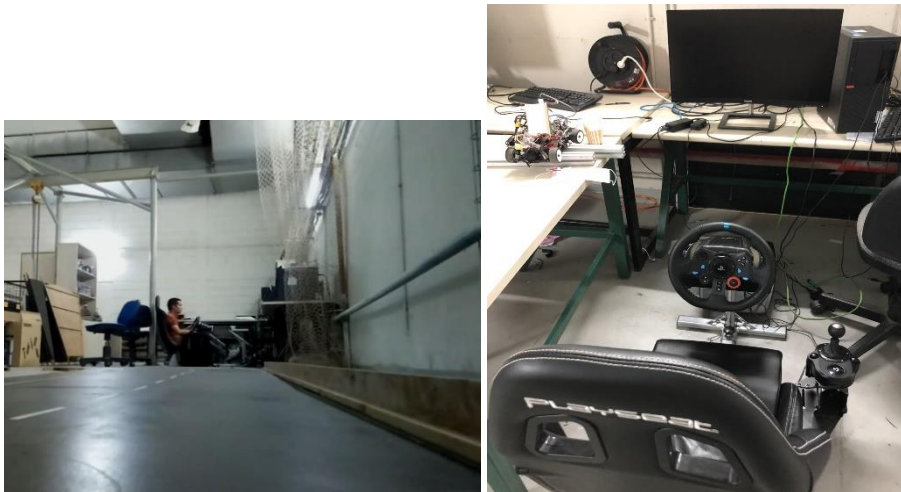
Figure 1. Testbed's onboard camera view (left) and experimental driving station (right).

The participants driving the CARLA driving simulator utilized the same driving station as the testbed. The driving interface of CARLA can be found at: https://www.youtube.com/watch?v=5SZKsZMiTWE. For further information on CARLA, please refer to Dosovitskiy et al. (2017). Figures 2-3 present CARLA's driving interface as well as the map of the route in CARLA and the map of the route at the lab (testbed).

Figure 2. CARLA's driving interface.

Figure 3. Driving map: CARLA (left) and the testbed (right).

Questionnaires

This study used three questionnaires. The first is a demographics questionnaire that included the following details: age, gender, self-declared driving experience, driving side in origin country (in case participants are from Common Wealth's countries), and any knowledge of teleoperation. This questionnaire was administered before training and driving.

The second questionnaire is the NASA-TLX to assess participants' mental workload during their driving task on each platform. The average value of these six questions (about mental, physical, temporal, task performance, effort, and frustration) represents the task load: higher means the task load is more burdensome.

The third questionnaire was the Psychological Feeling Questionnaire consisting of the following self-evaluation dimensions: the sensation of realism, environment's reactivity, presence, stress level, and vehicle/activities mastery. Milleville-Pennel and

Charron (2015) utilized this questionnaire to compare between driving a simulator and driving a vehicle in the real world. The question number is treated as an independent variable with seven levels (1 to 7).

Experimental design

This study includes a 2x2 within-subjects design. Each participant experienced driving in the two platforms in two consecutive sessions (counterbalanced). Each driving session in each platform included two phases: training and testing.

The testing phase starts when a participant is "ready" after the training phase. After each session, participants completed the NASA-TLX and psychological feelings questionnaires. During each session, the vehicle dynamics were recorded, including the steering command, acceleration, and velocity. The total elapsed times (only when the vehicle was moving) in both sessions (including "restarts" if the participant had to do so).

Procedure

Regardless of the platform, participants were asked to drive as they would normally drive in similar real-world situations. When a participant finished the first session and filled out the relevant questionnaires, he/she was given the instructions for the second session (different platform). At the end of the second session, participants filled out the relevant questionnaires for the second time, and the experiment ended.

The observation from the pilot study showed that some experienced drivers misused the CARLA driving simulator and drove freely in the environment without obeying the study's instructions to drive normally. They occasionally ran into objects to see what a crash looked like since there is no penalty/reward for driving performance, this curiosity could have occurred during the experiment and jeopardized its validity. To avoid this curiosity and unwanted driving behaviour, before the experiment began, they were told that if they wished to experience the simulated environment <u>but not during the experiment</u>, they were welcome to come to the lab at a different time and drive freely. This procedure assured that during the experiment, participants would follow the instructions. Two participants (one male and one female with a maximum level of experience (5) returned to the lab after the experiment and drove freely for half an hour.

Results

We used the generalized linear mixed models (GLMM) framework to perform the statistical analyses (SPSS version 28). The independent variables that were included in the initial models in all analyses as fixed effects were the self-declared level of experience (1-no experience to 5-highly experienced), gender, age, platform (testbed or simulator), order (drive on the simulator first and the testbed second or vice versa). Participants were included in all models as a random effect. In addition, the two-way interactions between order and all other main effects were also included. The statistical analyses were conducted at a significance level of 0.05. Non-significant interaction effects were removed from the model using the backward elimination

procedure. Post hoc pairwise contrast analyses were conducted whenever necessary, and the Bonferroni procedure was applied to correct alpha for multiple comparisons.

For the dependent variable of driving duration (both training and testing sessions), a natural logarithm was applied before the analysis. Then, we applied a linear regression model within the GLMM framework. The estimated means of training and testing durations are displayed in the time domain after re-conversion to original values.

For modelling the dependent variable CVs of speed, a $\text{Logit}(\ln\left(\frac{CV}{1-CV}\right))$ was applied (i.e., CV) before applying the Linear regression within the GLMM framework. For modelling the dependent variable CVs of steering, a natural logarithm was applied to the steering CV values before applying the linear-regression within the GLMM framework.

Table 1. Summary of Statistical Analyses

Dependent variable	*Significant fixed effects in the final models and estimated means (EM)*
Training phase duration (s)	Platform* LOE [$F_{3,16}$=22.5, p=.01]: LOE 1: simulator (EM=55s) < testbed (EM=102s); LOE 5: simulator (EM=329.6s) > testbed (EM=63.5s) (Figure 4).
CV of speed (%)	Platform [$F_{1,19}$=10.7, p=.004]: simulator (EM=35%) > testbed (EM=27%). Age [$F_{1,19}$=16.3, p < .01]: older people, higher CV of speed (β=0.09).
Testing phase duration (s)	LOE [$F_{3,8}$=6.7, p=.01]: LOE 5 (EM=25.3s) < LOE 1 (EM=70.0s), 2 (EM=64.3s), 4 (EM=75.1s). Gender [$F_{1,8}$=7.5, p=.025]: Male (EM=37s) < Female (EM=81.4s). Age ($F_{1,8}$=4.3, p=.07): older people, longer testing duration (β=0.079).
CV of steering (%)	Platform [$F_{1,22}$=314.7, p <.001]: simulator (EM=180%) > testbed (EM=60%). Age [$F_{1,22}$=5.8, p=.025]: older people, smaller CV of steering (β= -0.03). LOE [$F_{3,22}$=5.1, p=.008]: LOE 5 (EM=88%) <LOE 1 (EM=128%).
Psychological questionnaire	Interaction between platform type and question number [$F_{7,195}$=2.2, p=.034]: Question 1* (p_{adj}=.09): (simulator (EM=3.4) < testbed (EM=4)) marginally significant. Question 3* (p_{adj}=.02): (simulator (EM=3) < testbed (EM=4)).
NASA TLX	Platform [$F_{1,172}$=9.9, p=.002]: simulator (EM=3.2) < testbed (EM=4.2).

*Note. Q1: Did you feel that the driving situations were realistic when compared with the sort of situations you usually encounter on the road? Q3: How would you rate the feeling that you were «there» in the driving environment?

The psychological questionnaire included seven self-evaluation questions, and each participant responded on a 5 points Likert scale for each question. We applied a linear regression within the generalized linear models (GLM) framework with question numbers as repeated measures (participants were included as a random effect). For the NASA TLX questionnaire, we applied a linear regression within the GLM

framework with question numbers as repeated measures (participants were included as a random effect).

Finally, Table 1 includes a summary of all statistical analyses. We did not include order-related significant effects in Table 1 as they are less relevant to the hypotheses. The order variable was included to control the experimental design only.
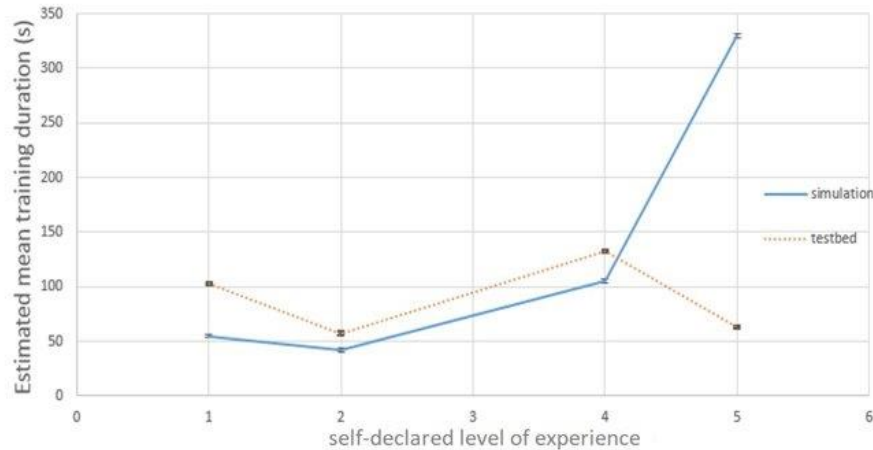
Figure 4. A significant interaction between LOE and the platform.

Discussion

This work has two goals: (1) To develop a driving testbed that reflects physical driving but in a safe and controlled lab-like environment, and (2) To compare a driving simulator platform (CARLA; Dosovitskiy et al., 2017), and a self-developed remote-driving testbed across various measures of performance and perceptions. The first goal was achieved and reported elsewhere (Le et al., 2021). To address the second goal, we generated four hypotheses. The discussion section is organized around the hypotheses, and each hypothesis is discussed in turn.

The evidence of responsibility/ risk aversion

As H1 is accepted only for LOE 1 (and marginally significant for LOE 2 and LOE 4), the participants with little driving experience needed more time to prepare for the experiment at the testbed than at the simulator. Since these participants had similar self-declared LOE, the responsibility/risk aversion can be seen as one of the factors affecting their prolonged training time at the testbed, which exposed more physical risks to them during the training session compared to the simulator, especially since they do not have much actual driving experience.

The rejection of H1 at LOE 5 (post hoc analysis) can imply the participants are sufficiently experienced in driving and ready to start the actual drive with a short training duration while they were enjoying driving in the simulator without any risk, which prolonged their training time at the simulator (note that we did not give any

reminder regarding the training purpose, or to stop them from their driving passion). Consequently, there is an alternative factor that can manipulate the training duration.

H2 was also confirmed, showing that the CV of steering commands and speed were higher in the driving simulator than in the testbed. The confirmation of H2 further supports the argument of the differences in risk aversion generated by the two platforms when the participants "speeded" and "steered" more at the simulator than the testbed. Regarding the steering behaviour differences between the two platforms may also be attributed to the application of a novel torque feedback algorithm implemented in the testbed (Le et al., 2021) but not in the driving simulator, where there was only a constant torque feedback.

The negative correlation between the LOE and the amount of time it took participants to complete the testing phase, which had a fixed length, when driving in the testbed but not in the driving simulator (somehow is the absolute speed), may also imply that the risk perception at the testbed is varied according to the LOE (drivers with less experience drove much slower) while there is no, or very little, perceived risk at the simulator where all LOEs finished the test driving at about the same amount of time (drivers chose speeds regardless of driving experience).

Sense of Realism and Presence based on the psychological questionnaire

In the third hypothesis, it was speculated that participants would have a better sense of realism in the driving testbed than in the CARLA driving simulator. Our study found that the sense of realism was significantly (marginal) higher in the testbed than in CARLA and that the feedback provided by the environment was more realistic in the testbed than in CARLA. These results partially confirm our hypothesis. Comparing our findings with those of Milleville-Pennel and Charron (2015), who used a real vehicle, we found weaker relations between realism and actual driving (question 1). These differences can be a result of several reasons. (1) the statistical approach differences (Milleville applied a t-test for every pair), (2) the perceptual difference between the simulator and the testbed is smaller than that of a simulator and a driving school's real vehicle and the driving station. In other words, Milleville-Pennel and Charron used two separate driving stations (a simulation room and the school's vehicle), while in our experiment, the testbed and simulator shared the same driving station, potentially eliminating any perceptual divergence regarding the driving station.

Nevertheless, our findings are partially consistent with a similar study that used a real vehicle (Milleville-Pennel & Charron, 2015) and are thus promising. We believe that increasing the sample size would have resulted in more statistical power.

We found the perceived presence was higher at the testbed than at the simulator, while Milleville did not see a difference in presence between the driving schools' real vehicle and the simulator (i.e., question 3). This, again, can be a result of the fact that in our study, we used the same driving station for both platforms, which might have affected the interpretation of the term "there" in Q3 to the graphical display instead of the whole system as in Milleville's experiment.

In Q2, we asked the participants, "How would you rate the reactivity of the environment in response to your actions." The difference in perceived environment's reactivity (Q2) could not be found in our experiment, meaning the participants did not report any significant difference between the platforms in that sense, while it was significant in Milleville-Pennel and Charron's (2015). We speculate that we did not find similar differences due to either the small sample size or because the remote vehicle's electric motor reacted differently than the engine of actual vehicles. This issue should be further explored in future studies.

Crash - The source of the higher workload at the testbed

The crash (or the aversion of crash) that could damage the remote driving vehicle and require a long time for repair, is potentially the source of risk participants associated with the testbed. Note that when a participant crashes in the testbed, it not only costs the participants more time (to complete the experiment), but it also causes them to feel guilty, as most of them shared personally after the experiment. Driving in the testbed becomes challenging (or higher workload) as H4 is validated, particularly for the inexperienced drivers who crashed during the training.

Conclusion

The concern of crashes was found to be a natural source of risk in the testbed, a conclusion supported by four validated hypotheses, H1, H2, H3, and H4. The validation of H3 or the realism is perceived higher at the testbed than the virtual simulator, which runs on a high-end graphical engine (Unreal 4) as CARLA, which is consistent with the findings of Ranney (2011), who predicted that - increased fidelity of the driving experience does not address the problem of poor psychological fidelity, but the perception of risk is the genuine source of realism.

The testbed brings advantages over the simulator and is more economical than (real) vehicle-based experiments. It is better than the simulator in the aspects of physical and psychological fidelity, and it allows more realistic experiments of mixed traffic situations. Nevertheless, it is less flexible than the simulator, and it is harder to keep similar driving conditions for different experiments as a driving simulator allows. Adding more objects (e.g., other vehicles, pedestrians, etc.) or objective measurement that can be done easily in the simulator, becomes more challenging (and costly) in the testbed. However, on the good side, most of the critical safety issues related to experiments with real vehicles disappear while a genuine risk associated with one's actions is maintained. The results of this study indicate that the developed testbed has the potential to serve as an alternative driving simulator, e.g., to test the human factor as well as the performance of autonomous vehicles in mixed traffic scenarios.

Acknowledgments

References

De Winter, J., van Leeuwen, P. M., & Happee, R. (2012, August). Advantages and disadvantages of driving simulators: A discussion. *In Proceedings of measuring behavior* (Vol. 2012, p. 8th).

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017, October). CARLA: An open urban driving simulator. In *Conference on robot learning* (pp. 1-16). PMLR.

Kemeny, A., & Panerai, F. (2003). Evaluating perception in driving simulation experiments. *Trends in cognitive sciences, 7*, 31-37.

Le, Q. S., Arogeti, S., & Borowsky, A. (2021). Remote driving testbed with force feedback based on slip angle estimation. In *2021 European Conference on Mobile Robots (ECMR)* (pp. 1-8). IEEE.

Milleville-Pennel, I., & Charron, C. (2015). Driving for real or on a fixed-base simulator: is it so different? An explorative study. *Presence, 24*, 74-79.

Punzo, V., & Ciuffo, B. (2010). Integration of driving and traffic simulation: Issues and first solutions. *IEEE transactions on intelligent transportation systems, 12*, 354-363.

Ranney, T.A. (2011). Psychological fidelity: Perception of risk. In *Proceedings of the Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL: CRC Press

Shabalina, K., Sagitov, A., Su, K. L., Hsia, K.H., & Magid, E. (2019). Avrora unior car-like robot in gazebo environment. In *International Conference on artificial life and robotics*, 116-119.

Taheri, S.M., Matsushita, K., & Sasaki, M. (2017). Virtual reality driving simulation for measuring driver behavior and characteristics. *Journal of transportation technologies, 7*, 123-132.

Promoting sustainable mobility by behaviour-based design: A pilot study

Giorgia Tempestini & Francesco Di Nocera
Sapienza University of Rome, Italy

Abstract

The use of private transportation has negatively impacted the environment and the citizens' quality of life, health, and sociality. Although it is impossible to isolate a single behaviour to be considered as the cause of such a broad and complex outcome, several behaviours have contributed to the emergence of negative outcomes. That is why it is important to analyse behavioural patterns to modify problem behaviours in favour of others that are more beneficial. The present research aims to investigate whether choices can be guided by adopting a behavioural strategy. Specifically, avoiding unexpected delays (negative reinforcement) could increase the adoption of sustainable mobility behaviour. This goal has been pursued by designing a board game that participants were asked to play in a small group. The purpose of the game is to reach destinations in the shortest possible time. Participants choose means of transportation among driving, public transport, walking or cycling. During the game, sustainable behaviour is negatively reinforced with a variable ratio (VR5) reinforcement schedule. Results show that the use of negative reinforcement is a viable strategy to encourage people to make more sustainable choices: cycling has been adopted more frequently and this result could have important implications for designing navigation apps delivering this type of information.

Introduction

Human behaviour is causing the deterioration of the environment, including climate change and the loss of global biodiversity. Although it is not possible to isolate a single behaviour to be considered as the cause of such a large and complex outcome, there are several behaviours that have contributed to changing environmental conditions. By considering the negative effects of human behaviour, researchers have focused on analysing certain behaviours that should be implemented for improved environmental conditions. It is shared in the literature the term "pro-environmental behaviours" to mean all those behaviours that generate environmental benefits (e.g., recycling) and the avoidance of those behaviours that harm it (e.g., avoiding air travel) (Lange & Dewitte, 2019). Pro-environmental behaviours are defined precisely by the consequences they have on the environment itself.

Thus, to reduce these problems, it is necessary to understand what aspects discriminate between individuals' choices to consume or conserve resources, decisions to dispose of waste properly, promote environmental causes, or make sustainable mobility choices.

Increasingly congested and polluted cities have made urban mobility one of the most critical factors for cities and citizens' well-being. Managing people's mobility has become one of the most difficult challenges for the environment, making sustainability a determining factor for the future of mobility.

Mobility is responsible for negative consequences at different levels, impacting not only the environment, for example, through greenhouse gas emissions, but also, more generally, citizens' quality of life and their health and sociality (Khreis et al., 2017). Urban dwellers are constantly exposed to high levels of air and noise pollution, injuries, or deaths from traffic accidents (Nieuwenhuijsen & Khreis, 2019), and negative health impacts from a lack of green spaces, replaced by urban heat islands (Gago et al., 2013). Although the concept of healthy cities has been included within policy agendas since 1988, and the European Healthy Cities Agenda 2014-2018 prioritises the transportation sector, there is a need to address the issue to mitigate the negative impacts caused by this sector.

Considering the limits set by the WHO, Rome exceeds the maximum values with very high percentages. Rome exceeded with 65% of the maximum value for Particulate Matter (PM) = 10, 123% for PM = 2.5, and 187% for Nitrogen Dioxide (NO_2). For this reason, this research project could be beneficial when we consider a city as busy and pollution-intensive as Rome.

Gamification for sustainable mobility

Gamification is frequently invoked as an effective way to influence people to achieve goals. Gamification refers to using a game-like approach to achieve learning goals in non-game contexts (Deterding et al., 2011). The principle behind this methodology is to use the dynamics and mechanics of gaming, such as accumulating points, achieving levels, obtaining rewards, and displaying badges. In the literature, it is possible to find several studies that have specifically investigated the positive relationship between gamification/gaming and mobility (e.g., Kazhamiakin et al, 2005; Merugu et al., 2009).

