# Understanding Human Behaviour in Complex Systems

de Waard, Dick ; Toffetti, Antonella; Pietrantoni, Luca; Franke, Thomas; Petiot, Jean-François; Dumas, Cédric; Botzer, Assaf; Onnasch, Linda; Milleville, Isabelle; Mars, Franck

# *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference*

*Understanding Human Behaviour in Complex Systems*

*Edited by*
***Dick de Waard, Antonella Toffetti, Luca Pietrantoni, Thomas Franke, Jean-François Petiot, Cédric Dumas, Assaf Botzer, Linda Onnasch, Isabelle Milleville, and Franck Mars***
**ISSN 2333-4959 (online)**

*Please refer to contributions as follows:*
[Authors] (2020), [Title]. 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. pagenumbers). Downloaded from http://hfes-europe.org (ISSN 2333-4959)


*Available as open access*

*Published by HFES*

# Contents

# Disentangling the enigmatic slowing effect of microgravity on sensorimotor performance

*Bernhard Weber, Martin Stelzer, & Cornelia Riecke*
*German Aerospace Center, Institute of Robotics and Mechatronics, Oberpfaffenhofen, Germany*

**Abstract**

The success of many space missions depends on astronauts' performance. Yet, prior research documented that sensorimotor performance is impaired in microgravity, e.g. aimed arm movements are slowed down and are less accurate. Several explanatory approaches for this phenomenon have been discussed, such as distorted proprioception or stress-related attentional deficits. In the current work, sensorimotor performance was investigated during aimed joystick-controlled motions in a simulation. The task included rapid as well as fine matching motions. Results of two different studies were compared: 1) a study utilising a dual-task paradigm to investigate the impact of attentional distraction ($N$ = 19) and 2) a study investigating the impact of microgravity during spaceflight ($N$ = 3). In both studies, an overall slowing effect was found. However, results diverged when comparing feedforward vs. feedback-controlled parts of aiming. Reduced attentional resources mainly affected feedforward control, which was reflected in significantly longer response times and longer rapid motion times. Microgravity, however, did not affect response times at all, but rapid aiming times as well as fine matching times substantially increased. These findings provide evidence that impaired attention is not the main trigger behind the slowing effect, but rather it is distorted proprioception which impairs feedback-controlled motions.

## Introduction

Space agencies around the world are planning crewed lunar and Mars missions to be realised within the next decade (International Space Exploration Coordination Group, 2018). Apart from the enormous technological challenges, these human space exploration missions would also critically depend on human capabilities and performance. It has been shown, however, that adaptation to the adverse space environment is challenging - even for astronauts who passed a hard selection and training process before starting their mission. Spaceflight has a substantial impact on human physiology (e.g. cardiovascular, vestibular and sensorimotor systems), sleep and circadian rhythms are disturbed, and psychological stressors such as isolation, confinement, high workload, etc. additionally compromise astronauts' well-being and performance (see Kanas & Manzey, 2008 for an overview).

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Furthermore, many basic functions like spatial orientation, oculomotor control, posture and locomotion (see Lackner & DiZio, 2000) as well as mass discrimination (Ross et al. 1986; Ross and Reschke, 1982) are affected by microgravity. Prior research repeatedly documented that human motor performance is also degraded in microgravity (see Bock, 1998; Lackner & DiZio, 2000). Impairments have been found across different task paradigms like aiming (e.g. Bock et al., 2001), tracking (e.g. Manzey et al., 1993) and force production (e.g. Mierau & Girgenrath, 2010). When performing rapid aiming movements in weightlessness, a general slowing-down effect was found, i.e. peak accelerations decreased and motion times increased accordingly (Berger et al., 1997; Bock et al., 2001; Crevecoeur et al., 2010; Mechtcheriakov et al., 2002; Newman & Lathan, 1999; Ross, 1991; Sangals et al., 1999). Moreover, positional accuracy in tracking tasks decreases (Bock et al., 2003, Manzey et al., 1993, 1995, 2000) and studies on isometric force production reported less accurate force regulation in weightlessness (Mierau, et al., 2008; Mierau & Girgenrath, 2010).

Several explanatory approaches for the substantial deterioration of basic and indispensable sensorimotor skills in microgravity have been proposed. Frequently, researchers explain their findings by disturbed proprioception in altered gravity conditions (e.g. Bock et al., 1992, 1998; Fisk et al., 1993, Manzey et al., 2000). According to this approach, muscle spindle activity which is crucial for proprioception is altered by the weightlessness of the body and limbs (e.g. Lackner & DiZio, 2000). Consequently, the sensorimotor system is in a state of "sensorimotor discordance" (Bock, 1998) and has to adapt to the lack of valid proprioceptive feedback. Corrective motor responses would be delayed due to additional information processing. The general slowing-down effect for aiming tasks and time-delayed correction initiation during tracking (Manzey et al., 2000) support this notion. Moreover, weightlessness effects were stronger in dual-task performance compared to single-task performance in the early mission phase (Manzey et al., 2000) or during parabolic flight (Bock et al., 2003), providing evidence for higher resource demands in the initial phase of adaptation to microgravity.

However, the impaired proprioception approach is not sufficient to explain the performance decrement in the early and late phases of the 20-days mission reported by Manzey and his colleagues (1995, 2000) during tracking tasks. The performance losses in the later phase were explained by prolonged work and the cumulative impact of general stressors of the mission. While higher cognitive functions (memory, reasoning etc.) are seemingly not impaired by spaceflight, attentional selectivity affects performance in weightlessness as revealed in dual-task paradigms (Bock et al., 2003; Fowler et al., 2008; Manzey et al. 1993, 1995).

Still, the specific contributions and relevance of both mechanisms to the overall microgravity effects on sensorimotor performance are difficult to determine and researchers attributed their results either to distorted proprioception (e.g. Bock, 1998), cognitive load (e.g. Fowler, 2008) or both processes (e.g. Manzey, 2000). Most studies investigating the degradation of sensorimotor performance in space utilised aiming (arm movement or device control), arm tracking, or unstable, compensatory tracking (joystick controlled) as experimental paradigms. Like any voluntary motion task, these tasks require feedforward motion planning as well as feedback-controlled

motion sequences, while the relative contribution of both control types is contingent on task demands. During rapid, aimed arm movement a major part of the movement has to be planned as a pre-programmed forward model that is corrected and updated by feedback loops integrating afferent information in the course of motion execution. During motor tasks requiring slow and precise closed-loop motions (e.g. tracking) the major part of motion control is based on visual and proprioceptive feedback (Desmurget & Grafton, 2000). Although optimal motion control relies on feedforward as well as feedback processes, they are two distinct mechanisms which are controlled by different brain structures. While cortical structures (e.g. primary motor cortex) have been identified to be mainly responsible for feedforward processes, subcortical structures (e.g. cerebellar regions) are associated with feedback control, as reported by Seidler and colleagues (2004), who analysed fMRI recordings during joystick controlled aiming tasks. In their study, the activation of these brain regions was moderated by task difficulty, i.e. cortical activity was positively correlated with increasing target size and subcortical activity was negatively correlated with target size.

Distinguishing these two basic functions of motor control seems a promising approach to better understand the mechanisms behind sensorimotor performance losses in space. Provided that distorted proprioception is the main trigger of performance decrements, then it is obvious that the feedback-controlled parts of motion should be mainly affected. On the contrary, a potential attentional deficit should mainly interfere with feedforward control. Johansen-Berg and Matthews (2002), for instance, could show that attention distraction (counting back in threes as the secondary task) affects the activity in the motor cortical areas including the primary motor cortex when performing the primary target acquisition task. In another dual-task experiment, Taylor and Thoroughman (2007) also found evidence that corrective movements (i.e. feedback control) were not affected when performing arm reaching tasks with a manipulandum that introduced random perturbations. However, the secondary task (auditory discrimination task) did interfere with adjustments of the feedforward model.

Based on this evidence and these considerations we designed an experimental aiming task, allowing a discrimination of feedforward and feedback controlled motor performance. In the present work, this experimental paradigm is pre-tested under terrestrial conditions to identify the impact of attentional distraction on performance during rapid, open-loop aiming and subsequent slow, terminal corrective adjustments. In a next step, the same aiming task is performed by cosmonauts in terrestrial and mission sessions on-board the ISS (2 weeks in space) to determine the effects of spaceflight.

An overall increase of aiming times is expected when attention is distracted as well as during spaceflight. More specifically, however, it is hypothesised that:

> H1: *Feedforward control is mainly affected by attentional distraction while feedback control is mainly affected by distorted proprioception during spaceflight.*

Thus, performance losses due to attentional deficits should primarily result in increased reaction times and rapid motion times (Fowler et al., 2000, Fowler et al.,

2008). Performance losses due to proprioceptive deficits should be evident for fine motion times as reported by Fisk and colleagues (1993).

**Methods**

*Study 1: The Effects of Attentional Distraction*

*Sample.* Nineteen subjects (5 females, 14 males; *M* = 24.6 (2.5) years of age) voluntarily participated in the study after having signed an informed consent document.

*Apparatus.* Participants were seated at a table, in front of a notebook (Lenovo T61P-6457) with a 15.4" TFT display showing the experimental GUI. The space qualified Joystick "Kontur-2" developed at the German Aerospace Center (Riecke et al., 2016, workspace of ±20° in each axis, angular resolution of $3.18° \cdot 10^{-3}$, see Fig. 1, left), was connected to the computer. For the present experiment, an upward motion scaling of 1:2 was implemented, i.e. the required experimental workspace was fully covered with joystick deflections of ±10° for both axes. Data were recorded with a sampling rate of 100 Hz.

*Experimental Tasks.*

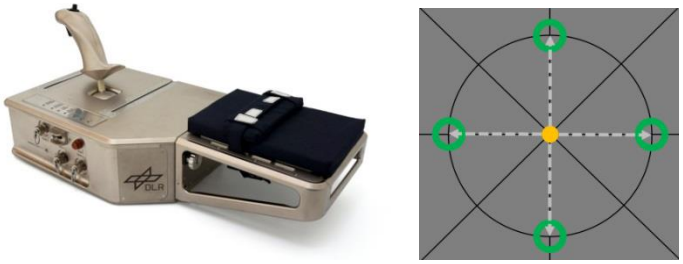


*Figure 1. Joystick "Kontur-2" (left); Experimental GUI with cursor at starting position and the four different target positions (right).*

Primary Aiming Task: The experimental GUI showed black crosshairs on a grey background (see Fig. 1, right). The aiming trials were started by moving the black cursor exactly to the crosshair's center. Upon reaching the center, the cursor turned green and a countdown was displayed on the screen. After holding the position for 2s the cursor turned orange and a green target ring was displayed at one of the four different target positions (see Fig. 1, right). The cursor had to be brought to the center of the target ring as quickly as possible and the final position had to be held for 0.5 sec. Subsequently, the next trial was started and subjects moved back to the centre of the crosshairs. Please note that the order of the four target positions was randomly chosen to avoid anticipatory movements.

Secondary Counting Task: During the aiming tasks, subjects had to count forwards in intervals of seven starting with 12 up to 103 and then backwards again (12-19-26-33-…-103-96-89-82-….12). An acoustic signal (metronome sound) prompted the subjects to speak the next number aloud every 4 seconds.

*Experimental Design.* A within-subject design was utilised with all subjects completing a single-task condition (aiming task only) and dual-task condition (aiming and counting) while the order of both conditions was counter-balanced across subjects.

*Procedure.* Chair height was individually adjusted by the participants so that their right arm rested comfortably on the joystick's padded arm support. For reasons of standardisation, subjects also attached a strap around the right elbow, ensuring that arm orientation and position was comparable across participants but still allowing free motion in the required range of motion. Participants read the instructions that were displayed on the monitor. The two experimental conditions (single vs dual-task) were presented in a sequence, separated by a short break of 2–3 min. In each condition, two aiming trials were performed for training, and then the experimental trials were started. After having completed these trials, subjects were asked to rate their perceived workload ("Please rate your overall workload during the last task", adapted from the OWS scale, Vidulich & Tsang, 1987; 20-point bipolar scale ranging from "very low" to "very high").

*Study 2: The Effects of Spaceflight*

*Sample.* The subjects were three male cosmonauts (42, 45, and 53 yrs.; two of them with space mission experience).

*Apparatus.* The same joystick was installed on board of the Russian Zvezda service module of the ISS (see Figure 2). Body stabilisation was realised by rails on the module "bottom" and an additional grip for the left hand. The experimental GUI window was displayed on the 15.4" TFT display of the notebook (same as in Study 1).

*Figure 2. Cosmonaut Andrei Borisenko at the experimental workstation on board the ISS.*

*Experimental Design and Procedure.* All of the three cosmonauts performed the same aiming tasks as in Study 1 (without a secondary task) during a pre-mission training session three months before their mission launch, on-board the ISS (exactly two weeks after Soyuz docking) and during a post-mission session, two weeks after having finished their half-year space missions. The procedure (instruction, experimental workflow and questionnaire) was similar to the procedure in Study 1.

*Data analysis.* Reaction times, rapid motion times and fine motion times were calculated for each aiming trial. Reaction time was defined as the time from task start until exceeding a pre-defined threshold velocity (in contrast to the positional threshold approach the authors utilised in a prior study; Weber et al., 2018). Rapid motion time was the time from exceeding the threshold velocity until the center of the cursor touched the green target ring. Fine motion time was the remaining time until target and cursor centers were precisely matched and constantly held for 0.5 sec. These temporal variables were averaged across all of the four targets. For Study 1 the single and dual-task conditions were compared using paired t-tests. Additionally, the effect sizes were calculated using Hedges' *g*. In Study 2, only effect sizes were determined due to the small sample size. Results of both terrestrial conditions (pre- and post-mission) were averaged and utilised as a comparison baseline for mission session.

**Results**

Study 1. Performing paired t-tests on the average reaction times and rapid motion times revealed a significant increase in the dual-task compared to the single task condition (for both conditions, $p < .05$; see Table 1). A large effect was evident for reaction time ($g$ =.82) and a moderate effect for rapid motion time ($g$ =.68). No significant difference was found for fine motion times. Finally, the subjective workload rating was significantly increased in the dual-task condition ($p$ <.001).

The number of counting errors during the secondary task and the reaction as well as rapid motion times were positively correlated ($r_{RT}(19) = .50$; $p$ <.05 and $r_{RMT}(19) = .51$; $p$ <.05). Seemingly, no task switching occurred, but both primary and secondary task were influenced simultaneously.

Study 2. A quite different result pattern was found in Study 2, comparing terrestrial conditions (1g) and microgravity (µg) conditions during spaceflight. When comparing both conditions, large effect sizes were evident for rapid motion ($g$ =.80) and fine motion times ($g$ =1.08). Regarding workload ratings, a small effect of microgravity ($g$ =.27) was found, i.e. workload increased marginally.

*Table 1: Performance Measures (M (SD), paired t-tests and Hedges' g for Study 1 and 2*

| **Study 1** (n =19) | | **Terrestrial Dual-Task Experiment** | | | |
|---|---|---|---|---|---|
| **Measures** | | **Single Task** | **Dual Task** | Sign. (t-test) | Effect Size g |
| Reaction Time | [s] | 0.139 (0.064) | 0.303 (0.271) | $p < .05$ | **0.82** |
| Rapid Motion Time | [s] | 0.545 (0.167) | 1.242 (1.419) | $p < .05$ | **0.68** |
| Fine Motion Time | [s] | 2.467 (0.969) | 2.164 (1.139) | *n.s.* | 0.28 |
| Overall Workload | [1-20] | 6.3 (4.0) | 11.5 (4.1) | $p < .001$ | **1.27** |
| **Study 2** (n = 3) | | **Space Flight Experiment** | | | |
| **Measures** | | **1g** | **µg** | | Effect Size g |
| Reaction Time | [s] | 0.220 (0.077) | 0.216 (0.010) | | 0.06 |
| Rapid Motion Time | [s] | 0.394 (0.046) | 0.503 (0.148) | | **0.80** |
| Fine Motion Time | [s] | 2.351 (0.232) | 3.020 (0.663) | | **1.08** |
| Overall Workload | [1-20] | 4.3 (2.08) | 5.0 (2.00) | | 0.27 |

**Discussion**

The slowing of aimed arm movements in microgravity has been repeatedly documented by researchers since the early 1990s. However, this phenomenon remained enigmatic due to the substantially altered working conditions of spaceflight and multiple potential mechanisms triggering such sensorimotor performance losses. In prior research, two explanations for the slowing effect of microgravity have been discussed: distorted proprioception due to the lack of a gravitational force and attentional selectivity due to general mission-related workload. In the current paper, a simple joystick-controlled aiming task was utilised to explore the effects of reduced attentional resources and spaceflight on feedforward and feedback-controlled parts of motion.

It was hypothesised that decreased attentional capacity would mainly affect feedforward control and deficient proprioception would mainly affect feedback-controlled motions. Indeed, two substantially divergent result patterns are evident for both studies: When performing a concurrent counting task, motion planning and the early feedforward controlled aiming motion are significantly disturbed as reflected by increased reaction and rapid motion times compared to the single-task condition. No significant effect emerges for the feedback-controlled fine motion section. In contrast, the cosmonauts did not show any additional delay of reaction times in microgravity compared to the terrestrial baseline condition, but rapid motion and fine motion time increase. Note that the overall effect pattern is diametrically opposed. Reducing attentional resources has the strongest effect on motion initialisation, but disappears towards the end of motion. Regarding the impact of microgravity, the inverse pattern emerges: the effect increases the more feedback is required for motion plan corrections. Altogether, this confirms the formulated hypothesis and provides evidence that – in this case – a proprioceptive deficit is the main trigger behind the slowing effect of microgravity. The subjective ratings additionally provide further evidence that, in the present study, increased workload is not a plausible explanation for slowed aiming motions in microgravity.

Although a stronger impact of attentional distraction was expected for the rapid motion times, a similar slowing effect occurred during spaceflight. This result might be explained by the fact that the rapid, open-loop arm motion is not exclusively executed on basis of pre-planned forward models, but also integrates feedback during the ongoing motion. In line with this notion, Bock et al. (2001) also reported no effect of microgravity on aimed arm motions in the initial 80ms, but motions increasingly slowed down towards the end positions. Indeed, the minimal delay of proprioceptive feedback loops ranges between 80 and 100ms. Thus, internal feedback loops refine the initial motion plan even during rapid arm motions (Seidler et al., 2004).

Additional analyses of the aiming trajectories recorded in Study 2 also revealed that cosmonauts show very irregular and unstable motion paths when moving their arm in the sagittal plane (i.e. vertical motion axis in the experimental GUI) in microgravity. The occurrence of this direction-specific effect (anisotropy) might also be an indicator of a proprioceptive deficit as documented in studies investigating aiming motions of patients without proprioception caused by large-fiber sensory neuropathy (e.g. Ghez et al., 1990).

One major limitation of the current study is that no dual-task condition was implemented in Study 2, which actually was an integral part of a series of experiments pursuing a different research agenda. Thus, the question how attentional and proprioceptive processes interact during spaceflight cannot be answered with the present work. It is well conceivable, for instance, that a mismatch of internal motion models and afferent information also leads to increased attention demands as reported by Ingram and colleagues (2000).

The comparison of two studies investigating attention distraction and microgravity effects on basic aiming tasks provides evidence that distorted proprioception seems to be the main mechanism underlying the slowing of voluntary aiming motions at least in the early phase of a space mission (two weeks in space). The question still is whether the terrestrial performance can be reached again after having completed the initial adaptation to the space environment. A recent study of the authors (Weber et al., 2019) investigating the effects of spaceflight on performance during a real telerobotic aiming task, provides evidence that performance is degraded even after six weeks of space travel, seemingly due to an altered motion strategy. For human space missions to be successful it is imperative to identify effective measures to attenuate these performance losses, e.g. by providing haptic assistance as part of the human-machine interface, or intention-detection concepts.

**References**

Berger, M., Mescheriakov, S., Molokanova, E., Lechner-Steinleitner, S., Seguer, N., & Kozlovskaya, I. (1997). Pointing arm movements in short-and long-term spaceflights. *Aviation, Space, and Environmental Medicine, 68,* 781-787.

Bock, O. (1998). Problems of sensorimotor coordination in weightlessness. *Brain research reviews, 28,* 155-160.

Bock, O., Howard, I.P., Money, K.E., & Arnold, K.E. (1992). Accuracy of aimed arm movements in changed gravity. *Aviation, Space, and Environmental Medicine, 63,* 994-998.

Bock, O., Abeele, S., & Eversheim, U. (2003). Sensorimotor performance and computational demand during short-term exposure to microgravity. *Aviation, Space, and Environmental Medicine, 74,* 1256-1262.

Bock, O., Fowler, B., & Comfort, D. (2001). Human sensorimotor coordination during spaceflight: an analysis of pointing and tracking responses during the "Neurolab" Space Shuttle mission. *Aviation, Space, and Environmental Medicine, 72,* 877-883.

Crevecoeur, F., McIntyre, J., Thonnard, J.L., & Lefèvre, P. (2010). Movement stability under uncertain internal models of dynamics. *Journal of Neurophysiology, 104,* 1301-1313.

Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences, 4,* 423-431.

Fisk, J., Lackner, J.R., & DiZio, P. (1993). Gravitoinertial force level influences arm movement control. *Journal of Neurophysiology, 69,* 504-511.

Fowler, B., Meehan, S., & Singhal, A. (2008). Perceptual-motor performance and associated kinematics in space. *Human Factors, 50,* 879-892.

Fowler, B., Comfort, D., & Bock, O. (2000). A review of cognitive and perceptual-motor performance in space. *Aviation, Space, and Environmental Medicine.*

Ghez, C., Ghilardi, M.F., Christakos, C.N. & Cooper, S.E. (1990). Roles of proprioceptive input in the programming of arm trajectories. In *Cold Spring Harbor Symposia on Quantitative Biology, 69, Volume 69* (pp. 837-847). Cold Spring Harbor Laboratory.

Ingram, H.A., Van Donkelaar, P., Cole, J., Vercher, J.L., Gauthier, G.M., & Miall, R.C. (2000). The role of proprioception and attention in a visuomotor adaptation task. *Experimental Brain Research, 132,* 114-126.

International Space Exploration Coordination Group (2018). *Global Exploration Roadmap (3rd edition)*, Retrieved from (26.09.2019) http://www.globalspace exploration.org

Johansen-Berg, H., & Matthews, P. (2002). Attention to movement modulates activity in sensori-motor areas, including primary motor cortex. *Experimental Brain Research, 142,* 13-24.

Kanas, N., & Manzey, D. (2008). *Space Psychology and Psychiatry* (Vol. 22). Springer Science & Business Media.

Lackner, J.R. & DiZio, P. (2000). Human orientation and movement control in weightless and artificial gravity environments. *Experimental Brain Research, 130,* 2-26.

Manzey, D., Lorenz, B., Schiewe, A., Finell, G., & Thiele, G. (1993). Behavioral aspects of human adaptation to space analyses of cognitive and psychomotor performance in space during an 8-day space mission. *The Clinical Investigator, 71,* 725-731.

Manzey, D., Lorenz, B., Schiewe, A., Finell, G., & Thiele, G. (1995). Dual-task performance in space: results from a single-case study during a short-term space mission. *Human Factors, 37,* 667-681.

Manzey, D., Lorenz, B., Heuer, H., & Sangals, J. (2000). Impairments of manual tracking performance during spaceflight: more converging evidence from a 20-day space mission. *Ergonomics, 43,* 589-609.

Mechtcheriakov, S., Berger, M., Molokanova, E., Holzmueller, G., Wirtenberger, W., Lechner-Steinleitner, S., ... & Gerstenbrand, F. (2002). Slowing of human arm movements during weightlessness: the role of vision. *European Journal of Applied Physiology, 87*, 576-583.

Mierau, A., Girgenrath, M. & Bock, O. (2008). Isometric force production during changed-Gz episodes of parabolic flight. *European Journal of Applied Physiology, 102*, 313-318.

Mierau, A., & Girgenrath, M. (2010). Exaggerated force production in altered Gz-levels during parabolic flight: The role of computational resources allocation. *Ergonomics, 53*, 278-285.

Newman, D. J., & Lathan, C. E. (1999). Memory processes and motor control in extreme environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 29(3)*, 387-394.

Riecke, C., Artigas, J. Balachandran, R., Bayer, R., Beyer, A., Brunner, B., Buchner, H., Gumpert, T., Gruber, R., Hacker, F., Landzettel, K., Plank, G., Schätzle, S., Sedlmayr, H.-J., Seitz, N., Steinmetz, B.-M., Stelzer, M., Vogel, J., Weber, B., Willberg, B., & Albu-Schäffer, A. (2016). *KONTUR-2 Mission: The DLR Force Feedback Joystick for Space Telemanipulation from the ISS.* i-SAIRAS Conference 2016, Beijing, China.

Ross, H.E., Brodie, E.E., & Benson, A.J. (1986). Mass-discrimination in weightlessness and readaptation to earth's gravity. *Experimental Brain Research, 64*, 358-366.

Ross, H.E. (1991). Motor skills under varied gravitoinertial force in parabolic flight. *Acta Astronautica, 23*, 85-95.

Ross, H.E., & Reschke, M.F. (1982). Mass estimation and discrimination during brief periods of zero gravity. *Perception & Psychophysics, 31*, 429-436.

Sangals, J., Heuer, H., Manzey, D., & Lorenz, B. (1999). Changed visuomotor transformations during and after prolonged microgravity. *Experimental Brain Research, 129*, 378-390.

Seidler, R.D., Noll, D.C., & Thiers, G. (2004). Feedforward and feedback processes in motor control. *Neuroimage, 22*, 1775-1783.

Taylor, J.A., & Thoroughman, K.A. (2007). Divided attention impairs human motor adaptation but not feedback control. *Journal of Neurophysiology, 98*, 317-326.

Vidulich, M.A., & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors Society Annual Meeting, Volume 31(9)* (pp. 1057-1061). Los Angeles, CA: SAGE Publications.

Weber, B., Schätzle, S., & Riecke, C. (2018). Comparing the effects of space flight and water immersion on sensorimotor performance. In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference* (pp. 57-66). Available from http://hfes-europe.org.

Weber, B., Balachandran, R., Riecke, C., Freek, S. & Stelzer, M. (2019). *Teleoperating Robots from the International Space Station: Microgravity Effects on Performance with Force Feedback.* International Conference on Intelligent Robots and Systems (IROS) 2019, Macau, China.

# Divided attention and visual anticipation in natural aviation scenes: The evaluation of pilot's experience

*Jason A.M. Khoury, Colin Blättler, & Ludovic Fabre*
*CREA -Centre de Recherche de l'Ecole de l'Air*
*French Air Force Academy Research Centre*
*France*

**Abstract**

The present study aims to investigate whether spatial representation bias can be used to assess the trainee's air skills. Spatial representations contribute in large part to the development of situational awareness (Endsley, 1996), making it a key factor in aviation performance and safety. Blättler et al (2011) have shown that a memory displacement of spatial representation is larger among pilots than novices. The purpose of this study was to provide evidence that spatial representation bias can discriminate novice from experienced pilots. Furthermore, several studies showed that not all the processes underlying displacement are automatic (Hayes & Freyd, 2002). The second objective of this study was to test whether experts share the same sensitivity to divided attention as novices in a task measuring displacement, since the expert's automation makes processes specific to his activities more resistant to the effect of the dual task (Froger, Blättler, Dubois, Camachon, & Bonnardel, 2018; Strobach, Frensch & Schubert, 2008). This study was conducted to explore these questions in an experiment with 19 experienced glider pilots from the French Air Force and 25 novices. Participants were shown dynamic real-world landing scenes in ego-motion (Thornton & Hayes, 2004) during a representational momentum (RM) task. Gaze fixations data were also recorded to explore their potential relationship with spatial memory bias. This study provides evidence that spatial representation bias can discriminate novices from experienced pilots who only have a few hours of training.

**Introduction**

Spatial representation is crucial when flying an aircraft. Situational awareness, which includes anticipation and is based on spatial representation, is a key element of air safety. However, it is difficult to objectively evaluate the evolution of performance in spatial representation during student training. The objective of this study was to test whether a process underlying spatial representation was sensitive enough to be an appropriate measurement and analysis tool. The experiment performed here evaluated the spatial representation of natural glider landing scenes by experienced pilots and novices.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Understanding spatial representation is a major challenge since it is the result of the influence of multiple factors. Its understanding is essential for actors in the aeronautics world (industries, training schools, etc.) to design both human-system interaction interfaces and ad hoc training. It must de facto be studied through a rigorous protocol. A special case for studying spatial representation is that of the processes that underlie "Representational Momentum" (RM) (Freyd & Finke, 1984). Because of its properties, described below, this work is part of understanding how the cognitive system succeeds in learning to cope with complex dynamic visual situations. Representational momentum refers to a memory displacement for the final position of a previously viewed moving target in the direction of the target's motion. Finke, Freyd and Shyi (1986) suggested that the properties of such a memory displacement could help observers anticipate the future positions of moving objects. In the rest of the article, the term "displacement" will be used to refer to a displacement of the spatial position in memory of a moving object or scene.

The variables that influence the direction and amplitude of displacement act in a similar way to the physical principles of movement. That is why studying displacement is a way of studying how the physical principles of movement are incorporated into mental representations. One of the experimental protocols (*figure 1*) conventionally used to show a displacement is that of Hubbard and Bharucha (1988). The authors presented participants with a target that moved continuously and linearly (to the left or right and up or down). After a few moments of animation, the target disappeared unexpectedly. As soon as the target disappeared, participants clicked on the place where they thought the target had disappeared. The results showed that participants recalled the position of the target, at the time of its disappearance, not at its exact location, but a little further in the direction of the target's trajectory. They suggested that, like a moving object that does not immediately stop but continues along its path under its own momentum, spatial representation does the same and shifts the last perceived spatial position in the direction of the motion.

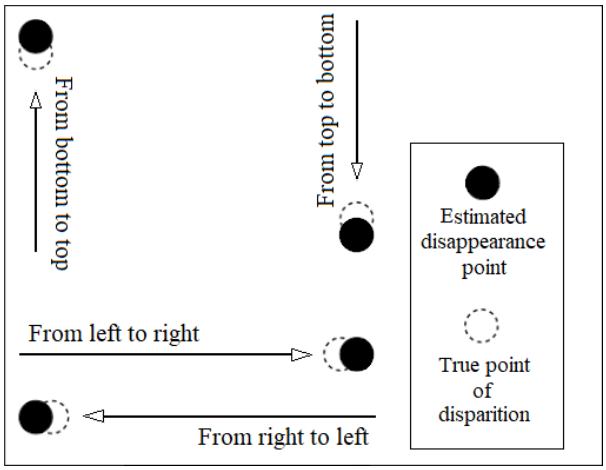



*Figure 1. Material and results adapted from Hubbard & Bharucha (1988)*

The distance between the actual disappearance position and the one recalled by the participants can vary in magnitude depending, for example, on the speed of a target's movement. The higher the speed, the greater the magnitude of the displacement (Freyd & Finke, 1985; Hubbard & Bharucha, 1988; de sá Teixeira, Hecht, & Oliveira, 2013). The analogies between physical motion and displacement are also spatio-temporal in nature. Freyd and Johnson (1987) varied the time between the disappearance of a moving target and the latency with which the participant gave his response (from 10ms to 900ms). The results obtained showed an increase in displacement magnitude with the increase in encoding latency. This corresponds to what would happen physically, as the movement of an object lasts for a few moments if nothing prevents it. But it should be noted that when latency exceeded a certain threshold, in this case 300 ms, this effect decreased as latency increased. This decrease after 300 ms suggests that the evolution of the displacement is similar to the movement that an object would actually have, namely stopping of movement over time. This similarity between real movement and displacement makes the latter a dynamic representation. Taken together, these results suggest that displacement is based on a spatio-temporal coherence similar to that of physical principles. Overall, displacement is described in terms of dynamic representations and thus, by analogy to real-world dynamics, Hubbard (2010) conventionalized it as the "momentum metaphor", suggesting as said earlier that the principles of momentum are indeed incorporated into mental representation.

The plurality of analogies from the physical world has motivated the prolific development of research protocols and since the 1980s, a significant number of variables that modulate displacement have been investigated (see Hubbard, 2005b, 2018 for reviews). While some variables foster the development of a displacement in the direction of perceived movement e.g., speed (Freyd & Finke, 1985; Hubbard & Bharucha, 1988; de sá Teixeira, Hecht, & Oliveira, 2013), downward motion (Hubbard, 1990; Hubbard & Bharucha, 1988), and high contrast (Hubbard & Ruppel, 2014), others foster a displacement in another direction e.g., representational gravity (de sá Teixeira, 2014; de sá Teixeira & Hecht, 2014; Hubbard, 1995b, 2005b; Motes, Hubbard, Courtney, & Rypma, 2008), reduce the magnitude of the displacement e.g., representational friction (Hubbard, 1995a, 1995b), or promote a displacement in the opposite direction of movement e.g., surrounding context (Hubbard, 1993), and memory averaging (see for example Hubbard, 1996). Thus, outside the laboratory, there is a set of different variables, with diverse, congruent or opposite influences, which are co-articulated and induce a result which is the spatial representation of a scene. For example, Hubbard and Bharucha (1988) showed that the position of a target moving in a straight line is recalled further in the direction of movement but also lower. Many replicates (Hubbard, 1990, 1995b, 1997, 2001) have determined that this result of a combination of a forward displacement effect and the effect of implicit knowledge of gravity (representational gravity) results in a downward displacement. In this vein, Hubbard (1995a; 2010) proposed a model that reflects this multiplicity of influences. In his "vector addition" model, each type of influence is matched by a vector that codes for the direction and magnitude of displacement. "Such vectors can be broadly construed as corresponding to magnitudes and directions of activation within a network architecture that preserve functional mapping between physical space and represented space" (Hubbard, 2010, p. 352). While many studies have

massively contributed to determining low-level influences (target shape, surrounding context, etc.), more recent studies show that displacement is also modulated by cognitive factors such as the expertise of observers and the allocation of attention resources.

Blättler, Ferrari, Didierjean and Marmèche (2011) showed an effect of expertise on displacement in the aeronautical context. In their study the authors adjusted the Thornton and Hayes (2004) protocol. Dynamic simulated aircraft landing scenes were presented to participants who were either total novices to aeronautics or expert pilots (over 3000 hours of flight experience). The scenes were interrupted by the display of a black screen lasting 125 ms and then resumed in one of three conditions: a shift forward (with respect to the aircraft's direction of motion), a shift backward (in the direction opposite to the plane's motion), or no shift (i.e., at exactly the same point as before the interruption: the same-resumption condition). In the shift conditions, the size of the forward and backward shifts was manipulated (125 ms, 250 ms, 375 ms, and 500 ms). Participants had to compare the last image seen before the cut to the first image seen after the cut and decide whether the scene had shifted backward or forward. The results showed that only the expert pilots produced a forward displacement, while among the novices no displacement (either forward or backward) was obtained. After successive studies increasing the accuracy of the measurement, a significant displacement was obtained in the novices. The magnitude of the displacement was so short in the novices that it could not be observed with the accuracy measurement used to detect a displacement among the experts in the first study. This expertise effect resulted in an increase in the amplitude of the displacement in the direction of the perceived movement.

Similar results have been obtained in the automobile context (Blättler et al., 2010; Blättler et al., 2012, 2013; Didierjean, Ferrari & Blättler, 2014) and in the sports context (Hiroki, Mori, Ikudome, Unenaka, & Imanaka, 2014; Jin et al., 2017; Chen, Belleri, Cesari, 2019; Gorman, 2015; Anderson, Gottwald, & Lawrence, 2019). Thus, the effect of expertise seems robust. However, the way in which expertise is manifested is not clearly established. Furthemore, the literature (see for review Gegentfurtner, Lehtinen & Säljö, 2011; Peißl, Wickens & Baruah, 2018; Reingold, Charness, Pomplun & Stampe, 2001; Ziv, 2016) show that systematic eye movement differences between experts and novices occur. Therefore, in accordance with the first objective of the current study, eye tracking data were collected, as part of an exploratory attempt to gain insight into the manifestation of the experience in the displacement.

Another way in which the effect of expertise could manifest itself in the processes underlying the displacement is through the effect of automation of cognitive procedures. Hayes and Freyd (2002) showed that not all the processes underlying displacement are automatic (see also, Joordens, Spalek, Ramzy & Duijn, 2004). However, since the constitutive process of expertise development is automation (Logan, 1988), it is conceivable that the processes underlying the displacement if it shares the same property may gradually become automatic. Thus, the more experienced an individual is, the more automated specific processes of his activity are. This automation makes it more resistant to the effect of the dual task (Froger, Blättler,

Dubois, Camachon, & Bonnardel, 2018; Strobach, Frensch & Schubert, 2008). Experiments on divided attention (Hayes & Freyd, 2002; Joordens et al., 2004) show an increase in the amplitude of forward displacement when attention is divided during perception of the moving target. If the processes underlying displacement share the same properties as those associated with automation, the displacement of experienced individuals should be less sensitive to the dual task effect than that of novices. The second aim of this study was therefore to test whether experts share the same sensitivity to divided attention as novices in a task measuring displacement.

In summary, the first purpose of this experiment was to determine whether displacement can be an index that would be sensitive enough to assess the progress of student pilots. The assumption is that experienced pilots will produce a greater displacement in the direction of movement than novices. Complementary to this goal, the eye tracking was used to explore the link between this displacement and gaze fixations of the experienced pilots. The second objective was to evaluate whether the processes underlying the displacement are sensitive to the automation process conventionally observed during the development of expertise. The hypothesis is that experienced pilots will be less sensitive than novices to a disturbance caused by a dual task.

## Method

### *Participants*

Forty-four participants were recruited for the study, drawn from two distinct skill levels: an experienced glider pilot group ($n$ = 19) with 78.16 flying hours on average ($SD$ = *177*) and an average age of 23 years ($SD$ = 5), and a second experimental group ($n$ = 25) composed entirely of novices ($M_{age}$ = 27 years, $SD$ = 8). All participated were volunteers, had normal or corrected vision and were naive to the specific purpose of the study.

### *Material*

Following Blättler et al. (2011), 10 video sequences (*figure 2)* inside a Centrair Marianne C201B glider were used (24 frames/s). Each landing scene was filmed from the pilot's perspective (i.e., first-person view, with a small part of the cockpit visible and no view of the instruments). To ensure that the inclination, angle and approach speed were the same for all scenes or to ensure that all approaches were consistent compared to an optimal approach, an instructor was present on all flights.

*Figure 2. Scene example with, the left to the right: -250 ms. 0 ms and +250 ms condition.*

The speed chosen for the landing was a standard speed for a glider (i.e., the distance a glider travels in 125 ms is about 3.125 meters at a speed of 90 km/h - 87.1 km/h without wind for an optimal run). The test stimuli were displayed on a Dell Precision 7710 laptop computer (17.3 in. screen, refreshment 60 Hz, resolution 1920 x 1080). The participants were positioned 60 cm from the screen. Each scene (all of which had a different landing scenario) was used to make nine videos. Each of these nine videos was followed by a perceptual interruption (interstimulus interval, ISI) lasting 250 ms. After the cut, the trial resumed in one of nine conditions (*Figure 3*) that differed in the magnitude of the shift of the image (-250 ms, -187 ms, -125 ms, -62 ms, 0 ms, +62 ms, +125 ms, +187 ms, +250 ms). There was a total of 90 different videos (10 scenes x 9 shifts = 90).

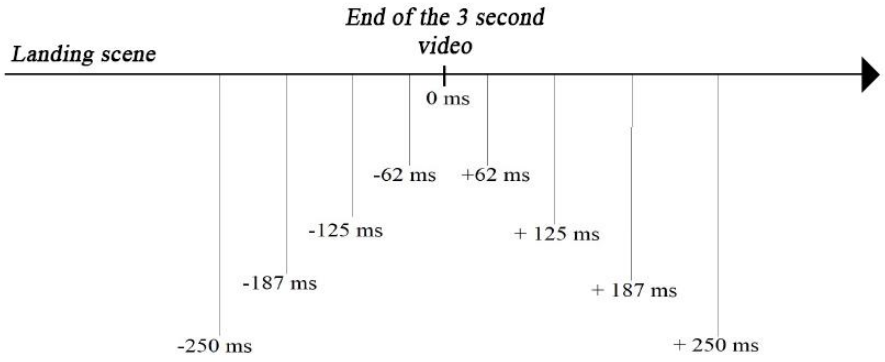



*Figure 3. Landing scene and conditions in accordance with Blättler et al., (2011).*

Eye position data were captured by an eye-tracker Tobii Pro X3 with a sampling rate of 120 Hz. The analyses used to examine the data were based on static exploratory areas to collect information on participants' eye movements and fixations.

*Procedure*

Each trial (i.e., video stimuli) was displayed on the computer monitor for 3 seconds, followed by the 250 ms ISI. After the perceptual interruption, the trial was resumed with an image from one of the nine conditions. In the same-resumption condition (i.e., "no shift condition"), the video started up at exactly the same point as before the cut (a comparison between the two images shows that they are identical).

In the forward-shift condition, the trial started after a forward shift of +62 ms, +125 ms, +187 ms, or +250 ms. In the backward-shift condition, the trial resumed with an image corresponding to -62 ms, -125 ms, -187 ms, or -250 ms. From the moment the test started (i.e., when the image appeared) the participant had 15 seconds to respond. If he answered, or if the 15 seconds had elapsed, a black fixation cross on a white screen appeared for 2 seconds, followed by a new trial *(Figure 4).*

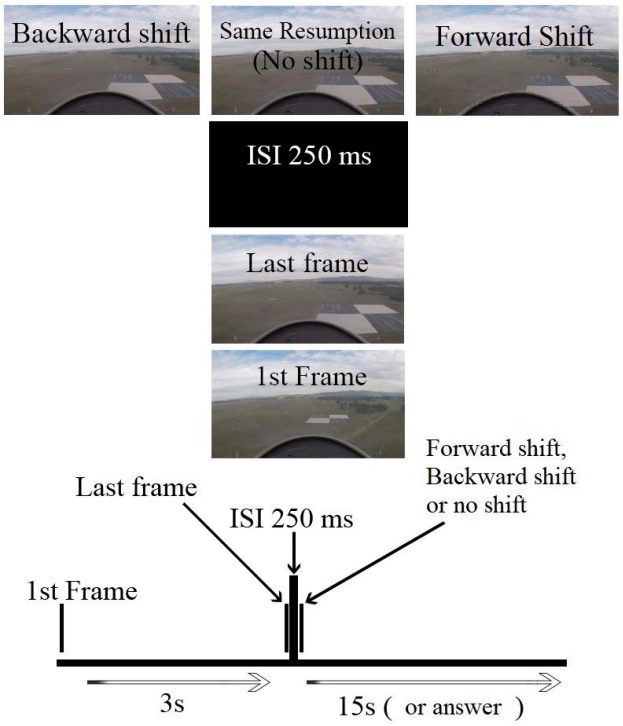



*Figure 4. Material (top) and procedure (bottom). The video began with 3 s of a landing scene. Then a cut occurred with an interstimulus interval (ISI) of 250 ms. After the cut, the video resumed with a backward shift (upper left: backward shift of 250 ms), no shif (upper middle), or a forward shift (upper right: forward shift of 250 ms).*

The experiment was conducted in two successive phases; a task familiarization phase, followed by the experimental phase. Before the familiarization phase, the experimenter gave the participants the following instructions.

*In the full attention condition:*
Participants had to compare the last image seen before the cut to the first image seen after the cut and decide whether the scene had shifted backward or forward. In line with previous studies, note that no information about the existence of same resumptions was given to the participants. Indeed, the PSE's measure is showing the point of maximal uncertainty, in this particular design, if the possibility of same resumption is not introduced to the participants. That way participants must answer according to their representations and not according to their knowledge of possible answers. After reading the instructions, the participants became familiar with the task by completing 14 practice trials (7 in the divided attention condition, 7 in the full attention condition) on two scenes that were not used in the experimental phase. Then the experimental phase began. In this phase, 10 scenes were used, each giving nine resumption conditions. This made 90 trials (10 * 9), which were presented in a random order to all participants.

*In the divided attention condition:*
Participants performed the primary task as described in the first condition while simultaneously listening via headphones to an auditory recording of a continuous stream of four randomized individually presented digits during each landing scene. They were instructed to monitor this recording for the occurrence of even digits (2, 4, 6 and/or 8), and to mentally keep track of the number of times that such runs had occurred to recall it. It should be noted that the presentations of the one to four even digit runs were not linked to the visual presentation of stimuli in any systematic way. This test condition showed the same clips as those displayed in the full attention condition. The clips were presented in a random order.

**Results**

An analysis of RM magnitude was used to assess the magnitude of shifts and to compute the point of subjective equality (PSE) for each participant. This point is the theoretical value of the stimulus that the participant considers to be subjectively equal to the standard. It indicates the point of maximum uncertainty. This measure was computed by fitting the distributions of the percentages of each participant. Each PSE was calculated from this curve by taking all the responses of that participant into account. A positive PSE (i.e., significantly above zero) indicated a forward displacement (FD). A negative PSE (i.e., significantly below zero) indicated a backward displacement (BD) (see *Figure 5* for the PSE mean by group).

*Table 1. PSE descriptive data. Full attention condition (FA); Divided attention (DA).*

| **Descriptive** | ***Novices FA*** | ***Novices DA*** | ***Pilots FA*** | ***Pilots DA*** |
|---|---|---|---|---|
| *N* | 25 | 25 | 19 | 19 |
| *Mean* | -34.40 | -59.36 | -16.68 | 3.342 |
| *SD* | 52.62 | 77.05 | 44.28 | 72.03 |

An analysis of variance (ANOVA) was conducted with experience as a between-groups factor (novices vs experienced pilots) and attention as a within-group factor (full attention vs divided attention). The experience factor was significant, $F(1,42) = 6.133$, $MSE = 34911$, $p < .05$. Novices' mean PSE was significantly lower than that of the experienced glider pilots. The attention effect was not significant, $F(1, 42) = 0.056$, $MSE = 131.6$, $p > .1$. The interaction between experience and attention was significant, $F(1, 42) = 4.658$, $MSE = 10925.7$, $p < .05$.

Hence, subsequent t-test comparisons were made. The analyses showed that the means of experienced glider pilots in FA, t(18)=-1.642, p =.118 and DA, t(18)=0.202, p=.842 were not significantly different from zero, while they were significantly different from zero for novices in both, FA, t(24) = -3.269, p = .003, and in DA, t(24)=-3.852,p <.001. Moreover, while there was no significant difference between FA and DA for experienced glider pilots, novices' mean PSE in FA was significantly larger than the novices' mean PSE in DA, t(42)=2.029,p =.027. Hence, the pattern of the interaction in Figure 5 demonstrates that backward displacement was larger for novices in DA than in FA. Conversely, there were no backward displacement in DA or FA for experienced pilots. Therefore, the interaction shows that experience modulates the effect of attention allocation in the displacement process. Furthermore, in both FA and

DA, the experienced pilots' mean PSE was significantly superior to the novices' mean PSE ,t(42)=1.795, p=.045 and t(42)=3.559, p =.001, respectively.

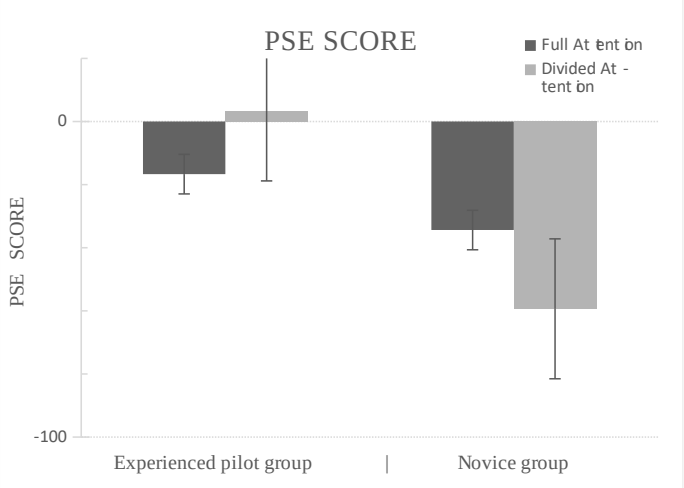



*Figure 5. PSE mean in Full Attention (FA) and Divided Attention (DA) for each experience group (Novice vs Experienced pilot).*

To assess the validity of the divided attention condition, the average success rate of participants in the dual task was measured. The average success rate of participants in the dual task was 93.55%. The mean success rate was 94.60% (SD =3.75) for the experienced glider pilots and 92.67% (SD=6.75) for the novices. The experienced pilot's mean PSE was significantly inferior to one hundred, t(9)=4.557, p<.01. The novices' success rate was also significantly inferior to one hundred, t(11)=-3.765, p<.01. The experienced pilots' mean success rate was not significantly superior to the novices' mean success rate, t(20)= -0.798, p =.223. The results did not show any ceiling effect.

*Eye fixation data*

Eye tracking data were recorded for twenty-two of the forty-four participants: 10 in the experienced glider pilot group with 125.8 flying hours on average (*SD = 238*) and an average age of 24 years (*SD* = 6.5), and 12 in the experimental group of novices ($M_{age}$ = 27 years, *SD* = 6.5). We computed fixation duration in seconds on two main areas of interest; the upper part and the lower part of the screen.

Expert pilots (French air force instructors) on the one hand tend to describe their visual behaviour as having a tendency to look as far as possible along the runway or beyond when flying. Secondly, the instruction of students follows this rule which has been established on the basis of the experience of these same instructors. As no data were available, we decided to explore this subjectively recalled behaviour by separating the screen during the experiment into these two main areas. The software used and the eye tracking device made it possible to monitor the time of fixation of the gazes in these areas. Thus, the scenes were divided into two equal areas of interest, (1) the

"upper part" (0x,0y; 1920x, 540y) and (2) the "lower part" (0x,540y; 1920x,1080y). The analyses were based on the average fixation duration in seconds. As a way to explore the link between information-gathering strategy and forward displacement it was decided to use correlation. Our assumptions include only experienced glider pilots because novices did not recall any flight experience, and therefore should not be affected by the type of gaze behaviour they employ.

*Table 2. Eye fixations Descriptive data. Full attention condition (FA); Divided attention (DA).*

| **Descriptive data Upper-part of the screen (s)** | ***Novices FA*** | ***Novices DA*** | ***Pilots FA*** | ***Pilots DA*** |
|---|---|---|---|---|
| *N* | 12 | 12 | 10 | 10 |
| *Mean* | 0.358 | 0.331 | 0.426 | 0.368 |
| *Std. Deviation* | 0.257 | 0.393 | 0.382 | 0.379 |
| **Descriptive data Lower-part of the screen (s)** | ***Novices FA*** | ***Novices DA*** | ***Pilots FA*** | ***Pilots DA*** |
| *N* | 12 | 12 | 10 | 10 |
| *Mean* | 1.504 | 1.53 | 1.658 | 1.743 |
| *Std. Deviation* | 0.416 | 0.456 | 0.496 | 0.487 |

Correlation analysis full attention (FA) trial block:

Experienced pilot's fixation data for the upper part were positively correlated to PSE, rs = 0.697, df=9, p =.016. Meaning that when pilots were looking at the upper part they recorded higher PSE score. Also, fixations on the upper part of the screen were positively correlated with the number of flying hours, r = 0.568, df = 9, p = 0.043. This measurement shows that pilots with the most flying experience were those who were looking at the upper part of the screen the most.

Experienced pilot's fixation data for the lower part were negatively correlated to PSE, rs = -0.564, df=9,p = 0.048. This indicates that when pilots were looking at the lower part they recorded lower PSE score. Also, fixations on the lower part of the screen were negatively correlated with the number of flying hours, r=-0.576, df=9, p = 0.041. This measure shows that pilots with less flying experience were those who were looking at the lower part of the screen the most.

Correlation analysis divided attention (DA) trial block:

No correlation in divided attention was reported, either among pilots or novices. No correlation between the number of flying hours and eye fixations was found.

**Discussion**

The displacement of the spatial representation of experienced pilots and novices, whose attention was divided, was evaluated for real dynamic scenes of glider landing. The first objective was to assess whether this protocol is sufficiently accurate to be used as a tool to assess the evolution of student pilots' skills as well as to explore the relationship between experienced pilot's visual features and the spatial memory bias. The second objective was to evaluate whether the processes underlying the displacement are sensitive to the automation process conventionally observed during the development of expertise.

Our findings are in line with the literature (Blättler et al., 2010; Blättler et al., 2012, 2013; Didierjean et al., 2014; Hiroki et al., 2014; Jin et al., 2017; Chen et al., 2019; Gorman, 2015; Anderson et al., 2019), indicating that there is an experience effect within the displacement process, here for natural dynamic glider landing scenes. It was found that novices have a significantly greater backward displacement than glider pilots even though, on average, the pilots only have 78 flight hours compare to 3000 hours for the expert participants of Blättler et al. (2011). These results are consistent with the possibility of using such a protocol to evaluate the evolution of student pilots' skills during their training. However, the fact that no group has any forward displacement should put this interpretation into perspective. According to Hubbard's (2010) vector addition model, it can be concluded that the device used here includes a "backward" factor that influences all groups. Thus, future studies will have to determine what this influence is in order to control it.

The results obtained when attention is divided are in line with those of Gorman et al. (2018). Experienced pilots did not show sensitivity to the division of attention on displacement, while for novices the division of attention acted as a "backward" influence. It is currently impossible to conclude on the automation of the processes underlying spatial representation, but in this particular situation, it appears that there is an automation process that induces a reduction in the "backward" shift effect among experienced pilots, even if it is not yet highlighted. In these terms, the use of this dual-task method, which modulates the direction and amplitude of the displacement, is an additional tool for evaluating performance evolution of student pilots during their training.

The results obtained with gaze fixations present a link between gaze fixations and displacement in individuals who are familiar with the scene and are free to explore it visually when their attention is not divided. These results explore a gap between the research about the expert's ocular behaviour and the expert's anticipation, whereas Gorman's study (2018) suggests that differences in displacement of spatial representation are unlikely to be related to differences in visual behaviours. Second, these data show an effect of the division of attention among experienced pilots. This effect might point to a sensitivity of experienced pilots to the division of attention that can be mapped into measures other than displacement. Further studies exploring more directly the link between a particular position in the scene and spatial memory bias should be made before eye tracking data might be used as a complementary tool to evaluate the evolution of the performance of student pilots during their training.

In conclusion, this study contributes to a better understanding of spatial representation in aviation and of pilots' visual interaction with a real-world environment. Our results have confirmed that trainees can be evaluated with the use of displacement measurement. Since gaze fixations also proved useful as a complementary index of pilots' anticipatory behaviours and experience, the use of eye tracking technology in addition to other data recording might assist in the comprehension and application of better training for situational awareness. Finally, the use of this evaluation methodology is expected to be useful in reducing the cost of training. Indeed, it should provide a way to assess the efficiency of simulation training (by evaluating anticipation scores) especially during critical phases as in landing scenarios.

**References**

Anderson, D.N., Gottwald, V.M., & Lawrence, G. (2019). Representational Momentum in the Expertise Context: Support for the Theory of Event Coding as an Explanation for Action Anticipation. *Frontiers in psychology, 10*, 1838.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience, 15*, 600-609.

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2011). Representational momentum in aviation. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1569-1577.

Blättler, C., Ferrari, V., Didierjean, A., & Marmèche, E. (2012). Role of expertise and action in motion extrapolation from real road scenes. *Visual cognition, 20*, 988-1001.

Blättler, C., Ferrari, V., Didierjean, A., Van Elslande, P., & Marmèche, E. (2010). Can expertise modulate representational momentum? *Visual Cognition, 18*, 1253-1273.

Chen, Y.H., Belleri, R., & Cesari, P. (2019). Representational momentum in adolescent dancers. *Psychological research*, 1-8.

De Sa Teixeira, N.A., & Hecht, H. (2014). The dynamic representation of gravity is suspended when the idiotropic vector is misaligned with gravity. *Journal of Vestibular Research, 24*, 267-279.

De Sá Teixeira, N.A., Hecht, H., & Oliveira, A.M. (2013). The representational dynamics of remembered projectile locations. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1690-1699.

De Sá Teixeira, N., & Oliveira, A.M. (2014). Spatial and foveal biases, not perceived mass or heaviness, explain the effect of target size on representational momentum and representational gravity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1664-1679.

Didierjean, A., Ferrari, V., & Blättler, C. (2014). Role of knowledge in motion extrapolation: The relevance of an approach contrasting experts and novices. In *Psychology of Learning and Motivation* (Vol. 61, pp. 215-235). Academic Press.

Finke, R.A., Freyd, J.J., & Shyi, G.C. (1986). Implied velocity and acceleration induce transformations of visual memory. *Journal of Experimental Psychology: General, 115*, 175-188.

Freyd, J.J., & Finke, R.A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 126-132.

Freyd, J.J., & Finke, R.A. (1985). A velocity effect for representational momentum. *Bulletin of the Psychonomic Society, 23*, 443-446.

Freyd, J.J., & Johnson, J.Q. (1987). Probing the time course of representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 259-268.

Froger, G., Blättler, C., Dubois, E., Camachon, C., & Bonnardel, N. (2018). Time-Interval Emphasis in an Aeronautical Dual-Task Context: A Countermeasure to Task Absorption. *Human Factors, 60*, 936-946.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*, 523-552.

Gorman, A. D., Abernethy, B., & Farrow, D. (2015). Evidence of different underlying processes in pattern recall and decision-making. *The Quarterly Journal of Experimental Psychology, 68*, 1813-1831.