Gamification has been frequently used to promote sustainable behaviours. Following the principles that characterise this methodology, several products have been implemented, including team competitions, digital games, smartphone apps, data collection apps, and board games (see Douglas & Brauer, 2021 for reviews on games and apps for sustainability).

Merugu and collaborators (2009), for example, used gamification with the goal of reducing commute times and increasing commuter travel comfort, while also attempting to reduce congestion, fuel consumption and pollution. To incentivize commuters to travel at times with less road congestion, an incentive program (INSTANT - Infosys-Stanford Traffic) was devised in which credits accumulated by each commuter based on the time of arrival at their destination were then exchanged for a monetary backup reinforcer. The more credit a commuter had, the higher the reward amount he or she could win and the higher the chance of winning a reward. The data showed an increase in the number of commuters arriving at the recommended times to avoid road congestion (almost double for each time slot),

showing the effectiveness of the incentive program designed by the researchers. Another example of gamification to implement sustainable mobility behaviours is that of the "Mordor Shaper App" (Olszewski & Turek, 2020), in which participants, divided into groups, had to achieve goals assigned to them within an augmented reality. The participants' main goal was to transform the Mordor neighbourhood into an inhabitant-friendly Smart City.

The two examples above are helpful in understanding how applicable the methodology of gamification can be to implement target behaviours to promote all those behaviours that can improve environmental, economic and social conditions. As reported by Di Nocera and Tempestini (2022), long before the term gamification made its way into the literature, this dynamic was well known (and still it is) under the name "token economy" and "it is important to emphasise that the complexity of the model underlying behaviour analysis (or functional analysis) requires conceptual, terminological, and methodological rigor" (p. 251).

Although digital products and services to implement desirable behaviours are becoming common, they rarely represent the output of a documented research activity and are limited to the common sense implementation of badges or points that can be eventually exchanged with rewards (e.g., credit, discounts). Research is often limited to testing the potential efficacy of such solutions (e.g., Olszewski & Turek, 2020; Lee et al, 2020).

Pilot Study

Purpose of the current study

The pilot study reported here is a first step of a project grounded in the research area of Behavioural Design (although we prefer to use the expression Behaviour-Based Design), which is concerned with studying behaviours to modify them and make them desirable for the individual, the environment, and society (Khadilkar & Cash, 2019; 2020). This study aims to test if it is possible to guide sustainable choices through behavioural intervention. The hypothesis is that introducing aversive stimulation (i.e., delay) associated with unsustainable means of transportation will act as negative reinforcement when alternative and more sustainable means of transport are chosen. To achieve the research goal, a board game was created to observe how people made decisions related to urban mobility.

Material and Methods

Participants

Five individuals (mean age = 28; st.dev.= 3.94; all females) participated in this pilot study. The only inclusion criterion for participation in the experiment concerned the use of a car as preferred means for commuting. Participants had a driving license since more than 6 years (mean = 9.40; st.dev. = 4.33). All participants authorized the treatment of the data collected for research purposes.

Material

Routes. One-hundred routes in the city of Rome with the same starting point and a variable ending point were created (Figure 1). The routes had similar driving time. Times for other means of transportation were derived by randomly adding variable minutes to the initial driving time (Table 1).

Table 1. Mean and Standard Deviation of route times for each means of transportation.

	Driving	**Public transport**	**Walking**	**Cycling**
Time - Mean	17.44	22.19	42.13	19.47
Time - St. Dev	3.816	4.56	8.69	4.10
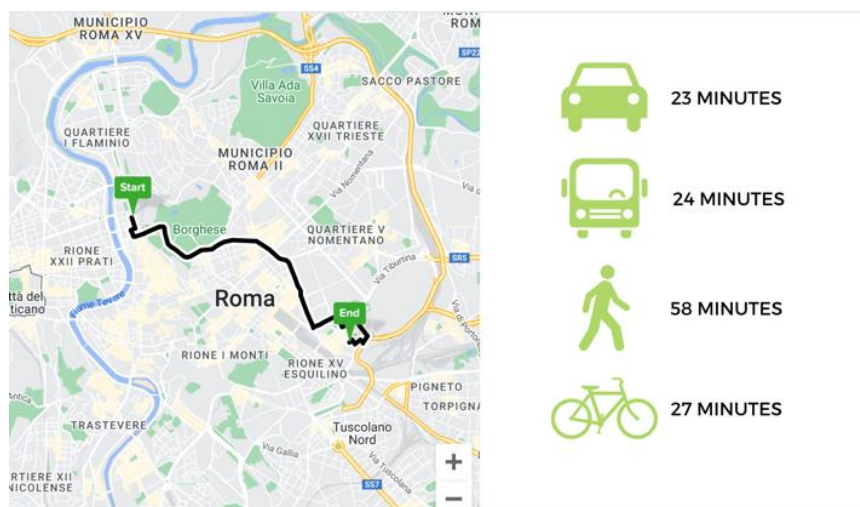
Figure 1. Sample slide with route and times for different means of transportation (driving, public transport, walking, cycling).

Scenario Cards. Identical stacks of 100 "Scenario Cards" were created (Figure 2). A stack was assigned to each subject. The cards could be neutral (no unexpected event) or "aversive". In the latter case the card reported an unexpected event adding an additional travel time ranging from +15 to +30 minutes. The position of the aversive cards into the stack was determined as a variable ratio schedule (VR5) with presentation of the aversive card in variable order (five on average). Aversive stimulation was therefore presented in 25% of the cases.

Individual Reporting Sheet. The individual reporting sheet is designed to collect information about choices (driving, public transport, walking, cycling) and route time (for each round and cumulative). Participants filled the sheet autonomously to increase the engagement.
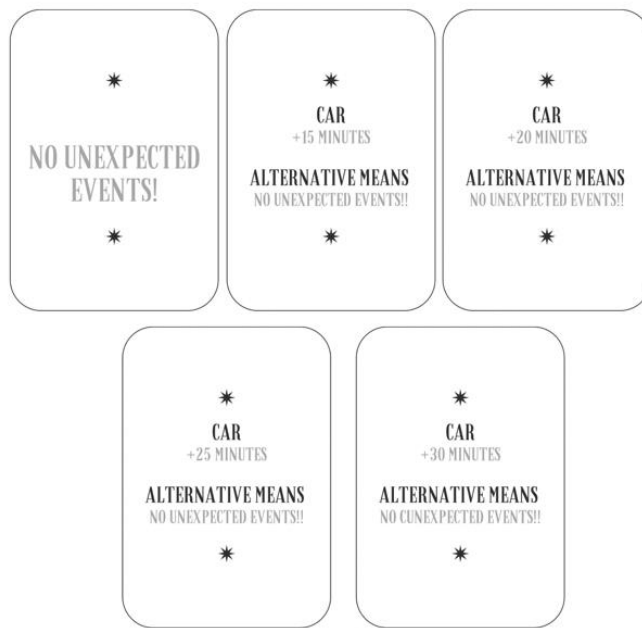
Figure 2. Type of cards included into the stack.

Procedure

Participants were asked to play a board game whose scope was to reach 100 destinations (one for each round) in the metropolitan area of Rome in the shortest time. Each round started always from the same point and only the destination was different. To get from point A to point B, the participant had to choose between four means: driving, public transportation, walking, and cycling. The winner of the round is who reaches the destination in the shortest time, the winner of the game is who reaches all the destinations in the shortest (cumulated) time.

For each round of the game, the scenario was projected on a screen, and after choosing the mean, each participant collected a "Scenario card" in which the presence or absence of an unexpected event was indicated, resulting in a potential increase in time if driving was chosen. Therefore, sustainable behaviour was negatively reinforced (by avoidance of the less advantageous driving choice).

Data Inspection

During the game, driving was selected in 25% of the rounds, public transport in 21% of the rounds, and cycling in 54% of the rounds. Walking was never chosen as a means for accomplishing the task. Subjects did not show a stereotyped choice but rather changed the means of transportation according to the dynamics of the game (about 60% of the time). On average, the sustainable behaviour was negatively reinforced 21 times and the driving was punished 4 times (over 25 occasions). Average number of

each selected choice and average route duration were computed for bins of ten rounds. Figure 3 shows that two different processes are most probably taking place. During the first half of the game subjects were clearly attempting to reduce the route time as demonstrated by the linear decrease in Figure 3, whereas in the second half of the game they changed their strategy. This reflected in the choice of the transport type (Figure 4): in the first half of the game cycling was chosen more frequently for avoiding both driving occasional delays (aversive stimulation) and systematic longer route times of the public transport (and for this reason its contribution will not be taken into consideration in the following inspection).
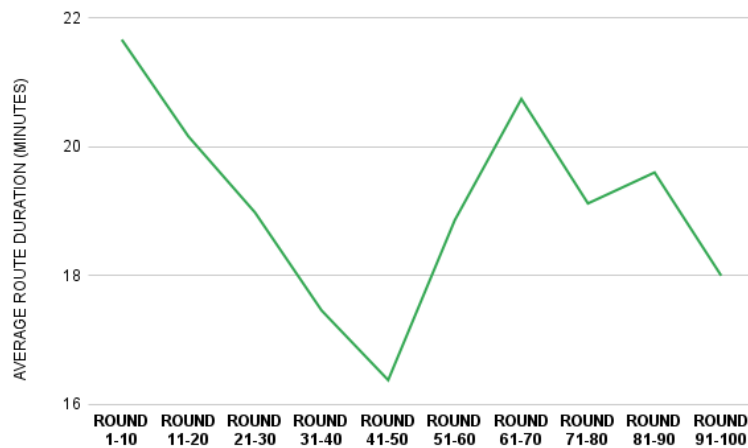

Figure 3. Average route duration (all rounds).
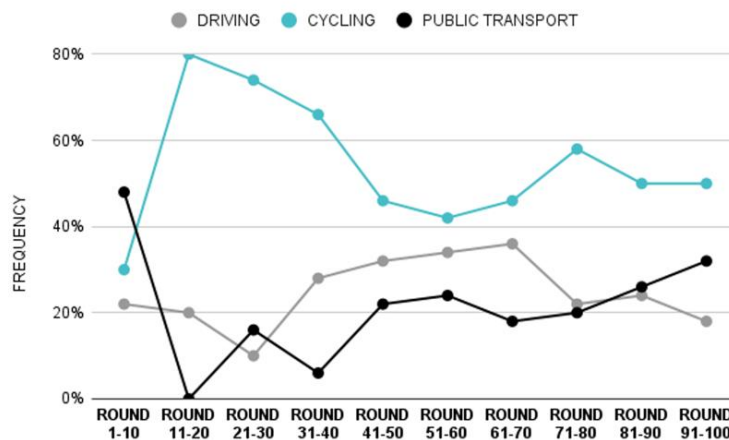
Figure 4. Choice frequency (all rounds).

Only in the second half of the game the behaviour of the subjects stabilized, and a more realistic pattern of choice became visible (Figure 5). The inspection of the

second half of the game only, showed that on average the sustainable behaviour was negatively reinforced in 80% of the cases. That reflects in a clear increment of the cycling choice with respect to driving.
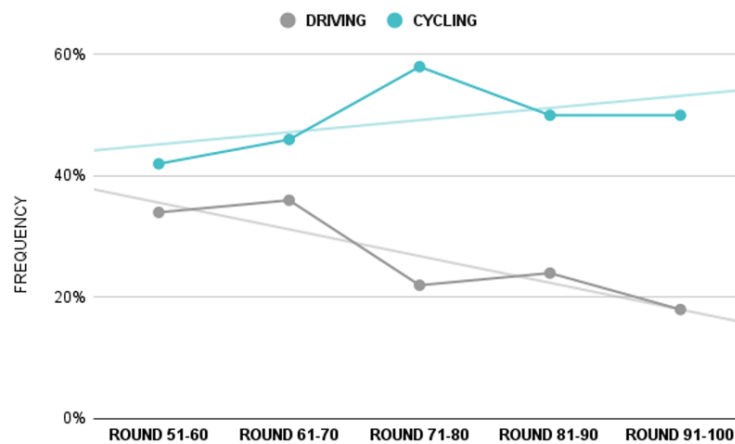
Figure 5. Choice frequency (second half of the game).

Discussion

Human behaviour is the most relevant factor to consider when talking about sustainability. Every behaviour we engage in affects our environment: littering, cooking, moving both within our own city and across different nations. Human beings have a great responsibility to the environment in which they live. It is difficult, however, to steer these behaviours toward sustainability, especially if we cannot isolate the variables that influence these behaviours. The present pilot study examined trends in the behaviours of a limited number of subjects who share a tendency to use a private vehicle for city commuting, manipulating the time variable of reaching the destination and considering in a plausible way the unforeseen events that may arise (e.g., traffic, time for finding a parking spot). The objective was to investigate whether avoiding the unexpected delay (negative reinforcement) could increase the adoption of sustainable mobility behaviour. To achieve this goal, five participants were asked to participate in a board game in which they had to choose the means of transportation to reach point B from point A, with the goal of completing all tasks in the shortest possible time relative to others.

The results showed that the use of negative reinforcement is a viable strategy for encouraging people to make sustainable choices. After an initial phase in which subjects attempted to guess the dynamics of the game (and, likely, the goal of the experiment), the second half of the game is characterized by a stabilization of choices that shows a symmetrical pattern for cycling and driving.

This pilot study was a necessary step for addressing the use of a behavioural intervention in this specific domain. Moreover, choosing to use an experimental approach based on a board game allowed to easily implement a behavioural approach,

which avoids self-report measures that may be unreliable when subjects are asked to answer questions about environmental behaviours, hypothetical scenarios, or intentions (Lange, 2022). The main issues are those involving the subject's inability to accurately recall the frequency of certain behaviours, to realize which behaviours are considered undesirable (Gifford, 2014), and, finally, social desirability bias (Vesely & Klöckner, 2020).

The approach used here is not without its limitations. First, one single group of players is not enough, and a real experiment should rely on many sessions with several groups. Second, the group here was limited because of the known differences between males and females in spatial cognition (Havet et al., 2021; Nazareth et al., 2019). Homogeneity was a priority in this pilot study, but heterogeneity will be a priority in the future. Third, the board game might be considered a microworld to test specific hypotheses, but it cannot compare with the complexity of a similar choice in the real world. People decide to get somewhere using a car instead of a more sustainable means of transportation for several reasons, including the logistics of their daily lives (e.g., taking children to school after work or doing groceries). That is, time to get there is not the only type of reinforcer.

Another critical limitation concerns the absence of a baseline to understand the behaviour in the absence of reinforcement. However, that is something we cannot control: the game's goal is to reach destinations in a shorter time than others, so time is a crucial factor. Route duration associated with driving was always less than that of all other means of transportation. Therefore, the theoretical baseline is 100% for driving and 0% for the other options.

In general, future studies should add more complexity to the game. For example, specifying the nature of the delay (e.g., traffic, accidents, strikes, parking). The choice of one means over another can be influenced by many factors, such as habit or logistical reasons. Considering the hectic nature of city life, we often do not stop short of reasoning about the consequences on the environment and society of our mobility choices. The consequences of our choices on the environment are too long-term for them to change our decisions for sustainability. A delay in reaching a goal, on the other hand, represents a more immediate consequence that impacts the daily lives of every individual. For this reason, considering the results of this study, it is helpful to consider reporting unexpected events that may arise when making decisions.

References

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining "gamification". *In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15). Publisher? Website?

Di Nocera, F. & Tempestini, G. (2022). Getting Rid of the Usability/Security Trade-Off: A Behavioral Approach. *Journal of Cybersecurity and Privacy*, *2*, 245-256.

Douglas, B.D. & Brauer, M. (2021). Gamification to prevent climate change: A review of games and apps for sustainability. *Current opinion in psychology*, *42*, 89-94.

Gago, E.J., Roldan, J., Pacheco-Torres, R., & Ordóñez, J. (2013). The city and urban heat islands: A review of strategies to mitigate adverse effects. *Renewable and sustainable energy reviews*, *25*, 749-758.

Gifford, R. (2014). Environmental psychology matters. *Annual Review of Psychology*, *65*, 541–579.

Hamari, J., Koivisto, J., & Sarsa, H. (2014,). Does gamification work?--a literature review of empirical studies on gamification. *In 2014 47th Hawaii international conference on system sciences* (pp. 3025-3034). IEEE. Location/publisher

Havet, N., Bayart, C., & Bonnel, P. (2021). Why do gender differences in daily mobility behaviours persist among workers? *Transportation research part A: policy and practice*, *145*, 34-48.

Kazhamiakin, R., Marconi, A., Perillo, M., Pistore, M., Valetto, G., Piras, L., & Perri, N. (2015, October). Using gamification to incentivize sustainable urban mobility. In 2015 IEEE first international smart cities conference (ISC2) (pp. 1-6). IEEE.

Khadilkar, P.R. & Cash, P. (2019). Understanding behavioural design: Integrating process and cognitive perspectives. *In Proceedings of the Design Society: International Conference on Engineering Design* (Vol. 1, No. 1, pp. 1863-1872). Cambridge University Press.

Khadilkar, P.R. & Cash, P. (2020). Understanding behavioural design: barriers and enablers. *Journal of Engineering Design*, *31*, 508-529.

Khreis, H., Kelly, C., Tate, J., Parslow, R., Lucas, K., & Nieuwenhuijsen, M. (2017). Exposure to traffic-related air pollution and risk of development of childhood asthma: a systematic review and meta-analysis. *Environment international*, *100*, 1-31.

Lange, F. (2022). Behavioral paradigms for studying pro-environmental behavior: A systematic review. *Behavior Research Methods*, ISSUE?, 1-23.

Lange, F. & Dewitte, S. (2019). Measuring pro-environmental behavior: Review and recommendations. *Journal of Environmental Psychology, 63*, 92-100.

Merugu, D., Prabhakar, B.S., & Rama, N. (2009, July). An incentive mechanism for decongesting the roads: A pilot program in Bangalore. *In Proceedings of ACM NetEcon workshop. Webaddress?*

Nazareth, A., Huang, X., Voyer, D., & Newcombe, N. (2019). A meta-analysis of sex differences in human navigation skills. *Psychonomic bulletin & review*, *26*, 1503-1528.

Nieuwenhuijsen, M. & Khreis, H. (2019). Integrating human health into the urban development and transport planning agenda: a summary and final conclusion (pp. 707-718). Springer International Publishing.

Olszewski, R. & Turek, A. (2020). The mordor shaper—the warsaw participatory experiment using gamification. *In ICT Systems and Sustainability: Proceedings of ICT4SD 2019*, Volume 1 (pp. 255-264). Singapore: Springer Singapore.

Vesely, S. & Klöckner, C.A. (2020). Social desirability in environmental psychology research: Three meta-analyses. *Frontiers in Psychology*, *11*, 1395.

Investigation of anthropomorphic system design features for sense of agency in automation technologies

Tina Frenzel & Josephine Halama,
Chemnitz University of Technology,
Germany

Abstract

As a substantial part of conscious mind, sense of agency (SoA) describes the subjective experience of control over actions and sensory outcomes. Since the interaction with automation technologies changes the users' role from operator to observer and actions are rather caused by another agent than self-directed, the question arises how SoA can be maintained to prevent out-of-the-loop performance problems and encourage a positive user experience. The online experiment ($N = 52$) investigated how anthropomorphic features in system design affect users' SoA in automated processes. Therefore, participants observed other-directed actions and SoA was implicitly measured by temporal binding of action initiation and outcome in an interval estimation task. Manipulating anthropomorphic features (within-subjects) flanking the automated actions of a non-embodied agent, four conditions were compared: no cue vs. auditory cue 1 (computer-generated voice) vs. auditory cue 2 (human voice) vs. visual cue (human hand). Controlling for observers' affinity for technology, the ANCOVA indicated anthropomophic cues to support SoA ($\eta_p^2 = .06$), except for human-voice-based speech. In conclusion, these findings highlight the potential of appropriate anthropomorphic features in system design to promote SoA during interactions with automation technologies and, furthermore, identify directions for further research.

Introduction

The core human need for control, a psychological mechanism for creating safe environments, is also reflected in interactions with modern technologies. The extent to which we feel in control and perceive ourselves as agents, our *sense of agency* (SoA), can influence the willingness to continuously interact with a system and determine user experience (Berberian, 2019; Limerick et al., 2014). In particular, when using automated systems, such as self-driving cars, in which humans act as controllers, balance between assistance and SoA needs to be considered in order to avoid out-of-the-loop performance problems and, thus, errors and accidents (Endsley, 2016; Parasuraman et al., 2000).