Gorman, A.D., Abernethy, B., & Farrow, D. (2018). Reduced attentional focus and the influence on expert anticipatory perception. *Attention, Perception, & Psychophysics, 80*, 166-176.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K.R. (2006). Memory modulates color appearance. *Nature neuroscience, 9*, 1367–1368.

Hayes, A.E., & Freyd, J.J. (2002). Representational momentum when attention is divided. *Visual Cognition, 9*, 8-27.

Hubbard, T.L., & Bharucha, J.J. (1988). Judged displacement in apparent vertical and horizontal motion. *Perception & Psychophysics, 44*, 211-221.

Hubbard, T.L. (1990). Cognitive representation of linear motion: Possible direction and gravity effects in judged displacement. *Memory & Cognition, 18*, 299-309.

Hubbard, T.L. (1993). The effect of context on visual representational momentum. *Memory & Cognition, 21*, 103-114.

Hubbard, T.L. (1995a). Environmental invariants in the representation of motion: Implied dynamics and representational momentum, gravity, friction, and centripetal force. *Psychonomic Bulletin & Review, 2*, 322-338.

Hubbard, T.L. (1995b). Cognitive representation of motion: Evidence for friction and gravity analogues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 241-254.

Hubbard, T.L. (1996). Displacement in depth: Representational momentum and boundary extension. *Psychological Research, 59*, 33-47.

Hubbard, T.L. (1997). Target size and displacement along the axis of implied gravitational attraction: Effects of implied weight and evidence of representational gravity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1484-1493.

Hubbard, T.L. (2001). The effect of height in the picture plane on the forward displacement of ascending and descending targets. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 55*, 325-329.

Hubbard, T.L. (2005a). An effect of target orientation on representational momentum. *Paidéia (Ribeirão Preto), 15*, 207-216.

Hubbard, T.L. (2005b). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review, 12*, 822-851.

Hubbard, T.L. (2010). Approaches to representational momentum: Theories and models. *Space and time in perception and action*, 338-365.

Hubbard, T.L., & Ruppel, S.E. (2014). An effect of contrast and luminance on visual representational momentum for location. *Perception, 43*, 754-766.

Hubbard, T.L. (Ed.). (2018). *Spatial Biases in Perception and Cognition*. Cambridge University Press.

Jin, H., Wang, P., Fang, Z., Di, X., Ye, Z.E., Xu, G., & Rao, H. (2017). Effects of Badminton Expertise on Representational Momentum: A Combination of Cross-Sectional and Longitudinal Studies. *Frontiers in psychology, 8*, 1526.

Joordens, S., Spalek, T.M., Razmy, S., & Van Duijn, M. (2004). A Clockwork Orange: Compensation opposing momentum in memory for location. *Memory & Cognition, 32*, 39-50.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological review, 95*, 492-527.

Motes, M.A., Hubbard, T.L., Courtney, J.R., & Rypma, B. (2008). A principal components analysis of dynamic spatial memory biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1076-1083.

Nakamoto, H., Mori, S., Ikudome, S., Unenaka, S., & Imanaka, K. (2015). Effects of sport expertise on representational momentum during timing control. *Attention, Perception, & Psychophysics*, *77*, 961-971.

Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-tracking measures in aviation: a selective literature review. *The International Journal of Aerospace Psychology, 28*, 98-112.

Reingold, E.M., Charness, N., Pomplun, M., & Stampe, D.M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science, 12*, 48-55.

Strobach, T.I.L.O., Frensch, P.A., & Schubert, T.O.R.S.T.E.N. (2008). The temporal stability of skilled dual-task performance. In *Cognitive Science 2007*. Proceedings of the 8th Annual Conference of the Cognitive Science Society of Germany. Saarbrücken.

Thornton, I., & Hayes, A. (2004). Anticipating action in complex scenes. *Visual Cognition, 11*, 341-370.

Ziv, G. (2016). Gaze behavior and visual attention: A review of eye tracking studies in aviation. *The International Journal of Aviation Psychology, 26*, 75-104.

# Predicting self-assessment of the out-of-the-loop phenomenon from visual strategies during highly automated driving

*Damien Schnebelen[1], Camilo Charron[1,2], & Franck Mars[1]*
*[1]Centrale Nantes, CNRS, LS2N, Nantes, [2]Université Rennes 2*
*France*

**Abstract**

During highly automated driving, drivers do not physically control the vehicle anymore, but they still have to monitor the driving scene. This is particularly true for SAE level 3 (SAE International, 2016), as they need to be able to react quickly and safely to a take-over request. Without such an (even partial) monitoring, drivers are considered out-of-the-loop (OOTL) and safety may be compromised. This OOTL phenomenon may be particularly important for long automated driving periods. The current study aimed at scrutinizing driver's visual behaviour for a long period of highly automated driving (18 minutes). Intersections between gaze and 13 areas of interest (AOI) were analysed, considering both static (percentage of time gaze spent in one single AOI) and dynamic (transitions from one AOI to another) patterns. Then, a prediction of the self-reported OOTL level (subjective assessment) from gaze behaviour was performed using Partial Least Squares (PLS) regression models. The outputs of the PLS regressions allowed defining visual strategies associated with good monitoring of the driving scene and paved the way for an online estimation of the OOTL phenomenon based on driver's spontaneous visual behaviour.

**Introduction**

In manual driving, drivers must gather information about the driving scene and the vehicle (perceptual process), interpret this information (cognitive process) and act appropriately (motor process), which in turn generate information. However, with the imminent deployment of highly automated vehicles on the roads (between 2020 and 2030 depending on the organization (Chan, 2017)), where the operational driving task is performed by automation, drivers are likely to become supervisors of the driving scene. In this case, the perceptual-motor loop is neutralized, which has consequences on perception and cognition (Mole et al., 2019). This is referred to the out-of-the-loop (OOTL) phenomenon.

In automated driving, the OOTL phenomenon was investigated by comparing the driver's behaviour during automated and manual driving. In terms of gaze behaviour, automated driving leads to greater horizontal dispersion (Louw & Merat, 2017; Mackenzie & Harris, 2015), and a decrease of the percentage of glances to the road centre (Louw et al., 2015; Mackenzie & Harris, 2015). Similarly, in curve driving,

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

automated driving has been shown to enhance long-term anticipation (through look-ahead fixations) to the detriment of the short-term anticipation used to guide the vehicle (Mars & Navarro, 2012; Schnebelen et al., 2019).

The consequences of the OOTL phenomenon were also observed during level 3 automated driving , where drivers had to take control of the vehicle when automation required it. Indeed, in response to a critical case, drivers had longer reaction times in automated driving than in manual driving (Feldhütter et al., 2017; Neubauer et al., 2012; Saxby et al., 2013; Zeeb et al., 2015; Zeeb et al., 2017). Such changes in driver behaviour during takeover have been attributed to drivers being more OOTL during automated driving.

Drivers' performance during takeover is also affected by the duration of automation, with higher reactions times after a prolonged period of automation than after a short drive (Bourrelly et al., 2019; Feldhütter et al., 2017). Feldhütter et al. (2017) have shown, for instance, that a 20-minutes' drive in automated mode is sufficient to increase the reaction time to a takeover request. Drivers experienced mind wandering, distracting themselves from the supervision task, which impaired the perceptual and cognitive processing of information.

Recently, Merat et al. (2019) proposed an operational definition of the OOTL concept. It relies on two aspects: To be out-of-the-loop, drivers must not have physical control of the vehicle (no motor process), and must not monitor the driving scene (perception/cognition process). When the driver is in manual control, he is considered to be in-the-loop. An intermediate state, the on-the-loop (OTL) level, has been introduced to designate situations in which the driver correctly monitors the driving situation during autonomous driving. Thus, estimating the driver's ability to manage imminent takeover situations is a matter of determining whether the driver is OOTL or OTL based on the observation of his/her monitoring of the situation. However, the question of how to model and quantify what constitutes proper monitoring of the driving scene remains open.

Two principal issues were addressed in the present study:

- What is a good monitoring of the driving situation? In other words, can we identify the gaze behaviour characteristic of OOTL drivers?
- Is it possible to predict the driver's OOTL state from the observation of spontaneous gaze strategies?

In the current study, participants experienced an 18-min drive of automated driving (similar to Fleurette et al., 2017) without any non-driving activities to perform. The assessment of the OOTL state was based on the self-reported time of mind wandering during the drive. The driver's gaze behaviour was analysed considering 13 areas of interest, using static (percent of time on each AOI) and dynamic (transitions matrix from and to each AOI) patterns.

## Material and method

### *Participants*

This study involved 12 participants (N = 12; 3 females; 9 males), with a mean age of 21.4 years (SD = 5.34). Most of them were students from Centrale Nantes. They held a valid driver's licence (average driving experience: 9950 km/year, SD = 5500) and signed written informed consent to participate in this study.

### *Experimental device*

The experiment took place on a driving simulator (Figure 1), consisting in 3 screens (120° Field of View), with one additional screen for the HMI. The eye tracker (SmartEye Pro v5.9) computed gaze intersections with the screens at 20 Hz.

Most of the road was a 40 km two-lane dual carriageway, with a speed limit of 130 km/h in accordance with French regulations. Occasional changes in road geometry (temporary 3-lane traffic flow; highway exits; slope variation) and speed limits (130 km/h to 110 km/h) have been included to make driving less monotonous. In both directions on the highway, traffic was fluid, with 8 overtaking situations.


*Figure 1. Driving Simulator Setup.*

### *Procedure*

After a presentation of the driving simulator and a short drive in manual driving mode, participants were trained to activate (pressing a button) and disactivate (pressing the button, pedals or steering wheel) the automated mode. Instructions corresponding to a level 3 (SAE) automated driving were given: Automated driving was available only for a portion of road, and drivers had to take over the system when required (auditory + visual signals). Then, they experienced 4 takeover situations, with relatively long

(45 s; 2 situations) or short (8 s, 2 situations) time-to-collision. No collision occurred during the training session.

Then, the experiment proper started. Participants activated the automated driving mode just before entering the highway. Gaze data were recorded as soon as the vehicle was correctly inserted in the lane and reached 130 km/h. No major driving events appeared for the first 15 min on the highway to let the driver enough time to become out-of-the-loop. The driver did not perform any secondary task during that time. A critical case occurred at the 18$^{th}$ minute, and the scenario ended thirty seconds after. Participants were then asked to report on a continuous Likert scale the proportion of time spent thinking at something else than the driving task throughout the trial. Since this paper focuses on the link between gaze behaviour and the OOTL scores, the results on the critical case will not be presented here.

*Data structure and annotations*

*Definition of the OOTL score Y*
The evaluation of the percentage of time spent thinking about something else than the driving task may be considered as a self-assessment of the OOTL phenomenon. In that sense, the higher the percentage was, the more drivers estimated they were out-of-the-loop. Percentages for all participants were stored in a vector with 12 elements, named OOTL score and denoted Y.

*Definition of the matrix of gaze behaviour X*
The driving scene was divided into 13 areas of interest (AOI) (see figure 2):

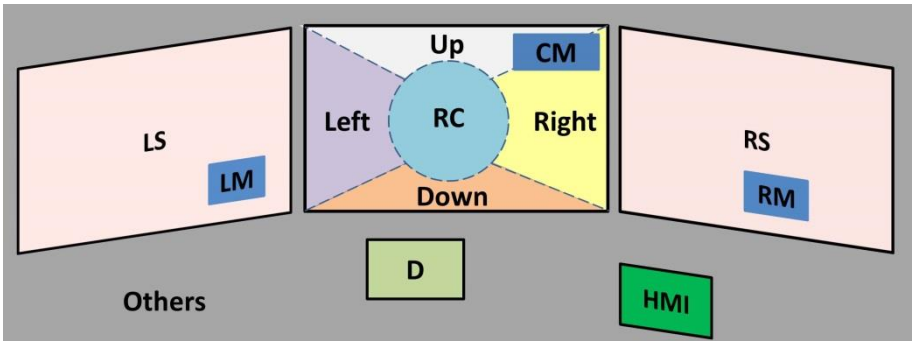



*Figure 2. Division of the driving environment into 13 areas of interest.*

- The central screen contained six areas: The central mirror (area CM), the road centre (RC), defined as a circular area of 8° radius in front of the driver, and 4 additional areas defined relatively to the road centre (Up, Left, Down, Right). The Percentage Road Centre (PRC) defined as the proportion of time spent in RC has been introduced by Victor (2005). A decrease of the PRC was found to be a good indicator of distraction during driving, as drivers reduced this time when visually or auditory distracted (Victor et al., 2005)
- Each peripheral screen contained two areas: The lateral mirror (LM, RM) and the remaining peripheral scene (LS, RS)

- The dashboard (D) and the HMI (HMI) All gaze data directed outside of all the previous areas were regrouped in area Others.

Drivers gaze behaviours for each participant were considered in this study as the combination of static (percentage of time in one AOI) and dynamic (transitions matrix between AOIs) patterns. Thus, a vector of 182 numerical indicators (= 13x13 transitions + 13 percentage of time on each AOI) summarizes gaze behaviour for one participant. When considering all participants, the matrix of gaze behaviour was named X and its size was 12 (participants) x 182 (visual indicators).

Due to the small number of observations (12) compared to the number of visual indicators (182), we used the PLS regression to predict the OOTL score from gaze behaviour. This method performed a decomposition of X and Y in orthogonal components in order to explain the maximum of the variance of Y. The components actually reflect the underlying structure of the prediction model.

*Data analysis*

Two sequential stages composed the analysis (Figure 3):

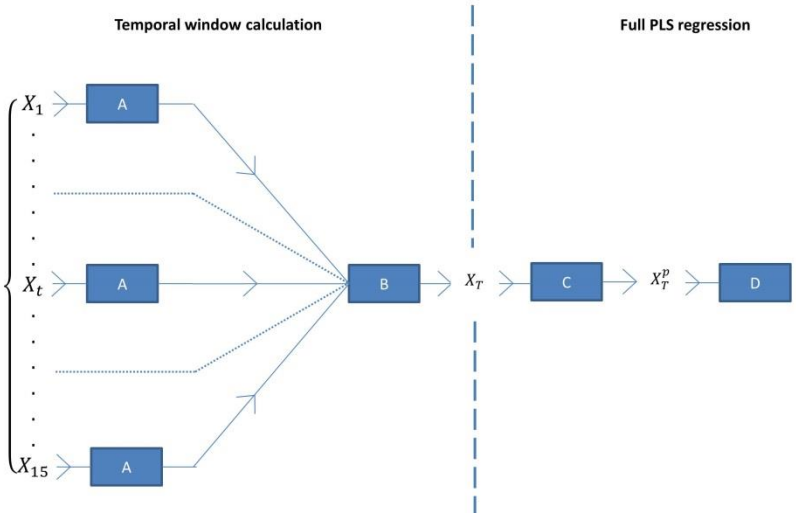



*Figure 3. Multi-step approach for data analysis.*

- The first one (steps A and B) focused on selecting the best time window (T) to predict the OOTL score. To do so, 15 matrixes of gaze behaviour were computed and labelled $X_t$. It differed by the time on which visual indicators were computed, that varied from 1 to 15 minutes.
- The selection was then based on the most stable (over time) and accurate (in terms of percent of variance explained) model of prediction. The second one (steps C & D) consisted of predicting the OOTL score using $X_T$ as predictors and the PLS regression model. After reducing the dimension of X to increase

prediction power (step C), the model was tested using the training and the validation data set (step D).

The details of data analysis are presented in the results section.

## Results

### *OOTL Score*

The OOTL scores (Figure 4) showed large variations between participants (range ~= 75%). The median score was 43%. Even in the absence of a secondary task, some participants (9 to 12) declared that they spent 80% of the time thinking at something else than the driving task.

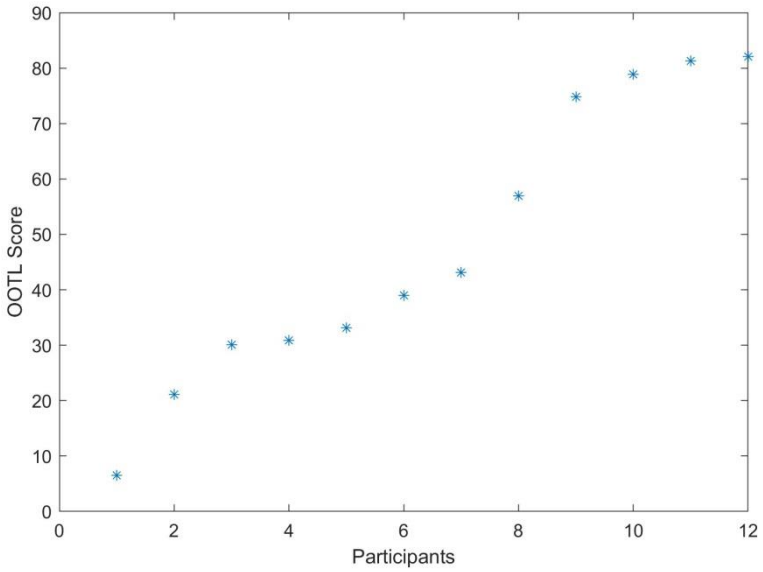



*Figure 4. OOTL scores reported by the participants.*

### *Time window selection*

On the first step (A), the optimal number of components for each matrix $X_t$ was obtained by minimizing the mean square error of prediction. This number of components, reflecting the structure of the prediction model, actually changed depending on the integration window (figure 5), but reached a stability level for time windows higher than 9 minutes. The most appropriate temporal window, labelled T, was selected (step B) as the one maximizing the variance of the OOTL score explained, among the stable models. All subsequent analysis referred to the matrix of gaze behaviour computed over T = 11 minutes of automated driving.

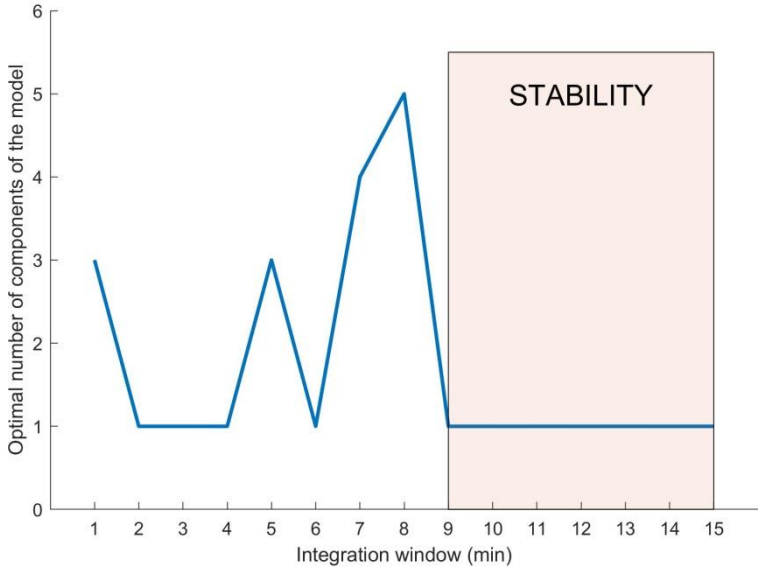



*Figure 5. Optimal number of components of the PLS regression as a function of the integration window. The figure shows that model stability was achieved from the 9th minute.*

*Reduction of the number of visual indicators*

After selecting the most appropriate time window, the prediction model explained 62.53% of the variance of Y, using the 182 visual indicators. Then, the aim was to reduce the number of visual indicators by selecting only the most relevant visual indicators.

The PLS regression is a linear model: the variable to be estimated ($\hat{Y}$) and the predictor ($X_T$) are linked by a matrix of coefficients C: $\hat{Y} = C * X_T$. The relevant indicators were determined by the absolute magnitude of their coefficient: If the magnitude was close to zero, the contribution to the prediction was negligible. On the contrary, a high magnitude indicated a very important indicator for the prediction.

In practice, the coefficients magnitudes were compared with an increasing threshold value. A new regression model was computed for each partial matrix (i.e. a matrix comprising only those indicators whose coefficient amplitude exceeded the threshold value). The threshold was increased by step of 0.005 until the percentage of variance of Y explained by the partial model stopped increasing. With our data, the maximum of explained variance was 85.64%, with only 8 visual indicators (Figure 6).

On these 8 indicators, 5 contributed to an increase of the OOTL score (in red on Figure 6): Taking the eyes off the central mirror to look away from the driving scene, taking the eyes off the road centre area to look down or away from the driving scene, spending too much time in the down area. By contrast, 3 indicators contributed to a reduction of the OOTL score (green arrows on Figure 6): Redirecting the gaze to the

road centre or to the left side of the driving scene from any area outside the driving scene, take your eyes off the road centre to check the left rear-view mirror.

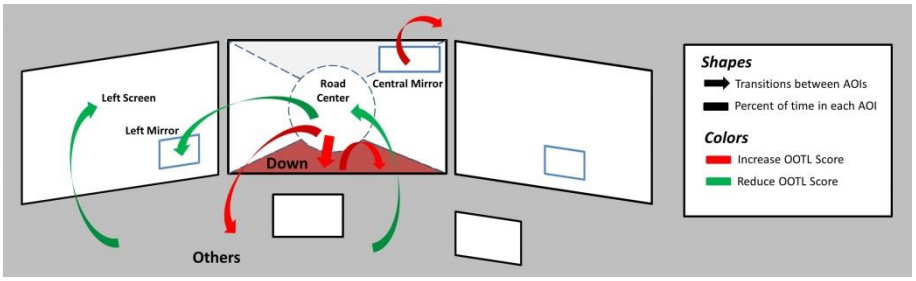



*Figure 6. Visual indicators relevant for OOTL score prediction.*

*Final prediction of the OOTL score*

A final PLS model (step D) was computed to predict the OOTL score from the best partial matrix (containing the 8 visual indicators relevant for the prediction). The prediction of the model compared to real values of the OOTL score is presented on Figure 7.

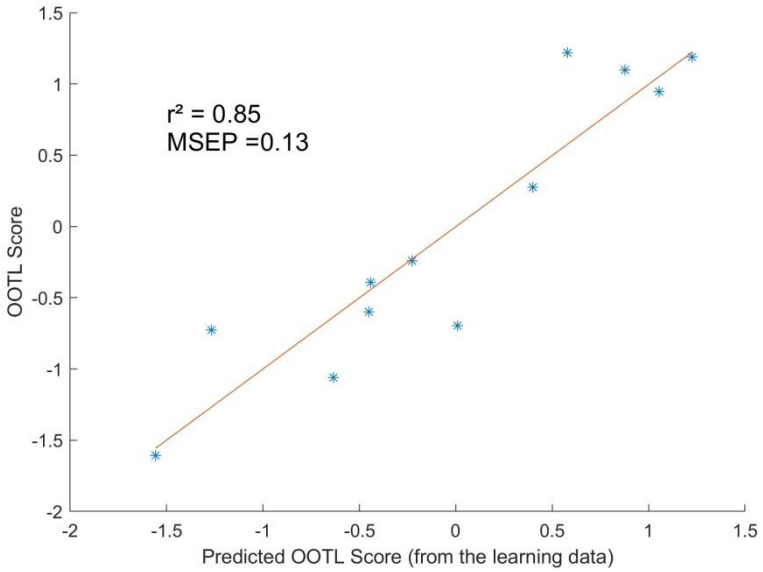



*Figure 7. Correlation plot between the OOTL score and the prediction of the OOTL score by PLS regression.*

The PLS regression performed a good estimation of the OOTL score, with a low mean square error of prediction (0.13) and a significant positive correlation between the estimated and real values ($r = 0.92$, $p<0.01$).

**Discussion**

During automated driving, the OOTL phenomenon results from an incorrect monitoring of the driving situation (Merat et al., 2019). The alternative state, namely being OTL (on-the-loop), corresponds to passive drivers who satisfactorily monitor their driving environment. However, a more precise definition of what constitutes a great monitoring of the environment is still needed to distinguish OTL from OOTL drivers. This study investigated this issue in a highway driving context with the analysis of drivers gaze behaviour, with both static (percent of time in AOI) and dynamic (transitions between AOI) patterns. The methods consisted in using PLS regressions to identify the most characteristic elements of the gaze behaviour of OTL and OOTL drivers. The multi-step approach began with 182 visual indicators as an input matrix, and retained in the end only 8 relevant elements to predict an accurate OOTL score.

The results revealed that drivers with a lower OOTL score made more transitions from the road centre to the left mirror. After spending time looking at area unrelated to driving ("others" area), they returned more frequently to the road (road centre area) or to the left screen where they could monitor traffic. Conversely, drivers with higher OOTL scores made more transitions from the road centre to areas irrelevant to driving. They spent more time and made multiple fixations in the lower part of the front screen.

These findings may be interpreted in terms of the adequacy of the driver's gaze strategy to maintain good situation awareness (Endsleigh, 1995) in autonomous mode. Situation Awareness (SA) during automated driving actually involved three levels: Perception, Comprehension and Projection (Merat et al, 2019). In the current study, OTL drivers remained dynamically aware of their surrounding by regularly checking the left lane and mirror. This certainly have helped to anticipate future hazards. They also remained attentive to the road well-ahead in time. In other words, these gaze strategies allowed to perceive, comprehend and project on the future state of the driving situation in an appropriate way, i.e. to have a good enough SA. On the other hand, the OOTL drivers' gaze was more strongly attracted by irrelevant information inside or outside the simulator. Even when looking at the driving scene, the driver favoured the road immediately in front of them (down area), suggesting a lack of visual anticipation.

In the current study, PLS regressions appear to be a relevant approach to predict the driver's state from spontaneous gaze behaviour. Indeed, PLS regressions allowed finding one optimal temporal window, reducing the dimensions of the matrix of gaze behaviour from 182 to 8 relevant elements, but also indicated whether they contributed to increase or decrease the OOTL score. Then, the prediction of the OOTL score given by the model was accurate with a strong correlation between the predicted and the real values. However, a validation step (i.e. testing the model with another set of gaze behaviour data) is required to confirm the results presented here.

In the current study, the OOTL score could be predicted from the driver's spontaneous strategies over 11 minutes of automated driving. For further research, it may be interesting to apply this model on shorter durations of automated driving, and to apply similar methods to other driving contexts.

## Conclusion

The current study used PLS regression to satisfactory predict driver's state from their visual monitoring of the driving situation. The analysis of gaze behaviour proved that an appropriate gaze strategy for being on the loop requires to get information on the oncoming traffic as well as interleaving glances on the road centre. To provide a more accurate detection of the OOTL phenomenon during automated driving, the analysis of gaze behaviour might be coupled with other approaches, for example by incorporating physiological measurements or the analysis of the driver's posture in the diagnosis.

## References

Bourrelly, A., de Naurois, C.J., Zran, A., Rampillon, F., Vercher, J.-L., & Bourdin, C. (2019). Long automated driving phase affects take-over performance. *IET Intelligent Transport Systems*, *13*, 1249 - 1255.

Chan, C.-Y. (2017). Advancements, prospects, and impacts of automated driving systems. *International Journal of Transportation Science and Technology*, *6*, 208-216. https://doi.org/10.1016/j.ijtst.2017.07.008.

Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Feldhütter, A., Gold, C., Schneider, S., & Bengler, K. (2017). How the duration of automated driving influences take-over performance and gaze behaviour. In *Advances in Ergonomic Design of Systems, Products and Processes* (p. 309–318). Springer.

Louw, T, Kountouriotis, G., Carsten, O., & Merat, N. (2015). Driver Inattention During Vehicle Automation : How Does Driver Engagement Affect Resumption Of Control? *New South Wales*, 16.

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

Mackenzie, A.K., & Harris, J.M. (2015). Eye movements and hazard perception in active and passive driving. *Visual Cognition*, *23*, 736–757.

Mars, F., & Navarro, J. (2012). Where we look when we drive with or without active steering wheel control. *PLoS One*, *7*(8), e43858.

Merat, N., Seppelt, B., Louw, T. et al. (2019). The "out-of-the-loop" concept in automated driving : Proposed definition, measures and implications. *Cognition, Technology & Work 21*, 87–98. https://doi.org/10.1007/s10111-018-0525-8.

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

Neubauer, C., Matthews, G., Langheim, L., & Saxby, D. (2012). Fatigue and voluntary utilization of automation in simulated driving. *Human Factors*, *54*, 734–746.

SAE International. (2016). *Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems (J3016)*.

Saxby, D.J., Matthews, G., Warm, J.S., Hitchcock, E.M., & Neubauer, C. (2013). Active and Passive Fatigue in Simulated Driving : Discriminating Styles of Workload Regulation and Their Safety Impacts. *Journal of Experimental Psychology. Applied*, *19*, 287-300. https://doi.org/10.1037/a0034386.

Schnebelen, D., Lappi, O., Mole, C., Pekkanen, J., & Mars, F. (2019). Looking at the Road When Driving Around Bends : Influence of Vehicle Automation and Speed. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01699.

Victor, T. (2005). *Keeping eye and mind on the road* (PhD Thesis). Acta Universitatis Upsaliensis.

Victor, T.W., Harbluk, J.L., & Engström, J.A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F*, *8*, 167-190. https://doi.org/10.1016/j.trf.2005.04.014.

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

Zeeb, K., Härtel, M., Buchner, A., & Schrauf, M. (2017). Why is steering not the same as braking? The impact of non-driving related tasks on lateral and longitudinal driver interventions during conditionally automated driving. *Transportation Research Part F*, *50*, 65-79. https://doi.org/10.1016/j.trf.2017.07.008.

# Task load of professional drivers during level 2 and 3 automated driving

*Hans-Joachim Bieg[1], Constantina Daniilidou[2], Britta Michel[2], & Anna Sprung[2]*
*[1]Robert-Bosch GmbH, [2]MAN Truck & Bus SE*
*Germany*

**Abstract**

As level 2 automated driving systems (SAE partial automation) become more elaborate, the similarity to a level 3 system (SAE conditional automation), from a driver's perspective, is gradually increasing. We examined differences in driver behaviour concerning level 2 and 3 automation in a driving simulator experiment with 31 professional truck drivers. All drivers received specific instructions concerning differences in the driver's role in both automation levels. Despite this, drivers had difficulties in adapting their behaviour to the different demands of level 2 vs. level 3 driving. An analysis of driver reactions shows potentially critical lapses in attention during level 2 drives, when drivers were performing an engaging non-driving related task while driving. A comparison of drivers' gaze distributions suggests that these lapses are likely due to a de-prioritisation of on-road glances during task performance. These results highlight the difficulties that may accompany improvements of level 2 automation performance and underline the need for measures to assist drivers in adapting their behaviour accordingly.

**Introduction and previous work**

Advancing sensor technology and signal processing methods lead to a gradual improvement of automation performance in automated level 2 (SAE 2016) vehicles, resulting in fewer driving mistakes that vindicate the driver's supervisory role. From a layman's perspective, well-functioning level 2 systems more and more seem like level 3 systems (Campbell et al., 2018). These systems seemingly need no supervision, despite the fact that the driver is still considered a crucial safety factor by its designers (SAE 2016).

In both partial (SAE level 2) and conditional automation (SAE level 3) the driver's main task can be described in terms of a vigilance task (Davies & Parasuraman, 1982): In partial automation drivers monitor longitudinal and lateral control to detect and respond to silent automation failures. In conditional and higher automation modes, drivers detect and respond to requests to intervene, which are issued by the automated system when it approaches a system boundary. System designers may adjust the saliency of requests to intervene such that the signal detection task in level 3

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

automation and higher becomes relatively easy. For example, relevant design guidelines mandate the use of multimodal warnings and offer advice on colour, symbolism, and warning tones or messages (e.g., Campbell et al., 2018). No such control over signal saliency is available for level 2 driving: silent failures may take on the form of lane drifts or non-reactions (NTSB 2017).

Previous work suggests that drivers may find it difficult to appreciate the demands of a (well-functioning) partially automated vehicle. For example, Omae et al. (2005) and Llaneras et al. (2013) found that drivers were more likely to engage in non-driving related tasks that restricted their monitoring ability such as interacting with a handheld electronic device. Such results may potentially be explained by assuming that drivers lacked exact information about the automated system's capabilities or their monitoring duties and may be combated by appropriate instructions (Campbell et al., 2018).

The presented work directly compares the behaviour of instructed, professional truck drivers to examine whether the drivers are able to adjust to the differential demands of level 2 and level 3 automation.

## Materials and methods

The study was conducted in MAN's fixed-base, high-fidelity driving simulator with professional truck drivers.

### *Participants*

Of 32 participants, one aborted the experiment whose data is excluded in the following. The remaining 31 participants of the study (all males, M=42.5 years, SD=14.6, range=22-70 years) were in possession of a valid driver's license for trucks or busses (German C/CE or D/DE license, first issued on average 20 years ago, SD=13 years). Most of the drivers were currently working full-time as professional truck or bus drivers (mainly long distance), 10 of the drivers were working in part-time. Half of the drivers reported a yearly mileage of more than 100.000 km, 11 participants a mileage between 10.000 up to 100.000 km and 4 participants between 500 and 10.000 km.

### *Procedure*

Upon arriving, participants were informed about the nature and duration of the experiment as well as the safety instructions for the simulator. All participants provided written informed consent before testing and received monetary compensation for their participation. Participants were equipped with electrodes for measuring heart rate and skin conductance. After these preparations, an initial questionnaire was filled out and participants were instructed how to perform the non-driving related task in the experiment – a quiz task.

The experiment consisted of a familiarisation drive and two test drives (*L2*, *L3*, counterbalanced) in the simulator. In the familiarisation drive, drivers received instructions on how to operate the vehicle and automation and were familiarised with

the Driving Activity Load Index (DALI) questionnaire. Each test drive was preceded by a short period to obtain a baseline for the physiological measures followed by 50 minutes of automated driving in each automation condition. The *L2* test drive was designed to resemble a drive with a partially automated vehicle (SAE level 2), the *L3* to resemble a drive with conditional automation (SAE level 3). During both test drives, participants experienced three non-driving related task conditions in randomized order: no task (*none*), an auditory quiz (*auditory*), and a visual-manual quiz (*visual*) for 10 minutes. After each condition, participants were asked to fill out the DALI questionnaire. A brief break was made in between both test drives.

*Automation levels*

The automation levels were implemented as follows: a peripheral detection/vigilance task (PVT) was embedded within the driving period as a proxy for a silent automation failure (partial, *L2*) or take-over request (high, *L3*). The PVT comprised of a small green rectangle that randomly appeared on the simulator screen (see Figure 1). In the high automation condition (*L3*), a short sound and a bright blue LED above the steering wheel (take-over cue) announced the appearance of the rectangle 10 seconds in advance. During partial automation (*L2)* no such cue was presented. Participants were instructed to respond to the PVT by pressing a button on the steering wheel as quickly as possible and to prioritize this task over others.

To avoid potential confounds, missed PVT prompts did not result in any feedback to the driver and no particularly arousing or aversive situations, (i.e., potential crashes due to inappropriate take-over behaviour) was presented in the experiment. The rather abstract implementation of the automation HMI and of the required responses made sure that differences between the two automation conditions were reduced to the core distinguishing differences of both vigilance tasks.

*Figure 1. Top: peripheral vigilance task (PVT) in the L2 condition. Bottom: PVT in the L3 condition with the take-over LED near the steering wheel.*

*Non-driving related Tasks*

The non-driving related task chosen in the current study is based on a quiz (Petermann-Stock et al., 2013). The quiz consisted of 240 questions covering the fields of common knowledge, proverbs, movies and TV shows, sports, geography, cars and trucks. The questions of the original quiz were adapted to reduce the skill level required and to cater to the targeted audience of truck drivers (e.g., by selecting trucking specific questions). For each question, three possible answers were presented whereas only one of them was correct. Two versions of the quiz were presented to engage the driver into tasks with similar characteristics to (hands-free) telephone conversation or using an electronic handheld device: In the *auditory* condition, the question as well as the answers were read to the participant. The participant was asked to provide a verbal answer. In the *visual*-manual condition, questions were presented visually on a tablet computer. In order to reveal a possible answer, the participant had to touch the screen where an indicator (A, B, C) was presented (Figure 2).

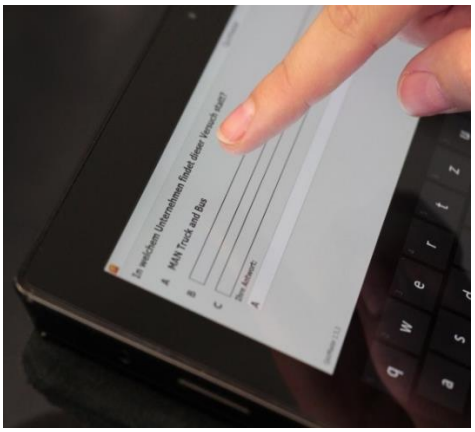



*Figure 2. Visual-manual quiz.*

*Driving simulator & roadway*

The study was conducted in the static driving simulator of MAN Truck & Bus AG. This simulator provides a 180° visual simulation as well as acoustic simulation of motor sounds and other vehicles. It is a mock-up of a full size TGX cabin built with aluminium profiles. The simulation software used was SILAB (Würzburg Institute for Traffic Sciences GmbH) which allows for the recording of vehicular data (e.g., velocity) as well as the integration of the physiological measurement equipment. The roadway implemented for the two test drives consisted of a 2-lane-highway with steady, but little traffic.

*Eye-tracking*

Two infrared video cameras (ON Semiconductor PYTHON1300, 1.3 MP) were installed in the simulator cabin near the A-pillar and centre console. Images were recorded throughout the experiment at a rate of 60 Hz. Processing of the images occurred off-line. Gaze direction information was computed using proprietary eye-tracking software (SmartEye embedded SDK v0.8.2). From this, gaze heading and pitch angles were computed. Both values were normalized using the mean gaze heading and pitch angles that were computed for each participant for the L2 automation drive without a non-driving related task. A road-centre region was defined with +/- 20° eccentricity and gaze information was classified as “on road” or “off road” based on this definition and the recorded, normalized heading and pitch angles. From this, the percentage of road centre gazes (PRC) for each condition was computed.

*Self-assessment of workload*

A self-assessment of workload was conducted using the Driving Activity Load Index (DALI). This questionnaire is designed for the assessment of workload during driving and addresses different factors such as perceptual load, mental workload and the driver’s state (Pauzié, 2008). All participants were asked to fill out the questionnaire during both test drives after each condition.

*Physiology*

In addition, participants' skin conductance and heart rate were recorded during the experiment, the analysis of which is omitted in the present paper.

*Data analysis*

Statistical data analysis was conducted using IBM SPSS (version 24.0) and R (version 3.5.2). If applicable, data was analysed using repeated measures ANOVAs with the factors automation level (*L2, L3*) and task (*none, visual, auditory*). For post-hoc analyses, t-tests for repeated measures were performed. Results of the ANOVA were corrected according to Greenhouse-Geisser whenever the Mauchly test of sphericity indicated heterogeneity of covariance. In the case of a violation of requirements, non-parametric ANOVAs (Friedman) and Wilcoxon Tests were used. Findings were considered statistically significant at $p < 0.05$.

**Results**

*Monitoring ability*

Participants' primary task consisted of an abstract detection task (PVT), which captured the core differences in terms of signal saliency between the two automation conditions (*L2/L3*). Figure 3 depicts the number of correct detection responses and missed signals per condition (three signals were presented per condition). Detection ability significantly differs between automation levels and groups ($\chi^2(31)=80.22$, $p < 0.001$). In *L3* most participants [77 – 90%] manage to react to all three out of three stimuli (3/3) which is significantly higher than in *L2* ($Z = [-4.48;-2.36]$, all $p < 0.05$). Here, only 11-63% of the participants are able to react to all three presented stimuli. In *L2*, there are significant differences between the three conditions. When drivers are engaged in the *visual*-manual task, fewer signals are detected in comparison to the *auditory* task condition ($Z = -3.27$, $p < 0.01$) or when participants performed no (*none*) non-driving related task ($Z = -4.10$, $p < 0.001$). The difference between the *none* and *auditory* condition is not significant ($Z = -1.70$, $p = 0.09$). In *L3*, differences are also found for the *auditory* and *visual* task condition ($Z = -2.26$, $p < 0.05$).

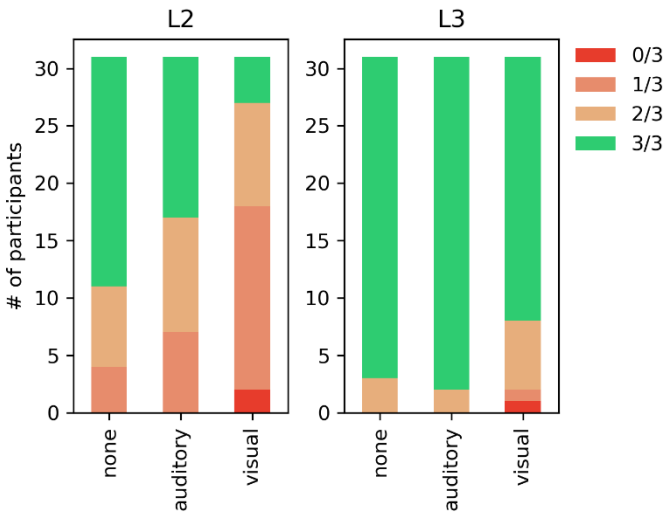



*Figure 3. Number of correct responses to the detection task (PVT).*

*Gaze behaviour*

Gaze behaviour was analysed by computing a percentage road centre (PRC) statistic per participant and drive based on angular gaze information. This analysis was performed for a subset of the 31 participants. Eight participants were excluded from this analysis either because of incomplete video recordings, issues with eye-tracker calibration and accuracy or because drivers partially made use of reading glasses. The following section presents the results for the data from the remaining 23 participants.

The analysis shows a significant difference between automation levels ($F(1,22) = 41.8$, $p < 0.01$), revealing that drivers are glancing at the roadway less frequently during *L3* (Figure 4). The analysis also shows a significant main effect of the type of non-driving related task ($F(2,44) = 268.6$, $p < 0.01$). Bonferroni corrected post-hoc comparisons corroborate the assumption that drivers' gaze is away from the road much more frequently during the *visual*-manual task condition in *L3* in comparison to no activity (mean of difference = 0.34, $t(22) = 14.9$, $p < 0.01$). Importantly, drivers' gaze is also off-road more frequently during the *visual*-manual task condition during *L2* (mean of difference = 0.36, $t(22) = 18.9$, $p < 0.01$). The comparison of PRCs in the *visual*-manual condition between *L3* and *L2* shows a small (mean of differences = 0.092) but significant difference ($t(22) = 4.2$, $p < 0.01$).

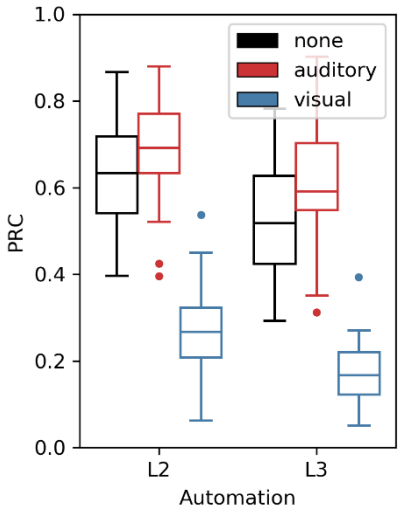



*Figure 4. Percentage road centre (PRC) distributions.*

*Self-assessment of workload*

For this analysis the data from one participant, who only partially completed all DALI questionnaires was removed. In general, the participants' ratings range in the lower end of the DALI scale. Significant differences in self-assessed workload are found between automation levels ($F(1,30) = 7.29$, $p < 0.05$) as well as between tasks ($F(2,60) = 11.15$, $p < 0.01$). Post-hoc tests show that workload is considered higher in the *visual*-manual task condition compared to the *none* condition across both automation conditions ($t(30) = [-4.50;-2.65]$, all $p < 0.05$). In addition, this condition is rated significantly lower in workload during *L3* automation ($t(30) = 4.0$, $p < 0.01$).

Individual DALI factor results are shown in Figure 6. Apart from an overall muting effect on all DALI factors, drivers' responses in *L3* in particular show a reduction of the stress factor, which in this automation condition seems to result in even lower stress ratings as the *none* condition. Note that the tactile factor was omitted in the present experiment.

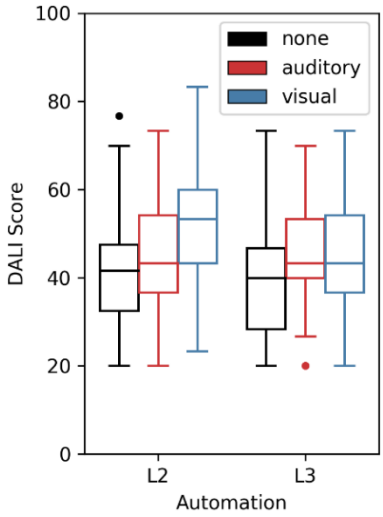



*Figure 5. Subjective workload assessment results (DALI questionnaire score).*

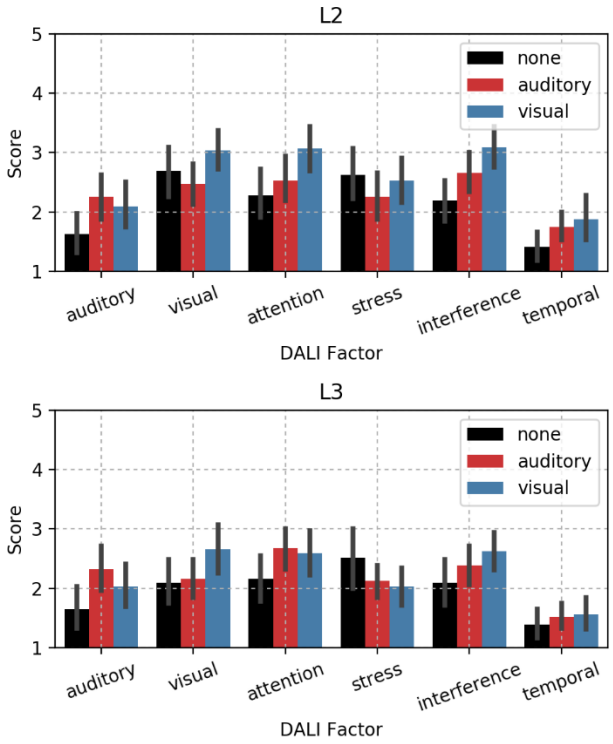



*Figure 6. DALI factors.*

**Discussion**

The present study investigated differences in professional truck drivers behaviour between level 2 (*L2*) and level 3 (*L3*) automated driving. Even though drivers received explicit instructions regarding their core, driving-related task, namely system supervision in *L2* and reaction to take-over requests in *L3*, and despite explicit requests by the experimenter to prioritise this task, drivers showed marked lapses in monitoring performance during *L2* – in particularly when they were engaged in a visual-manual non-driving related activity. For example, only two of the 31 drivers correctly responded to all three PVT (peripheral vigilance task) prompts that were presented as a proxy for a silent automation failure during the *L2* condition when they were also engaged in the visual-manual quiz task.

Failures to respond to PVT prompts (i.e., take-over requests) were also observed in *L3* driving, albeit at a much smaller frequency. In interpreting the *L3* response proportions, it must be noted that the system did not escalate the PVT prompt using e.g., additional or louder warning tones when drivers did not respond, as would be the case and easily feasible in a more comprehensive automation HMI (e.g., Llaneras et al., 2017).

PVT targets were solely presented visually in the *L2* condition, requiring participants to adapt their visual scanning behaviour: In this condition, participants should have prioritised the monitoring task in comparison to the L3 condition. Although the comparison of on-road glance distributions showed that drivers were looking at the road more frequently during L2, the difference to glance proportions in L3 was very small and comparatively low (ca. 30 % on average). This was the case despite participants noting the differential demands of both automation conditions as per the DALI questionnaire, where the visual task received lower workload ratings during L3 in comparison to L2.

Together, these findings highlight the fact that drivers seem to have difficulties in prioritizing their monitoring activity and non-driving related task adequately – despite clear instructions by the experimenter regarding the expected priorities. Reasons for this may be found in a lack of motivation regarding the primary monitoring task since the study was conducted in a driving simulator and not in a real vehicle. Yet, studies in real vehicles (e.g., Omae et al., 2005) and recent incidents in real traffic with *L2* vehicles (e.g., NTSB, 2016) suggest that drivers' priorities may be similarly misguided. Granted, the latter results and incidents were observed for non-professional drivers and it stands to reason whether professional drivers exhibit similar behaviour in a real vehicle.

Instead of factors that pertain to drivers' motivation due to the simulator setting, we suggest that the present results may be partially explained by a lack of drivers' self-awareness regarding their monitoring behaviour when performing a particularly engaging non-driving related activity. For example, time perception is known to be malleable by task characteristics (e.g., Hart et al., 1979), potentially skewing subjects' perception of the time spent with one or the other task in a dual-task scenario.

Secondly, drivers may exhibit an incomplete understanding of what constitutes necessary monitoring performance, e.g., unrealistic beliefs about failure frequencies or detection ability. Such intuitions are hard to gather from instructions but are typically acquired through interactive experience and in particularly consequences of one's action or inaction. In the present experiment, consequences (e.g., crashes) of failing to monitor properly were not presented purposely for other reasons, but it is also expected that an absence of performance feedback regarding the monitoring task is realistic. Technical advancements will gradually decrease failure rates of automated systems. In a well-working L2 vehicle, opportunities for a reinforcement of proper monitoring behaviour will thus become rarer. Ideally, such reinforcement is provided in terms of positive reinforcement, e.g. a system failure that is compensated for by a successful intervention by the driver. With decreasing failure rates, however, potential failures may unfortunately even lead to more fatal outcomes because drivers are encountering them unprepared.

Unfortunately, past research has shown that drivers are more likely to take up non-driving related activities while driving (monitoring) an automated vehicle, presumably simply to combat boredom (e.g., Omae et al., 2005, see also review by Cunningham & Regan, 2017). Taken together, these observations may be of relevance for the designers of automated vehicles and vehicle HMIs. For example, designers may strive to implement in-vehicle systems that offer and encourage safe non-driving related activities (i.e., auditory-verbal activities) or that facilitate transforming unsafe activities into safe activities, for example, by offering services to integrate a driver's mobile devices into the vehicle's infotainment system (e.g., by "pairing" a device, see NHTSA, 2013). Another possibility is the introduction of on-line driver monitoring and warning systems (e.g., Llaneras et al., 2017). These systems could for example be employed to raise drivers' awareness concerning their monitoring duties and appropriateness of their glance behaviour. Such approaches may be of particular relevance in vehicles that offer multiple automation levels (i.e., *L2* and *L3*), to assist drivers in adapting to the respective automation requirements.

**Acknowledgments**

This research was supported by the German Federal Ministry of Economics and Technology (Bundesministerium für Wirtschaft und Energie, BMWi) through the TANGO project (www.projekt-tango-trucks.com).

**References**

Campbell, J.L., Brown, J.L., Graving, J. S., Richard, C.M., Lichty, M.G., Bacon, L.P., Morgan, J.F., Li, H., Williams, D.N., & Sanquist, T. (2018). *Human Factors Design Guidance for Level 2 and Level 3 Automated Driving Concepts* (Report DOT HS 812 555). Columbus, Ohio, USA: Battelle Memorial Institute.

Cunningham, M.L. & Regan, M.A. (2017). Driver distraction and inattention in the realm of automated driving. *IET Intelligent Transport Systems, 12,* 407-413.

Davies, D.R. & Parasuraman, R. (1982). *The psychology of vigilance*. London: Academic Press.

Hart, S.G., McPherson, D., & Loomis, L. (1978). Time estimation as a secondary task to measure workload: summary of research (N79-15634). In *14th Annual Conference on Manual Control* (pp. 693-712). Mountain View, California: NASA Ames Research Center.

Llaneras, R.E., Salinger, J., & Green, C.A. (2013) Human factors issues associated with limited ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. In *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 92-98). Iowa City, Iowa: University of Iowa.

Llaneras, R.E., Cannon, B.R., & Green, C.A. (2017). Strategies to Assist Drivers in Remaining Attentive While Under Partially Automated Driving: Verification of Human--Machine Interface Concepts. *Transportation Research Record*, *2663*, 20-26.

NHTSA (2013). Visual-Manual NHTSA Driver Distraction Guidelines for Portable and Aftermarket Devices, Docket No. NHTSA-2013-0137. *Federal Register*, 81, 87656-87683.

NTSB (2017). *Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck near Williston, Florida May 7, 2016* (Report NTSB/HAR-17/02). Washington, D.C., USA: NTSB.

Omae, M., Hashimoto, N., Sugamoto, T., & Shimizu, H. (2005). Measurement of driver's reaction time to failure of steering controller during automatic driving. *Review of automotive engineering*, *26*, 213-215.

Pauzié, A. (2008) A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, *2*, 315-322.

Petermann-Stock, I., Hackenberg, L., Muhr, T., & Mergl, C. (2013). Wie lange braucht der Fahrer? Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt. *6. Tagung Fahrerassistenzsysteme*. Munich, Germany: TU Munich, Lehrstuhl für Fahrzeugtechnik (FTM).

SAE (2016). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems* (Report J3016). Warrendale, Pennsylvania, USA: SAE International.

# Driving with an L3 – motorway chauffeur: How do drivers use their driving time?

*Johanna Wörle & Barbara Metz*
*Würzburg Institute for Traffic Sciences GmbH, Veitshöchheim*
*Germany*

**Abstract**

Advances in the technology of automated driving (AD) raises the question how AD might change driving in general. Especially the option for users to engage in other activities is seen as a major benefit. The aim of the presented study was to investigate which non-driving related activities (NDRAs) drivers want to engage in during conditionally automated driving and what proportion of the driving time they spend on these activities. In a driving simulator study, N=31 drivers used an L3-motorway chauffeur during six driving sessions which took place at six different days. Drivers were free to bring whatever they want to engage in during the drives and to use the AD function as they liked. Handling of the system, drivers' state and drivers' engagement with self-chosen side tasks was continuously annotated by the experimenter for all drives. After every drive, evaluation and acceptance of the system was assessed with a questionnaire. Drivers spend an average of 80% of the time the AD function was active on NDRAs. Only when they were fatigued this number decreased. The time spend on activities that involved both hands increased over the drives. By far the most popular activity was smartphone use. The relevance of the study findings is interpreted with regard to safety and societal benefits.

**Introduction**

Automated driving is expected to yield benefits such as an increased travel comfort and a more productive use of travel time. When reaching the level of conditional automation, i.e. level 3 according to the SAE classification (SAE, 2018), drivers will not be required to monitor the system and are allowed to engage in secondary activities. Users want to spend the time travelling in an automated vehicle for activities such as private communication, route information, eating and drinking, entertainment, work, wellness and sleep (Dungs et al., 2016). The engagement in such side-activities or secondary activities is widely investigated in human factors research in terms of their distractive potential or the ability to take over the driving task when being engaged in side-tasks. This "new role" of the driver in highly automated driving is subject to investigation in many research projects.

The work presented here is part of the research project L3Pilot (https://www.l3pilot.eu/). Two assessment areas in the L3Pilot project are potential safety impacts as well as socio-economic impacts of automated driving. For both of

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

these impact areas, drivers' engagement in secondary tasks is relevant. High distraction due to side activities can cause drivers to react slower to take-over requests and thus provoke safety-critical situations. On a socio-economic level, when using a highly automated driving system, travel time could be used for work or otherwise being productive and thus create a societal profit. For both evaluation areas, it is important to know what kind of activities drivers engage in and for how long they execute the activities.

*The distractive potential of side tasks*

The ability of drivers to respond to a take-over request (TOR) highly depends on the driver state before the TOR. The driver might, for instance, be fatigued or distracted and thus not immediately be ready to take over.

The German consortium research project Ko-HAF investigated drivers' ability to take-over control from automated driving when being engaged in different non-driving related tasks (NDRTs). Befelein et al. (2017) showed that the type of NDRT has an impact on take-over times and the subjective criticality of take-over situations. For highly motivating tasks such as playing Tetris® take-over times were prolonged.

In a Wizard-of-Oz driving study simulating a SAE level 3 vehicle, drivers experienced take-over situations when being engaged in natural NDRTs (Naujoks et al. 2019). The tasks were chosen such that different workload areas were addressed: Drivers were listening to an audio book (auditory workload), executed a search task where they had to turn around and reach for a bag at the central console (motoric workload), read a magazine (motoric, visual and cognitive workload) and played Tetris® on a tablet (motoric, visual and cognitive workload). Take-over times were longest in the search task and the reading task and the take-overs were subjectively evaluated as being more critical by the drivers. The authors conclude that tasks that involve a motoric component and tasks that require the driver to turn away from the driving scene require longer take-over times.

In a meta-analysis of 129 studies with SAE level 2 or higher, Zhang et al. (2019) found side-tasks which involve hand-held devices as well as visual-motor tasks to increase reaction times to a TOR by 1.33 seconds and 0.29 seconds. When drivers had their eyes closed before the TOR, reaction times were increased by 1.19 seconds.

Monotonous NDRTs can impact drivers' take-over performance such that drivers get fatigued by the tasks and react slower to a TOR due to their fatigue (Jarosch et al. 2019). On the other hand, an activating task can have a positive impact in that respect compared to executing no side-task (Vogelpohl et al. 2018).

It can thus be concluded that NDRTs can have a negative impact on take-over performance especially when drivers engage in motoric side tasks. On the other hand, the engagement in side tasks can keep the driver activated and prevent them from becoming fatigued.

*The use of travel time in automated driving*

The use of travel time is also of interest in terms of productivity. While drivers are not occupied by executing the driving task, they have time for other activities like e.g. in public transportation. In a survey on rail commuters, reading for leisure, window gazing and people watching, text messages and phone calls, working, studying, listening to music and checking emails were among the most popular activities during the rail travel (Lyons et al. 2013). This might be transferable to the automated driving context, because – like in public transportation – the driver is rather a passenger.