In the case of automated driving, as the systems' autonomy increases, so does the risk of drivers no longer perceiving themselves as agents. With increasing levels of automation, drivers have to manage fewer active tasks (e.g. steering, braking) such as

those which they were engaged in at lower levels of automation, and, hence, exert less influence on the situation. On the one hand, this loss of control can cause unpleasant feelings (e.g. helplessness, frustration, discomfort). On the other hand, it becomes problematic for safety when emergency situations occur in which human takeover is required, for example, because the technical system fails or cannot cope with unforeseen situational demands (e.g. Berberian et al., 2012; Zanatto et al., 2021). For drivers to still be able to react appropriately and efficiently in these takeover situations, an unimpaired SoA is required.

Taking into account a human-centred development of automated technologies, this paper aims to investigate to what extent anthropomorphic system features promote users' perceived control during interactions with autonomously-acting systems and, thus, examine whether this represents one possible effective solution to strengthen users' SoA in monitoring tasks.

Background

Sense of agency: Theoretical and methodological background

SoA refers to the subjective feeling of control over actions resulting from the perception of being the agent who is able to control events in the environment (e.g. Haggard & Chambon, 2012; Wen et al., 2019). According to the *cue integration concept*, which has received increasing scientific recognition in recent years, SoA emerges as part of a neurocognitive process. In said process, different cues, including internal sensorimotor signals and external situational information, are weighted based on their reliability and then integrated to derive valid inferences about the origin of actions and, thus, the causal attribution of action outcomes (Moore & Fletcher, 2012).

Phenomenologically, two components of SoA which require different types of measurement methods, can be distinguished from one another (e.g. Dewey & Knoblich, 2014; Haggard, 2017). Explicit measures using ratings capture the *judgement of agency*, as a conscious component of SoA, whereas implicit methods capture the *feeling of agency* as a preconscious facet. The majority of implicit methods are grounded in the basic assumptions of the forward model (Blakemore et al., 2001), which assumes that individuals develop mental motor action plans for intended actions from which predictions about sensory consequences can be made (Wolpert et al., 1995; Wolpert & Ghahramani, 2000). After executing the action, the actual sensory event is compared with the predicted one and, if it matches, is perceived as a consequence of one's own action, thus, leading to the perception of SoA (Blakemore et al., 2001; Wolpert & Ghahramani, 2000).

The most popular method for implicitly capturing SoA is the *Libet Clock paradigm*, in which subjects are instructed to observe a hand rotating on a clock only to then estimate at which position on the clock the hand was located when a particular outcome occurred (Libet et al., 1983). A viable and robust alternative to the Libet clock, which also relies on the perception of time of actions that produce a particular outcome, is the *interval estimation procedure* (e.g. Coyle et al., 2012; Humphreys & Buehner, 2009). This paradigm offers the advantage that the subjective time interval between action and outcome is directly assessed. This allows for a broad use in a

variety of experimental scenarios and tasks, including tasks involving visual features. The technique is particularly well-suited to experimental designs in which binding effects are to be measured across different conditions (Coyle et al., 2012). Both methods are based on so-called binding effects, i.e. the perceived time of action onset is moved closer to that of the sensory consequences during intentional actions. The time interval in between is, thus, perceived as shortened (Haggard et al., 2002). This temporal binding effects serve as indicators for greater SoA.

Sense of agency in observer situations

Automation technologies pose a challenge to the human experience of control due to the arising ambiguity of agency and responsibility for action outcomes. If SoA is not stabilised sufficiently, serious consequences can arise, especially regarding highly automated systems (e.g. automated vehicles, process monitoring in manufacturing automation), in which users have to take on monitoring tasks. Humphreys and Bühner (2009) showed by means of intentional binding that self-initiated actions lead to a stronger SoA than actions that are merely observed by the user and do not contain any intentional or causal action.

Several studies investigated which factors can modulate the SoA in observation-only settings. Using the Libet Clock, Wohlschläger et al. (2003) observed that the activation of our proprioceptive system, which usually accompanies self-directed actions, positively affects SoA. However, this effect of sensory feedback could be overridden by the assigned intentionality of the agent in the case of other-directed actions (e.g. machine-directed actions). When the agent is perceived as intentional, higher control was experienced in observational actions, regardless of whether it was a biological or technical co-agent. Obhi and Hall (2011) also emphasise the relevance of beliefs about co-agents for the emergence of SoA in joint-actions. When people collaborate with another biological agent, a 'we-identity' emerges at a pre-reflective level. However, this does not arise if we believe a machine co-initiated the action and triggered the sensory consequences. Furthermore, Poonian and Cunnington (2013) were able to show in intentional binding-based experiments that features of the action configuration can also determine the emergence of SoA. Comparing different observational actions, they provided the participants with videos showing either a human hand triggering a sound by pressing a button or videos showing the same action without a visible hand. The existence of the visual cue produced higher SoA.

Taken together, these findings indicate that our perception of (co-)agents is highly relevant for the emergence of SoA in non-operant situations. The identified determinants (attribution of intentionality, humanoid physical appearance) suggest the understanding of co-agents as human-like entities letting us experience greater control. This implies a human factor must exist which strengthens our conscious action awareness and, thus, modulates our SoA. Therefore, the following section elaborates the possibility of using this human factor as part of the design of technical systems to promote SoA in interaction with automated technologies.

Anthropomorphism

According to the *Computers Are Social Actors* (CASA) paradigm, humans subconsciously treat computers as real persons and automatically apply human-human social rules (e.g. application of norms, assignment of traits and intentions) in interactions, even when they are aware of interacting with machines (Reeves & Nass, 1996; Nass & Moon, 2000). In addition to the initially addressed interactions of users with media agents in the form of desktop computers, the CASA framework has been validated in recent studies for a broad range of contexts including more complex technologies, such as embodied agents and voice-based navigational systems (Gambino et al., 2020). *Anthropomorphism*, the transfer of human characteristics to non-human entities, reinforces the previously described behaviour (Gambino et al., 2020). The more humans attribute a humanlike mind to a machine, the more they employ scripts that are usually observed in human-human interactions (Nass & Moon, 2000). The tendency to anthropomorphise entities can be supported by a wide range of anthropomorphic features (e.g. morphology, behavioural cues; Schroeder & Epley, 2016). In this study, anthropomorphism will be considered as a key concept for strengthening SoA in interactions with automation technologies.

In the context of Human-Robot interaction (HRI), the application of anthropomorphic features to robotic systems has been well studied and is known to have positive effects on interaction qualities such as user acceptance and trust (Duffy, 2003; Kiesler et al., 2008; Waytz et al., 2014). Furthermore, evidence in HRI suggests an opportunity to increase the users' SoA by modulating the design of a technical interaction partner (Barlas, 2019; Roselli et al., 2022). Barlas (2019) demonstrated that the human-like appearance of artificial co-agents was positively correlated with SoA. Anthropomorphic features help to co-represent the intentions and action plans of artificial agents (Obhi & Hall, 2011; Wohlschläger et al., 2003). For instance, Sahaï et al. (2017) found that humanised systems such as robots incorporating biological movements helped anticipate machine-directed actions, which directly led to increased SoA in the interaction. Most research addresses the physical, humanoid appearance of artificial co-agents and provides evidence supporting the efficacy of visual cues. However, the transferability of these findings to artificial systems which are not, or only to a limited extent, suitable for resembling humans in visual aspects (e.g. non-embodied in-vehicle intelligent agents) is rather restricted.

In those cases, non-visual cues need to be used to increase the agents' human-likeness. Voice has been identified as an effective feature to promote anthropomorphism in various contexts (e.g. Festerling & Siraj, 2021; Waytz et al., 2014). However, studies comparing embodied agents integrating visual cues with non-embodied voice-only agents show inconsistent findings regarding users' perception and experience: While Williams et al. (2013) reported that the embodied robot was perceived as friendlier than the voice-only agent, Dong et al. (2020) observed embodied robot agents to be perceived less likeable, less comfortable, and less competent than voice-only agents. Research considering the impact of speech-based cues on SoA is limited. Although some studies concerning conversational interfaces showed that embodied agents with additional humanoid visual cues (e.g. avatars) support SoA more than voice-only agents (e.g. Appel et. al. 2012; Von der Pütten 2010), it remains disputed whether

these conclusions apply to the experience of control in monitoring tasks, and how speech-based cues should optimally be designed to foster SoA.

Research question and hypotheses

This study aims to investigate whether humanising automated systems by incorporating anthropomorphic features is an effective way to stabilise the SoA in interactions with automated systems to avoid negative consequences of perceived losses of control. The broader research question at hand is: How does the anthropomorphisation of automated technologies influence the users' perceiced control? In this particular context, the extent to which speech-based anthropomorphic cues affect the SoA during the observation of an automated process was compared to absent cues and conventional visual anthropomorphic cues often used in HRI. Therefore, participants performed an interval estimation task based on videos showing other-directed actions. To test the potential influence of anthropomorphic cues in system design, the observed actions could be flanked by different cues (Table 1).

Table 1. Overview of experimental conditions (in ascending order according to the predicted level of induced SoA).

Condition	Label	Description	Cue Modality
(a)	No cue	No additional cue flanking the automated process	-
(b)	Computer voice	Computer-generated speech-based cue announcing action initiation	auditory
(c)	Human voice	Human-voice speech-based cue announcing action initiation	auditory
(d)	Visual cue	Human hand visibly initiating action by a button press	visual

Considering preliminary work, the anthropomorphisation of autonomously acting systems (b-d) will support users' SoA in the monitoring task compared to non-anthropomorphised systems (a). More precisely, adding speech-based cues was assumed to promote an attribution of human mind to the automated agent and, thus, help strengthen users' SoA while monitoring automated processes. Thereby, a computer-generated voice should be a less effective design feature than a human one due to the lack of paralinguistic information. Following HRI research, adding humanoid visual cues to the other-generated action should provide the strongest anthropomorphic feature and, hence, the strongest SoA, since observing a human agent performing an action should provoke a stronger attribution of intentionality and, additionally, allow to build up a precise sensorimotor representation of the observed action (Khalighinejad et al., 2016; Roselli et al., 2022). Therefore, this condition should provide a reference for the effectiveness of speech-based anthropomorphic features.

Method

Design

In a repeated measures design, the effects of the anthropomorphism of the automated system on the users' SoA were investigated by means of intentional binding. Within-subjects, the anthropomorphisation level was varied by adding different anthropomorphic features (no cue vs. computer voice vs. human voice vs. visual cue) to the autonomously operating system (see Table 1 above). To avoid confounding influences of individual characteristics, the subject's affinity for technology, the individual level of positive affects towards technology and its use, was recorded as a control variable, as it can be assumed that this can positively influence the SoA (Bufton, 2017).

Participants

In total, $N = 52$ individuals (51.9% male; age: $M = 30.10$ years, $SD = 13.78$ years) participated in the study. The sample was mainly academic (56% students, 75% reported having at least a high school diploma) and had an above-average affinity for technology ($M = 3.78$, $SD = 1.09$). The latter was also reflected in the frequency of computer usage (90% indicated to use it daily or 5-6 times per week). Participants were invited to the study via a study-related mailing list and personal invitations. Persons with impaired vision or hearing could not participate. Participation was not monetarily compensated. The a priori power analysis with G*Power (version: 3.1.9.7) resulted in a required minimum sample size estimation of 24 participants ($p = .05$, $f = .25$, power = .80).

Materials

The online experiment was conducted via LimeSurvey 3.0. Videos were presented in which an action (production of a sound signal) was either performed by technical (a-c) or human agents (d). The users of the automated system took on the positions of observers. The videos which were created with Adobe Premiere Pro 2020 showed a conventional computer workstation from a frontal perspective consisting of a computer screen and keyboard placed in front of it on a desk (Figure 1).

Figure 1. Presented video set-up (screenshot). Left image: no cue (a) and auditory cue conditions (b-c); right image: visual cue condition (d).

Each video presentation followed a fixed sequence (Figure 2) starting with a 3-second visual countdown followed by the varied action scenario. In each scenario, two non-speech-based acoustic signals sounded one after the other, marking the initiation and the result of the observed automated action. The duration of the time interval between action initiation (first sound) and action outcome (second sound) was varied between the videos. For each experimental condition, two repetitions per interval length (500 ms, 1000 ms, 1500 ms) were presented.

While in condition (a) no additional anthropomorphic feature was embedded, the automated action was accompanied by different anthropomorphic features in the other conditions. In conditions (b) and (c), auditory cues announced the action initiation commenting "I am now pressing the trigger" in either a computer-generated or a human male voice. The free software Text To Speech[•] (selected voice: Michael) was used to generate the synthetic male voice which was comparable to the human voice used in terms of pitch and timbre. In condition (d), instead of the audio commentary, a visual cue was displayed in the form of a human hand, which allowed the participants to follow the start of the action on the basis of the displayed movement to the keyboard up to the visible keystroke.
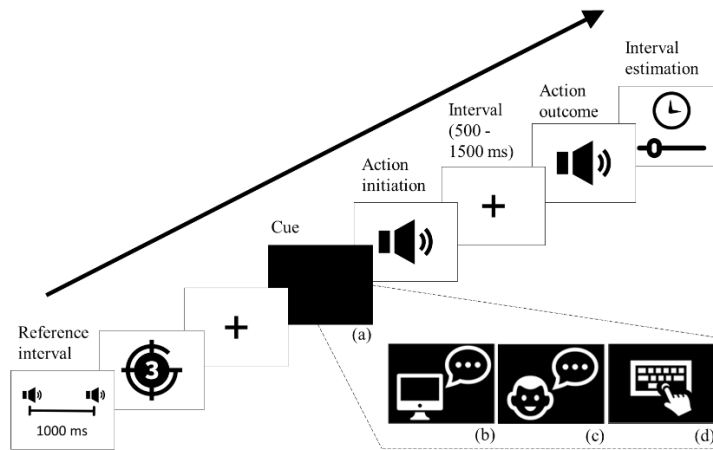
Figure 2. Schematic illustration of one exemplary task trial regarding the experimental conditions: (a) no cue, (b) computer voice, (c) human voice, (d) visual cue.

The individual affinity for technology was assessed via the Affinity for Technology Interaction (ATI) scale (Franke et al., 2019). This captures agreement towards 9 statements (e.g. "I like testing the functions of new technical systems") on a Likert scale from 1 (completely disagree) to 6 (completely agree). For the manipulation check, the anthropomorphism subscale of the Godspeed questionnaire was used (Bartneck et al., 2009). The original scale contains 5 semantic differential items (e.g. "fake|natural") to be rated on a 5-point Likert scale. In this study, the item regarding

movement was excluded, since this could not be transferred to the application context of the automated systems considered.

Task and procedure

After the participants had been informed about and consented to their participation, demographic and personal data (e.g. affinity for technology) were collected via questionnaires. Then an audio check was carried out, which allowed to identify participants invalidly taking part without audio output and to exclude them from the analysis. Afterwards, the task instruction was given. The task required the participants to estimate the time interval between the two acoustic signals presented in a short video sequence and to quantify its duration on a slider (range from 0 ms to 2000 ms). To provide a reference for their time estimates, an interval with a duration of 1000 ms was acoustically presented before each trial. To familiarise themselves with the task, the participants completed an exemplary task run (without cue) in a brief training phase. Then the experimental phase followed, consisting of 24 randomly presented trials. For each of the four conditions, 6 trials (2 repetitions x 3 different interval lengths) were presented. After the interval estimation tasks had been completed, the participants were asked to assess the conditions regarding their perceived anthropomorphism and were given the opportunity to give feedback. The procedure took about 20 minutes.

Data Preparation and Analysis

The total sample comprised of 54 complete data sets, of which $n = 2$ subjects were excluded from analysis due to either conspicuous response tendencies (consistent interval estimates of 0 ms) or missing values in interval estimates. Checking for extreme outliers (1.5 IQR) did not require further exclusions. The analysis is therefore based on a final sample of $N = 52$ valid cases. To be able to analyse the interval estimates as indirect measures of SoA, estimates of 0 ms were recorded as missing values. Subsequently, for each condition, a mean value for the interval estimate was calculated from the 6 trials and, then, the difference between the actual interval lengths and these mean estimates was computed. Hereafter, these difference values are referred to as reproduction errors (RE). With positive RE, the interval length was underestimated, i.e. perceived as shortened. Negative RE indicate overestimations of the duration, i.e. the duration between action initiation and action outcome was experienced as longer than it actually was. Regarding anthropomorphism ratings (manipulation check), for each condition a mean value was formed from the four items of the Godspeed scale. Reliability analyses revealed good to excellent internal consistencies (Cronbach's $\alpha_{(a)} = .87$, Cronbach's $\alpha_{(b)} = .89$, Cronbach's $\alpha_{(c)} = .82$, Cronbach's $\alpha_{(d)} = .90$). All analyses were conducted in IBM SPSS Statistics (version: 28.0.0.0) at a significance level of $p < .05$. The covariate (affinity for technology) was centred (Schneider et al., 2015).

Results

Manipulation Check

To check whether the conditions differed from each other regarding perceived anthropomorphism in the expected order, a repeated-measures ANOVA was carried out. This indicated significant differences between the experimental conditions ($F(3, 147) = 23.24$, $p < .001$, $\eta_p^2 = .32$). As expected, participants rated anthropomorphism of the no cue condition lowest ($M = 1.90$, $SD = 0.91$), followed by the computer voice ($M = 2.75$, $SD = 1.14$) and human voice condition ($M = 3.23$, $SD = 0.92$). Perceived anthropomorphism was rated highest in the visual cue condition ($M = 3.45$, $SD = 1.17$). The Bonferroni-adjusted post-hoc analysis (Table 2) revealed significant differences between all conditions ($p < .001$) except for the two auditory cue conditions as well as the conditions human voice and visual cue. These results indicate successful manipulation of anthropomorphism via the selected anthropomorphic cues.

Table 2. Test statistics (mean differences, p-values and 95% confidence intervals) of Bonferroni-adjusted post-hoc analysis for anthropomorphism ratings in different experimental conditions.

				95% CI for mean differences	
Condition i	Condition j	$M_i - M_j$	p	lower bound	upper bound
No cue	Computer voice	-0.845	<.001	-1.357	-0.333
	Human voice	-1.325	<.001	-1.854	-0.796
	Visual cue	-1.550	<.001	-2.168	-0.932
Computer voice	Human voice	-0.480	.138	-1.042	0.082
	Visual cue	-0.705	.010	-1.289	-0.121
Human voice	Visual cue	-0.225	1.000	-0.732	0.282

Effects of anthropomorphic cues on sense of agency

To analyse differences in SoA related to the anthropomorphic cues implemented, a repeated-measures ANCOVA was conducted. RE was included as the dependent variable in the analysis, while the different degrees of anthropomorphisation represented the independent variable. Affinity for technology was controlled as a covariate. Due to the violation of sphericity, a Huynh-Feldt correction was applied. The ANCOVA showed a significant effect of anthropomorphism on SoA ($F(2.52, 126.19) = 3.23$, $p = .032$, $\eta_p^2 = .06$). Although the Bonferroni-corrected multiple comparisons did not indicate any significant differences between the experimental conditions, the RE increased with increasing anthropomorphism of the automated system (no cue: $M = 126.00$, $SD = 166.44$; computer voice: $M = 135.04$, $SD = 175.35$; visual cue: $M = 160.94$, $SD = 205.59$), with the exception of the human voice condition (Figure 3). The latter had the lowest RE ($M = 103.00$, $SD = 196.63$), indicating the lowest temporal binding. All RE were positive values and, thus, represented subjectively shortened interval durations. Affinity for technology did not significantly affect SoA.
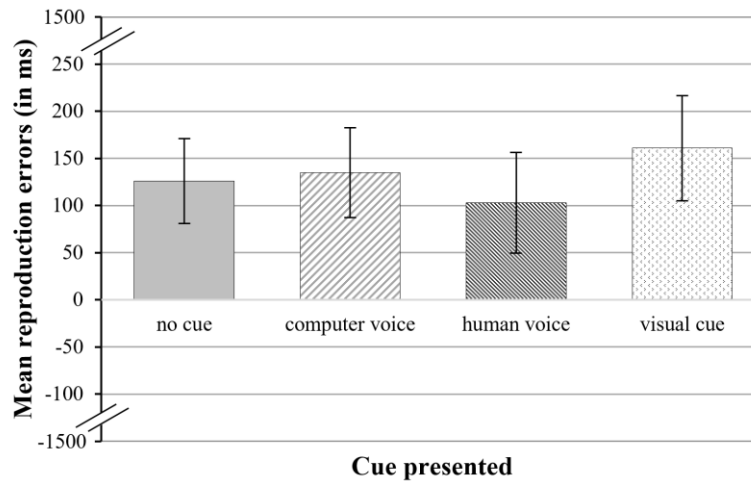
Figure 3. Mean reproduction errors (in milliseconds) in different experimental conditions. Error bars indicate the 95% confidence interval.