In an internet-survey, 5000 respondents from 109 countries were asked what secondary activity they would be willing to engage in while using a highly automated driving system. Most frequently chosen options were listening to the radio, interacting with other passengers, observing, eating, phoning and mailing (Kyriakidis et al. 2015). It should be noted that many of these activities are executed in manual driving as well. It was also found that, not surprisingly, the higher the automation level, the more drivers would be willing to engage in side activities.

Another survey yielded similar results. 1500 respondents from the USA, Japan and Germany stated private communication, route information, eating and drinking, online information search, passive entertainment , shopping, organization, work and wellness (in that order) as the main activities they would execute if their vehicle would operate in level 3 automated mode (Dungs et al., 2016). In a follow-up survey respondents stated “sleeping and relaxing” as the most desired activity followed by “working and being productive”, “eating and drinking”, “entertainment” and “beauty, wellness and fitness” (Becker et al., 2018).

A variety of side activities can be expected from drivers during automated driving. The aim of this study was to investigate what activities drivers engage in during an automated drive and what proportion of their travel time they use for side activities.

**Method**

N = 31 participants (mean age = 37, sd = 11.75) completed 6 drives in a high-fidelity driving simulator (see Figure 1). The simulator runs with the simulation software Silab® (WIVW GmbH, Veitshöchheim, Germany). The participants always drove on a simulated highway and had an L3 motorway chauffeur (L3MC) available. In all drives, drivers were free to use the L3MC as they liked, meaning they could activate and deactivate it and engage in NDRTs as they wished. They were instructed that they could use the function as they like but that they need to be able to take back control if requested by the system. For the description of the system and the responsibility of the driver, the wording of the German Road Transport Law on the driver’s responsibility when using an L3 automated driving system (BMJV, 2017) was used in the instruction.

Throughout all drives, the experimenter continuously coded via a tablet application if the driver was engaged in secondary tasks. The coding on the tablet was saved synchronized with the rest of the data in one data log. Furthermore, subjective evaluation of the motorway chauffeur as well as drivers opinion on potential NDRTs

was assessed with a questionnaire developed within L3Pilot (see Metz, Rösener, Louw, Aittoniemi, Bjorvatn, Wörle et al. in prep.).

*Figure 1: High-fidelity motion-base driving simulator from the outside (left) and from the inside (right)*

*Tested function*

The L3MC was implemented according to the "average" function tested in the L3Pilot project in the on-road driving tests. The system had a speed range of 0 – 130 km/h. It adopted the driven speed to the surrounding traffic as well as to speed limits along the road. The upper limit of the supported speed range was 130 km/h. This means that on sections with no speed limit, the system kept a speed of 130 km/h. The system was able to execute lane changes automatically and as a consequence was able to overtake slower vehicles. System limits were exits from and entrances to motorways, construction sites, sections with bad or missing lane markings and heavy rain. If a system limit was reached, the system issued a TOR with a take-over time of 15 seconds (for a reference see Griffon, Sauvaget, Geronimi, Bolovinou, & Brouwer, 2019).

*Experimental procedure*

Drivers were invited to participate in a study on long-term effects of an L3MC on user behaviour. The study consisted of six driving sessions. For an overview see Table 1.

Before every session, drivers were asked to bring with them any items they would plan to use during an automated drive (e.g. smartphone, newspaper). At the beginning of the 1st session, they were informed about the study and gave their informed consent. Then, they completed an extensive pre-drive questionnaire (L3Pilot pre-questionnaire). After that, every driver completed an introductory drive where they learned the system handling and where they experienced the behaviour of the vehicle at a TOR. Then, drivers completed their first 35-minute drive with the system. After the drive they filled in an extensive post-drive questionnaire (L3Pilot post-drive questionnaire).

The following sessions all started with a short version of the pre-questionnaire. Then the drivers completed their test drives. During the six sessions, driving situations and environment differed with regard to traffic density (e.g. with and without traffic jam),

frequency and reasons for TORs and length and reason of sections outside ODD (e.g. construction site, highway intersection, heavy rain). After the drives, a short version of the post-drive questionnaire was filled in. Only in the 6th session after the test drive, all drivers completed the full version of the post-drive questionnaire. Then they were compensated for their participation.

*Table 1: Overview of study procedure*

| *Session* | *Procedure* |
|---|---|
| 1 | Full pre-drive questionnaire<br>Introductory drive<br>35 minutes' drive on motorway<br>Full post-drive questionnaire |
| 2 | 35 minutes' drive on motorway<br>Short post-drive questionnaire |
| 3 | 1,5 hours' drive on motorway<br>Short post-drive questionnaire |
| 4 | 35 minutes' drive on motorway<br>Short post-drive questionnaire |
| 5 | 1,5 hours' drive on motorway<br>Short post-drive questionnaire |
| 6 | 35 minutes' drive on motorway<br>Full post-drive questionnaire |

In all drives, the participants were instructed to use the system as they would use it in their real life. They were free to activate or deactivate the system and to attend to self-chosen NDRTs. The 3rd and the 5th drive differed from the other 4 drives because they were longer and more monotonous. During one of the two drives, the drivers were sleep deprived, meaning that the drive started at 6 am and drivers had been instructed to sleep a maximum of 4 hours the night before the drive. The order of those two drives was balanced across drivers. To avoid that effects of driver state are mingled with effects of repeated usage, the session without sleep deprivation is always presented as 3rd session and the session with sleep deprivation as 5th session.

*Analysed parameters*

During all sessions a variety of parameters were logged, including questionnaire data, data from the driving simulator, eye tracking data and information coded by the experimenter. It was coded whether participants were engaged in NDRTs, whether the NDRT actively involved the driver's hands (manual distraction, e.g. through browsing on a smartphone, holding food) and whether drivers closed their eyes for a longer time. From this coded data, the proportion of time with active L3MC spent on NDRTs, spent on NDRTs with active involvement of the hands and spent with closed eyes were analysed. Furthermore, it was coded which types of NDRTs were actually executed during the drives.

Before the first session, drivers rated how frequently they engage with various NDRTs in manual driving. After the sixth session, they rated how frequently they would engage in various NDRTs if they would be driving with L3MC. After each session, they filled in a short questionnaire assessing their evaluation of the L3MC. For statistical testing, ANOVAS with a within-subject design were calculated.

**Results**

Already during the first drive with L3MC drivers spent about 70% of time with the system active on various NDRTs. There was a large variability between N=2 drivers who did not engage in any NDRT at all and N=8, who spent more than 90% of the driving time on NDRTs. In the following sessions, all drivers used at least 10% of driving time for NDRTs or closing the eyes; on average about 80% of time was spent on NDRTs or closed eyes. There was a significant effect of session on the proportion of time spent on NDRTs ($F(5, 145)=5.3386$, $p=.00016$) which was caused by a drop in session 5 – drive with sleep deprivation - from about 80% of time to 60%. The drop went hand in hand with an increase of driving time with closed eyes from 0% in drives that were not monotonous to 27% on average during the drive with sleep deprivation.

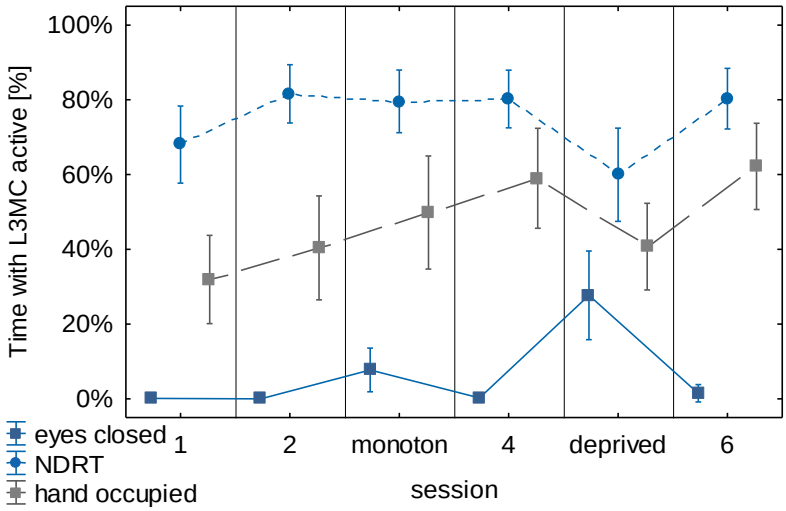



*Figure 2: Proportion of time driving with activated L3 ADF that was spent on NDRTs. The graph shows means and 95%-interval of confidentiality.*

A more detailed analysis showed a change in the type of NDRT with repeated usage: there was a significant rise of time spent on tasks that actively involved the hands ($F(5, 145)=4.4653$, $p=.00082$) from 30% of driving in the first session to 60% of time in the sixth session. The increase of time spent on NDRTs involving the hands was reflected in the answers given to the questionnaire item „ I would use the time the system was active to do other activities.” Already after the first session, there was a strong agreement with the statement and agreement significantly rose further in the following sessions (see figure 3, $F(5, 125)=5.0505$, $p=.00030$).

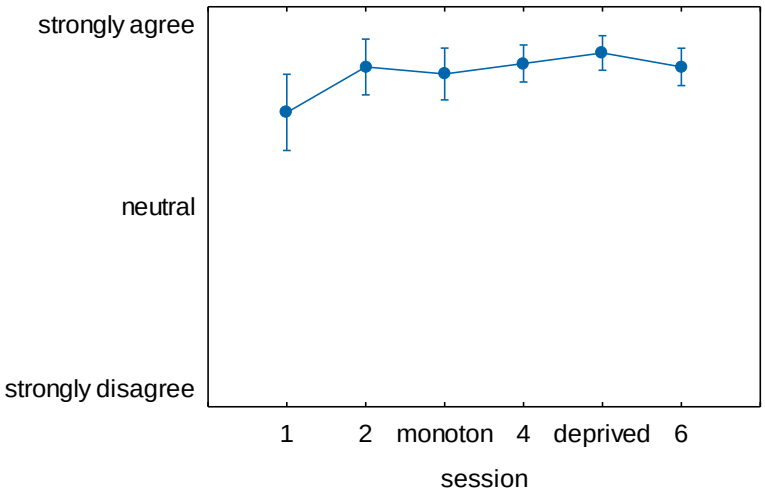



*Figure 3: Subjective agreement with the statement „ I would use the time the system was active to do other activities." The graph shows means and 95%-interval of confidentiality.*

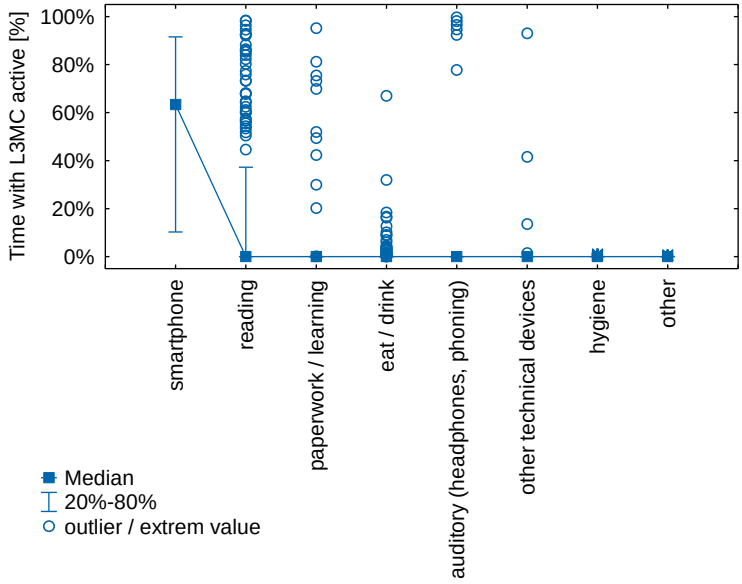



*Figure 4: Proportion of time spent on different NDRTs during the time driving with L3ADF active. The graph shows median, 20% until 80% interval and outliers.*

Analysis of the types of NDRTs actually done showed that drivers mostly attended to their smartphones with on average 60% of driving time over all drivers and sessions. The next frequent type of NDRT was reading (this included papers, magazines, books and e-readers). This type of NDRT was done less often but if it occurred, drivers sometimes spent more or less the whole drive reading. The same was the case for

doing paper works and listening e.g. to music over headphones. N=5 out of 31 drivers attended to paper work during at least one drive. The rest of the sample never used their driving time in the experimental sessions for work related tasks.

Figure 5 shows that in the questionnaire the order of various NDTRs based on their rated frequency remained in large parts the same between manual driving and assumed driving the L3MC. Drivers would attend most frequently to auditory tasks like listening to music or audiobooks followed by interaction with a passenger. The biggest difference between manual driving and potential driving with L3MC occurred for all NDRTs related to a smartphone (calling, texting, apps, internet, social media). Drivers expected that they would attend to those NDRTs way more frequently if they had the system available. The frequency of doing no NDRTs was expected to be lower with L3MC. NDRTs related to work were expected to be done on average every now and then. N=1 driver stated that he / she would do work tasks very frequently, 30% stated that they would work frequently and another 30% at least every now and then.

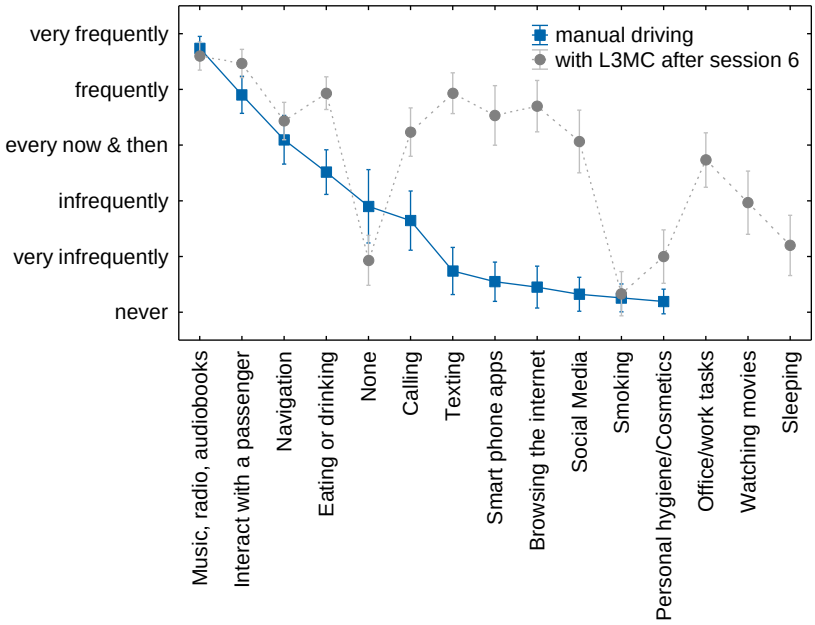



*Figure 5. Subjective evaluation of the frequency with which various NDRTs are done during manual driving and would be done while driving the L3MC.*

**Discussion**

Subjective ratings as well as actually measured driver behaviour show an increase of willingness to engage in NDRTs with repeated usage of the L3MC. It has to be noted however, that both subjective as well as objective measures started from an already high level in the first session and raised up to 80% of driving time spent on NDRTs during the following sessions. The main variation in how drivers spent their time with L3MC active can be explained by the manipulation of drivers' state. When being

fatigued, drivers use less time to engage actively in NDTRs, instead they choose to close their eyes and use the time in the vehicle to relax and rest or even to sleep.

It needs to be noted that in the instruction given to the drivers it was emphasized that they need to be ready to take control back if required by the L3MC in case of a TOR and that drivers experienced various TORs during all sessions. Nevertheless, they decided to use the driving time for resting when being tired. 50% of the sample stated that they would never sleep when driving with the system, but the other half of the sample can imagine to sleep at least sometimes, 10% would even sleep very frequently when driving with the system. This result supports the worry that drivers might misuse L3 systems to doze or sleep although this is clearly outside the allowed usage of L3-systems.

The two most frequent NDRTs in manual driving and also during hypothetical driving with L3MC could not systematically be studied within the presented experiment: neither was a radio or music system available in the simulator nor was a passenger present during the sessions. Nevertheless, since these two tasks are probably the two most common side tasks in manual driving, there is no reason to doubt that drivers would attend to them while driving with an L3MC. Compared to manual driving, all NDRTs related to a smartphone are rated as being much more frequent when driving with an L3 system. The ranking of potential NDRTs from the questionnaires is in line with what is known from the literature. For instance Kyriakidis et al. (2015) report that listening to music, interaction with passenger and eating and drinking were listed as the most likely side tasks in highly automated driving.

This result from the questionnaires is supported by objective data: smartphone usage was the most frequent NDRT in the study. Also quite frequently drivers used the time in the vehicle to read (a task not included in the questionnaire). Sixteen percent of the sample used the driving time with the function active for work related tasks. Compared to the results of the questionnaires, it seems that drivers used the driving time in the experiment less frequently for work than they imagine they would do in real life. In the questionnaire, in total 63% of the sample stated that they would work while driving with the system at least every now and then if not more frequently. This figure fits the 65% of the sample, who stated in the pre-study questionnaire that they could do part of their work while travelling. This result is of special interest for researchers who evaluate the potential benefit of L3 systems for society. One potential benefit of L3 systems is that driving time can be used for new tasks and is no longer occupied with driving. The monetary value used in cost-benefits analyses for the spared driving time differs between time used for work and time used for leisure. Based on our results it can be assumed that drivers would use L3 systems to work in the car but that most of the time would be spent on leisure activities, like reading, listening to music or using the smartphone.

**Acknowledgement**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723051. The sole responsibility of this publication lies with the authors. The authors would like to thank all partners within L3Pilot for their cooperation and valuable contribution.

**References**

Becker, T., Herrmann, F., Duwe, D., Stegmüller, S., Röckle, F., & Niko, U. (2018). *Enabling the Value of Time.* Retrieved from https://www.iao.fraunhofer.de/langen/index.php?option=com_content&view=article&id=1389&Itemid=1&lang=de

Befelein, D., Naujoks, F., & Neukum, A. (2017). Driver takeovers at system boundaries of conditionally automated driving as a function of naturalistic non-driving-related tasks - a preliminary study. Presented at the *59th Conference of Experimental Psychologists*, Dresden, Germany.

BMJV. (2017). Deutsches Straßenverkehrsgesetz § 1b (Vol. Achte Änderung pp. 1648-1650). Bonn: Bundesanzeiger Verlag.

Dungs, J., Herrmann, F., Duwe, D., Schmidt, A., Stegmüller, S., Gaydoul, R., Peters, P.L., Soh, M. (2016). *The value of time. Potential for user-centered services offered by autonomous driving.* Retrieved from https://www.iao.fraunhofer.de/lang-en/images/iao-news/studie-value_of_time_EN.pdf

Griffon, T., Sauvaget, J.-L., Geronimi, S., Bolovinou, A. & Brouwer, R. (2019). *Deliverable D4.1 Description and taxonomy of automated driving functions.* Deliverabel D4.1 of the L3Pilot project.

Jarosch, O., Bellem, H., & Bengler, K. (2019). Effects of Task-Induced Fatigue in Prolonged Conditional Automated Driving. *Human factors, 61*, 1186-1199, doi: 10.1177/0018720818816226

Kyriakidis, M., Happee, R., & de Winter, J. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour, 32*, 127-140.

Lyons, G., Jain, J., Susilo, Y., & Atkins, S. (2013). Comparing rail passengers' travel time use in Great Britain between 2004 and 2010. *Mobilities, 8*, 560-579.

Metz, B., Rösener, C., Louw, T., Aittoniemi, E., Bjorvatn, A., Wörle, J. et al. (in prep.). *Deliverable D3.3 - Evaluation methods.* Deliverable D3.3 of the L3Pilot project.

Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical State Transitions During Conditionally Automated Driving on German Freeways: Effects of Non–Driving Related Tasks on Takeover Time and Takeover Quality. *Human factors, 61*, 596-613.

SAE. (2018). T*axonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (J3016). Society of Automobile Engineers.

Vogelpohl, T., Kühn, M., Hummel, T., & Vollrath, M. (2018). Asleep at the automated wheel—Sleepiness and fatigue during highly automated driving. *Accident Analysis & Prevention, 126*, 70-84.

Zhang, B., de Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation research part F: traffic psychology and behaviour, 64*, 285-307.

# The Renaissance of Wizard of Oz (WoOz) – Using the WoOz methodology to prototype automated vehicles

*Klaus Bengler[1], Kamil Omozik[2], & Andrea Isabell Müller[1]*
*[1]Technical University of Munich, Chair of Ergonomics, [2]BMW Group Germany*

**Abstract**

The strong increase in momentum behind the development of automated systems is leading to a change in paradigm with regard to the distribution of control in human-machine interaction. Therefore, in the context of automated driving, it is necessary to explore fundamental questions such as the interaction between driver and vehicle. However, the underlying automated driving functions are still under development and thus can only be used for studies to a limited extend. From a technical point of view, the introduction of automated systems results in an increased proportion of probabilistic components. Due to the resulting non-determined behaviour of the automation, it is difficult to perform studies in a systematic manner. A suitable method to study the effects of such "intelligent" probabilistic systems are Wizard of Oz (WoOz) setups, where a human simulates the behaviour of the system. The results obtained through WoOz studies are promising, but considering the system behaviour reproduced by the driving wizard researchers apply the method in different ways. Furthermore, there seems to be a lack of systematics regarding the experimental procedure, ethics and the guarantee of scientific quality. This article evaluates and systematizes published experimental approaches and proposes a specification language for the driving wizard's behaviour.

## Introduction

The introduction of automated vehicles is leading to fundamental changes in the relation between vehicles, users and other traffic participants. To analyse this change in relation real automated vehicles can only be used to a limited extend, since the underlying driving functions are still under development. At the same time, developers of the technical system need input on human abilities and restrictions in interaction, which cannot simply be transferred from other domains like aviation or process control. Gasser et al. (2015) give a detailed overview of relevant questions related to level 3 automated driving. Additionally, the rise in automated driving functions within the vehicle system leads to an increased proportion of probabilistic components. However, from the perspective of human factors research, a more deterministic behaviour of the technical system, i.e. the automation, is necessary to evaluate human-machine interaction, since investigations could suffer from random effects in scene interpretation, environmental influences or surrounding traffic behaviour.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

A comparable situation was given in the area of human-computer interaction when, for example, intelligent tutoring systems and speech or gesture recognition were mature to be introduced but had to be evaluated in a systematic way. Here, the Wizard of Oz (WoOz) paradigm was applied with great success and enabled research on human-machine interaction in parallel to technical development. Within the automotive research community, WoOz vehicles are also an established method for analysing the effects of "intelligent" probabilistic systems that have not been fully developed yet, such as automated vehicles.

**Exemplary application of WoOz studies**

WoOz studies are used when complex systems have to be evaluated prior to becoming available. The systems are simulated by humans, the so-called wizards (Fraser & Gilbert, 1991), in a hidden manner. Ideally, this causes users to believe that they are interacting with the real technical systems rather than a simulated one (Bernsen et al., 1994). John F. Kelley invented the WoOz paradigm in 1975 to simulate a not yet functional speech recognition system (Green & Wei-Haas, 1985). Further studies have followed using the WoOz paradigm to simulate natural language recognition systems, such as Kelley (1983) simulating a software assistant to support users when interacting with a digital calendar programme or Gould et al. (1983) simulating a "listening typewriter". From the early 1990s on the WoOz paradigm was also used to prototype multi-modal recognition systems. Hauptmann (1989) simulated a graphics programme that could be used to edit images through speech and gestural input, while Robbe et al. (1997) simulated a spatial planning programme that could likewise be controlled through speech and gestural input.

In the automotive sector, the WoOz methodology is commonly used to design user interfaces (Pettersson & Ju, 2017), such as a multi-modal recognition system to control non-driving related vehicle functions (Stecher et al., 2018). However, the WoOz methodology can also be used to simulate automated vehicles. In this case, so-called driving wizards (Baltodano et al., 2015), simulate the automation by driving the vehicle hidden from participants (Coelingh et al. 2018). When simulating natural language or multi-modal recognition systems, the wizards do not sit in the same room as the participants and the system to be simulated (Hauptmann, 1989; Stecher et al., 2018). However, when simulating automated vehicles, driving wizards act as part of the test tool and are located within the test tool (Müller et al., 2019) allowing them to experience their actions in the same way as the participants.

In 2006 Kiss et al. (2006) developed a WoOz vehicle for the first time to simulate driver assistance systems in real traffic conditions. In the same year, Schomerus et al. (2006) developed the theatre-system technique, which is set in a driving simulator and represents a special case: the deception used in WoOz studies can deliberately be lifted so that researchers can directly get in touch with participants (Schomerus et al. 2006). Fuelled by the development of the Ghost Driver methodology (Rothenbücher et al. 2015) and the RRADS vehicle setup (Baltodano et al., 2015), the WoOz paradigm is currently becoming more used to simulate automated vehicles in real traffic.

*Common construction forms of WoOz vehicles*

Studies involving vehicle occupants as participants require complex vehicle setups to create the illusion of an automated vehicle. All these setups have in common that usually a participant, a driving wizard and an interaction wizard occupy the vehicle. The interaction wizard typically also acts as the investigator. The classification by Manstetten et al. (2019) does not cover all published WoOz vehicle setups. Therefore, a more systematized approach is proposed in the following.

WoOz vehicle setups used for occupant studies can be divided into setups where the participant is seated in the front row or in the back row. Vehicle setups, where the participant is seated in the back (see Figure 1), are typically used for simulating level 5 automation (Karjanto et al., 2018; Sandhaus & Hornecker, 2018; Sherry et al., 2018). The driving wizard operates the vehicle by using the serial driver workplace. An opaque partition obscures the vehicle controls and the driving wizard. Visibility to the front of the vehicle for participants can be realised by mounting a TV displaying a video of the environment (Karjanto et al., 2018) or by not covering the area between the headrests and the vehicle roof (Sherry et al., 2018).

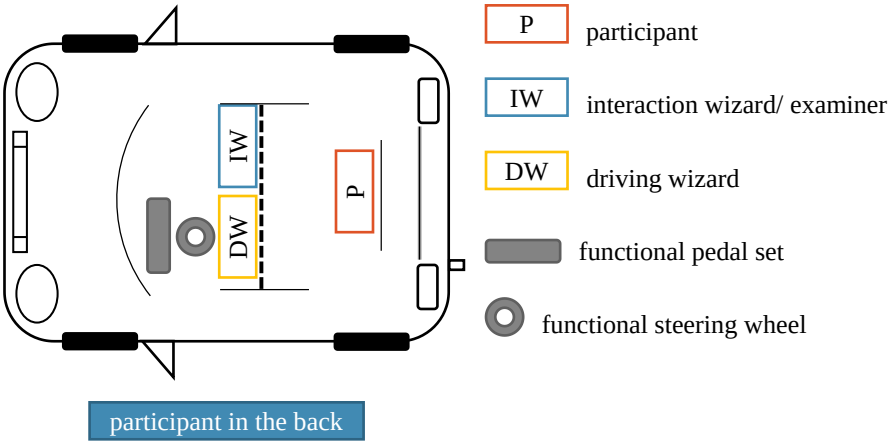



*Figure 1: Common WoOz vehicle setup where participants are seated in the back row*

In case of participants sitting in the front row four different WoOz vehicle setups could be identified (see Figure 2). These can be divided into setups, where participants can drive the vehicle (Figure 2 bottom row) and ones where they cannot (Figure 2 top row).

Setups where participants cannot drive the vehicle should be used for studying level 4 or level 5 automation since Requests to Intervene (RtI) typically cannot be represented. One of these setups is based on a left-hand drive vehicle (see Figure 2 top left). To ensure the disbelief, the driving wizard and vehicle controls are hidden using a partition between driving wizard and participant.

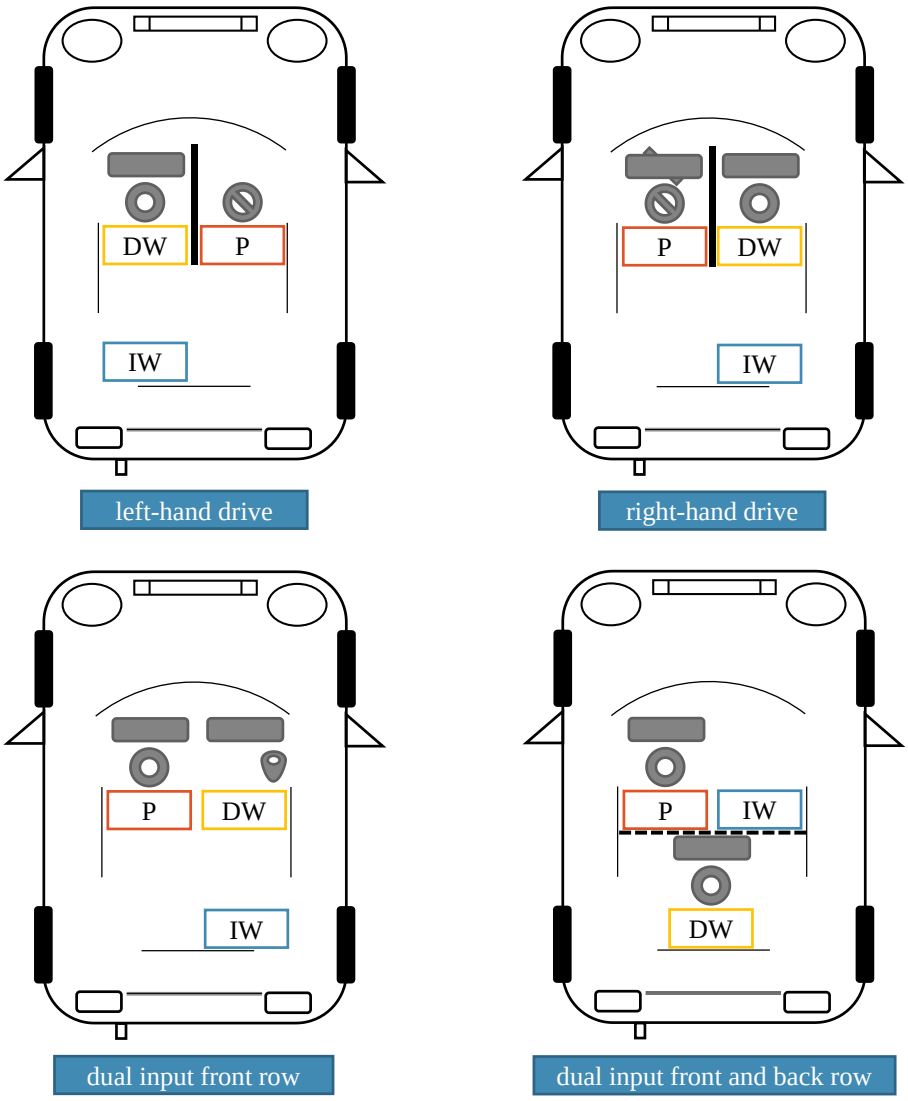



*Figure 2. Common WoOz vehicle setups where participants are seated in the front row*

Furthermore, the participant's seat is equipped with a non-functional steering wheel. Baltodano et al. (2015) developed this setup called RRADS (Real Road Autonomous Driving Simulator). Another setup is based on a right-hand drive vehicle (see Figure 2 top right). The seating position of participants on the (in most countries) usual driver's side acts as a strong cue, that participants are not only passengers. To intensify this feeling, the driving wizard and vehicle controls are concealed using either a partition (Wang et al., 2017), a curtain (Weinbeer et al., 2017) or a hat with covers on the right side (Rittger et al., 2017). To simulate a level 3 automation, Wang et al. (2017) invented the Marionette system, where the driving wizard reproduces the exact input that participants perform using dummy control elements. Weinbeer et al. (2017) attached three displays to the dashboard that represented the highway lanes to simulate an RtI to which participants had to react using dummy control elements.

In vehicle setups where participants are capable of driving themselves, they are always provided with the serial vehicle control elements, whereas the driving wizard uses a retrofitted driving environment. The driving wizard can sit either in the front row (see Figure 2 bottom left) or in the backrow (see Figure 2 bottom right). These setups can be used to simulate automation levels 2 to 4. Level 5 can be simulated with certain limitations since the serial driving workplace provides a strong cue of needing to control the vehicle at some point.

A dual front row input can be realised by providing the driving wizard with another set of pedals and a hidden steering device integrated into the right door (Naujoks et al., 2019). In this case, there is no visibility barrier to ensure that the vehicle is always either controlled by the driving wizard or the participant during simulated RtIs. When using a retrofitted steering wheel as a steering option for driving wizards, a visibility barrier is installed to improve the illusion of an automated vehicle. However, to ensure safe transfers of control during RtIs, the driving wizard must be provided with a display of the current state of vehicle control (Sportillo et al., 2019). For WoOz vehicles where the driving wizard is seated in the back, a semi-transparent glass, that allows the driving wizard to view through the windscreen, separates the driving wizard and the participant (Jarosch et al., 2019). As a special feature of this vehicle setup, participants can sit completely by themselves in the front of the vehicle (Osz et al., 2018).

*Simulating automated driving behaviour*

To simulate automated driving behaviour, the driving wizard must be able to consistently reproduce an automated driving style. For this reason, it is advisable to define the automated driving style and instruct driving wizards accordingly. The most obvious instruction for driving wizards is to let them drive similar to their idea of how automated vehicles will behave (Wang et al., 2017). Moreover, it is possible to instruct driving wizards in a metaphoric way, e.g. by telling them to drive "similar to a professional limo driver" or to achieve a smooth and conservative driving style (Baltodano et al., 2015). Additionally, a qualitative description of the intended driving style can be used. Possible parameters to be defined include the accelerating and decelerating behaviour (Baltodano et al., 2015), the stopping behaviour (Ekman et al., 2019), the distance to surrounding traffic (Ekman et al., 2019), the choice of lane (Naujoks et al., 2019), the lane change behaviour (Weinbeer et al., 2018), the choice

of gear (Ekman et al., 2019) as well as the position within a lane (Ekman et al., 2019). The most detailed way of instructing driving wizards is to specify driving strategies of automated vehicles by quantitative parameter sets. These can refer to the maximum velocity (Jarosch et al., 2019; Naujoks et al., 2019; Omozik et al., 2019; Weinbeer et al., 2018), a maximum lateral acceleration (Karjanto et al., 2018) or permitted ranges for longitudinal acceleration and deceleration (Ekman et al., 2019). To realize the predefined driving behaviour, Adaptive Cruise Control (ACC) and Lane Keeping Assistant can be used (Rittger et al., 2017). However, one must be aware that this holds the risk of unintentionally simulating a state-of-the-art system (Weinbeer et al., 2018).

**Methodology and good practices**

Through the presence of a human wizard, a variety of cognitively demanding tasks, that have not been implemented yet, can be simulated (Bernsen et al., 1994) to realise novel systems fast and without technical development (Kiss et al., 2006). The WoOz methodology allows for a timely user feedback as well as observations in a natural environment and is cost-efficient (Stevens et al., 2019). Compared with existing automated systems, the WoOz paradigm allows for less constrained experiments by using improvisation through the wizard, but also more systematically constrained experiments by omitting the limitations of an automated system (Osz et al., 2018). Since the later technical realisation of the system is unclear (Stevens et al., 2019), one methodological risk is to simulate the technical system in an idealistic way or to insert human deficits into the simulation of machine-like behaviour. Furthermore, the wizard is in a feedback loop with the surrounding traffic system. Compared to other WoOz realizations this is a novelty. In general, it seems challenging to ensure the scientific quality of results achieved using the WoOz paradigm. In this context, Müller et al. (2019) identified the following main methodological challenges related to WoOz:

1. Participants must be under the impression that they are interacting with a real automated vehicle.
2. The simulated automated vehicle must behave as if it were a real automated vehicle.
3. One driving wizard must be able to reproduce the pre-defined driving style at different times.
4. Different driving wizards must be able to reproduce the pre-defined driving style.

As a result, when using the WoOz methodology, not only hypotheses considering the research questions have to be tested, but also considering the comparability of test drives and the believability of the illusion. Therefore, to ensure reliable and comparable results, WoOz requires investigators to record, analyse and report additional data compared to other research paradigms (Dillmann et al., 2019). These include the driving dynamics produced by the driving wizard, the speed and location of surrounding traffic participants, as well as the interface output displayed to participants. Interviews and video recordings are useful to evaluate how participants experienced the illusion created through WoOz (Maulsby et al. 1993). Moreover, not all kinds of research questions can or should be answered by employing the WoOz methodology. It should not be used for research questions where an input by the driver

triggers a system reaction and when an introduction of a specific system in the market has to be decided. In addition, take-over situations always have to be manageable and therefore cannot be examined in situations with high urgency (Feldhütter et al., 2017).

**Need for research**

The authors propose an "inverted" Turing Test methodology to be able to validate different driving wizards related to the research question under investigation and in relation to the automation system under development. Furthermore, a taxonomy is needed to describe the wizard's driving and decision behaviour in a qualitative and quantitative way. As it seems challenging to instruct and for the driving wizard to monitor quantitative values while driving, a qualitative description seems reasonable to instruct wizards in a metaphoric way on driving style and strategic behaviour on the manoeuvring level. A quantitative description of the wizard's driving behaviour is necessary to enable a quantitative comparison between different data sets of one or more wizards and with the automated system under investigation. It is informative to compare different data sets using average values. However, a more differentiated view on values gathered before, during and after certain manoeuvres seems to be necessary. For this comparison, several metrics seem suitable, such as the minimum time to collision (TTC_min), the frequency of lane changes, the minimum gap size of a lane change as well as metrics quantifying the cooperation with other road users. Besides objective data regarding the manoeuvres, it also seems necessary to describe the environment at the time of the manoeuvre. This could be traffic density, number of road lanes and time of day as well as weather and road conditions. Currently, there is no criteria available to decide systematically between data sets produced by different driving wizards in similar contexts or by the same driving wizard in differing contexts. This problem is well known from field operational tests and naturalistic driving studies. Systematic comparison and selection criteria should be checked for a potential transfer. Additionally, it seems necessary to compare the simulated driving behaviour created by driving wizards with that of a real automated vehicle.

**Conclusion**

The WoOz paradigm was invented in 1975 to prototype natural language recognition systems. Nowadays it is becoming increasingly more popular to simulate automated vehicles on real roads. Typical WoOz vehicle setups were identified, including a setup where participants are seated in the back, setups based on left-hand drive as well as right-hand drive vehicles and two setups where both the participant and the driving wizard can drive the vehicle. The identified strategies to instruct driving wizards can be divided into metaphoric, qualitative and quantitative instructions. Lastly, strengths and weaknesses of the WoOz paradigm were discussed, possible fields of application were evaluated and a further need for research to improve the scientific quality of WoOz studies was determined.

**References**

Baltodano, S., Sibi, S., Martelaro, N., Gowda, N., & Ju, W. (2015). The RRADS Platform: A Real Road Automated Driving Simulator. In *Proceedings of the*

*7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 281-288). New York: ACM.

Bernsen, N.O., Dybkjær, H., & Dybkjær, L. (1994). Wizard of Oz Prototyping: When and How? In *CCI Working Papers in Cognitive Science and HCI*.

Coelingh, E., Nilsson, J., & Buffum, J. (2018). Driving tests for self-driving cars. *IEEE Spectrum*, *55*, 40-45.

Dillmann, J., den Hartigh, R., Kurpiers, C., Raisch, F., de Waard, D., & Cox, R. (2019). Intentional Dynamics in Conditionally Automated Driving. In *International Conference On Perception And Action*.

Ekman, F., Johansson, M., Bligård, L.-O., Karlsson, M., & Strömberg, H. (2019). Exploring automated vehicle driving styles as a source of trust information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *65*, 268-279.

Feldhütter, A., Hecht, T., & Bengler, K. (2017). *Fahrerspezifische Aspekte beim hochautomatisierten Fahren* (Report FE 82.0628/2015). Munich, Germany: Technical University of Munich, Chair of Ergonomics.

Fraser, N.M., & Gilbert, G.N. (1991). Simulating speech systems. *Computer Speech and Language*, *5*, 81-99.

Gasser, T.M., Schmidt, E.A., Bengler, K., Chiellino, U., Diederichs, F., Eckstein, L., Flemisch, F., Fraedrich, E., Fuchs, E., Gustke, M., Hoyer, R., Hüttinger, M., Jipp, M., Köster, F., Kühn, M., Lenz, B., Lotz-Keens, C., Maurer, M., Meurer, M., Meuresch, S., Müller, N., Reitter, C., Reschka, A., Riegelhuth, G., Ritter, J., Siedersberger, K.-H., Stankowitz, W., Trimpop, R., & Zeeb, E. (2015). Report on the Need for Research: Round Table on Automated Driving - Research Working Group.

Gould, J.D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM*, *26*(4), 295-308.

Green, P., & Wei-Haas, L. (1985). *The Wizard of Oz: A Tool for Rapid development of User Interfaces* (Report UMTRI-85-27). Ann Arbor, USA: University of Michigan, Transportation Research Institute.

Hauptmann, A.G. (1989). Speech and Gesture for Graphic Image Manipulation. In *Proceedings of the ACM CHI'89 Human Factors in Computing Systems Conference* (pp. 241-245). New York, ACM.

Jarosch, O., Paradies, S., Feiner, D., & Bengler, K. (2019). Effects of non-driving related tasks in prolonged conditional automated driving – A Wizard of Oz on-road approach in real traffic environment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 65, 292-305.

Karjanto, J., Yusof, N. M., Wang, C., Terken, J., Delbressine, F., & Rauterberg, M. (2018). The effect of peripheral visual feedforward system in enhancing situation awareness and mitigating motion sickness in fully automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 58, 678-692.

Kelley, J.F. (1983). An empirical methodology for Writing User-Friendly Natural Language computer applications. In Janda, A (Ed.), *CHI '83 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 193-196). New York, ACM.

Kiss, M., Schmidt, G., & Babbel, E. (2006). *Das Wizard of Oz Fahrzeug: Rapid Prototyping und Usability Testing von zukünftigen Fahrerassistenzsystemen.* VW Konzernforschung.

Manstetten, D., Marberger, C., & Beruscha, F. (2019). Wizard-of-Oz Experiments in Real Traffic - Can They Restart Human Factors? In *9. Darmstädter Kolloquium „mensch + fahrzeug"* (pp. 21-31).

Maulsby, D., Greenberg, S., & Mander, R. (1993). Prototyping an Intelligent Agent Through Wizard of Oz. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 277-284). New York: ACM.

Müller, A.I., Weinbeer, V., & Bengler, K. (2019). Using the Wizard of Oz Paradigm to Prototype Automated Vehicles: Methodological Challenges. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (pp. 181-186). New York: ACM.

Naujoks, F., Purucker, C., Wiedemann, K., & Marberger, C. (2019). Noncritical State Transitions During Conditionally Automated Driving on German Freeways: Effects of Non–Driving Related Tasks on Takeover Time and Takeover Quality. *Human Factors*, 61, 596-613.

Omozik, K., Yang, Y., Kuntermann, I., Hergeth, S., & Bengler, K. (2019). How long does it take to relax? Observation of driver behaviour during real-world conditionally automated driving. In *Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 245-251).

Osz, K., Rydström, A., Fors, V., Pink, S., & Broström, R. (2018). Building Collaborative Test Practices: Design Ethnography and WOz in Autonomous Driving Research. *Interaction Design and Architecture(s) Journal*, 12-20.

Pettersson, I., & Ju, W. (2017). Design Techniques for Exploring Automotive Interaction in the Drive towards Automation. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (pp. 147-260). New York: ACM.

Rittger, L., Wiedemann, K., Schmidt, G., & Green, C.A. (2017). HMI for anticipation of upcoming curvature in automated lateral control. In M. Lienkamp (Ed.), *8. Tagung Fahrerassistenz, Einführung hochautomatisiertes Fahren*.

Robbe, S., Carbonell, N., & Dauchy, P. (1997). Constrained vs spontaneous speech and gestures for interacting with computers: A comparative empirical study. In S. Howard, J. Hammond, & G. Lindgaard (Eds.), *Human-Computer Interaction. IFIP — The International Federation for Information Processing* (pp. 445–452). Boston: Springer.

Rothenbücher, D., Mok, B., Li, J., Ju, W., & Sirkin, D. (2015). Ghost Driver: A Platform for Investigating Interactions Between Pedestrians and Driverless Vehicles. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 44-49). New York: ACM.

Sandhaus, H., & Hornecker, E. (2018). A WOZ Study of Feedforward Information on an Ambient Display in Automated Cars. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings* (pp. 90-92). New York: ACM.

Schomerus, J., Flemisch, F., Kelsch, J., Schieben, A., & Schmuntzsch, U. (2006). Erwartungsbasierte Gestaltung mit der Theatersystem-/ Wizard-Of-Oz-Technik am Beispiel eines haptischen Assistenzsystems. In *AAET 2006 Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel* (pp. 209-225).

Sherry, J., Beckwith, R., Esme, A.A., & Tanriover, C. (2018). Getting things done in an automated vehicle. In *Social Robots in the Wild Workshop at the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction*.

Sportillo, D., Paljic, A., & Ojeda, L. (2019). On-Road Evaluation of Automated Driving Training. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 182-190). Piscataway: IEEE.

Stecher, M., Michel, B., & Zimmermann, A. (2018). The Benefit of Touchless Gesture Control: An Empirical Evaluation of Commercial Vehicle-Related Use Cases. In N.A. Stanton (Ed.), *Advances in Human Aspects of Transportation. AHFE 2017. Advances in Intelligent Systems and Computing* (pp. 383–394). Cham: Springer.

Stevens, G., Meurer, J., Pakusch, C., & Bossauer, P. (2019). Investigating Car Futures from Different Angles: An Overview of Methods Used to Study Human Factors of Autonomous Driving. In *Mensch und Computer 2019 – Workshopband* (pp. 400–409). Bonn: Gesellschaft für Informatik e.V.

Wang, P., Sibi, S., Mok, B., & Ju, W. (2017). Marionette: Enabling On-Road Wizard-of-Oz Automated Driving Studies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 234-243). New York: ACM.

Weinbeer, V., Baur, C., Radlmayr, J., Bill, J.-S., Muhr, T., & Bengler, K. (2017). Highly automated driving: How to get the driver drowsy and how does drowsiness influence various take-over aspects? In M. Lienkamp (Ed.), *8. Tagung Fahrerassistenz, Einführung hochautomatisiertes Fahren*.

Weinbeer, V., Muhr, T., & Bengler, K. (2018). Automated Driving: The Potential of Non-driving Related Tasks to Manage Driver Drowsiness. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics* (pp. 179-188). Cham: Springer.

# Does driving experience matter? Influence of trajectory behaviour on drivers' trust, acceptance and perceived safety in automated driving

*Patrick Rossner & Angelika C. Bullinger*
*Chemnitz University of Technology*
*Germany*

**Abstract**

Currently, trajectory behaviour as one part of the driving style of an automated car is mostly implemented as a lane-centric position. However, drivers show quite different preferences, especially in combination with oncoming traffic. A driving simulator study was conducted to investigate seemingly natural reactive driving trajectories on rural roads. 53 subjects, 30 experienced and 23 inexpereinced drivers, tested a static and a reactive trajectory behaviour. There were twelve oncoming traffic scenarios with vehicle variations in type (trucks or cars), quantity (one or two in a row) and position (with or without lateral offset to the road centre) in balanced order. Results show that reactive trajectory behaviour leads to significantly higher acceptance, trust and subjectively experienced driving performance among experienced drivers. Smaller lane width (2.75 m) and oncoming trucks result in lower perceived safety. Lateral offset to the road centre and the number of oncoming vehicles lead to lower safety ratings. Interestingly, for the group of inexperienced drivers, no significant differences between the experimental conditions could be found. Driving experience can hence be stated as being linked to driving style preferences in automated driving. Results help to design an accepted, preferred and trustfully trajectory behaviour for automated vehicles.

**State of literature and knowledge**

Sensory and algorithmic developments enable an increasing implementation of automation in the automotive sector. Ergonomic studies on highly automated driving constitute essential aspects for a later acceptance and use of highly automated vehicles (Banks & Stanton, 2015; Elbanhawi et al., 2015). In addition to studies on driving task transfer or out-of-the-loop issues, there is not yet sufficient knowledge on how people want to be driven in a highly automated vehicle (Gasser, 2013; Radlmayr & Bengler, 2015; Siebert et al., 2013). First insights show that preferences regarding the perception and rating of driving styles are widely spread. Many subjects prefer their own or a very similar driving style and reject other driving styles that include e.g. very high acceleration and deceleration rates or small longitudinal and lateral distances to other road users (Festner, 2016; Griesche & Nicolay, 2016). Studies show that swift, anticipatory, safe and seemingly natural driving styles are prioritized (Bellem et al., 2016; Hartwich et al., 2015). In existing literature, trajectory behaviour as one part of the driving style is mostly implemented as a lane-centric position of the

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

vehicle in the lane. From a technical point of view this is a justifiable and logical conclusion, but drivers show quite different preferences, especially in curves and in case of oncoming traffic (Bellem et al., 2017; Lex et al., 2017). In manual driving, subjects cut left and right curves and react on oncoming traffic by moving to the right edge of the lane. When meeting heavy traffic, subjects' reactions are even greater (Dijksterhuis et al., 2012; Mecheri et al., 2017; Schlag & Voigt, 2015). The implementation of this behaviour into an automated driving style includes high potential to improve the driving experience in an automated car. Previous studies (Rossner & Bullinger, 2018; Rossner & Bullinger, 2019) with experienced drivers showed tendencies to higher perceived safety, significantly higher driving comfort and driving joy as well as preferences for a seemingly natural reactive trajectory behaviour based on manual driving. Type of the oncoming traffic as well as lane width had an influence on perceived safety. A small lane width and oncoming trucks resulted in lower perceived safety. There was an effect of quantity and position of oncoming traffic, too. Vehicles with a lateral offset to the road centre led to lower safety ratings as well as more approaching vehicles. However, the question of driving experience's influence has not yet been explored. To gain insights in the importance of driving experience, an experiment has been set up parallel to previous research, but with inexperienced drivers who have not yet developed an individual driving style. Results are compared between the different user groups and an outlook on further studies is provided.

## Method and variables

The aim of the study was to investigate seemingly natural reactive driving trajectories on rural roads in an oncoming traffic scenario to better understand people's preferences regarding driving styles. A fixed-based driving simulator (Fig. 1) with an adjustable automated driving function was used to conduct a within-subject design experiment. 53 subjects, 30 experienced and 23 inexpereinced drivers, tested a static and a reactive trajectory behaviour on the most common lane widths in Germany: 2.75 m and 3.00 m. This resulted in four experimental conditions that were presented in randomized order to minimize potential systematic biases. All subjects of the experienced group were at least 25 years old and had a minimum driving experience of 2.000 km last year and 10.000 km over the last five years. The inexperienced drivers had no to a few hours driving experience (see Table 1 for details). The static trajectory behaviour kept the car in the centre of the lane throughout the whole experiment whereas the reactive trajectory behaviour moved to the right edge of the lane when meeting oncoming traffic.

*Figure 1. Driving simulator with instructor centre (left) and an exemplary subject (right)*

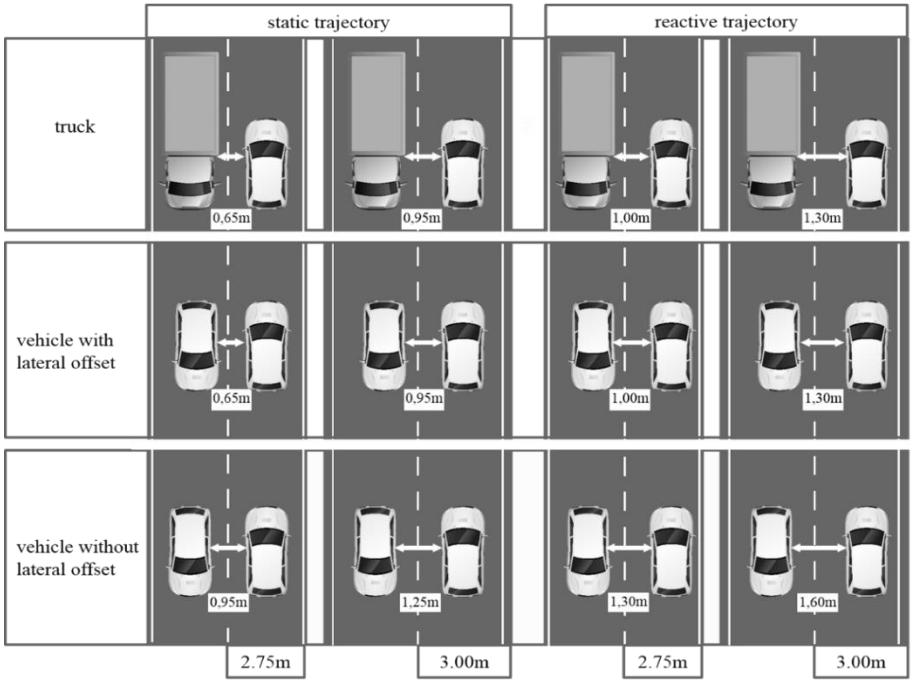



*Figure 2. Variations of oncoming traffic, resultant lateral distances to the ego-vehicle on two different lane widths and in two different trajectory behaviour models*

There were twelve oncoming traffic scenarios that varied in type (trucks and cars), quantity (one or two in a row) and position (cars in the middle of the oncoming lane and cars with lateral offset to the road centre) in balanced order – see Fig. 2. The participants were required to observe the driving as a passenger of an automated car. During the drive subjects' main feedback tool was an online handset control to measure perceived safety as shown in Fig. 3. This tool provides information about the occurrence of safety concerns in each location of the track and could be recorded in sync with video, eye-tracking, physiological or driving data (Hartwich et al., 2015). After each experimental condition subjects filled in questionnaires regarding acceptance (Van der Laan et al., 1997), trust in automation (Jian et al., 2000) and subjectively experienced driving performance (Voß & Schwalm, 2017) and were interviewed at the end of the study.

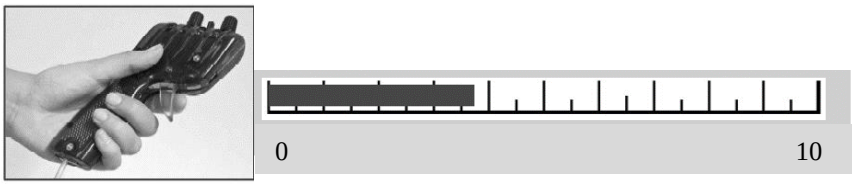



*Figure 3. Handset control (left) and visual feedback (right) to measure perceived safety while driving highly automated. Higher values indicate higher perceived safety.*

*Table 1. Subjects characteristics*

| | Number | Age | | Driver's licence holding [years] | | Mileage last five years [km] | |
|---|---|---|---|---|---|---|---|
| | | M | SD | M | SD | M | SD |
| Experienced drivers | | | | | | | |
| female | 12 | 29.8 | 7.9 | 10.6 | 4.2 | 40,083 | 32,745 |
| male | 18 | 30.9 | 6.8 | 11.9 | 6.1 | 68,333 | 43,661 |
| total | 30 | 30.4 | 7.1 | 11.3 | 5.3 | 54,208 | 41,501 |
| Inexperienced drivers | | | | | | | |
| female | 14 | 16.8 | 0.4 | / | / | / | / |
| male | 9 | 16.9 | 0.3 | / | / | / | / |
| total | 23 | 16.8 | 0.4 | / | / | / | / |

## Results

Ratings of acceptance, trust and subjectively experienced driving performance were compared performing two-factor ANOVAs with repeated measurements including lane width and trajectory behaviour. Fig. 4 shows the mean values of the dependent variables for all four drives, whereas Table 2 describes the overall and interaction effects of the two independent variables.

*Acceptance*

Within-subject tests show no difference for the usefulness scale, but significantly lower satisfaction ratings for the static trajectory behaviour, $F(1, 29) = 8.038$, $p = .008$, $\eta_p^2 = .217$, and for the 2.75 m lane condition, $F(1, 29) = 5.193$, $p = .030$, $\eta_p^2 = .152$, for experienced drivers. As seen in Figure 4, subjects tend to differentiate more between trajectory behaviours on the 2.75 m lane condition. No interaction effect between lane width and trajectory behaviour is found (Table 2). No significant differences between the experimental conditions are found for inexperienced drivers.

*Trust*

Within-subject tests show significantly lower trust ratings for the 2.75 m lane condition, $F(1, 29) = 12.103$, $p = .002$, $\eta_p^2 = .294$, and the static trajectory behaviour, $F(1, 29) = 10.587$, $p = .003$, $\eta_p^2 = .267$, for experienced drivers. As seen in Figure 4, subjects tend to differentiate more between trajectory behaviours on the 2.75 m lane condition. No interaction effect between lane width and trajectory behaviour is found (Table 2). No significant differences between the experimental conditions are found for inexperienced drivers.