Discussion

In the present study, the influence of anthropomorphisation of automated processes on SoA was investigated. The implementation of two speech-based cues was compared to conditions either using conventional visual cues or no additional anthropomorphic features. SoA was measured indirectly via binding effects, enabling RE to provide information about the perceived level of control. RE increased between conditions in the following order: human voice, no cue, computer voice and visual cue. Affinity for technology was not found to have an impact on SoA in this study. This implies that the differences in control perception can be solely attributed to the degree of anthropomorphism, independent of individual disposition. Nevertheless, this relationship should be replicated with the help of a more heterogeneous sample.

In the visual cue condition, the interval durations were most strongly underestimated, which suggests that the participants experienced the highest level of control. It can be assumed that the visual cue implemented in this study, i.e. the human hand, promoted the perception of a human-controlled action execution and, in turn, strengthened the users' SoA. To determine the driving factors (e. g. physical appearance, biological movement), responsible for the effect, follow-up studies should investigate visual cues from both biological and technical agents.

It cannot be completely ruled out that the modality in which the stimuli were presented also had an influence on the estimation of the time intervals via different pathways. Since the visual presentation provided a timeline for the event, this might have allowed the participants to better prepare for the event than the acoustic announcement of the action. However, since the Interval Estimation Task does not represent a classical reaction time task, with time intervals being estimated based on the subjective impression of their length, this should not have any significant influence

on the time estimates. The same should apply to the different responsiveness to stimuli of different modalities. While experimental reaction time tasks have shown acoustic stimuli to be transmitted much faster to the motor cortex and visual stimuli take much longer to process resulting in increased reaction times (e.g. Jain et al., 2015, Shelton & Kumar, 2010), this should not affect the results in the present task, as the time intervals in all conditions were characterised by unimodal signals. As described, not every context allows the integration of visual anthropomorphic features, requiring the implementation of cues refering to other stimulus modalities. Considering the auditory cues, a surprising pattern emerged: Contrary to expectations, the implementation of speech did not lead to a strengthened SoA in both auditory cue conditions. Although the human voice condition was rated second highest with regard to the level of perceived anthropomorphism, the use of human-voice based speech led to a reduction of the SoA during observational action to below the level of the no cue condition. On the other hand, the inclusion of computer-generated speech led to a longer perceived interval duration compared to no cue, which indicates an enhanced SoA when monitoring the externally controlled action, even though the participants considered the latter to be less anthropomorphic. Thus, the use of speech-based cues appears to not be generally ineffective, but their effectiveness seems to depend largely on the design of the features. One possible reason why users felt less control during the observation of the autonomous action announced by a human voice could be the *uncanny valley effect* (Clark et al., 2021; Mori, 1970). This effect describes the phenomenon that technologies are experienced as irritating and uncanny if their design is especially human-like, because this can create a mismatch between the user's expectations and the perceived capabilities of the system (Männistö-Funk & Sihvonen, 2018). The effect has mostly been studied for visual stimuli in HRI, but it can also occur for auditory stimuli (Clark et al., 2021). Thus, in relation to the human-voice-based speech used in the present study, the absence of a physical human appearance may have counteracted the users' expectations towards the involved agent, which increased the uncanniness of the observed action. The computer-generated voice did not trigger this effect, as it presumably provoked different expectations regarding the underlying agent. However, as research on voice perception often refers to the design of embodied interaction partners, for a more comprehensive understanding, extensive research on voice perception in relation to automated processes is required (Clark et al., 2021; Männistö-Funk & Sihvonen, 2018). Also of interest is how multimodal anthropomorphic features interact with each other.

In general, there are some research gaps regarding the design of voice-based characteristics. In the present study, the voices investigated were male. Since many voice-based agentic systems, such as navigation devices or intelligent personal assistants (e.g. Google Assistant), often use female computer-generated voices as a default options, this could have been contrary to the users' habits. Accordingly, this might have led to a lower internal representation of the observed action in general and, consequently, to a lower SoA. Furthermore, studies demonstrated inconsistent findings on preferences and perceptions regarding the gender of voice-based agents (e.g. Dong et al., 2020; Nass & Moon, 2000). Besides the gender of the speaker, a consideration of concrete paralinguistic properties (e.g. pitch, pace) as well as the perception of voices as a function of interindividual properties (e.g. gender of the

listener) is relevant to draw conclusions about the design of an appropriate speech-based HMI (Do et al., 2022).

Instead of a one-fits-all solution, there is evidence that the context, the specific technical application and the fit between the characteristics of the user and those of the voice interface are particularly relevant (Cambre & Kulkarni, 2019). Hence, the individualisation of voice assistants (i.e. adaptation to individual preferences, expectations and demographic characteristics) plays a decisive role (Schmitt et al., 2021). In their review, Schmitt et al. (2021) provide several examples demonstrating that acoustical mimicry can have beneficial effects on system interaction and perception and, thus, can be considered a possible strategy for the design of voice-based interactions. For example, various studies have shown that the correspondence between the accent of the voice-based agent and that of the user results in greater trust and a preference for these agents (Cowan et al., 2016; Tamagawa et al., 2011). Similar results were observed for gender: Same-gender voices were evaluated more positively (Eyssell et al., 2012; Obinali, 2019), and experienced as less demanding (Truschin et al., 2014). Although the effects of specific voice-based characteristics on SoA have not yet been investigated, it seems reasonable to assume that the human experience of control could also benefit from adaptation to individual needs and preferences, and matching user and system characteristics could lead to greater identification with the system. However, it also remains an open question at which level of congruence uncanny valley effects overshadow the positive impact and create a negative user experience.

Delegating autonomy to automation technologies rises significant issues in HMI, since this could decrease humans' feeling of control. Seeking a solution to this challenge, anthropomorphic features in system design, including speech-based cues, bear great potential. However, the mindless mimicry of human-like features in the design of automated technologies might not be an appropriate method to strengthen users' SoA, since this study provided potential evidence for the verbal uncanny valley effects suggesting perceptual tension leading to a more-negative user experience in non-embodied speech-based agents.

Providing automated processes with adequate system information can help to hold users in the loop and enhance their feelings of control during the use of these systems. For example, in the context of autonomous driving, informing users about upcoming driving manoeuvres with the help of an in-vehicle assistance system may be a useful strategy for a successful interaction design that enhances users' SoA. Particularly in tasks where the sensory input to be processed by a user is primarily visual, acoustic cues outperform visual cues as they make information processing less demanding and interaction safer (e.g. Barón & Green, 2006; Ei-Wen Lo & Green, 2013; Horn & Gehlert, 2019). In this regard, speech-based input is intuitive and provides information more appropriately than unspecific sounds (Brandenburg & Epple, 2019). However, the results of this study show that in the design of voice-based agents special consideration should be given not to shape voices that sound too human-like, as this can irritate the users and cause undesired effects.

References

Appel, J., Von der Pütten, A., Krämer, N.C., & Gratch, J. (2012). Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction, 2012*, 1-10. https://doi.org/10.1155/2012/324694

Barlas, Z. (2019). When robots tell you what to do: Sense of agency in human- and robot-guided actions. *Consciousness and Cognition, 75*. https://doi.org/10.1016/j.concog.2019.102819

Barón, A., & Green, P. (2006). *Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review* (Report UMTRI-2006-5). Ann Arbor, MI, USA: University of Michigan Transportation Research Institute.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Berberian, B. (2019). Man-Machine teaming: a problem of Agency. *IFAC-PapersOnLine, 51*(34), 118–123.
https://doi.org/10.1016/j.ifacol.2019.01.049

Berberian, B., Sarrazin, J.C., Le Blaye, P., & Haggard, P. (2012). Automation technology and sense of control: A window on human agency. *PLoS ONE, 7(*3), e34075. https://doi.org/10.1371/journal.pone.0034075

Blakemore, S.J., Frith, C.D. & Wolpert, D.M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport 12*(9). 179-188.

Brandenburg, S., & Epple, S. (2019). Drivers' individual design preferences of takeover requests in highly automated driving. *i-com, 18*(2), 167-178.

Bufton, G.M. (2017). *The effects of automated technology on the experience of work* [Master' thesis]. Georgia Institute of Technology.
https://smartech.gatech.edu/handle/1853/58318

Cambre, J., & Kulkarni, C. (2019). One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. In *Proceedings of the ACM on Human-Computer Interaction* (pp. 1-19). New York, NY, USA: ACM. https://doi.org/10.1145/3359325

Clark, L., Ofemile, A., Cowan, B.R. (2021). Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech. In B. Weiss, J. Trouvain, M. Barkat-Defradas, and J.J. Ohala (Eds.), *Voice Attractiveness. Prosody, Phonology and Phonetics* (pp. 317-330). Singapore: Springer Nature. https://doi.org/10.1007/978-981-15-6627-1_17

Cowan, B. R., Gannon, D., Walsh, J., Kinneen, J., O'keefe, E., & Xie, L. (2016). Towards understanding how speech output affects navigation system credibility. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 2805–2812). New York, NY, USA: ACM. https://doi.org/10.1145/2851581.2892469

Coyle, D., Moore, J., Kristensson, P.O., Fletcher, P.C., & Blackwell, A.F. (2012). I did that! Measuring users' experience of agency in their own actions. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2025-2034). New York, NY, USA: ACM. https://doi.org/10.1145/2207676.2208350

Dewey, J.A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PLoS ONE, 9*(10), e110118. https://doi.org/10.1371/journal.pone.0110118

Do, T.D., McMahan, R.P., & Wisniewski, P.J. (2022). A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-11). New York, NY, USA: ACM. https://doi.org/10.1145/3491102.3517564

Dong, J., Lawson, E., Olsen, J., & Jeon, M. (2020). Female Voice Agents in Fully Autonomous Vehicles Are Not Only More Likeable and Comfortable, But Also More Competent. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1033-1037). Los Angeles, CA, USA: Sage. 10.1177/1071181320641248.

Duffy, B.R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Ei-Wen Lo, V., & Green, P.A. (2013). Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature. *International Journal of Vehicular Technology, 2013*, 1-13. http://dx.doi.org/10.1155/2013/924170

Endsley, M.R. (2016). *Designing for Situation Awareness*. Boca Raton, FL, USA: CRC Press. https://doi.org/10.1201/b11371

Eyssel, F., Kuchenbrandt, D., Bobinger, S., De Ruiter, L., & Hegel, F. (2012). "If you sound like me, you must be more human": On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 125–126). New York, NY, USA: ACM. https://doi.org/10.1145/2157689.2157717

Festerling, J., & Siraj, I. (2022). Anthropomorphizing Technology: A Conceptual Review of Anthropomorphism Research and How it Relates to Children's Engagements with Digital Voice Assistants. *Integrative Psychological and Behavioral Science, 56*, 709–738. https://doi.org/10.1007/s12124-021-09668-y

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction, 35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Gambino, A., Fox, J., & Ratan, R.A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication, 1*, 71-85. https://doi.org/10.30658/hmc.1.5

Haggard, P. (2017). Sense of agency in the human brain. Nature Reviews *Neuroscience, 18*(4), 197–208. https://doi.org/10.1038/nrn.2017.14

Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology, 22*(10), R390–R392. https://doi.org/10.1016/j.cub.2012.02.040

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382–385. https://doi.org/10.1038/nn827

Horn, H.-P., & Gehlert, T. (2019). *Marktüberblick über technische Maßnahmen zur Vermeidung von Ablenkung* (Report 62). Berlin, DE: Gesamtverband der Deutschen Versicherungswirtschaft e.V. Unfallforschung der Versicherer.

Humphreys, G.R., & Buehner, M.J. (2009). Magnitude Estimation Reveals Temporal Binding at Super-Second Intervals. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1542–1549. https://doi.org/10.1037/a0014492

Jain, A., Bansal, R., Kumar, A., & Singh, K. D. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research, 5(*2), 124. https://doi.org/10.4103%2F2229-516X.157168

Khalighinejad, N., Bahrami, B., Caspar, E.A., & Haggard, P. (2016). Social transmission of experience of agency: An experimental study. *Frontiers in Psychology, 7*, 1315. https://doi.org/10.3389/fpsyg.2016.01315

Kiesler, S., Powers, A., Fussell, S.R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*(2), 169–181. https://doi.org/10.1521/soco.2008.26.2.169

Libet, B., Gleason, C.A., Wright, E.W., & Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain: A Journal of Neurology, 106*(3), 623–642. https://doi.org/10.1093/brain/106.3.623

Limerick, H., Coyle, D., & Moore, J.W. (2014). The experience of agency in human-computer interactions: a review. Frontiers in Human Neuroscience, 8(AUG). https://doi.org/10.3389/fnhum.2014.00643

Männistö-Funk, T., & Sihvonen, T. (2018). Voices from the Uncanny Valley. *Digital Culture & Society*, *4*(1), 45–64. https://doi.org/10.14361/dcs-2018-0105

Moore, J.W., & Fletcher, P.C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition, 21*(1), 59–68. https://doi.org/10.1016/j.concog.2011.08.010

Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33–35.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81–103. https://doi.org/10.1111/0022-4537.00153

Obhi, S.S., & Hall, P. (2011). Sense of agency and intentional binding in joint action. *Experimental Brain Research, 211*, 655–662. https://doi.org/10.1007/s00221-011-2675-2/figures/1

Obinali, C. (2019). The Perception of Gender in Voice Assistants. In *Proceedings of the 22nd Southern Association for Information Systems Conference.* https://aisel.aisnet.org/sais2019/39

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 30*, 286–297. https://doi.org/10.1109/3468.844354

Poonian, S.K., & Cunnington, R. (2013). Intentional binding in self-made and observed actions. *Experimental Brain Research, 229*, 419–427. https://doi.org/10.1007/s00221-013-3505-5

Reeves, B., & Nass, C.I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge, UK: Cambridge University Press.

Roselli, C., Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions. *Scientific Reports, 12*(1), 1–7. https://doi.org/10.1038/s41598-022-18151-6

Sahaï, A., Pacherie, E., Grynszpan, O., & Berberian, B. (2017). Predictive mechanisms are not involved the same way during human-human vs. Human-machine interactions: A review. *Frontiers in Neurorobotics, 11*(OCT), 52. https://doi.org/10.3389/fnbot.2017.00052

Schmitt, A., Zierau, N., Janson, A., & Leimeister, J. M. (2021). Voice as a Contemporary Frontier of Interaction Design. In *Proceedings of the European Conference on Information Systems (ECIS). – Virtual.* https://ssrn.com/abstract=3910609

Schneider, B.A., Avivi-Reich, M., & Mozuraitis, M. (2015). A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Frontiers in Psychology, 6*(APR), 474. https://doi.org/10.3389/fpsyg.2015.00474

Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General, 145*, 1427–1437. https://doi.org/10.1037/xge0000214

Shelton, J., & Kumar, G. P. (2010). Comparison between auditory and visual simple reaction times. *Neuroscience and medicine, 1*(01), 30-32. http://dx.doi.org/10.4236/nm.2010.11004

Tamagawa, R., Watson, C. I., Kuo, I. H., Macdonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics, 3*(3), 253–262. https://doi.org/10.1007/s12369-011-0100-4

Truschin, S., Schermann, M., Goswami, S., & Krcmar, H. (2014). Designing interfaces for multiplegoal environments: Experimental insights from in-vehicle speech interfaces. *ACM Transactions on Computer-Human Interaction, 21*(1), 1-24. https://doi.org/10.1145/2544066

Von der Pütten, A.M., Krämer, N.C., Gratch, J., & Kang, S.H. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior, 26*, 1641-1650. https://doi.org/10.1016/j.chb.2010.06.012

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wen, W., Kuroki, Y., & Asama, H. (2019). The Sense of Agency in Driving Automation. Frontiers in Psychology, 10, 2691. https://doi.org/10.3389/fpsyg.2019.02691

Williams, K.J., Peters, J.C., & Breazeal, C.L. (2013). Towards leveraging the driver's mobile device for an intelligent, sociable in-car robotic assistant. In *2013 IEEE intelligent vehicles symposium (IV)* (pp. 369-376). Gold Coast, QLD, Australia: IEEE.

Wohlschläger, A., Engbert, K., & Haggard, P. (2003). Intentionality as a constituting condition for the own self - And other selves. *Consciousness and Cognition, 12*(4), 708–716. https://doi.org/10.1016/S1053-8100(03)00083-7

Wolpert, D.M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience, 3*, 1212–1217. https://doi.org/10.1038/81497

Wolpert, D. M., Ghahramani, Z., & Jordan, M.I. (1995). An internal model for sensorimotor integration. *Science, 269*, 1880–1882. https://doi.org/10.1126/science.7569931

Zanatto, D., Chattington, M., & Noyes, J. (2021). Sense of Agency in Human-Machine Interaction. In *Advances in Neuroergonomics and Cognitive Engineering. AHFE 2021. Lecture Notes in Networks and Systems* (pp. 353–360). Basel, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-80285-1_41

HMI design for a stress and time critical paediatric emergency scenario: PediAppRREST

Maria Giulia Losi[1], Buse Tezçi[1], Francesco Tesauri[1], & Francesco Corazza[2], &*
Silvia Bressan[2]
[1]RE:LAB s.r.l, Reggio Emilia, Italy, [2]Pediatric Emergency Division, Department of
Women's and Children's Health, University of Padova, Italy

Abstract

Medical support apps can be helpful training tools to minimize deviations from international guidelines in time-critical emergency scenarios. However, not all apps have been tested for usability and workload, and a lack of proper HMI design guidelines has been a significant limitation. To address these issues, an interactive support app, called PediAppRREST, was developed for the management of paediatric cardiac arrest (PCA) using an iterative approach. After the first design stage, the app was tested in a previously published study and found to have good usability evaluation and ease on mental workload. Participants provided qualitative feedback and the results guided the second design stage. Following the refinement, the app is currently undergoing further testing to measure its effectiveness. The objective of this paper is to report the Human-Machine Interaction (HMI) principles derived during the design and testing stages and to collect them as a step toward systemized HMI guidelines for medical cognitive support apps. From an HMI perspective, timely and precise information, effective interaction flows and correct interaction modalities, as well as providing effective feedback were found to be essential. Considerations also addressed User Interface (UI) layout, shapes and buttons, language and readability, and colour selection.

Introduction

Emergency situations, such as resuscitation, require an effective intervention that consist of multiple factors including accurate assessment and diagnosis of the situation, effective communication and coordination among medical staff, a timely and accurate intervention and continuous monitoring of the status of the patient and revaluating the optimal care (Ehrler et al., 2018). In the case of Paediatric Cardiac Arrest (PCA) Cardiac, despite its rarity, the high mortality rates, complex management, and great societal impact (Knudson et al., 2012; Tibballs & Kinnet, 2006) further emphasize the importance of implementing fundamental interventions. Consequently, in order to train professionals in the management of PCA and improve the resuscitation outcomes, the Paediatric Advanced Life Support (PALS) course has been created by the American Heart Association (AHA) (Cheng et al., 2020), and moreover they periodically update the resuscitation PALS algorithms (de Caen et al.,

2015; Duff et al., 2020). On the other hand, studies demonstrated that, despite training, frequent errors, omissions, and delays occur, deviating from guideline recommendations in PCA management, thus resulting to patients' worse clinical outcomes (Corazza et al., 2021). To support intervention outcomes and complex PCA processes, different strategies and digital devices have been developed to improve the management of cardiac arrest, but they are specifically targeted at adult cardiac arrest management (Corazza et al., 2021). Some of these tools include apps for adult cardiopulmonary resuscitation (CPR) (Corrazza et al., 2021; Metelmann et al, 2018), technologies developed to improve the quality of chest compressions through audio/visual feedback (Kirkbright et al., 2014; Lin et al., 2018), as well as augmented reality glasses (Drummond et al., 2017). Most of the tools showed only partial improvements in the management of simulated cardiac arrest scenarios, and, up to date, a limited number of articles investigated the User Experience (UX), usability, perceived workload, and related concepts in medical field (Bitkina et al., 2020).

Within this framework, an audio-visual interactive app for tablets, namely PediAppRREST, was designed and developed to support the management of PCA and its usability in a high-fidelity simulation-based setting was tested. The App is designed to guide and train the team leader to perform resuscitation interventions in the sequence/timeline and modalities reported in the American Heart Association (AHA) PALS 2015 guidelines (de Caen et al., 2015).

HMI design

The PediAppRREST application was designed from March to October 2019, in its first version, following the PALS 2015 guidelines (Image 1) (de Caen et al., 2015). It underwent a pilot study in October-November 2019: a prospective non-randomised controlled pilot study with advanced simulation scenarios was conducted involving 48 paediatric residents, divided into teams of 3 members. Advanced simulations were carried out on the one hand without the help of the app, on the other hand with it (APP vs. non-APP conditions). On the basis of the results obtained, it can be stated that the App has a good usability profile and does not increase the workload of the operators. It seems to contribute to an overall reduction in errors by increasing adherence to international recommendations (Corazza et al., 2020).

Subsequently, starting from November 2019, the application was further refined and redesigned as a second version, incorporating the results of the pilot study and the update of PALS 2020 guidelines (Figure 1) (Cheng et al., 2020).
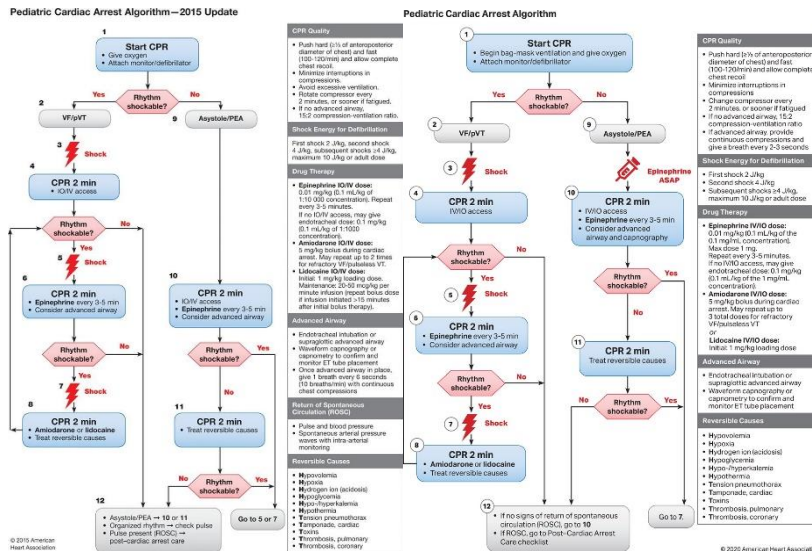
Figure 1. Pediatric cardiac arrest algorithms – 2015 (on the left) and 2020 (on the right).

The design approach started from the critical task analysis derived from a previous multicentric high-fidelity simulation study conducted by the University of Padova, in Padova, Firenze and Novara in 2018 (Stritoni et al., 2019) in which a PCA simulation scenario identified deviations from guidelines as well as frequent errors, omissions, and delays. These were derived and identified, as shown in the following tables:

Table 1. Most frequent errors emerged in PCA simulation.

Errors	n total = 27	%
Compression/ventilation ratio 30/2	4	15%
Wrong route of adrenaline administration	1	4%
Non-defibrillatory rhythm defibrillation	1	4%

Table 2. Most frequent omissions emerged in PCA simulation

Omissions	n total = 27	%
Rigid board placement never performed	20	74%
Search for reversible causes never performed	19	70%
Change to MCE never performed	9	33%
Emergency system activation never performed	9	33%
Adrenaline administration never performed	2	7%

Table 3. Most frequent delays emerged in PCA simulation

Delays	n total = 27	%
Δt positioning venous access/IO > 2 min	21	78%
Δt MCE change > 2 min	17	94%
Δt MCE from pulse assessment > 30 sec	4	15%

The analysis of most frequent errors, delays and omissions impacted both the User Interface (UI) layout and the UI navigation flow of the app. As a result, the UI layout has been outlined as follows:

Each screen of the application is structured in three zones: Zone 1, referred as Top Bar; Zone 2 which encompasses the Central Area, and Zone 3, referred as Bottom Area (Figure 2). This zoning approach assists a clear and organized layout for navigation and interaction within the application. In the Top Bar, from left to right a menu-log button, a 2-minute countdown clock for repeat rhythm check button, a button with countdown for medications, a metronome button and a total counter are displayed. The metronome button can be activated and deactivated by a tap on it. The metronome feature provides a sound guide to perform compressions at the recommended rate according to the number of operators available (100-120/min). The Central Area is the main zone where the actions to be performed (with buttons of same shape and colour) or a question with different choices (with buttons of different shapes and colours) are presented. Once the user chooses to tap on a button, the action is recorded in the action log and the flow of displayed prompts will progress following the choices of the user. In the Bottom Area, the "Rianimazione Cardio Polmonare" (RCP) and the Return of Spontaneous Circulation (ROSC) buttons are always available. The RCP button opens a recap of the characteristics of a high-quality cardio-pulmonary resuscitation (CPR), as reported by the PALS reference pocket card, whereas the ROSC button provides indication on the recommended management of a patient in case ROSC is achieved.
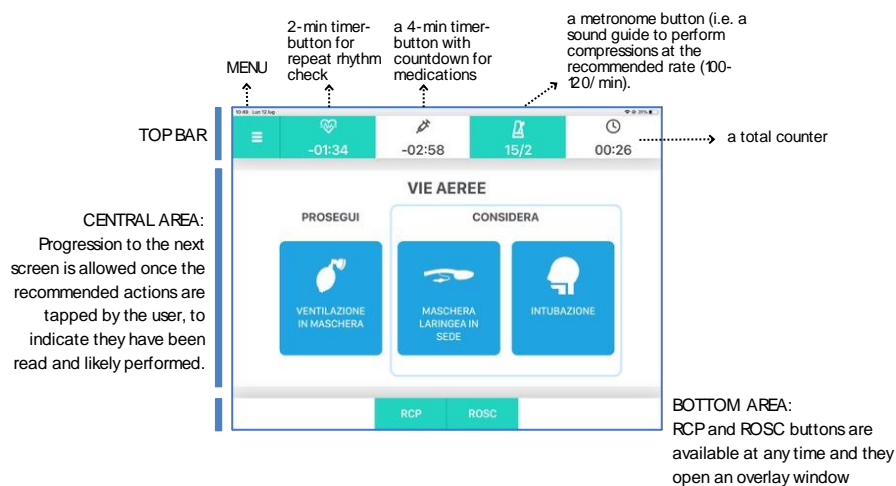
Figure 2. Image depicting the structure and the main elements of the PediAppRREST.

Designing digital UI and digital buttons for a critical scenario means designing for a context in which users experience an acute stress response that sharpens senses, intensifies focus, and drives fast, intuitive decision-making. This comes at a cost: one example is the loss of fine motor control (Swindler, 2021). Hence, some criteria might impact the use of digital interfaces meant for critical and emergency scenarios (see section User Experience / User Interface design criteria for detailed list of criteria used in PediAppRREST design).

User Interface navigation model

The conceptual UI navigation model of PediAppRREST is based on the Paediatric Cardiac Arrest Algorithm 2015 (Image 1) (de Caen et al., 2015). Keeping it as a reference both in the UX and UI design of the App had the objective, on one side of keeping the familiarity of the users with the mental model of the PALS pocket algorithm; and on the other side, of organizing UI items with respect to priority of actions to be coordinated by the team leader and carried out by the operators, minimizing potential errors and operators' workload. Another major contribution to the application development is a study conducted prior to the app development, in which the deviations that happen during PCA simulations when the simulation was conducted without a support application (Stirtoni et al., 2019). These allowed to define building the mental model of the application. Consequently, it was decided to create the application in checklist layout, that is the user is allowed to move on to the next intervention reported in the PALS algorithm (de Caen et al., 2015), only when the current intervention is performed by the team. This step is executed by the user by tapping on the action to report that the intervention recommended by algorithm.

User Experience / User Interface design criteria

Following establishing the mental model, UX/UI design phase was conducted applying the following seven criteria:

1. Timely Information. This criterion refers to providing timely presented information to user as the emergency scenario requires precision in information accessibility. Each screen gathers only the necessary information needed for each phase of the PALS algorithm, communicating it both visually and acoustically, with the aim of reducing the load on the team leader's working memory and relying on a multimodal communication. This design choice goes at the expense of offering a complete overview on the intervention phases, which is recap in a dedicated section (i.e. the log page) and it will be incorporated in a "(PALS) Flow" dedicated tab in the bottom bar in a future version of the app (Figure 3).
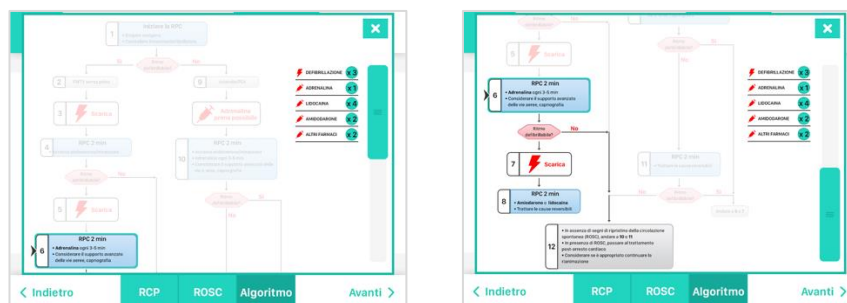
Figure 3. Flow ("Algoritmo" in Italian) tab in the Bottom Area of the PediAppRREST – version 2.

2. Priority. This criterion refers to the priority of information that will be presented to the user once at a time depending on the information hierarchies. Actions triggered

by timers have priority on other actions displayed on the screen (i.e., rhythm-check, medications) (Figure 4).

Figure 4. Timer activated pop-up.

3. Sequential or alternative choice. This criterion refers to presenting choices to the user in a most effective manner depending on the characteristics of the choices. Decisions that team leaders must take into consideration concern actions and choices to be performed sequentially or alternatively. In the first version of the app (Image 5), the buttons' shapes of the sequential actions had the same colour. One button next to the other collectively created a shape of an arrow pointing down, thus suggesting the idea of performing an action to move to the next one (Figure 5).

Figure 5. Sequential actions button shape and colour.

The alternative actions' buttons had the same shape, resembling a pointing down arrow on their own, but with a different colour, thus emphasizing the choice exclusivity (Figure 6).

Figure 6. Alternative actions button shape and colour.

In Version 1 of the app, each time the user (i.e., team coordinator) taps on a button, it remains visible with a pressed state and a checkmark, making it visible that the action has been performed and terminated (Figure 7).
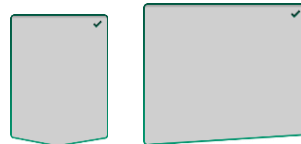
Figure 7. Button pressed state.

Figure 8. PediAppRREST UI layout – version 1.

In the second version of the application, the UI re-design has affected the colour coding, the UI button layout (Figure 9) as well as the update to new PALS guidelines (2020). In fact, sequential actions (e.g., A and B) have been differentiated in both shape and screen alignment on the screen, i.e., rectangular-shaped buttons were aligned vertically on the page, represented sequential actions, and whereas alternative choices (A or B) have been organized with square-shaped, buttons, were aligned horizontally on the screen (Figure 9). Colour coding was focused on a blue palette with a high text contrast, using different shades of blue to highlight alternative choices (e.g., shockable and not shockable flows). Green, purple, and orange colour palettes of the first version (Figure 8) were avoided, with the aim of avoiding confusion with wrong/right informative aspects or high priority of the actions to be performed.
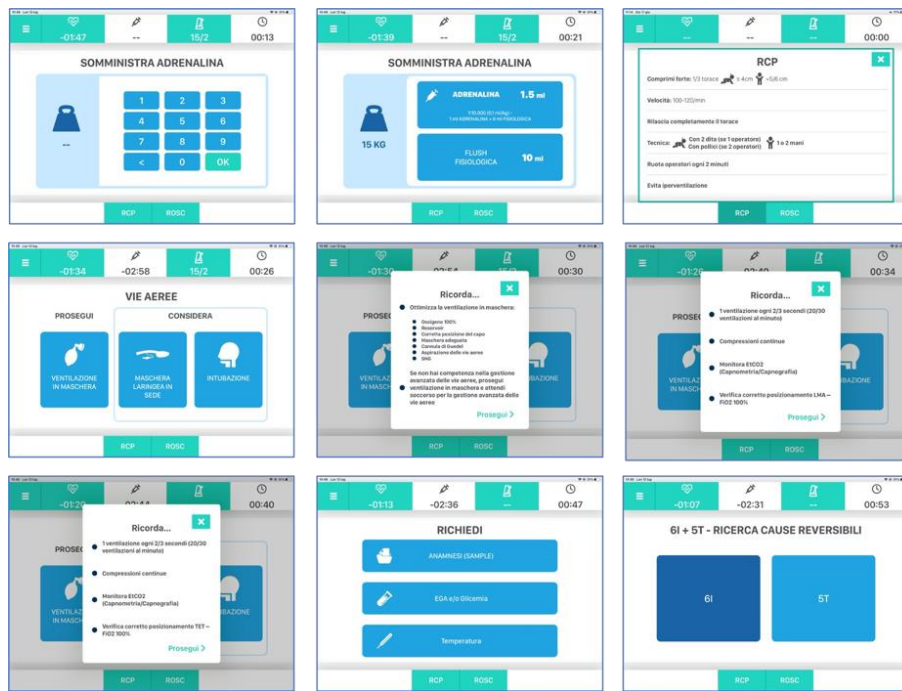
Figure 9. PediAppRREST UI layouts – version 2.

4. Information logging. This criterion refers to the ability of the app to save all actions performed by the team leader on the app. The actions done by the user are saved on the device in the "log" function in order to store workflow information on user choices that can be retrieved for any documentation and training purpose (Figure 10).

Figure 10. PediAppRREST Menu – Version 2.

5. *Reversing actions.* This criterion refers to giving the user the ability to undo the action. From the simulation studies it emerged that the Back/Undo action was sometimes causing a wrong guidance to the user in the correct flow of actions. Hence, it is inhibited in all the App's screens except when strictly necessary: in the choice of the correct flow (shockable and not shockable), in the reversible causes, and in the modification of the weight of the child that has an impact on the calculation of medication doses.

6. Multimodality. This criterion refers to the appropriate use of interaction modalities based on the situational necessities. The main modality through which the user can navigate the App is only touch gestures. Voice interaction has not been integrated due to the characteristics of the environment of a resuscitation room and of the emergency intervention, that would make recognizing and distinguishing vocal commands often mistaken. However, the App gives audio prompts, suggesting the user to perform the actions showed on the central part of the screen if the icon indicating an action is not tapped within five seconds. Audio prompts can be silenced and re-activated through the App menu.

7. Legibility, accessibility, and easy interaction. This criterion refers to providing users with a legible and clear User Interface that provides ease in interaction. This criterion includes increasing legibility and accessibility using correct word choice, providing a colour contrast ratio of at least a 7:1; providing linear decision path in order to not to overload the user with unnecessary number of choices; using short but very clear labels to buttons; using extra-large buttons to allow error margins in pointing and tapping and leaving an extra area for hand movements (Swindler, 2021).

Conclusions

PediAppRREST tablet app has been designed, developed, and iteratively assessed in advanced simulation scenarios with the aim of providing an interactive cognitive aid to Paediatric cardiac arrest (PCA) training scenarios. Advanced simulation has been fundamental in identifying user interaction improvements within the app design, with the aim of further enhancing its usability and minimizing the team leaders' workload.

The app initial version has been tested in a pilot study using standardized questionnaires, the User Experience Questionnaire (UEQ) and the National Aeronautics and Space Administration Raw-Task Load Index (NASA RTLX) questionnaire, respectively. It resulted into a good usability, in terms of efficiency, perspicuity, and stimulation, as well as user experience aspects such as attractiveness, stimulation, and novelty. Team leaders perceived workload was comparable between the 2 groups, with and without the App. The app also contributes to an overall reduction in errors by increasing adherence to international recommendations (PALS 2015). Qualitative feedbacks have been collected and further integrated as improvements and suggestions in a second release (Version 2 of the app).

Thanks to simulation studies with users, seven User Experience and User Interface design criteria have been identified, i.e., (1) timely information, (2) priority, (3) sequential or alternative choices, (4) information logging, (5) reversing actions, (6) multimodality, (7) legibility, accessibility, and easy interaction. They serve as achievements in UX/UI design phases attained through iterative design processes and they can be considered as initial steps towards establishing guidelines for cognitive-aid applications designed for emergency scenarios. By successfully addressing these criteria, it is possible to contribute to the development of user-centred designs that enhance usability and effectiveness in high-pressure situations. As the field of UX/UI design continues to advance, these provide valuable insights and principles for the design of future cognitive-aid applications, ensuring optimal support for emergency responders and improving outcomes in critical scenarios.

References

Bitkina, O.V., Kim, H.K., & Park, J. (2020). Usability and user experience of Medical Devices: An overview of the current state, analysis methodologies, and future challenges. *International Journal of Industrial Ergonomics*, 76, 102932. https://doi.org/10.1016/j.ergon.2020.102932

Cheng, A., Nadkarni, V.M., Mancini, M.B., Hunt, E.A., Sinz, E.H., Merchant, R.M., Donoghue, A., Duff, J.P., Eppich, W., Auerbach, M., Bigham, B.L., Blewer, A.L., Chan, P.S., & Bhanji, F. (2020). Part 6: resuscitation education science: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 142 (16_Suppl_2), S551-S579.

Corazza, F., Arpone, M., Snijders, D., Cheng, A., Stritoni, V., Ingrassia, P.L., De Luca, M., Tortorolo, L., Frigo, A.C., Da Dalt, L., & Bressan, S. (2021). PediAppRREST: Effectiveness of an interactive cognitive support tablet app in reducing deviations from guidelines in the management of paediatric cardiac arrest: Protocol for a simulation-based randomised controlled trial. *BMJ Open*, *11*. https://doi.org/10.1136/bmjopen-2020-047208

Corazza, F., Snijders, D., Arpone, M., Stritoni, V., Martinolli, F., Daverio, M., Losi, M.G., Soldi, L., Tesauri, F., Da Dalt, L., & Bressan, S. (2020). Development and usability of a novel interactive tablet app (PediAppRREST) to support the management of pediatric cardiac arrest: Pilot high-fidelity simulation-based study. *JMIR mHealth and uHealth*, *8*(10). https://doi.org/10.2196/19070

de Caen, A.R., Berg, M.D., Chameides, L., Gooden, C.K., Hickey, R.W., Scott, H.F., Sutton, R.M., Tijssen, J.A., Topjian, A., van der Jagt, É. W., Schexnayder, S. M., & Samson, R.A. (2015). Part 12: Pediatric advanced life support. *Circulation*, *132*(18_suppl_2). https://doi.org/10.1161/cir.0000000000000266

Drummond, D., Arnaud, C., Guedj, R., Duguet, A., de Suremain, N., & Petit, A. (2017). Google Glass for residents dealing with pediatric cardiopulmonary arrest. *Pediatric Critical Care Medicine*, *18*(2), pp. 120–127. https://doi.org/10.1097/pcc.0000000000000977

Duff, J.P., Topjian, A.A., Berg, M.D., Chan, M., Haskell, S.E., Joyner, B.L., Lasa, J.J., Ley, S.J., Raymond, T.T., Sutton, R.M., Hazinski, M.F., & Atkins, D.L. (2019). 2019 American Heart Association Focused Update on Pediatric Advanced Life Support: An Update to the American Heart Association guidelines for Cardiopulmonary Resuscitation and emergency cardiovascular care. *Circulation*, *140*(24). https://doi.org/10.1161/cir.0000000000000731

Ehrler, F., Del Zotto, M., Rouyer, F., Weinhold, T., Lovis, C., & Siebert, J. (2018). Design of InterFACE: a tool to improve collaborative work and decision making during rescucitation. In J. Mantas, M. Hercigonja-Szekeres, A. Kolokathi, Z. Sonicki, M. Crişan – Vida, K. Fišter, and M. Hägglund (Eds.), *Decision Support Systems and Education* (pp. 117-121). IOS Press.