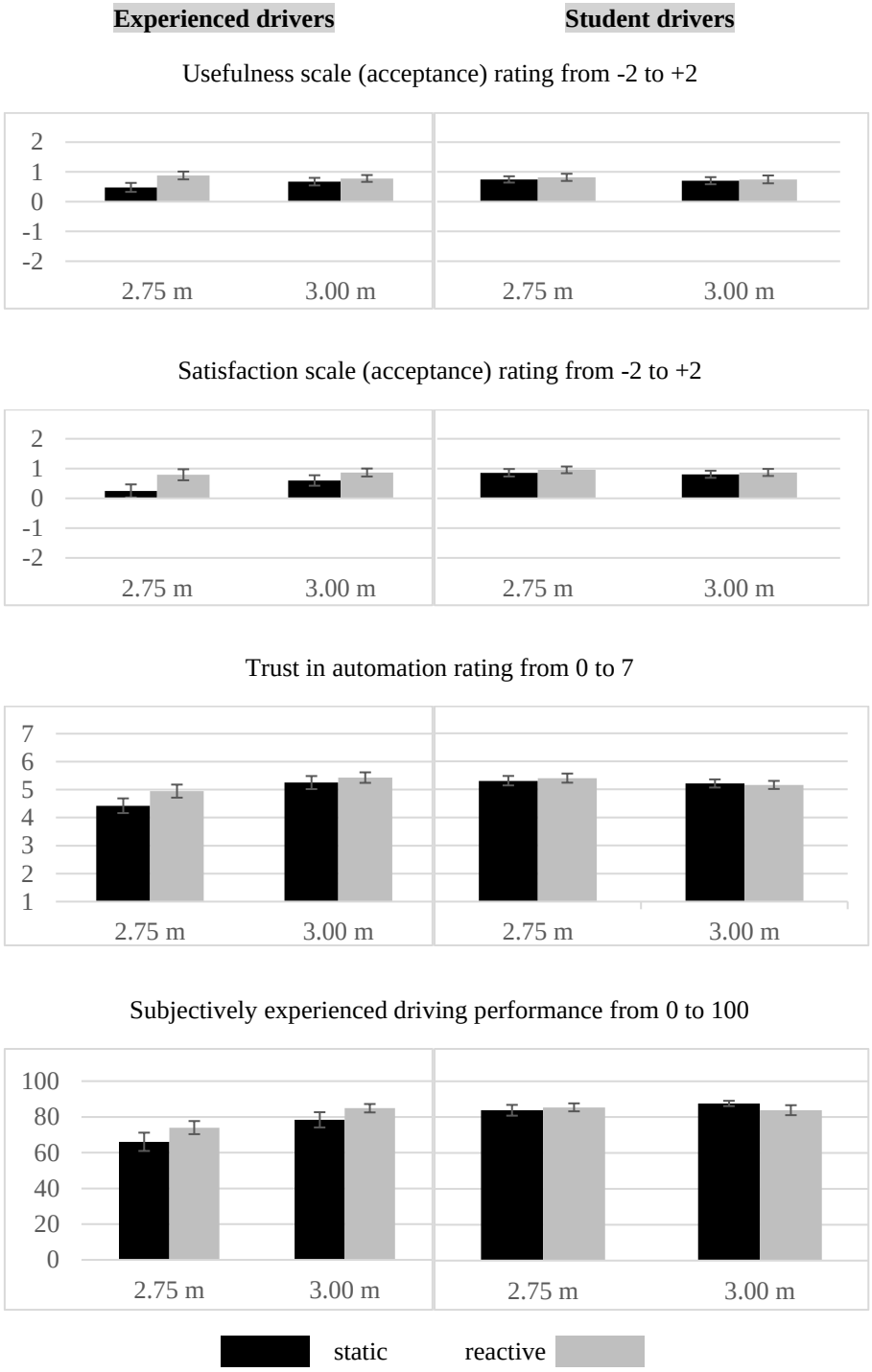



Figure 4. Mean values of acceptance (usefulness and satisfaction), trust and SEDP

*Subjectively Experienced Driving Behaviour (SEDP)*

Within-subject tests show significantly lower SEDP ratings for the 2.75 m lane condition, $F(1, 29) = 12.537$, $p = .001$, $\eta_p^2 = .302$, and the static trajectory behaviour, $F(1, 29) = 7.483$, $p = .011$, $\eta_p^2 = .205$, for experienced drivers. As seen in Figure 4, subjects differentiate between all four experimental conditions. No interaction effect between lane width and trajectory behaviour is found (Table 2). No significant differences between the experimental conditions are found for inexperienced drivers.

*Table 2. Results of two-factor ANOVASs with repeated measurements including lane width and trajectory beahaviour*

| Dep. variables | Independent variables | *F* | *p* | $\eta_p^2$ |
|---|---|---|---|---|
| Experienced driver | | | | |
| usefulness scale (acceptance) | Trajectory behaviour | 3.399 | .075 | .105 |
| | Lane width | 2.757 | .108 | .087 |
| | Trajectory behaviour x lane width | .454 | .506 | .015 |
| satisfaction scale (acceptance) | ***Trajectory behaviour*** | ***8.038*** | ***.008*** | ***.217*** |
| | ***Lane width*** | ***5.193*** | ***.030*** | ***.152*** |
| | Trajectory behaviour x lane width | 2.187 | .150 | .070 |
| trust | ***Trajectory behaviour*** | ***10.419*** | ***.003*** | ***.264*** |
| | ***Lane width*** | ***11.843*** | ***.002*** | ***.290*** |
| | Trajectory behaviour x lane width | 2.205 | .148 | .071 |
| SEDP | ***Trajectory behaviour*** | ***7.700*** | ***.010*** | ***.210*** |
| | ***Lane width*** | ***13.044*** | ***.001*** | ***.310*** |
| | Trajectory x lane width | .113 | .739 | .004 |
| Inexperienced drivers | | | | |
| usefulness scale | Trajectory behaviour | .627 | .437 | .028 |
| | Lane width | 1.980 | .173 | .083 |
| | Trajectory behaviour x lane width | .071 | .792 | .003 |
| satisfaction scale | Trajectory behaviour | .207 | .653 | .009 |
| | Lane width | .357 | .556 | .016 |
| | Trajectory behaviour x lane width | .842 | .369 | .037 |
| trust | Trajectory behaviour | 2.461 | .131 | .101 |
| | Lane width | .080 | .779 | .004 |
| | Trajectory behaviour x lane width | .626 | .437 | .028 |
| SEDP | Trajectory behaviour | .199 | .660 | .009 |
| | Lane width | .778 | .387 | .034 |
| | Trajectory x lane width | 3.458 | .076 | .136 |

*Handset control results*

For a detailed analysis, the handset control data was reversed and cumulated for all subjects to identify clusters that represent low perceived safety. Fig. 5 (experienced drivers) and Fig. 6 (inexperienced drivers) give an overview of the whole test route with its different types of oncoming traffic and show highlights for the absence of high perceived safety – hereafter stated as perceived safety concerns.

The graphs show the static and the reactive trajectory behaviour in comparison on 2.75 m (upper section) and 3.00 m (bottom section) lane width each. The maximum of perceived safety concerns is 300 (10 as maximum per subject x 30 subjects) for the group of experienced drivers and 230 (10 x 23) for the group of inexperienced drivers. For example, a data point of 80 can arise of 10 participants feeling complete unsafe or 20 people experiencing mid perceived safety.

Remarkably, the data for the inexperienced drivers shows no tendencies. When looking at the distribution of the descriptive data for experienced drivers, several tendencies of perceived safety concerns are able to be observed that are conform to the questionnaire results. Wider lanes and reactive trajectory behaviour lead to higher perceived safety. The feedback of the handset control set allows a more detailed and situation-specific analysis. Position, type and quantity of oncoming traffic do also have an influence on perceived safety (assumption based on descriptive data, inference statistical evaluation in progress):

1. More approaching vehicles lead to higher perceived safety concerns.
2. Oncoming traffic with lateral offset to the road centre leads to more perceived safety concerns than lane-centric oncoming traffic.
3. Heavy traffic (e.g. trucks in this experiment) lead to higher perceived safety concerns. Further analysis is going to include correlations between perceived safety concerns and number, type and positon of oncoming traffic as well as cross lane width evaluations.

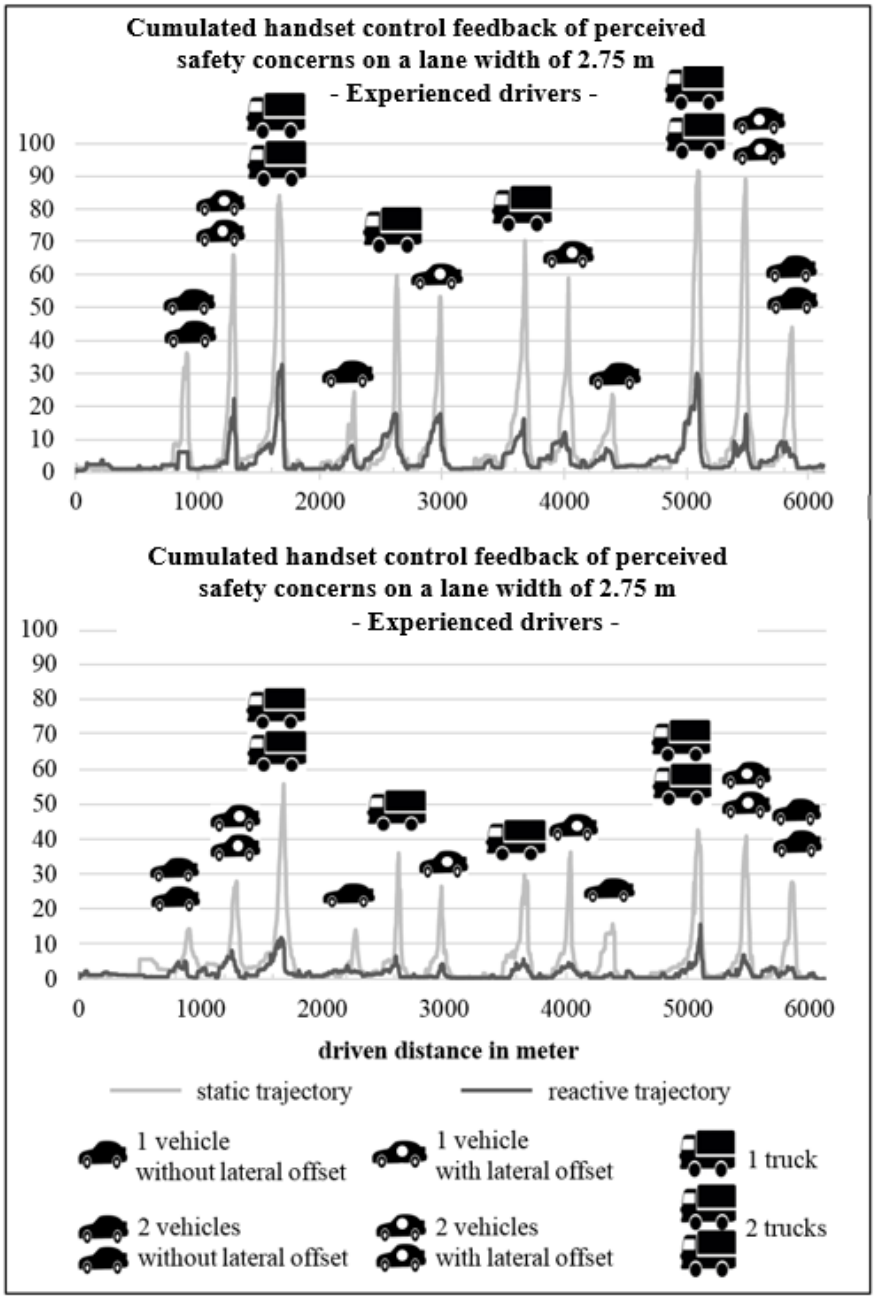



Figure 5. Cumulated handset control feedback of perceived safety concerns for experienced drivers

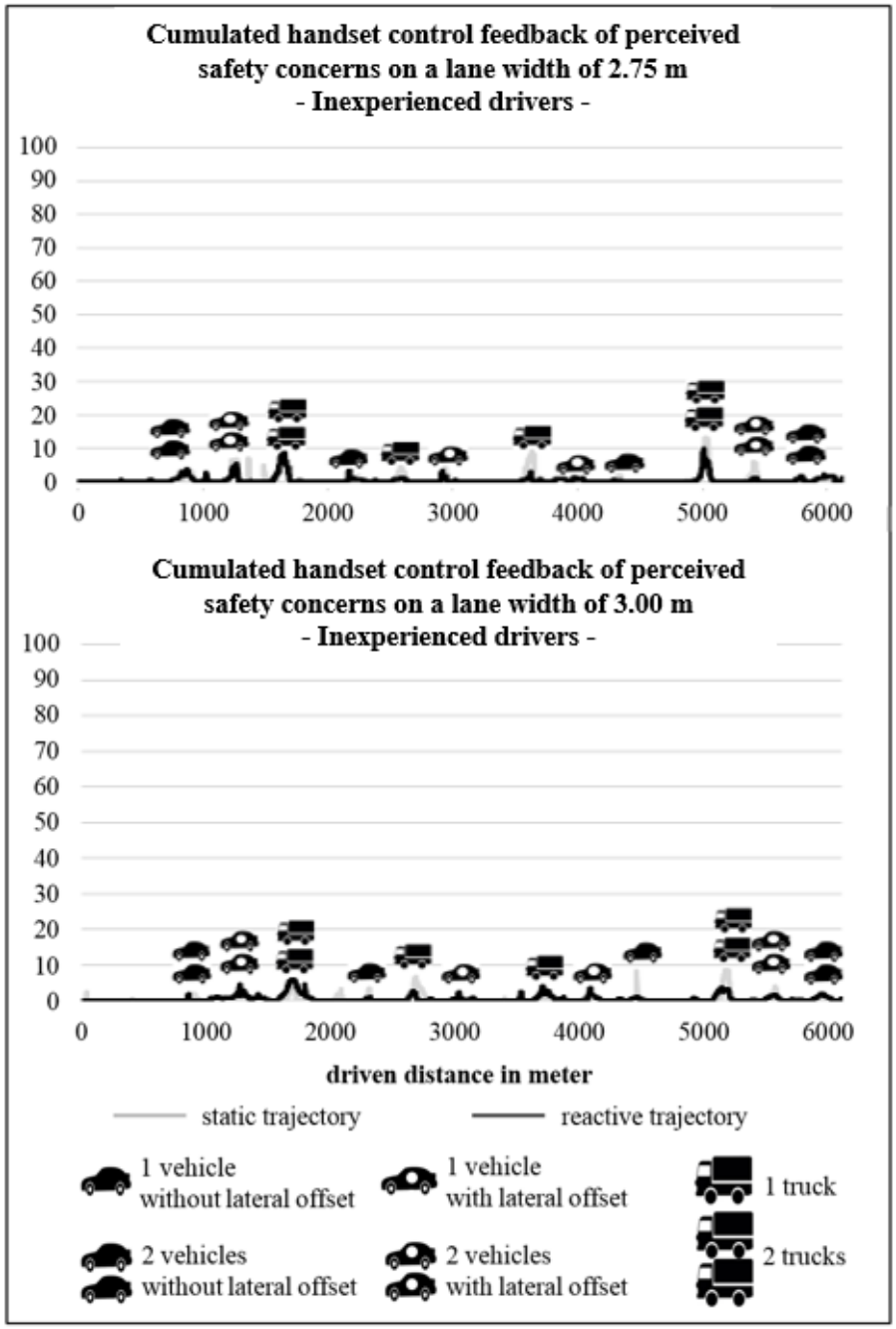



Figure 6. Cumulated handset control feedback of perceived safety concerns for inexperienced drivers

## Conclusion and outlook

The aim of the study was to investigate seemingly natural reactive driving trajectories on rural roads in an oncoming traffic scenario to better understand people's preferences regarding driving styles. The use of manual drivers' trajectories as basis for implementing highly automated driving trajectories showed high potential to increase perceived safety (Bellem et al., 2017; Lex et al., 2017; Rossner & Bullinger, 2018; Rossner & Bullinger, 2019). Data from the experienced drivers revealed significantly higher acceptance (only satisfaction scale), trust and SEDP for the reactive trajectory. We also identified traffic density, lateral position and type of oncoming vehicles as factors that influence perceived safety during automated driving. In order to better understand the impact of these different aspects, further inference statistical and correlation analysis should be conducted. Based on the results so far, it is concluded that factors which influence perceived safety in manual driving (Lex et al., 2017; Dijksterhuis et al., 2012; Mecheri et al., 2017; Schlag & Voigt, 2015) are also factors influencing perceived safety during highly automated driving. As drivers cannot react to oncoming traffic by shifting to the right edge of the lane, the automated vehicle has to do so to increase perceived safety and driving comfort of the passenger. Therefore, it seems most relevant to investigate manual trajectory behaviour in more detail to implement better reactive trajectories that include less negative side effects and lead to a better driving experience. For the inexperienced drivers, no effects for trajectory behaviour and lane width were found. A possible explanation is obviously the absence of driving experience which leads to the absence of a personal driving style. Without this personal driving style, preferences as a baseline against which the automated driving style can be compared, are missing. Additionally, witout driving experience critical driving situations can rather not be distinguished from uncritical driving situations. It is also possible that there exists more trust in automation within the group of inexperienced drivers. Another influencing factor might be the questionnaires that were developed to analyse the behaviour, attitude and perception of experienced and therefore older drivers. In sum, it can be concluded that the results provide an interesting outlook for the future when people may grow up with much more automation and devising driving styles will follow other paradigms than today. For the near future and thus for a set of experienced drivers, it is important to note that a positive driving experience has the potential to improve the acceptance of highly automated vehicles (Siebert et al., 2013; Hartwich et al., 2015) and therefore has both ergonomic and economic benefits.

## Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (research project: KomfoPilot, funding code: 16SV7690K). The sponsor had no role in the study design, the collection, analysis and interpretation of data, the writing of the report, or the submission of the paper for publication. We are very grateful to Konstantin Felbel, Marty Friedrich, Maximilian Hentschel and Maxine Börner for their assistance with data collection and analysis.

**References**

Banks, V.A., & Stanton, N.A. (2015). Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53*, 389-395.

Bellem, H., Schönenberg, T., Krems, J.F., & & Schrauf, M. (2016). Objective metrics of comfort: Developing a driving style for highly automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 41*, 45-54.

Bellem, H., Klüver, M., Schrauf, M., Schöner, H.-P., Hecht, H., & Krems, J.F. (2017). Can We Study Autonomous Driving Comfort in Moving-Base Driving Simulators? A Validation Study. *Human Factors, 59*, 442–456

Dijksterhuis, C., Stuiver, A., Mulder, B., Brookhuis, K.A., & De Waard, D. (2012). An adaptive driver support system: user experiences and driving performance in a simulator. *Human Factors, 54*, 772–785

Elbanhawi, M., Simic, M., & Jazar, R. (2015). In the Passenger Seat: Investigating Ride Comfort Measures in Autonomous Cars. *IEEE Intelligent Transportation Systems Magazine, 7*, 4–17

Festner, M., Baumann, H., & Schramm, D. (2016). Der Einfluss fahrfremder Tätigkeiten und Manöverlängsdynamik auf die Komfort- und Sicherheitswahrnehmung beim hochautomatisierten Fahren. *32nd VDI/VW-Gemeinschaftstagung Fahrerassistenz und automatisiertes Fahren*

Gasser, T.M. (2013). Herausforderung automatischen Fahrens und Forschungsschwerpunkte. *6. Tagung Fahrerassistenz*

Griesche, S., Nicolay, E., Assmann, D., Dotzauer, M., & Käthner, D. (2016). Should my car drive as I do? What kind of driving style do drivers prefer for the design of automated driving functions? *Contribution to 17th Braunschweiger Symposium Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel* (pp. 185-204)

Hartwich, F., Beggiato, M., Dettmann, A., & Krems, J.F. (2015). Drive me comfortable: Customized automated driving styles for younger and older drivers. *8. VDI-Tagung „Der Fahrer im 21. Jahrhundert“*

Jian, J.Y., Bisantz, A.M., & Drury, C.G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*, 53-71

Lex, C., Schabauer, M., Semmer, M., Magosi, Z., Eichberger, A., Koglbauer I., Holzinger, J., & Schlömicher, T. (2017). Objektive Erfassung und subjektive Bewertung menschlicher Trajektoriewahl in einer Naturalistic Driving Study. *VDI-Berichte Nr. 2311* (pp. 177-192)

Mecheri, S., Rosey, F., & Lobjois, R. (2017). The effects of lane width, shoulder width, and road cross-sectional reallocation on drivers' behavioral adaptations. *Accident; Analysis and Prevention, 104*, 65–73

Radlmayr, J., & Bengler, K. (2015). Literaturanalyse und Methodenauswahl zur Gestaltung von Systemen zum hochautomatisierten Fahren. *FAT-Schriftenreihe, 276*

Rossner, P., & Bullinger, A.C. (2018). Drive me naturally: Design and evaluation of trajectories for highly automated driving manoeuvres on rural roads. *Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference*

Rossner P., & Bullinger, A.C. (2019). Do You Shift or Not? Influence of Trajectory Behaviour on Perceived Safety During Automated Driving on Rural Roads. In H. Krömker (Eds.), *HCI in Mobility, Transport, and Automotive Systems* (pp. 245-254)

Schlag, B., & Voigt, J. (2015). Auswirkungen von Querschnittsgestaltung und laengsgerichtet Markierungen auf das Fahrverhalten auf Landstrassen. *Berichte der Bundesanstalt fuer Strassenwesen, Unterreihe Verkehrstechnik, 249*

Siebert, F., Oehl, M., Höger, R., & Pfister, H.R. (2013). Discomfort in Automated Driving – The Disco-Scale. In Proceedings of HCI International 2013, *Communications in Computer and Information Science, vol. 374* (pp. 337-341)

Van der Laan, J.D., Heino, A., De Waard, D. (1997). A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics. In *Transportation Research Part C: Emerging Technologies, 5*, 1–10.

Voß, G., & Schwalm, M. (2017). Bedeutung kompensativer Fahrerstrategien im Kontext automatisierter Fahrfunktionen. *Berichte der Bundesanstalt für Straßenwesen, Fahrzeugtechnik Heft F 118*

# Evaluation of different driving styles during conditionally automated highway driving

*Stephanie Cramer*[*,1], *Tabea Blenk*[*,1,2], *Martin Albert*[1], *& David Sauer*[1]
[1]*AUDI AG,* [2]*Elektronische Fahrwerksysteme GmbH, Germany*
[*]*These authors contributed equally to this work*

**Abstract**

Discomfort and well-being of the driver and/or the passengers during automated driving as well as their acceptance and trust in the automation system are important criteria considering the usage of automated driving vehicles. Thereby, the driving behaviour of the automated vehicle plays an important role. For this contribution, we implemented three driving styles, which differ only regarding the tactical driving behaviour on the manoeuvre level. Trajectory planning and control was identical. One driving style contained only lane following on the right lane without lane changes. The other two driving styles varied according to their lane change decision behaviour. To evaluate the aforementioned criteria of the driving styles, a driving study (N=31) was conducted in real traffic on a highway with a test vehicle in which vehicle guidance was performed by an automation system. The results reveal that the well-being of the drivers is not influenced by the driving style. On the contrary, trust and acceptance are influenced by the driving style. Overall, 97% of the participants would prefer a driving style including lane change manoeuvres. However, 61% had the highest feeling of safety while driving without lane changes.

## Introduction

Besides technical and legal questions, human-computer interaction is considered essential for the development of automated driving functions on all levels of automation which have been defined in the taxonomy for automated driving systems published by the Society of Automative Engineers (SAE, 2016), e.g. in Saffarian et al. (2012). So far, work in this domain mainly focused on concepts for the interaction between driver and automation (e.g. Albert et al., 2015; Flemisch, 2003; Flemisch et al., 2014; Hoc, 2000; Schreiber et al., 2009), control transitions and take-over requests (eg. Feldhütter et al., 2018; Gold, 2016; Gold et al., 2013; Petermann-Stock et al., 2013; Zeeb et al., 2015), or the design of human machine interfaces (e.g. Albert et al., 2015; Franz et al., 2012; Othersen, 2016). Furthermore, the way the vehicle behaves and its so called "driving style" is considered to have an important influence on trust, acceptance, and the experience of automated driving (Bellem et al., 2016; Elbanhawi et al., 2015; Festner et al., 2017; Oliveira et al., 2019). Following Griesche et al. (2016), the driving style is described by a set of parameters on the tactical and operational vehicle guidance layers, defined by Matthaei (2015).

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

However, there is no common knowledge about the precise configuration of the parameters that differentiate various driving styles. Most of the previous studies, comparing different driving styles during automated driving, focused on dynamic metrics such as velocity, longitudinal and lateral acceleration, jerk, and the duration of a lane change (Bellem et al., 2018; Festner et al., 2016; Hartwich et al., 2018; Lange et al. 2014). Regarding the accepted point in time at which the lane change should be initiated, research from a human factors perspective is sparse. Rossner and Bullinger (2019) compared three highly-automated driving styles during highway driving varying different factors. One of those factors, the inition time of the lane change manoeuvres, included the tactical lane change decision. Results show that people prefer a more comfortable driving style which is defined with a following distance to the leading vehicle of 2.9*s*, a maximum acceleration of 1.5$m/s^2$, a maximum deceleration of -2$m/s^2$, a duration of the lane change to the left of 9*s* and to the right of 8.5*s* and the distance to a leading vehicle with overtaking initiation of 130*m*. Nevertheless, by also varying these other factors, no conclusion can be made that the factor considering the initiation time of the lane change manoeuvres had the key influence on the perceived safety and comfort.

All the previous mentioned studies have in common that they were all conducted under simulated settings (Bellem et al., 2018; Rossner & Bullinger, 2019) or on test tracks (Festner et al., 2016; Festner et al., 2017; Hartwich et al., 2018; Lange et al., 2014) leaving aside important influences of real-world scenarios.

The aim of this study was to overcome these limitations and to investigate different driving styles differing on the tactical vehicle guidance in real-world highway driving. Main focus and, thus, an exploratory research question was if the driving style has an influence on the aforementioned metrics perceived comfort, personal well-being, trust, and acceptance. Moreover, it should be examined what the preferred driving style is considering well-being and safety.

## Method

### *Test setup and equipment*

The driving study took place on the three-lane German highway A9 between the highway exits Lenting and Holledau. The test vehicle was an Audi A7, year of construction 2010. A prototypical level 3 (SAE, 2016) automation system was implemented in the test vehicle which completely performed the lateral and longitudinal vehicle guidance. However, the test vehicle only drove on the right and middle lane of the highway due to safety reasons.

The participants were seated on the driver seat and were accompanied by two experimenters. One always sat on the passenger seat and was acting as a safety driver. This task was supported via a monitor containing information about the automation system, a second interior mirror, driving school mirrors as well as driving school pedals to be able to intervene in vehicle guidance in risky driving situations (referring to Cramer et al., 2018). This experimenter was able to adapt the driving function, for instance the target velocity or abort/initiate lane changes, only in exceptional cases if it was necessary. The second experimenter was seated in the back row and was

responsible for the questionnaires, functional variations, and providing the participants with instructions.

The participants received visual information about the activation status, the current manoeuvre, and surrounding obstacles in front of the vehicle in the instrument cluster display. Data recording included vehicle data, internal data of the automation system, audio recordings, front camera, as well as driver observation camera.

*Driving styles*

Three driving styles were implemented in the test vehicle. The functional realization on the operational layer of the automation system (according to Matthaei (2015)) was equal for all driving styles. The trajectory planning was based on the approach of Werling et al. (2010) including adaptions by Heil et al. (2016). The decisions on the tactical layer of the automation system (according to Matthaei (2015)), in this case executing lane changes, were different for the driving styles. Considering the first driving style, the vehicle was not performing any lane changes, and thus was only driving in the right lane of the highway. The other two driving styles performed lane changes. Their execution was implemented considering different aspects according to Ulbrich and Maurer (2015). The aspects of dynamic traffic were implemented based on a fuzzy logic (cf. Du and Swamy (2019) for basic principles about fuzzy logic). For the two driving styles with lane changes, the shape parameters of the membership function for the deceleration of the rear vehicle (cf. Ulbrich & Maurer, 2015) are varied: 0.6 and 0.9$m/s^2$ (dynamic driving style), or 0.38 and 0.63$m/s^2$ (cautious driving style). Moreover, the time gap for the rear vehicle (cf. Ulbrich & Maurer, 2015) differed between the cautious (2.0$s$) and the dynamic (0.5$s$) driving style. These parameters were selected with developers of the automated driving function. However, the two driving styles with lane changes were called *cautious* and *dynamic* to distinguish them, both represented defensive driving behaviour. This can further be seen in Figure 1, 2, and 3, which represent the timely distributions of the lateral and longitudinal accelerations as well as the velocity for each driving style. The amount of performed lane changes depending on the driving style is presented in Table 1. Lane change aborts occurred quite often. One main reason was the limited rear sensor range (approximately 150$m$).