Kirkbright, S., Finn, J., Tohira, H., Bremner, A., Jacobs, I., & Celenza, A. (2014). Audiovisual feedback device use by health care professionals during CPR: A systematic review and meta-analysis of randomised and non-randomised trials. *Resuscitation*, *85*, 460–471.

https://doi.org/10.1016/j.resuscitation.2013.12.012

Knudson, J.D., Neish, S. R., Cabrera, A. G., Lowry, A. W., Shamszad, P., Morales, D. L., Graves, D. E., Williams, E. A., & Rossano, J. W. (2012). Prevalence and outcomes of pediatric in-hospital cardiopulmonary resuscitation in the United States. *Critical Care Medicine*, *40*, 2940–2944. https://doi.org/10.1097/ccm.0b013e31825feb3f

Lin, Y., Cheng, A., Grant, V. J., Currie, G. R., & Hecker, K. G. (2018). Improving CPR quality with distributed practice and real-time feedback in pediatric healthcare providers – a randomized controlled trial. *Resuscitation*, *130*, 6–12. https://doi.org/10.1016/j.resuscitation.2018.06.025

Metelmann, B., Metelmann, C., Schuffert, L., Hahnenkamp, K., & Brinkrolf, P. (2018). Medical correctness and user friendliness of available apps for cardiopulmonary resuscitation: systematic search combined with guideline adherence and usability evaluation. *JMIR mHealth and uHealth, 6(11),* e9651.

Siebert, J. N., Lacroix, L., Cantais, A., Manzano, S., & Ehrler, F. (2020). The impact of a tablet APP on adherence to American heart association guidelines during simulated pediatric cardiopulmonary resuscitation: randomized controlled trial. *Journal of Medical Internet Research, 22(5),* e17792.

Stritoni, V., Martinolli, F., Daverio, M., Diogo, M., Binotti, M., Genoni, G, & Bressan, S. (2019) Gaps in the management of nonshockable paediatric cardiac arrest by paediatric residents in simulated scenario. In Proceedings of the Third European Pediatric Resuscitation and Emergency Medicine Conference (PREM). Ghent, Belgium p. 23-24.

Swindler K. (2021). *Life and Death Design. What life-saving technology can teach everyday UX designers*, New York: Rosenfeld Media

Tibballs, J., & Kinney, S. (2006). A prospective study of outcome of in-patient paediatric cardiopulmonary arrest. *Resuscitation*, *71*(3), 310–318. https://doi.org/10.1016/j.resuscitation.2006.05.009

Assessing physical demands using wearable smartwatches among male precast concrete construction workers

Oscar Arias, James Groehler, Mike Wolff, and Sang D. Choi
University of Wisconsin, Whitewater, USA

Abstract

Construction work is physically demanding and exposes workers' musculoskeletal system to high physical loads. Assessing physical activity levels at work as a proxy of the physical demands is essential to implement tailored interventions in physically demanding jobs. Our study aimed to characterise precast concrete workers' physical demands at and outside work. Direct heart rate (HR) measurements were used to estimate physical demands. A total of 21 precast concrete workers wore a smartwatch HR monitor for seven consecutive days. HR data was parsed in minutes associated with occupational and non-occupational physical activity. The percentage of HR reserve (%HRR) was used to determine physical intensity levels., The total amount of directly measured occupational physical activity (OPA) and physical activity outside work collected corresponded to 177.266 seconds and 162.694 seconds, respectively. The results indicate that precast concrete workers get a significant number of directly measured minutes of moderate OPA weekly. The total median number of minutes per week corresponds to 415 minutes. This study's findings highlight the importance of reducing physical demands among precast concrete workers and the potential use of smartwatches as a tool that can be implemented in the field to assess them.

Introduction

Construction workers are exposed to high physical demands at work. Their work activities involve frequent lifting, carrying heavy loads, bending or twisting of the back, adopting static postures, exposure to vibrations, and extreme weather conditions (Boschman et al., 2011). All of them have been related to the prevalence of musculoskeletal disorders (MSDs) (Arndt et al., 1996; Brenner & Ahern, 2000). Occupational physical activity (OPA) has been used to quantify physical work demands. Moderate or vigorous OPA levels are associated with an increased risk of work-related musculoskeletal disorders and cardiovascular disease mortality (Arndt et al., 1996; Gomez et al., 2022). On the contrary, moderate or vigorous intensity leisure time physical activity (LTPA) has been identified as a healthy lifestyle factor that confers positive health benefits to cardiovascular health (Clays et al., 2013; Holtermann et al., 2009; Lissner et al., 1996). This contradictory effect between OPA and LTPA is known as the "physical activity paradox." Factors such as frequency, intensity, duration, and recovery time seem to explain these contrasting effects (Coenen et al., 2018; Holtermann et al., 2018).

Precast concrete construction is a manufacturing method to produce the main structural, non-structural, and other components of a construction project in a facility and transport the concrete elements to the site for installation (Jang et al., 2022). Precast concrete workers participate in highly physically demanding tasks that can lead to work-related musculoskeletal disorders and pain in their upper back, shoulder, ankles/feet, and wrists/hands (Abas et al., 2020). A study on a group of precast concrete workers found that 15% were diagnosed with MSDS during the previous 12 months, and 57% had musculoskeletal pain in the lower back, shoulder, neck, and knee. Over 40% of the participants experienced musculoskeletal pain in the last seven days (Arias et al., 2023). Hence, the need to address physical demands and promote workers' health through integrated interventions in this working population to prevent MSDs and pain are paramount. Assessing the physical activity levels at work as a proxy of physical demands becomes essential in designing interventions to protect workers' health by limiting their physical demands to acceptable levels. Given the associated risk to workers' health, moderate or vigorous levels of physical demands should be restricted to a certain amount of time (e.g., risky if 40% HRR is maintained over 30–60 min, 60% HRR over 30 min, and 30% HRR over eight hours) (Wu & Wang, 2002).

This study aims to characterise directly measured physical demands in a convenience sample of precast concrete construction workers. The study results will serve as a baseline for designing tailored interventions to reduce physical demands in the precast concrete construction industry.

Methods

Participants

Precast concrete workers from the manufacturing area of a Wisconsin, USA facility participated in the study. These workers perform similar tasks daily, so collecting information for seven consecutive days was considered enough time to get a reliable estimate of their physical demands. They were at least 18 years old and did not take any medication that could modify their heart rate. A total of 21 male precast concrete workers with a median age of 34 years ranging from 19 to 54 years consented and enrolled between March 2022 and May 2022. They corresponded to 12.6% of the plant's precast concrete working population. Most participants were Hispanic (48%) and White (48%) males. The participants' trades/occupations were hollowcore labourer, foreman, wetcast labourer, maintenance, patcher/detailer, and quality control. They had a median body mass index (BMI) of 29, which corresponds to being overweight, and 7 years of working experience in the construction industry, working approximately 10 hours per day, more than five days per week. Their sociodemographic characteristics are shown in table 1.

Table 1. Sociodemographic and work characteristics

Sociodemographic and work characteristics		
Variable	n	%
Sex		
Male	21	100
Race		
Black	1	4
Hispanic	10	48
White	10	48
Trade/Occupation		
Hollowcore labourer	1	5
Foreman	3	14
Wetcast labourer	13	61
Maintenance	1	5
Patcher/detailer	1	5
Quality control	2	10
Smoking people*	6	32
	Median	Range
Age,(years)	34	19–54
Body mass index (BMI), kg/m^2*	29	22–35
Years in construction work*	7	0–25
Hours of work/day*	10	8–12
Days of work/week*	6	5–7

*Difference in subtotal population sample due to item nonresponse or missing

The primary precast concrete work was conducted inside the facility rather than at a job site. Different crews had to manually set-up formwork consisting of lifting, measuring, cutting, and placing side rails and bulkheads. This step included cleaning the bed with the use of scrappers and rags. Placing embedded material and pulling prestressed strands were also part of the formwork set-up. Process workers had to bend, kneel, twist, and lift multiple items throughout the set-up. Multiple subassembly groups welded and tied rebar reinforcing embedded material to be placed in the formwork before pouring the concrete. Once the form was set-up, the pouring took place using Tuckerbuilt machines. Employees had to rake, vibrate, and screed the concrete. Additional foam insulation was placed in certain elements, and lifting devices and other embedded items were placed. They then finished concrete using trowels, floats, and power trowel machines. Once the concrete was cured, it was stripped from the bed using indoor cranes. Next, pieces were detailed and patched after stripping using multiple hand tools, including torches, partner saws, scrappers, chisels, hammers, pry bars, and drills. Quality control checked the measurements, verified the concrete, and took samples throughout the process.

Participants gave written informed consent and had the right to leave the study at any time. All the study materials and protocols were approved by the University of Wisconsin-Whitewater review board (STUDY #: IRB-FY2021-2022-88).

Instrumentation and Procedures

The participants were invited to wear a smartwatch (Polar, Unite, Kempele, Finland) to continuously track their heart rate (HR) for seven consecutive days during working and non-working hours. The watch measured their heart rate every second using a photoplethysmography (PPG) sensor. The sensor quantified the volume variations of the blood circulation in the skin to determine heart rate (Castaneda et al., 2018). Each participant was instructed on how to wear the watch, initialise it, and record heart rate data. Workers also kept a daily log throughout the week of data collection, documenting the beginning and end of work shifts along the times they wore the smartwatch.

Data collected were stored in a relational database management system (Microsoft SQL Server). The heart rate data were parsed into two sets: (1) seconds of heart rate associated with work alone and (2) seconds of heart rate outside of work. For each set of parsed data, every heart rate measurement was assigned to a physical intensity level according to Norton's categories based on the percentage of heart rate reserve (%HRR). These categories included sedentary: <20% HRR, light: 20<40 %HRR, moderate: 40< 60 %HRR, vigorous: 60 < 85 %HRR, and high: ≥ 85 %HRR (Norton et al., 2010). The %HRR provided a relative measurement of physical demands accounting for individual differences (American College of Sports Medicine, 2018; Lounana et al., 2007). For its calculation, it considered the heart rate while resting, or resting heart rate (RHR), and the heart rate measured at work (HRworking). Then, it established the proportion of the maximum heart rate (HRMax) used at a given point. The HRmax was calculated using Tanaka's formula: HRmax = 208 − 0.7 × age (Tanaka et al., 2001). The relative heart rate reserve %HRR was calculated with the following formula:

$$\%HRR = \frac{HRworking - RHR}{HRmax - RHR} \times 100\%$$

The total number of weekly minutes accrued at each intensity category was calculated per participant. This information determined the amount of physical activity at and outside of work per week accrued at each physical activity level (e.g., total minutes at sedentary, light, moderate, and vigorous levels). We assumed a week of five eight-hour shifts to estimate the physical activity at work. Physical activity outside of work was calculated assuming eight hours of physical activity per day, seven days per week.

Results

Directly Measured Physical Activity

The total amount of directly measured occupational physical activity (OPA) and physical activity outside work collected for the study corresponded to 177.266 seconds and 162.694 seconds, respectively.. Each participant's directly measured heart rate data corresponding to OPA were adjusted to five working days of eight hours shifts per week. The heart rate data corresponding to physical activity outside

of work was adjusted to seven days per week and eight hours per day. The median number of moderate and vigorous OPA minutes per week corresponded to 415 and 21 min, respectively. The median moderate and vigorous physical activity outside work was 257 and 10 min per week, respectively (Table 2).

Table 2. Median heart rate and minutes of directly measured occupational physical activity at work and outside of work during a week (n=21)

Physical Activity	Median	Range
Occupational physical activity (OPA)		
Heart rate/min	94	48–192
Sedentary, min/week	357	1–930
Light, min/week	1429	303–1839
Moderate, min/week	415	37–1280
Vigorous, min/week	21	0–897
High, min/week	0	0–7
Physical activity outside of work		
Heart rate/min	86	48–183
Sedentary, min/week	1268	365–2594
Light, min/week	1526	747–2333
Moderate, min/week	257	18–985
Vigorous, min/week	10	0–160
High, min/week	0	0–9

Sedentary (<20 %HRR), light (20<40 %HRR), moderate (40<60 %HRR), vigorous (60<85 %HRR), and high (≥85 %HRR).

Discussion

This study aims to characterise the physical demands measured as the physical activity levels at work among a group of male precast concrete construction workers. The direct measures of OPA indicate that these workers accrue a significant median number of minutes (415 min) at moderate weekly levels. These results are higher than a previous study done among a group of commercial construction workers in Massachusetts (243 min) (Arias et al., 2015). Differences in the type of activities performed by commercial construction and precast concrete workers may explain the variation in the OPA levels accrued in these two groups of workers. Additionally, OPA levels were assessed using different equipment. In the Massachusetts study, an accelerometer was used to count steps. This device has some limitations in capturing physical activity associated with nonambulatory tasks (Matthew, 2005). In this study, using workers' heart rates allow us to estimate the individuals' OPA levels accounting for variations in heart rate levels due to individual characteristics and while performing nonambulatory tasks (American College of Sports Medicine, 2018; Lounana et al., 2007). Despite the differences between both studies, the results show that OPA levels are above the recommended guidelines for leisure time physical activity of 150 min per week (Piercy et al., 2018). The study results suggest that precast concrete workers may have an increased cardiovascular disease (CVD) risk

and all-cause mortality according to previous research studies (Holtermann et al., 2009, 2010).

Using a wearable electronic device (smartwatch) to track heart rate has great potential for characterising the physical demands of construction workers in the field without interfering with their daily activities. Heart rate is a physiological variable extensively used to quantify physical demands because of its association with cardiovascular loads (Åstrand et al., 2003). Using the %HRR allows us to estimate activity levels accounting for variations in heart rate due to individual characteristics, health conditions, and lifestyle factors. Quantifying occupational physical demands by measuring OPA at work is essential to protect workers from adverse health effects. Implementing this device to track workers' physical demands at work will allow the designing of ergonomic interventions to keep those demands within acceptable limits. Reducing those demands requires identifying highly physically demanding tasks and performing an ergonomic evaluation to propose effective interventions to reduce those high demands. Implementing this technology will also allow the assessment of the efficacy of the interventions by comparing the reduction of the OPA before and after an intervention. Given the risk to workers' health, moderate or vigorous levels of physical demands should be restricted to a certain amount of time (e.g., risky if 40% HRR is maintained over 30–60 min, 60% HRR over 30 min, and 30% HRR over eight hours) (Wu & Wang, 2002).

Our results suggest the importance of implementing intervention programs to reduce the negative effect on workers' health of high occupational physical demands. Reducing physical demands in the construction sector (limiting occupational demands to <30%HRR) may significantly impact workers' health and should be part of integrated approaches. A previous study in a group of commercial construction workers addressing working conditions by implementing an ergonomic program and promoting healthy lifestyles showed a significant reduction in physical demands (Peters et al., 2018).

There are some limitations of the study that require attention. The small sample size limits the generalizability of our findings to precast concrete workers with similar demographic characteristics and working practices. However, this group of workers performs similar tasks daily, and the information collected during one week provides a reliable estimator of the median physical demands in this group of workers. Another aspect that must be acknowledged is that the collected information allowed us to quantify only the workers' total physical demands at work. Designing and implementing interventions to reduce physical demands require addressing specific tasks associated with important ergonomic risk factors. However, high-risk tasks will be identified and evaluated in the future as part of an intervention program for these workers. The present study will serve as a baseline for comparisons while evaluating ergonomic interventions' efficacy in reducing the total physical work demands. It is also important to note that the data collected for the analysis only included heart rate data while at work. We did not consider heart rate data associated with commuting which may contribute to an increase in the median minutes at moderate levels of OPA. In this case, our results may underestimate the total OPA. However, our focus is to address factors associated with the high physical demands of the working tasks because those will be more likely to be addressed through ergonomic interventions

and may have a higher impact on the physical work demands than commuting. The results of our study are also limited to male workers. According to the Bureau of Labour Statistics, only 10.9 % of the construction workforce during 2022 corresponded to females (U.S. Bureau of Labor Statistics, 2023). Females' physiological response to the same physical demands may lead to a higher total number of minutes of moderate physical activity. Therefore, our results are limited only to male workers in the precast concrete industry.

Conclusions

This study directly measured the physical demands among precast concrete workers using smartwatches. These devices allow us to continuously monitor heart rate and categorise OPA and physical activity outside work. There is great potential for wearable heart monitoring technology for assessing the physical demands at and outside work among construction workers. The estimation of the number of minutes at different activity levels according to the %HRR will inform intervention programs focused on the reduction of physical demands and health promotion. The information collected through the smartwatches will serve as valuable input for implementing and improving ergonomic programs managing physical demands within acceptable limits in this particular group of construction workers.

Acknowledgements

References

Abas, N.H., Zamzam, M.S.S., & Mohamad, H. (2020). Work-related musculoskeletal disorders (WMSDs) survey among workers at two precast concrete plants in Johor, Malaysia. *Journal of Academic Research in Business and Social Sciences, 10*, 843-856.

American College of Sports Medicine. (2018). *ACSM's guidelines for exercise testing and prescription* (11th ed.): Wolters Kluwer.

Arias, O.E., Caban-Martinez, A.J., Umukoro, P.E., Okechukwu, C.A., & Dennerlein, J.T. (2015). Physical activity levels at work and outside of work among commercial construction workers. *Journal of Occupational and Environmental Medicine, 57*, 73.

Arias, O.E., Groehler, J., Wolff, M., & Choi, S.D. (2023). Assessment of musculoskeletal pain and physical demands using a wearable smartwatch heart monitor among precast concrete construction workers: A field case study. *Applied Sciences, 13*, 2347.

Arndt, V., Rothenbacher, D., Brenner, H., Fraisse, E., Zschenderlein, B., Daniel, U., et al. (1996). Older workers in the construction industry: Results of a routine health examination and a five year follow up. *Occupational Environmental Medicine, 53*(10), 686-691.

Åstrand, P.-O., Rodahl, K., Dahl, H., & StrÃ, S.B. (2003). *Textbook of work physiology: Physiological bases of exercise* (4th ed.). Champaign, USA: Human kinetics.

Boschman, J.S., van der Molen, H.F., Sluiter, J.K., & Frings-Dresen, M.H. (2011). Occupational demands and health effects for bricklayers and construction supervisors: A systematic review. *American Journal of Industrial Medicine, 54*, 55-77.

Brenner, H., & Ahern, W. (2000). Sickness absence and early retirement on health grounds in the construction industry in Ireland. *Occupational Environmental Medicine, 57*, 615-620.

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics, 4*, 195.

Clays, E., De Bacquer, D., Janssens, H., De Clercq, B., Casini, A., Braeckman, L., et al. (2013). The association between leisure time physical activity and coronary heart disease among men with different physical work demands: a prospective cohort study. *European Journal of Epidemiology, 28*, 241-247.

Coenen, P., Korshøj, M., Hallman, D.M., Huysmans, M.A., van der Beek, A.J., Straker, L.M., et al. (2018). Differences in heart rate reserve of similar physical activities during work and in leisure time – A study among Danish blue-collar workers. *Physiology & Behavior, 186*, 45-51.

Gomez, D.M., Coenen, P., Celis-Morales, C., Mota, J., Rodriguez-Artalejo, F., Matthews, C., et al. (2022). Lifetime high occupational physical activity and total and cause-specific mortality among 320 000 adults in the NIH-AARP study: A cohort study. *Occupational and Environmental Medicine, 79*, 147-154.

Holtermann, A., Krause, N., Van Der Beek, A.J., & Straker, L. (2018). The physical activity paradox: Six reasons why occupational physical activity (OPA) does not confer the cardiovascular health benefits that leisure time physical activity does. *British Journal of Sports Medicine, 52*, 149-150.

Holtermann, A., Mortensen, O.S., Burr, H., Sogaard, K., Gyntelberg, F., & Suadicani, P. (2009). The interplay between physical activity at work and during leisure time--risk of ischemic heart disease and all-cause mortality in middle-aged Caucasian men. *Scandinavian Journal of Work, Environment & Health, 35*, 466-474.

Holtermann, A., Mortensen, O.S., Burr, H., Sogaard, K., Gyntelberg, F., & Suadicani, P. (2010). Physical demands at work, physical fitness, and 30-year ischaemic heart disease and all-cause mortality in the Copenhagen Male Study. *Scandinavian Journal of Work, Environment & Health, 36*, 357-365.

Jang, Y., Son, J., & Hwang, S. (2022). Requirements analysis for development of off-site construction project management system: Focusing on precast concrete construction. *Buildings, 12*, 1499.

Lissner, L., Bengtsson, C., Bjorkelund, C., & Wedel, H. (1996). Physical activity levels and changes in relation to longevity. A prospective study of Swedish women. *American Journal of Epidemiology, 143*, 54-62.

Lounana, J., Campion, F., Noakes, T.D., & Medelli, J. (2007). Relationship between% HRmax,% HR reserve,% VO2max, and% VO2 reserve in elite cyclists. *Medicine and Science in Sports and Exercise, 39*, 350-357.

Matthew, C.E. (2005). Calibration of accelerometer output for adults. *Medicine & Science in Sports & Exercise, 37*(11 Suppl), S512-522.

Norton, K., Norton, L., & Sadgrove, D. (2010). Position statement on physical activity and exercise intensity terminology. *Journal of Science and Medicine in Sport, 13*, 496-502.

Piercy, K.L., Troiano, R.P., Ballard, R.M., Carlson, S.A., Fulton, J.E., Galuska, D.A., et al. (2018). The physical activity guidelines for Americans. *Jama, 320*, 2020-2028.

Tanaka, H., Monahan, K.D., & Seals, D.R. (2001). Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology, 37*, 153-156.

U.S. Bureau of Labor Statistics. (2023). Labor force statistics from the current population survey. from https://www.bls.gov/cps/cpsaat18.htm

Wu, H.-C., & Wang, M.-J.J. (2002). Relationship between maximum acceptable work time and physical workload. *Ergonomics, 45*, 280-289.

Effects of modified leg mechanics on cognitive performance and workload during dual-task walking

Norman Riedel, Michael Herzog, Thorsten Stein, & Barbara Deml
Karlsruhe Institute of Technology
Germany

Abstract

Mobile exoskeletons as assistive walking devices can modify body mechanics due to their own weight and restricted range of motion, becoming a potential physical and cognitive load for the user when the support is insufficient, or the power supply has failed. This study investigates the effect of modified leg mechanics on cognitive-motor interference in a controlled dual-task walking setting. Sixteen healthy young adults walked on a treadmill at their preferred walking speed with and without weights attached to their thighs and shanks while performing a visual-verbal Stroop test and a subtraction task. The dependent variables examined were performance on secondary tasks (correct response rates and dual-task effects) and perceived physical and cognitive workload (NASA-TLX). Results show a significant decrease in cognitive performance when walking with weights in the subtraction task, but not in the Stroop test. This suggests that walking with modified leg mechanics shares similar complex neural networks activated in particular during the subtraction task. Perceived cognitive workload increased for both tasks when walking with the weights. These results indicate that modified leg mechanics may impose a cognitive load. Additional analysis of the motion data may provide further insight into task prioritisation during walking with modified leg mechanics.

Introduction

Lower limb exoskeletons as assistive walking devices are being studied and developed for a wide variety of applications (see review in Young & Ferris, 2017). There are a few lower limb exoskeletons, especially in the field of rehabilitation, which are already being used in practical applications using predetermined trajectories (see review in Shi et al., 2019). The trajectories are collected from healthy persons normal gait data and restrict the user's motion accordingly to these trajectories. However, there is still a need for research and development before mobile exoskeletons for daily activities and highly dynamic applications successfully move from the laboratory environment to the field (Young & Ferris, 2017). As this transition gets closer, human factor aspects must receive greater attention in the development and evaluation of exoskeletons (Stirling et al., 2019). Davis et al. (2020) identified three key research areas to inform human-centred design of exoskeletons: user acceptance, physical and mental load (dual demands), and biomechanical effects (e.g., kinematics, kinetics). To date, most research has quantified the effects of lower limb exoskeletons using

biomechanical and physiological indicators. In contrast, the investigation of cognitive workload in human-exoskeleton interaction has barely been considered (Pinto-Fernandez et al., 2020). The analysis of cognitive workload in the context of human-exoskeleton interaction is crucial, because the user's cognitive abilities must be maintained such that operational activities can be performed appropriately (Stirling et al., 2020). A field study with soldiers demonstrated that wearing a lower limb exoskeleton resulted in slowed reaction times in a visual search task for some subjects (Bequette et al., 2018, 2020). This study provides preliminary hints that wearing an exoskeleton during early adaptation may place a cognitive load on the user. The authors suggest that some subjects showed increased cognitive workload due to the interaction with the mechanical properties (weight, bulk, range of motion) and some due to the actively applied assistance (actuators, control strategy).

Dual-task walking

There are different methods to assess cognitive workload. In the context of assistive wearable devices, dual-task paradigms and subjective assessments are predominantly used (Marchand et al., 2021). Dual-task paradigms are of interest in the study of human-exoskeleton interaction, as simultaneous cognitive and motor tasks have been shown to be interdependent (Al-Yahya et al., 2011; Woollacott & Shumway-Cook, 2002). In this context, the literature refers to cognitive-motor interference. Interferences occur when the capacity of limited cognitive resources is reached and is indicated by reductions in performance in the motor or cognitive task, or even in both tasks. The investigation of cognitive-motor interferences is widely used in clinical and epidemiological studies to investigate the influence of age-related factors and neurological diseases on cognitive and motor performance (Beurskens & Bock, 2012; Raffegeau et al., 2019). Motor control to maintain postural stability is thought to require more conscious attention in elderly than in healthy young adults (Lundin-Olsson et al., 1997). However, significant effects of dual task walking on motor or cognitive parameters have also been found in healthy young adults (Patel et al., 2014; Szturm et al., 2013; Yogev-Seligmann et al., 2010).

The extent of cognitive-motor interference is also determined by how the brain prioritises the individual tasks. The allocation of cognitive resources or task prioritisation depends on various factors, such as individual characteristics or task complexity (Kelly et al., 2012; Yogev-Seligmann et al., 2008, 2012). The traditional theory to explain task prioritisation in dual-task walking is the *posture first* principle, which postulates that healthy subjects prioritise the motor task over the cognitive task to avoid threats like falling when no specific instructions are given (Shumway-Cook et al., 1997). More recently, research suggests a more complex interplay of individual factors. According to the integrated model of task prioritisation of Yogev-Seligmann et al. (2012), two main factors contribute to the choice of the task prioritisation strategy. One factor is the *postural reserve* "that reflects the individual's capability to respond most effectively to a postural threat". The second factor is *hazard estimation* that involves different aspects of self-awareness such as the ability to estimate environmental hazards and being aware of self-limitations. These factors together with other factors such as expertise, personality and the nature of the secondary task determine the choice of the prioritisation strategy. Healthy young adults who have a

high postural reserve and high hazard estimation prioritise the cognitive task without reductions in gait performance (Yogev-Seligmann et al., 2012). However, more complex environments or motor tasks can demand the postural reserve, resulting in a shift of attention to the motor task to avoid potentially critical hazards. As a result, less cognitive resources are available to perform the secondary task, which can lead to a reduction in performance (Bequette et al., 2020). Studies using neural correlates support this hypothesis by reporting significant changes in brain activity with varying complexity of the cognitive task (Hill et al., 2013) or motor task (Reiser et al., 2019).

Present study

Using a dual-task walking paradigm, the present study investigates under controlled laboratory conditions the extent to which modified leg mechanics affect motor and cognitive performance while walking on a treadmill. Weight cuffs bilaterally attached to the thighs and shanks of the participants manipulate the mechanical properties and add complexity to the motor task.

The present paper shows preliminary results examining cognitive performance in secondary tasks and perceived workload. It was hypothesised that walking with modified leg mechanics demands the postural reserve and consequently reduces cognitive performance compared to normal walking (H1) and increases perceived cognitive (H2) and physical workload (H3) compared to normal walking or sitting.

Method

Participants

Sixteen healthy young adults (age: M = 24.1, SD = 3.4; height: M = 172.9 cm, SD = 8.8 cm; mass: M = 65.1 kg, SD = 10.4 kg; sex: 9 female, 7 male) were recruited among students of the Karlsruhe Institute of Technology. Participants completed a medical history screening and were excluded from the study if musculoskeletal, neurological, or cardiovascular disease or red-green weakness was present that could affect walking secondary task performance. Written informed consent was obtained in accordance with approved institutional review board procedures. The ethics committee of the Karlsruhe Institute of Technology approved the study.

Experimental procedures

In this experiment with a 3x3 within-subjects design, participants walked on a treadmill with and without weight cuffs bilaterally attached to the thigh and shank (*Motor Condition*: sitting, unloaded walking, loaded walking) and simultaneously performed cognitively demanding secondary tasks (*Cognitive Condition*: no secondary task, visual-verbal Stroop test [STR], descending subtraction task [SUB]). Table 1 gives an overview of the experimental conditions including four single task and four dual task conditions.

Table 1. Overview of experimental conditions. ST – single task; DT – dual task

	Sitting	Unloaded walking	Loaded walking
No secondary task	-	Control motor condition (ST_3)	Control motor condition (ST_4)
Stroop test (STR)	Control cognitive condition (ST_1)	DT_1	DT_3
Subtraction task (SUB)	Control cognitive condition (ST_2)	DT_2	DT_4

First, for familiarisation, the two secondary tasks were performed in a seated position. After a six-minute familiarisation period of walking on the treadmill (Meyer et al., 2019) and the determination of the preferred gait speed according to the procedure proposed by Jordan et al. (2007), the unloaded and loaded walking sessions were carried out in a counterbalanced order. Finally, the two secondary tasks were performed again in a seated position as single task control condition (Figure 1). These control conditions are needed to compare task performance under single and dual task conditions.
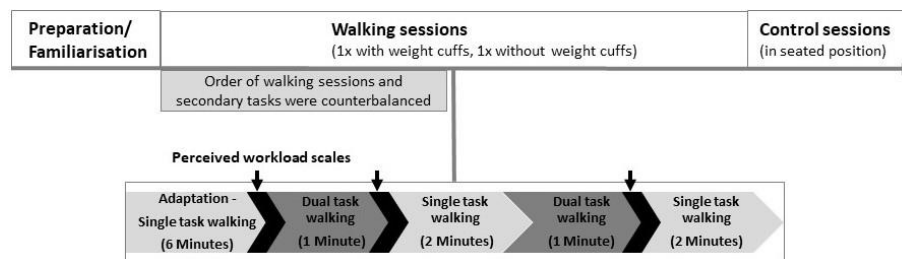

Figure 1. Experimental protocol. Black arrows indicate the times at which the perceived workload subscales were queried during the walking sessions.

Motor main task

As a motor task, participants walked on a treadmill (h/p/cosmos, model: saturn 300/100) with and without weight cuffs bilaterally attached to the thigh and shank, each weighing 2.25 kg (total weight: 9 kg). The individual preferred gait speed (M = 4.1 km/h, SD = 0.3 km/h) was kept constant for both sessions. A custom developed hip belt was used to attach the weight cuffs to the thigh. Two Velcro straps were attached to the hip belt on each side of the leg, into which the weight cuffs could be hooked. Figure 2 illustrates the experimental setup, the positions of the weight cuffs and shows how they were attached to a participant along with the hip belt. The Velcro straps could be adjusted in height so that the lower edge of the weight cuffs was positioned 10 cm above the knee joint axis for each participant. A safety harness secured participants while walking on the treadmill. Whole-body movements were recorded using an infrared camera system (Vicon Motion Systems Ltd.) with sixteen cameras. The marker setup includes 56 markers. In the present paper, motor performance is not investigated. However, extensive data sets are available and may be analysed further.
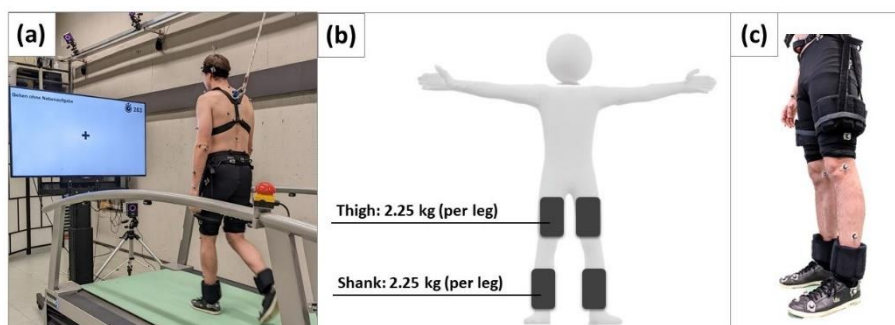
Figure 2. (a) Experimental setup. (b) Schematic representation of the positions of the weight cuffs. (c) Representation of the attachment of the weight cuffs to a participant.

Cognitive secondary task

In a meta-analysis, Al-Yahya et al. (2011) found that cognitive tasks disturb walking more when internal rather than external interfering factors are involved. Based on the type of mental processes required to perform the tasks, the authors established classifications. Two cognitive tasks from different classifications were used in this study: A visual-verbal version of the Stroop test (STR) (Stroop, 1935) as a decision-making task involving external interfering factors and a descending subtraction task (SUB) as a mental tracking task involving only internal interfering factors.

In STR, a 10x10-matrix of colour words (red, blue, green, yellow) with incongruent word and colour information was presented for 60 seconds (Figure 3). To avoid learning effects, there were five different matrices, which were presented in random order. Participants were instructed to name the respective font colour of the words as quickly as possible and without error. Participants started in the left top corner and continued column wise to the right. Cognitive performance was determined by the correct response rate (CRR). According to Galletly and Brauer (2005) this is calculated by multiplying the response rate (responses per second) and the accuracy (percent of correct responses).

In SUB, a random three-digit number between 201 and 999 was presented. The participants were asked to serially subtract the number 7 for 60 seconds starting with the presented number. The CRR was again used as the outcome measure. Both tasks were performed in a seated position (cognitive control condition) and during unloaded and loaded walking. For familiarisation, the participants first completed a 20-second test trial, followed by a 60-second training trial in a seated position for both tasks. While seated, the tasks were presented on a 22-inch monitor at a distance of approximately 80 cm from the participant. While walking on the treadmill, a 65-inch monitor was used at a distance of 240 cm from the participant. The monitor height was set so that the top edge was at the eye level of the participant. Participants' responses were recorded for analysis via a recording device (Sony, model: ICD-UX570) with a clip-on microphone (Phillips, model: LFH9173/00).

ROT	GELB	ROT	BLAU	ROT	ROT	GRÜN	GELB	BLAU	GRÜN
BLAU	BLAU	GRÜN	GELB	GELB	BLAU	GELB	ROT	ROT	ROT
ROT	ROT	ROT	GRÜN	BLAU	GELB	GRÜN	BLAU	ROT	BLAU
GRÜN	GELB	GRÜN	ROT	GRÜN	GRÜN	BLAU	GELB	GRÜN	GELB
GELB	GELB	ROT	BLAU	ROT	ROT	ROT	GRÜN	GELB	BLAU
ROT	GRÜN	GRÜN	GELB	GELB	GELB	GRÜN	GELB	GRÜN	GRÜN
BLAU	BLAU	GELB	ROT	ROT	GRÜN	BLAU	ROT	BLAU	GRÜN
GRÜN	ROT	BLAU	GRÜN	GRÜN	ROT	ROT	BLAU	GELB	BLAU
GELB	GELB	GRÜN	GELB	BLAU	GELB	GELB	GRÜN	ROT	GELB
ROT	GRÜN	ROT	BLAU	GRÜN	BLAU	BLAU	GELB	BLAU	ROT

Figure 3. Exemplary 10x10-matrix of colour words as used in the study.

Walking protocol

Figure 1 shows the protocol of a walking session. Both walking sessions (unloaded and loaded walking) started with a six-minute block of single task walking (motor control condition). This block controlled for possible adaptation effects to ensure that participants did not have to use cognitive resources to adapt to unfamiliar walking conditions. Noble and Prentice (2006) showed that adaptation is completed after 45-50 strides when walking with unilateral weights. This was followed by the first secondary task for 60 seconds. To counteract cognitive fatigue, a two-minute block of single task walking followed before the second secondary task was presented for 60 seconds. Sessions ended with another two-minute block of single task walking. In total, this protocol lasted 12 minutes each. The order of walking sessions and appearance of secondary tasks were counterbalanced to account for fatigue and learning effects. No specific instructions were provided regarding which task to prioritise.

Subjective measures

Immediately after each of the eight experimental conditions (Table 1), the two subscales *Mental Demand* and *Physical Demand* of the NASA-TLX (Hart & Staveland, 1988) were queried. Here, participants were presented with the subscale description along with the scale (0 – low demand, 100 – high demand) and had to verbally indicate the number that was appropriate for them. While walking on the treadmill, participants had 30 seconds per subscale to give a response. Figure 1 indicates the time of the queries with black arrows.

Dependent variables and statistics

The effects of the experimental conditions on cognitive performance and perceived workload were determined using the CRR and ratings of NASA-TLX subscales, respectively. To assess relative change of the CRR the dual task effects (DTE) were calculated (Kelly et al., 2010). Negative values represent a reduction under dual task conditions; positive values represent an improvement under dual task conditions. Since a lower CRR represents a reduction in task performance, the DTEs are calculated as follows:

$$DTE = \frac{(CRR_{dual\ task} - CRR_{single\ task})}{CRR_{single\ task}} \times 100\%$$ (1)

The Kolmogorov-Smirnov test was used to test the data for normal distribution. In the ratings on perceived workload, the assumption of normal distribution was violated in three out of twelve conditions. Since rmANOVAs are considered robust to violations of the normal distribution, the parametric tests were nevertheless used (Vasey & Thayer, 1987). Homogeneity of variances was tested using Levene's test based on the median and homogeneity of covariances was calculated by Box's test. Sphericity of the data was tested with the Mauchly test. When this assumption was violated, degrees of freedom were adjusted with the Greenhouse-Geisser correction. For all statistics, significance level was set a priori as $\alpha = .05$. Bonferroni correction was applied to post hoc comparisons. Effect sizes are given as partial eta squares with $\eta p^2 = .01$ indicating a small effect, $\eta p^2 = .06$ a medium effect and $\eta p^2 = .14$ a large effect (Cohen, 1988). Statistics were evaluated using SPSS 28 (IBM Statistics Armonk).

A 2x2x2-mixed-ANOVA with within-factors *Task Condition* (STR, SUB) and *Motor Condition* (Unloaded walking, loaded walking) and between-factor *Session Order* (Start with unloaded walking, start with loaded walking) was conducted to test differences in DTE (H1). Since the analysis of the descriptive data indicated order effects, this was exploratively included in the statistical model. The between-factor *Session Order* was used to investigate whether it makes a difference if the walking session is started with unloaded or loaded walking. Two 2x3-rmANOVAs were conducted with within-factors *Task Condition* (STR, SUB) and *Motor Condition* (Sitting, unloaded walking, loaded walking) to test differences of perceived cognitive (H2) and physical workload (H3).

Results

Table 2 shows absolute and relative values for cognitive performance variables and for perceived cognitive and physical workload in each single task and dual task condition.

Cognitive performance

Analysis of DTE showed no significant main effects of *Task Condition* ($F(1, 14) < 1$, $p = .974$, $\eta p^2 < .000$), *Walking Condition* ($F(1, 14) = 2.79$, $p = .117$, $\eta p^2 = .166$) and *Session Order* ($F(1, 14) = 2.20$, $p = .160$, $\eta p^2 = .136$).

There was a significant interaction effect between *Walking Condition* and *Task Condition* ($F(1, 14) = 5.65$, $p = .032$, $\eta p^2 = .287$). This indicates that the cognitive performance in the different walking conditions differed according to the type of task performed. Reviewing the interaction graph in Figure 4a, this suggests that cognitive performance in SUB decreases from unloaded to loaded walking, whereas cognitive performance in STR shows no differences from unloaded to loaded walking.

Table 2. Absolute and relative (%) measures of cognitive performance and perceived cognitive and physical workload. Values represent mean (standard deviation). DTE - dual task effects.

	Sitting	*Unloaded walking*	*Loaded walking*
Performance – Stroop-Test			
Correct response rate	1.200 (0.262)	1.155 (0.225)	1.169 (0.242)
Correct response rate DTE (%)	-	-3.09 (9.36)	-2.32 (5.90)
Starting w. unloaded walking (n=8)	-	-9.37 (2.47)	-4.05 (2.06)
Starting w. loaded walking (n=8)	-	3.19 (2.47)	-.58 (2.06)
Performance – Subtraction-Task			
Correct response rate	.277 (0.106)	.288 (0.113)	.251 (0.119)
Correct response rate DTE (%)	-	4.62 (19.02)	-9.75 (24.81)
Starting w. unloaded walking (n=8)	-	6.97 (6.91)	5.20 (7.11)
Starting w. loaded walking (n=8)	-	2.28 (6.91)	-24.70 (7.11)
Workload – Stroop-Test			
Cognitive workload	38.0 (23.6)	35.1 (25.1)	40.4 (28.2)
Physical workload	3.5 (4.7)	15.1 (8.6)	37.2 (21.6)
Workload – Subtraction-Task			
Cognitive workload	49.6 (24.9)	48.9 (22.8)	56.4 (25.6)
Physical workload	3.7 (5.0)	15.1 (7.3)	37.3 (22.7)
Workload – No secondary task			
Cognitive workload	-	6.7 (9.8)	8.5 (8.6)
Physical workload	-	14.1 (9.3)	37.1 (21.8)

There was a significant interaction effect between *Task Condition* and *Session Order* (F (1, 14) = 9.03, p = .009, ηp^2 = .392). This indicates that cognitive performance in STR and SUB differed depending on the session order. Figure 4b shows the mean DTE for each dual task condition for the group of participants who started with unloaded walking and the group of participants who started with loaded walking. The interaction graph revealed a disordinal interaction suggesting that starting with loaded walking strongly reduces cognitive performance for SUB and only slightly for STR. Starting with unloaded walking seems to have the opposite effect, suggesting a reduced cognitive performance in STR and a slightly increased cognitive performance in SUB.

The interaction effect between *Walking Condition* and *Session Order* showed a statistical trend (F (1, 14) = 4.43, p = .054, ηp^2 = .240). This indicates that cognitive performance during unloaded and loaded walking tend to differ depending on the session order (Figure 4b). The interaction graph revealed a disordinal interaction suggesting that starting with loaded walking decreases the cognitive performance during loaded walking and has only small positive effects on cognitive performance during unloaded walking. In contrast to this, starting with unloaded walking has only small effects on cognitive performance for both walking conditions.
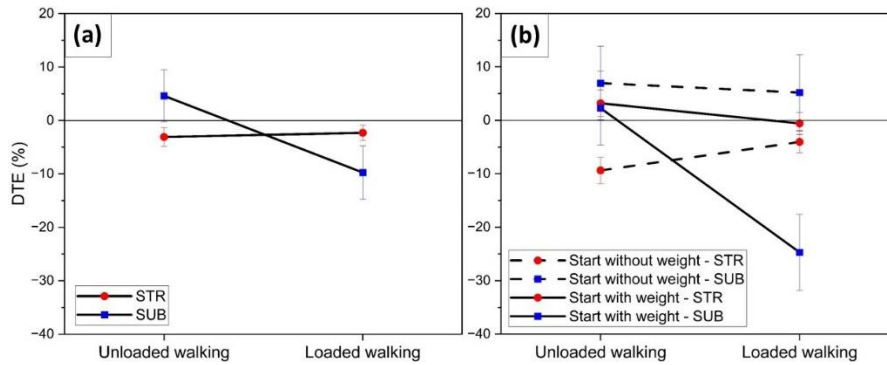
Figure 4. (a) Mean dual task effects (DTE) for STR and SUB in the unloaded and loaded walking condition. (b) Mean DTE for STR and SUB in the unloaded and loaded walking condition for the group of participants who started with unloaded walking (n=8) and the group of participants who started with loaded walking (n=8). Error bars reflect 95% CI.

The three-way interaction *Walking Condition* x *Task Condition* x *Session Order* was not significant (F (1, 14) = 1.60, p = .227, ηp^2 = .102).

Perceived cognitive workload

There was a significant main effect of *Task Condition* on the perceived cognitive workload (F (1, 15) = 9.257, p = .008, ηp^2 = .382). Reviewing the mean ratings indicated that the SUB was perceived as more cognitively demanding than STR (Figure 5a). There was also a significant main effect of the *Motor Condition* on the perceived cognitive workload (F (2, 30) = 4.155, p = .026, ηp^2 = .217). Post-hoc analysis revealed that perceived cognitive workload was not significantly different from control condition to unloaded walking (MDiff = 1.81, 95%-CI[-2.46, 6.08], p = .813) and to loaded walking (MDiff = -4.59, 95%-CI[-11.78, 2.59], p = .317). The difference of perceived cognitive workload between unloaded and loaded walking showed a statistical trend (MDiff = -6.41, 95%-CI[-13.07, 0.26], p = .062), suggesting a higher perceived cognitive workload in loaded compared to unloaded walking. There was no significant interaction effect between *Task Condition* and *Motor Condition* (F (2, 30) < 1, p = .474, ηp^2 = .049).

Perceived physical workload

There was no significant main effect of *Task Condition* on the perceived physical workload (F (1, 15) < 1, p = .871, ηp^2 = .002), indicating that the type of secondary task had no effect on the perceived physical workload (Figure 5b). There was a significant main effect of the *Motor Condition* on the perceived physical workload (F (1.091, 16.366) = 27.864, p < .001, ηp^2 = .650). Post-hoc analysis revealed that perceived physical workload significantly increased from control condition to unloaded walking (MDiff = -11.53, 95%-CI[-15.22, -7.85], p < .001) and to loaded walking (MDiff = -33.63, 95%-CI[-48.76, -18.49], p < .001). The differences between unloaded and loaded walking were also significant (MDiff = -22.09, 95%-CI[-36.70, -7.49], p = .003). There was no significant interaction effect between *Task Condition* and *Motor Condition* (F (2, 30) < 1, p = .991, ηp^2 = .001).
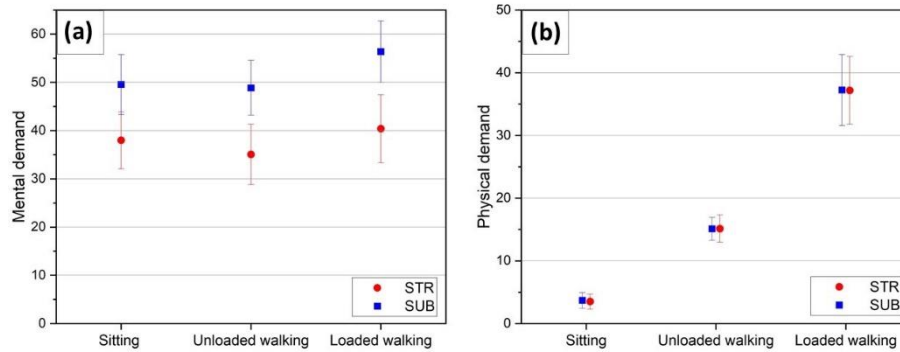
Figure 5. (a) Mean ratings of mental demand. (b) Mean ratings of physical demand. Error bars reflect 95% CI.

Discussion

The present paper investigated effects of modified leg mechanics on cognitive performance and perceived workload while walking on a treadmill using a dual-task paradigm. It was hypothesised that walking with modified leg mechanics reduces cognitive performance compared to normal walking (H1) and increases perceived cognitive (H2) and physical workload (H3) compared to normal walking or sitting. Additionally, possible order effects on cognitive performance were exploratively investigated.

Cognitive performance (H1)

Cognitive performance in the secondary tasks was assessed via the correct response rates. To account for relative changes from dual task walking compared to single task while sitting, the dual task effects were calculated. The results provide mixed support for the hypotheses H1. There is no simple main effect of modified leg mechanics on cognitive performance in the different dual task conditions. However, a significant interaction effect suggests that walking with modified leg mechanics decreased cognitive performance in SUB, but not in STR, indicating an increased cognitive load due to the added weights (Figure 4a). In agreement with the task prioritisation framework proposed by Yogev-Seligmann et al. (2012), this finding suggests that the type of secondary task and the complexity of the motor task as a threat to postural stability affects allocation of attention in healthy young adults. Walking with modified leg mechanics seems to threaten postural stability, so maintaining the stability of the otherwise largely automated process of walking requires conscious attention. This allocation of attention to walking could explain the reduction in performance in the secondary task. According to a meta-analysis by Al-Yahya et al. (2011), tasks that require memorizing information and simultaneously performing internal, mental processes, such as SUB, interfere stronger with gait performance than tasks involving external stimuli. Mental tracking tasks, such as SUB, appear to share similar complex neural networks to those activated during walking (Al-Yahya et al., 2011). In particular, the prefrontal cortex was found to be involved in locomotion and dual tasking (Hamacher et al., 2015; Holtzer et al., 2011). Hill et al. (2013) reported that walking while serially subtracting 7 increased the prefrontal cortex activity compared

to walking while counting backwards by 1 in young adults. This supports the results of the present paper that SUB required a significant amount of cognitive resources that may have interfered with loaded, but not with unloaded walking.

In fact, unloaded walking slightly improved performance in SUB compared to single task. Similar dual task benefits in normal walking were found in a previous study (Yogev-Seligmann et al., 2010). Practice effects can be excluded because all experimental conditions were counterbalanced, the starting number was randomised in SUB and the single task session was always performed after the dual task sessions. Therefore, the activity of (unloaded) walking itself may be the reason for the improved performance in SUB. According to the Yerkes-Dodson law (Yerkes & Dodson, 1908), performance increases with physiological or mental arousal, while performance decreases when the level of arousal is too low, as may be the case in the seated condition, or too high, as may be the case in the loaded walking condition.

As the present paper has not investigated motor dual task effects, no conclusive statements can be made about cognitive-motor interferences and task prioritisation strategies. For example, in agreement with the results of the present study, Patel et al. (2014) reported higher cognitive costs for the subtraction task compared to the Stroop test. However, they also reported higher motor costs for the Stroop test compared to the subtraction task. Additional analysis of the motion data may provide further insight into task prioritisation strategies.

The descriptive data of the cognitive performance showed that especially subjects who started with loaded walking showed reduced performance in SUB during loaded walking. For this reason, the between-subjects factor *Session Order* was included in the statistical model as an exploratory measure. In fact, interaction effects could be found that suggest an influence of the session order (Figure 4b). In particular, participants who started with the most complex dual task condition (loaded walking with SUB) showed cognitive performance reductions during loaded walking, while participants who started with unloaded walking even showed a little performance improvement. The interplay of novelty and complexity of the dual-task condition might have been perceived as an increased hazard for postural stability, which requires an intact hazard estimation. Interestingly, participants who started with unloaded walking showed slightly reduced dual task performance in STR while participants who started with loaded walking showed almost no change in dual task performance compared to single task. It is possible that different prioritisation strategies were adopted depending on the complexity of the motor task in the first attempt. Another explanation are the individual differences in the cognitive and motor abilities of the participants, which may mask the effects of the experimental manipulation due to the small sample in the present study (Bequette et al., 2020).

Perceived Workload (H2 & H3)

Perceived physical and cognitive workload were assessed with the respective subscales of the NASA-TLX in all experimental conditions. As hypothesised the perceived physical workload increased significantly from sitting to unloaded walking to loaded walking, validating the experimental manipulation. Perceived cognitive workload showed a statistical trend, indicating an increased cognitive workload in

loaded walking compared to unloaded walking. Bequette et al. (2020) reported similar results: Completing an obstacle course with a powered and unpowered exoskeleton was rated as significantly more cognitively demanding than completing the course without an exoskeleton. The course involved more complex motor tasks, which is presumably why the influence of the modified leg mechanics due to the exoskeleton on the perceived cognitive workload is stronger than in the present study.

Conclusion

The present paper suggests an increased cognitive workload during walking with modified leg mechanics in the early adaptation phase. However, cognitive performance reductions do not occur in general, but seem to be caused by an interplay of external factors (e.g., complexity of the motor/cognitive task, task order). The perceived cognitive workload also increases, although not significantly. The results highlight the relevance of assessing cognitive workload when evaluating exoskeletons and other wearable devices. In order to be able to make further statements about task prioritisation and attention allocation, motor performance must be evaluated in addition to cognitive performance.

Acknowledgement

References

Al-Yahya, E., Dawes, H., Smith, L., Dennis, A., Howells, K., & Cockburn, J. (2011). Cognitive motor interference while walking: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *35*, 715–728. https://doi.org/10.1016/j.neubiorev.2010.08.008

Bequette, B., Norton, A., Jones, E., & Stirling, L. (2018). The Effect of a Powered Lower-Body Exoskeleton on Physical and Cognitive Warfighter Performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *Volume 1* (pp. 1663–1667). https://doi.org/10.1177/1541931218621377

Bequette, B., Norton, A., Jones, E., & Stirling, L. (2020). Physical and Cognitive Load Effects Due to a Powered Lower-Body Exoskeleton. *Human Factors*, *62*, 411–423. https://doi.org/10.1177/0018720820907450

Beurskens, R., & Bock, O. (2012). Age-related deficits of dual-task walking: A review. *Neural Plasticity*, *2012*, 131608. https://doi.org/10.1155/2012/131608

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hoboken: Taylor and Francis.

Davis, K.G., Reid, C.R., Rempel, D.D., & Treaster, D. (2020). Introduction to the Human Factors Special Issue on User-Centered Design for Exoskeleton. *Human Factors*, *62*, 333–336. https://doi.org/10.1177/0018720820914312

Galletly, R., & Brauer, S.G. (2005). Does the type of concurrent task affect preferred and cued gait in people with Parkinson's disease? *Australian Journal of Physiotherapy*, *51*, 175–180. https://doi.org/10.1016/S0004-9514(05)70024-6

Hamacher, D., Herold, F., Wiegel, P., Hamacher, D., & Schega, L. (2015). Brain activity during walking: A systematic review. *Neuroscience and Biobehavioral Reviews*, *57*, 310–327. https://doi.org/10.1016/j.neubiorev.2015.08.002

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139-183.

Hill, A., Bohil, C., Lewis, J., & Neider, M. (2013). Prefrontal Cortex Activity During Walking While Multitasking. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Volume 57* (pp. 1224–1228). https://doi.org/10.1177/1541931213571272

Holtzer, R., Mahoney, J.R., Izzetoglu, M., Izzetoglu, K., Onaral, B., & Verghese, J. (2011). fNIRS study of walking and walking while talking in young and old individuals. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 66, 879-887.

Jordan, K., Challis, J.H., & Newell, K.M. (2007). Walking speed influences on gait cycle variability. *Gait & Posture*, *26*, 128–134. https://doi.org/10.1016/j.gaitpost.2006.08.010

Kelly, V.E., Eusterbrock, A.J., & Shumway-Cook, A [Anne] (2012). A review of dual-task walking deficits in people with Parkinson's disease: Motor and cognitive contributions, mechanisms, and clinical implications. *Parkinson's Disease*, *2012*. https://doi.org/10.1155/2012/918719

Kelly, V.E., Janke, A.A., & Shumway-Cook, A. (2010). Effects of instructed focus and task difficulty on concurrent walking and cognitive task performance in healthy young adults. *Experimental Brain Research*, *207*, 65–73. https://doi.org/10.1007/s00221-010-2429-6

Lundin-Olsson, L., Nyberg, L., & Gustafson, Y. (1997). "Stops walking when talking" as a predictor of falls in elderly people. *The Lancet*, *349*, 617. https://doi.org/10.1016/S0140-6736(97)24009-2

Marchand, C., Graaf, J.B. de, & Jarrassé, N. (2021). Measuring mental workload in assistive wearable devices: A review. *Journal of Neuroengineering and Rehabilitation*, *18*, 1-15. https://doi.org/10.1186/s12984-021-00953-w

Meyer, C., Killeen, T., Easthope, C.S., Curt, A., Bolliger, M., Linnebank, M., Zörner, B., & Filli, L. (2019). Familiarization with treadmill walking: How much is enough? *Scientific Reports*, *9*, 5232. https://doi.org/10.1038/s41598-019-41721-0

Noble, J.W., & Prentice, S.D. (2006). Adaptation to unilateral change in lower limb mechanical properties during human walking. *Experimental Brain Research*, *169*, 482–495. https://doi.org/10.1007/s00221-005-0162-3

Patel, P., Lamar, M., & Bhatt, T. (2014). Effect of type of cognitive task and walking speed on cognitive-motor interference during dual-task walking. *Neuroscience*, *260*, 140–148. https://doi.org/10.1016/j.neuroscience.2013.12.016

Pinto-Fernandez, D., Torricelli, D., Del Sanchez-Villamanan, M.C., Aller, F., Mombaur, K., Conti, R., Vitiello, N., Moreno, J.C., & Pons, J.L. (2020). Performance Evaluation of Lower Limb Exoskeletons: A Systematic Review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering : A Publication of the IEEE Engineering in Medicine and Biology Society*, *28*(7), 1573–1583. https://doi.org/10.1109/TNSRE.2020.2989481

Raffegeau, T.E., Krehbiel, L.M., Kang, N., Thijs, F.J., Altmann, L.J.P., Cauraugh, J.H., & Hass, C.J. (2019). A meta-analysis: Parkinson's disease and dual-task walking. *Parkinsonism & Related Disorders*, *62*, 28–35. https://doi.org/10.1016/j.parkreldis.2018.12.012

Reiser, J.E., Wascher, E., & Arnau, S. (2019). Recording mobile EEG in an outdoor environment reveals cognitive-motor interference dependent on movement complexity. *Scientific Reports*, *9*, 13086. https://doi.org/10.1038/s41598-019-49503-4

Shi, D. Zhang, W., Zhang, W., & Ding, X. (2019). A Review on Lower Limb Rehabilitation Exoskeleton Robots. *Chinese Journal of Mechanical Engineering*, *32*. https://doi.org/10.1186/s10033-019-0389-8

Shumway-Cook, A., Woollacott, M., Kerns, K.A., & Baldwin, M. (1997). The effects of two types of cognitive tasks on postural stability in older adults with and without a history of falls. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *52*, 232-240. https://doi.org/10.1093/gerona/52A.4.M232

Stirling, L., Kelty-Stephen, D., Fineman, R., Jones, M.L.H., Daniel Park, B.-K., Reed, M.P., Parham, J., & Choi, H.J. (2020). Static, Dynamic, and Cognitive Fit of Exosystems for the Human Operator. *Human Factors*, *62*, 424–440. https://doi.org/10.1177/0018720819896898

Stirling, L., Siu, H.C., Jones, E., & Duda, K. (2019). Human Factors Considerations for Enabling Functional Use of Exosystems in Operational Environments. *IEEE Systems Journal*, *13*, 1072–1083. https://doi.org/10.1109/JSYST.2018.2821689

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18, 643.

Szturm, T., Maharjan, P., Marotta, J.J., Shay, B., Shrestha, S., & Sakhalkar, V. (2013). The interacting effect of cognitive and motor task demands on performance of gait, balance and cognition in young adults. *Gait & Posture*, *38*, 596–602. https://doi.org/10.1016/j.gaitpost.2013.02.004

Vasey, M.W., & Thayer, J.F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, *24*, 479–486. https://doi.org/10.1111/j.1469-8986.1987.tb00324.x

Woollacott, M., & Shumway-Cook, A. (2002). Attention and the control of posture and gait: a review of an emerging area of research. *Gait & Posture*, *16*, 1–14. https://doi.org/10.1016/S0966-6362(01)00156-4

Yerkes, R.M., & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18, 459-482.

Yogev-Seligmann, G., Hausdorff, J.M., & Giladi, N. (2008). The role of executive function and attention in gait. *Movement Disorders: Official Journal of the Movement Disorder Society*, *23*, 329-342. https://doi.org/10.1002/mds.21720

Yogev-Seligmann, G., Hausdorff, J.M., & Giladi, N. (2012). Do we always prioritize balance when walking? Towards an integrated model of task prioritization. *Movement Disorders: Official Journal of the Movement Disorder Society*, *27*, 765–770. https://doi.org/10.1002/mds.24963

Yogev-Seligmann, G., Rotem-Galili, Y., Mirelman, A., Dickstein, R., Giladi, N., & Hausdorff, J.M. (2010). How does explicit prioritization alter walking during

dual-task performance? Effects of age and sex on gait speed and variability. *Physical Therapy*, *90*, 177–186. https://doi.org/10.2522/ptj.20090043

Young, A.J., & Ferris, D.P. (2016). State of the art and future directions for lower limb robotic exoskeletons. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25, 171-182.