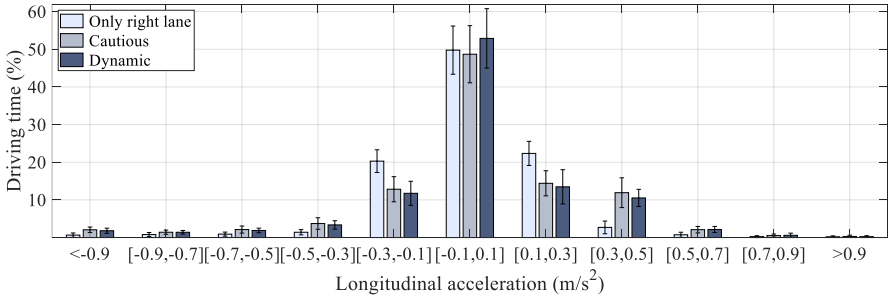



*Figure 1. Distribution (mean and standard deviation) of the longitudinal acceleration over the driving time for the three driving styles.*

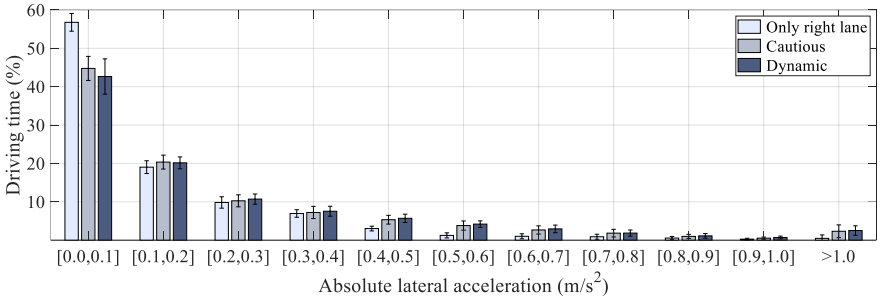



*Figure 2. Distribution (mean and standard deviation) of the absolute lateral acceleration over the driving time for the three driving styles.*

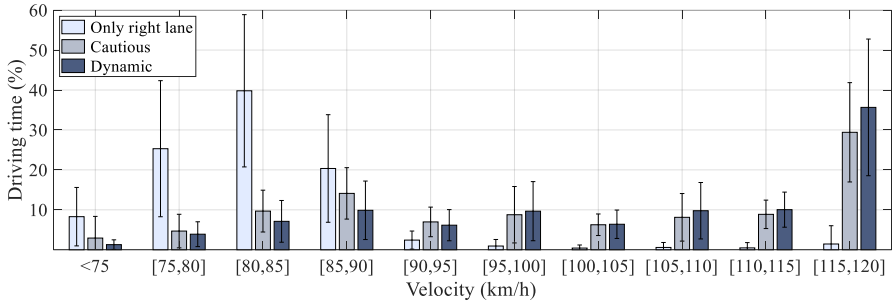



*Figure 3. Velocity distribution (mean and standard deviation) over the driving time for the three driving styles.*

*Table 1. Amount (mean (M) and standard deviation (SD)) of lane changes (LC) and lane change aborts depending on the driving style.*

| | *Cautious* | | | *Dynamic* | | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *LC abort* | *M* | *SD* | *LC abort* |
| Lane change left | 3.33 | 1.65 | 48.98% | 5.13 | 2.11 | 35.83% |
| Lane change abort left | 3.20 | 2.44 | | 2.87 | 1.59 | |
| Lane change right | 2.60 | 1.48 | 39.53% | 4.77 | 1.94 | 30.58% |
| Lane change abort right | 1.70 | 1.37 | | 2.10 | 1.37 | |

*Study design*

The driving study was conducted in German. At the beginning of the study, the participants received a verbal briefing on how to handle the test vehicle and what to expect during the driving study. Following, the participants drove manually on the highway and activated the automation system. The sequence of the driving study is presented in Figure 4.

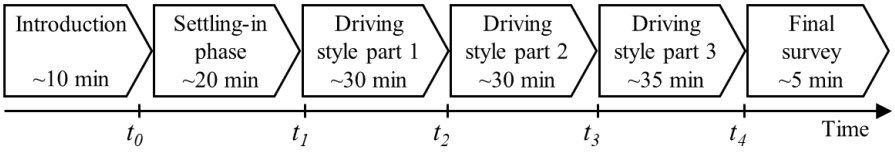



*Figure 4. Sequence of driving study.*

During the settling-in phase, for approximately the first 7 minutes, the automation system conducted no lane changes and started with these afterwards. Subsequently, the participants experienced the three driving styles in a randomized order. However, part 1 and 2 were a bit shorter as part 3 due to the fact that the turnaround at the highway exit was earlier. During the driving parts, the participants' task was to speak all their thoughts out loud (think-aloud method, Ericsson & Simon, 1980) about the driving behaviour of the automation system. The evaluation of the participants' comments is not part of this paper. At the end of every driving part, the participants answered a questionnaire about, for instance, trust and acceptance (cf. section results). Summing up, a short overall questionnaire was conducted.

*Processing and evaluation of the data*

The rating scales of the questionnaires were assumed as interval scaled variables because the answer scales were equidistant (Döring & Bortz, 2016). Furthermore, normal distribution of the data was expected if N>30 (Bortz & Schuster, 2010; Field, 2012). For data evaluation, a repeated measures analysis of variance (ANOVA) with following post-hoc analysis using Bonferroni correction was conducted for the dimensions well-being, comfort, trust, and acceptance. The data was corrected, if Mauchly's test for sphericity showed significance (Greenhouse-Geisser or Hunyh-Feldt correction ($\varepsilon$ >0.75)).

*Sample*

*N*=32 participants were available for this driving study, whereby one had to be excluded from data evaluation due to bad performance of the automation system induced by bad weather. The sample (*N*=31) had a mean age of 36.1 years (*SD*=11.9, *MIN*=22, *MAX* =65) and was a variation of professional background and gender (22.6% technical female, 25.8% technical male, 25.8% non-technical female, and 25.8% non-technical male). The median mileage per year was 15,001-20,000 *km* and the mean mileage per week was 265*km* (*SD*=203*km*) with on average 41% highway driving. 74% of the participants used adaptive cruise control, 77% lane keeping assistance, and 48% partially automated driving systems (e.g. traffic jam assistance) before.

**Results**

*Well-Being*

The well-being of the participants during the study was evaluated by the short version A of the German multidimensional state survey (MDBF, Steyer, et al., 1997). This short form has 12 items on a five-point rating scale from 1 ("not at all") to 5 ("very"), corresponding to the three bipolar dimensions *good-bad mood*, *awake-tired*, and *calm-nervous*. For every subscale, the values of the respective items were summed up leading to a value per subscale between 4 and 20, whereby a high value indicates a good mood, awakeness, and calmness and a low value a bad mood, tiredness, and nervousness. The participants were asked to rate their current well-being five times: in the beginning, after the settling-in phase, and after each driving style. No differences were found between the various times of measurement for either the

dimension *good-bad mood* ($F(2.46)=1.34$, $p=.268$, $f=.21$), *awake-tired* ($F(2.89)=2.45$, $p=.071$, $f=.29$), or *calm-nervous* ($F(2.92)=1.32$, $p=.273$, $f=.21$). All three subscales reached mean values between 15.8 and 18.5 out of a maximum of 20. Thus, the overall well-being of the participants during the experiment can be described as in a good, awake, and calm mood. The values for the mean (*M*) and standard deviation (*SD*) for all times of measurement and subscales can be found in Table 2.

*Table 2. Participants' mean ratings for the three dimensions of the MDBF*

| | *Beginning* | | *Settling-in phase* | | *Only right lane* | | *Cautious* | | *Dynamic* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Good-bad mood | 18.45 | 1.23 | 18.00 | 1.77 | 17.94 | 1.91 | 17.58 | 2.36 | 17.94 | 1.90 |
| Awake-tired | 16.81 | 2.07 | 17.00 | 1.79 | 15.84 | 3.01 | 16.35 | 2.67 | 16.65 | 2.48 |
| Calm-nervous | 16.39 | 2.62 | 16.48 | 2.11 | 17.35 | 2.76 | 16.55 | 2.80 | 17.00 | 2.07 |

*Comfort*

To survey driving comfort, the subscales *discomfort* and *comfort* of the questionnaire to measure driving comfort and enjoyment developed by Engelbrecht (2013) were used. Hereby, the rating scale was adapted to seven anchors from 1 (“does absolutely not apply”) to 7 (“does absolutely apply”). The participants were asked to rate the previous car ride after each driving condition. The sample of the subscale *comfort* was reduced due to a mistake in the questionnaire for the first participants. The ANOVA revealed no differences for the perceived *discomfort* ($F(1.52)=1.61$, $p=.214$, $f=.23$) and *comfort* ($F(1.35)=3.42$, $p=.063$, $f=.14$) between the three different driving styles. Overall, the experienced discomfort during the automated car ride was rated low (mean values around 2) and the comfort high (mean values around 5.50). The values for the mean (*M*) and standard deviation (*SD*) for each driving style and subscale can be found in Table 3.

*Table 3. Participants' mean ratings for their perceived comfort and discomfort for the three driving styles (scale: 1≙"does absolutely not apply" - 7≙"does absolutely apply").*

| | *Only right lane* | | *Cautious* | | *Dynamic* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Comfort (N = 23) | 5.85 | 1.14 | 5.21 | 1.20 | 5.73 | 0.84 |
| Discomfort | 1.78 | 0.99 | 2.07 | 1.90 | 1.77 | 0.79 |

*Trust*

To assess the trust in the automation the questionnaire of Körber (2018) was used which is divided into six subscales with a range from 1 (“strongly disagree”) to 5 (“strongly agree”). To determine the general trust in automation, the subscale *Propensity to Trust* was surveyed once before the study. In order to get the respective trust in the automation system of each driving style, the participants were asked to rate the corresponding items of the subscales *Reliability/Competence*, *Understanding/Predictability*, and *Trust in Automation* after each driving condition.

The evaluation of the *Propensity of Trust* scale showed a mean value of the sample of 3.56 (*SD*=.53). The applied ANOVA indicated significant differences between the driving styles for the three subscales *Reliability/Competence* ($F(1.67)=3.42$, $p=.049$, $f=.34$), *Understanding/Predictability* ($F(1.92)=10.90$, $p<.001$, $f=.60$), and *Trust in Automation* ($F(1.65)=5.43$, $p=.001$, $f=.43$). The following post hoc pairwise comparisons did not reveal any significant difference for the dimension *Reliability/Competence* ($p>.05$). Considering the subscale *Understanding/ Predictability*, results of the post hoc analysis showed that the participants ranked the driving style only using the right lane with higher understanding and predictability in comparison to the dynamic ($M_{1-3}=0.36$, $p=.019$) as well as the cautious driving style ($M_{1-2}=0.61$, $p=.001$). Furthermore, the participants showed less trust in automation during the cautious driving style compared to the driving style only using the right lane ($M_{1-2}=.44$, $p=.044$), and the dynamic driving style ($M_{2-3}=-0.32$, $p=.047$). The results are represented in Figure 5. and Table 4.

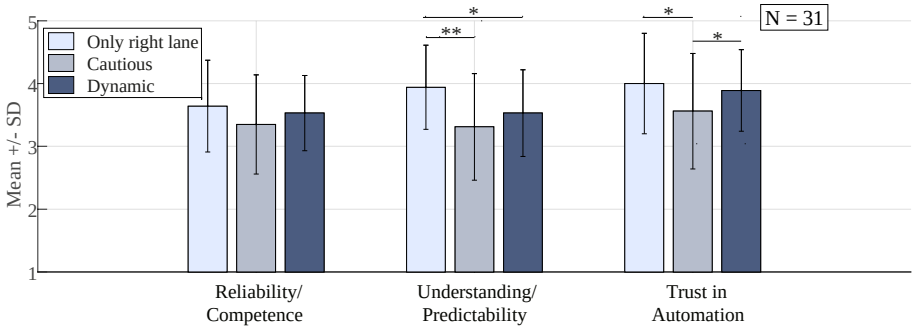



*Figure 5. Participants' mean ratings for three dimensions of the questionnaire of Körber (2018) for the three driving styles (scale: 1≙"strongly disagree" - 5≙"strongly agree"; *p<.05, **p<.01).*

*Table 4. Participants' mean ratings for three dimensions of the questionnaire of Körber (2018) for the three driving styles.*

| | *Only right lane* | | *Cautious* | | *Dynamic* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Reliability/Competence | 3.64 | 0.73 | 3.35 | 0.79 | 3.53 | 0.60 |
| Understanding/Predictability | 3.94 | 0.67 | 3.31 | 0.85 | 3.53 | 0.69 |
| Trust in Automation | 4.00 | 0.80 | 3.56 | 0.92 | 3.89 | 0.65 |

*Acceptance*

The acceptance of the driving style was evaluated by the questionnaire of Van der Laan et al. (1997) in the German version (Kondzior, n.d.). This questionnaire has nine items on a five-point rating scale from -2 to 2 in which the mean value of five items results in the *usefulness* scale (y-axis) and the mean value of the other four items in the *satisfying* scale (x-axis). The ANOVA revealed a significant difference between the driving styles for both the *usefulness* ($F(1.84)=5.03$, $p=.012$, $f=.41$) and *satisfying* scale ($F(1.95)=3.28$, $p=.046$, $f=.33$). Subsequently post hoc analysis showed a

significant higher usefulness for the dynamic driving style compared to the driving style only using the right lane ($M_{1-3}$=-.42, $p$=.009). No other post hoc pairwise comparison showed a significant effect ($p$>.05). The scores with positive mean values point out that all driving styles were seen as useful and satisfying (Figure 6). The values for the mean (*M*) and standard deviation (*SD*) for each driving style and the two subscales can be found in Table 5.

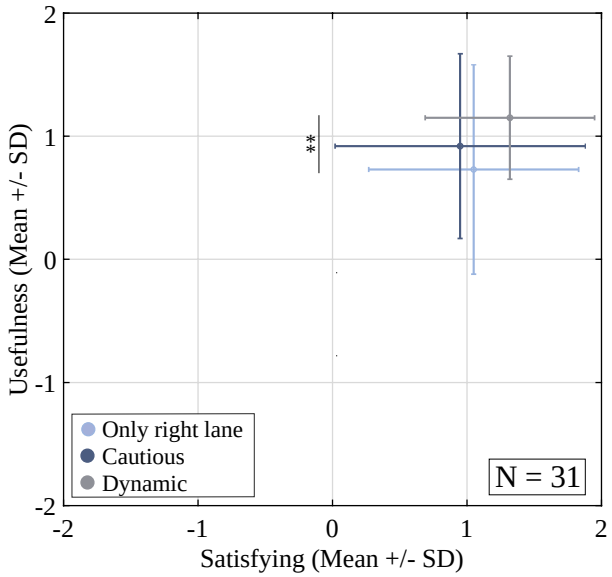



*Figure 6. Evaluation of acceptance of the three driving styles (scale: five-point semantic differential; **p<.01)*

*Table 5. Participants' mean ratings for the two dimensions usefulness and satisfying of the acceptance questionnaire of van der Laan (1997) for the three driving styles (scale: five-point semantic differential)*

| | *Only right lane* | | *Cautious* | | *Dynamic* | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Usefulness | 0.73 | 0.85 | 0.92 | 0.75 | 1.15 | 0.50 |
| Satisfying | 1.05 | 0.78 | 0.95 | 0.93 | 1.32 | 0.63 |

*Prioritisation*

After the participants had experienced all three driving styles, they were asked to choose one of them considering the following statements:

- During which car ride did you feel the best *well-being*?
- During which car ride did you feel the *safest*?
- Which car ride's driving style would you prefer for an automated vehicle driving on the highway?

For the factor well-being, nearly half of the participants (48.39%) preferred the dynamic driving style. Only four participants (12.90%) chose the driving style that was only using the right lane, and 12 (38.71%) the cautious driving style. In contrast, 19 participants (61.29%) indicated that they felt the safest during the driving style only using the right lane and only seven (22.58%) during the cautious driving style, and five (16.13%) during the dynamic driving style. For their overall prioritisation, 96.8% of the participants favoured a driving style including lane change manoeuvres (dynamic: 54.84%, cautious: 41.94%) and only one participant (3.23%) would prefer a driving style only using the right lane (Figure 7).

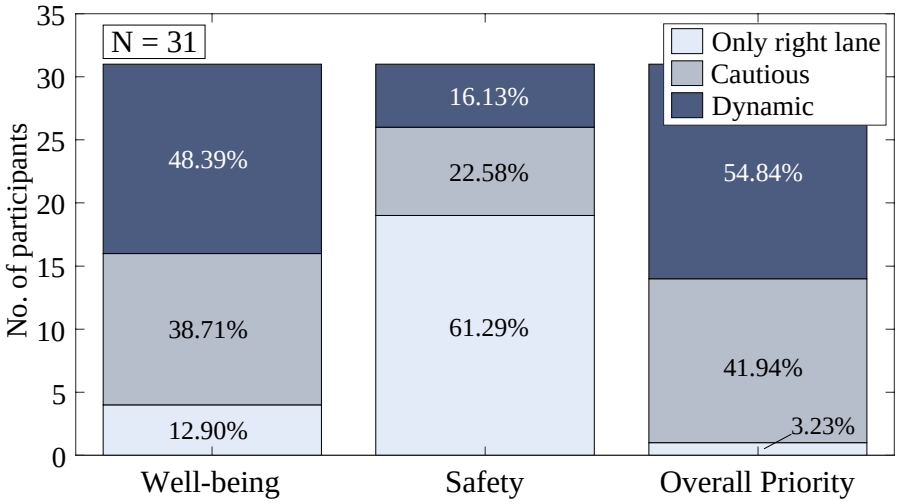



*Figure 7. Distribution of the preferred driving style considering well-being, safety, and an overall priority.*

**Conclusion and Discussion**

Three different driving styles for conditionally automated highway driving with varying lane change behaviour were evaluated. Overall, over 60% of the participants felt the safest during the driving style only using the right lane of the highway as well as rated this driving style as the most predictable and understandable. An explanation for this result is that the absence of lane changes leads to the higher predictability and feeling of safety. Moreover, the lower velocity could also have influenced the feeling of safety (Figure 3). In contrast to this, the driving style only using the right lane was considered as less useful than the dynamic driving style. The overall priority clearly showed that the majority preferred a driving style including lane changes as only one driver voted for the driving style only using the right lane. However, the driving style did not influence the well-being of the participants. This metric was always evaluated after the test drive when the vehicle was parked and, thus, might have influenced the real well-being while driving. Evaluating the latter metric while driving should be considered. During the cautious driving style, the participants reported less trust in automation than during the dynamic driving style. A presumable reason for that could be the higher number of aborted lane changes during the cautious driving style. Overall, even if there are differences between the three driving styles, the participants

always perceived high well-being and comfort as well as high trust and acceptance. Furthermore, results indicate that the dynamic driving style is overall preferred, even though ratings in trust and safety were higher during a driving style only using one lane of the highway.

As it is always important to have a look at real-world scenarios, this also has its limitations when it comes to the standardisation of the conditions. On a real highway among other vehicles, the behaviour of other drivers, the traffic, and the weather is not controllable as it is in simulated settings or on test tracks. Considering this, the study took place at the same times during the day to ensure similar traffic and it was avoided to drive when it was raining, but in real-world settings, some variances are not preventable. The study was voluntary, so most of the participants were interested in automated driving and not too anxious or sceptical about it. Consequently, this could have influenced the ratings.

Much more research is necessary in this field to design a driving style for automated highway driving. One aspect for instance could be the influence of the motivation of the car ride or non-driving related tasks. Both aspects could have an important impact on the perception of different driving styles during automated driving.

**Acknowledgments**

We would like to extend a big "thank you" to our colleagues, in particular Neha Lal, Sebastian Bayerl, Alexander Freier, and Frieder Gottmann for their assistance with the test vehicle software or supporting as a safety driver.

**References**

Albert, M., Lange, A., Schmidt, A., Wimmer, M., & Bengler, K. (2015). Automated Driving - Assessment of Interaction Concepts Under Real Driving Conditions. In *6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, AHFE 2015*

Bellem, H., Schönenberg, T., Krems, J.F. & Schrauf, M. (2016). Objective metrics of comfort: Developing a driving style for highly automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *41*, 45-54

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: Traffic Psychology and Behaviour*, *55*, 90-100.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. ed.). Berlin, Heidelberg: Springer.

Cramer, S., Kaup, I., & Siedersberger, K.-H. (2018). Comprehensibility and Perceptibility of Vehicle Pitch Motions as Feedback for the Driver During Partially Automated Driving. *IEEE Transactions on Intelligent Vehicles*, *4*, 3-13.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. ed.). Berlin, Heidelberg: Springer.

Du, K.-L. & Swamy, M.N.S. (2019). Introduction to Fuzzy Sets and Logic. In *Neural Networks and Statistical Learning*. London: Springer

Elbanhawi, M., Simic, M., & Jazar, R. (2015). In the Passenger Seat: Investigating Ride Comfort Measures in Autonomous Cars. *IEEE Intelligent Transportation Systems Magazine*, 7(3), 4-17.

Engelbrecht, A. (2013). *Fahrkomfort und Fahrspaß bei Einsatz von Fahrerassistenzsystemen*. Hamburg: Disserta-Verlag.

Ericsson, K.A., & Simon, H.A. (1980). Verbal Reports as Data. *Psychological Review, 87(3),* 215–251.

Feldhütter, A., Segler, C., & Bengler, K. (2018). Does Shifting Between Conditionally and Partially Automated Driving Lead to a Loss of Mode Awareness? In Stanton N. (Eds), *Advances in Human Aspects of Transportation. AHFE 2017*. Springer

Festner, M., Baumann, H. & Schramm, D. (2016). Der Einfluss fahrfremder Tätigkeiten und Manöverlängsdynamik auf die Komfort- und Sicherheitswahrnehmung beim hochautomatisierten Fahren. In *32. VDI/VW-Gemeinschaftstagung Fahrerassistenz und automatisiertes Fahren*.

Festner, M., Eicher, A., & Schramm, D. (2017). Beeinflussung der Komfort- und Sicherheitswahrnehmung beim hochautomatisierten Fahren durch fahrfremde Tätigkeiten und Spurwechseldynamik. In *11. Workshop Fahrerassistenzsysteme und automatisiertes Fahren.*

Field, A. (2012). *Discovering Statistics Using SPSS* (3. ed.). Los Angeles, USA: Sage.

Flemisch, F., Adams, C., Conway, S., Goodrich, K., Palmer, M., & Schutte, P. (2003). *The H-metaphor as a guideline for vehicle automation and interaction: NASA/TM, 2003-212672*.

Flemisch, F., Bengler, K., Bubb, H., Winner, H. & Bruder, R. (2014). Towards cooperative guidance and control of highly automated vehicles: H-Mode and Conduct-by-Wire. *Ergonomics, 57,* 343-360.

Franz, B., Kauer, M., Bruder, R., & Geyer, S. (2012). pieDrive - a New Driver-Vehicle Interaction Concept for Maneuver-Based Driving. In 2012 *IEEE Intelligent Vehicles Symposium (IV).*

Gold, C. (2016). *Modeling of Take-Over Performance in Highly Automated Vehicle Guidance*. PhD thesis. Technische Universität München.

Gold, C., Damböck, D., Lorenz, L. M. & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57 (1),* 1938-1942.

Griesche, S., Nicolay, E., Assmann, D., Dotzauer, M. & Käthner, D. (2016). Should my car drive as I do? What kind of driving style do drivers prefer for the design of automated driving functions? In *17. Braunschweiger Symposium – Automatisierungssysteme, Assistenzsysteme und eingebettete Systeme für Transportmittel (AAET).*

Hartwich, F., Beggiato, M., & Krems, J.F. (2018). Driving comfort, enjoyment and acceptance of automated driving – effects of drivers' age and driving style familiarity. *Ergonomics, 61(8),* 1017-1032.

Heil, T., Lange, A. & Cramer, S. (2016). Adaptive and Efficient Lane Change Path Planning for Automated Vehicles. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*.

Hoc, J.M. (2000). From human - machine interaction to human-machine cooperation. *Ergonomics, 43,* 833-843.

Kondzior, M. (n.d.). *Akzeptanzskala - Methode zur Erfassung der Akzeptanz eines Systems*. Retrieved from http://www.hfes-europe.org/accept/accept_de.htm.

Körber, M. (2018). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In: Bagnara S., Tartaglia R., Albolino S., Alexander T., Fujita Y. (Eds*) Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018).* Springer

Lange, A., Maas, M., Albert, M., Siedersberger, K.H., & Bengler, K. (2014). Automatisiertes Fahren-So komfortabel wie möglich, so dynamisch wie nötig. In *30. VDI-VW-Gemeinschaftstagung Fahrerassistenz und integrierte Sicherheit*.

Matthaei, R. (2015). *Wahrnehmungsgestützte Lokalisierung in fahrstreifengenauen Karten für Assistenzsysteme und automatisches Fahren in urbaner Umgebung*. PhD thesis. Technische Universität Braunschweig.

Oliveira, L., Proctor, K., Burns, C.G., & Birrell, S. (2019). Driving Style: How Should an Automated Vehicle Behave? *Information*, *10*(6), 219.

Othersen, I. (2016). *Vom Fahrer zum Denker und Teilzeitlenker*. PhD thesis. Technische Universität Braunschweig.

Petermann-Stock, I., Hackenberg, L., Muhr, T., & Mergl, C. (2013). Wie lange braucht der Fahrer? Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer automatisierten Staufahrt. In *6. Tagung Fahrerassistenz. Der Weg zum automatisierten Fahren*.

Rossner, P. & Bullinger, A.C. (2019). How Do You Want to be Driven? Investigation of Different Highly-Automated Driving Styles on a Highway Scenario. In In: Stanton N. (Eds) *Advances in Human Factors of Transportation. AHFE 2019* (pp. 36-43). Springer

SAE. (2016). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (2016-09 ed.) (No. J3016).

Saffarian, M., De Winter, J.C.F., & Happee, R. (2012). Automated Driving: Human-Factors Issues and Design Solutions. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56(1),* 2296–2300.

Schreiber, M., Kauer, M. & Bruder, R. (2009). Conduct by Wire – Maneuver Catalog for Semi-Autonomous Vehicle Guidance. In *IEEE Intelligent Vehicles Symposium (IV)* (pp. 1279-1284).

Steyer, R., Schwenkmezger, O., Notz, P. & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)*. Göttingen: Hogrefe.

Ulbrich, S. & Maurer, M. (2015). Situation Assessment in Tactical Lane Change Behavior Planning for Automated Vehicles. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*

Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A Simple Procedure for the Assessment of Acceptance of Advanced Transport Telematics. *Transportation Research Part C: Emerging Technologies*, *5*, 1–10.

Werling, M., Ziegler, J., Kammel, S. & Thrun, S. (2010). Optimal Trajectory Generation for Dynamic Street Scenarios in a Frenét Frame. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Zeeb, K., Buchner, A., & Schrauf, M. (2015). What determines the take-over time? An integrated model approach of driver takeover after automated driving. *Accident Analysis & Prevention*, *78*, 212–221.

# An adaptive assistance system for subjective critical driving situations: understanding the relationship between subjective and objective complexity

*Alexander Lotz[1], Nele Russwinkel[2], Thomas Wagner[1], & Enrico Wohlfarth[1]*
*[1]Daimler AG, [2]Technische Universität Berlin*
*Germany*

**Abstract**

Partial and conditional automated driving allows the driver to transfer responsibility to the vehicle. While assistance systems are designed to deal with aspects of the driving task, currently no assistance systems are available to predict driver behaviour for take-over when the vehicle is handling the driving task. This is important as drivers might interpret driving situations differently than an activated automation function. This can cause self-initiated take-overs leading to a reduction of trust in the system. In theory, if a prediction is robust, an assistance system could also adapt based on this prediction. A new subjective complexity model addressing these situations is introduced. The subjective complexity model learns situations in which individual drivers have previously self-initiated control of the driving task. Based on exemplary sideswipe manoeuvres, the system concept is explained and simulated with a training and test dataset. Upon introducing this system, a discussion is initated on the difference between objective and subjective situation complexity. A distinction is drawn between mathematical descriptions based on vehicular sensor data and human interpretation of the environment. The proposed system also functions as a carrier technology for further investigations between the differences of objective and subjective complexities.

**Introduction**

The driving task consists of many short-term decisions, e.g. steering to hold vehicle in lane, and long-term decisions, e.g. route navigation. Different factors need to be considered in order for the driver to successfully solve these tasks. Therefore, it is important to understand which aspects are considered complex. This can help to include, enhance or adjust assistance accordingly. At the same time, humans are decisively influenced by the environment, which they encounter and thereby confined in their range of interactions. With various developments in the field of automated driving, possibilities of directing attention away from the driving task will become possible. In Level 3 (SAE J3016, 2018) the driver can focus on non-driving related tasks, but needs to regain control of the vehicle when warned. Recent research regarding take-over behaviour has shown that environmental factors such as traffic density and time-budget (distance to objects) play a crucial role in successful take-over capability for Level 3 (Gold, et al., 2016; Lotz, et al., 2019; Zhang, et al., 2019). There is a large variety of definable environmental factors, making it difficult to

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

pinpoint isolated factors to varying driver take-over behaviour. When revisiting traffic density as an exemplary factor, other environmental factors such as time to collision, number of lanes and colour of vehicles can form interaction effects. Multiple isolated environmental factors can merge to form singular driving situations through the relation of several of these factors over time. Arbitrary measures, such as low or high traffic density, also make a comparison difficult. However, the driving environment can be measured with a variety of different sensors mounted on a vehicle. Based on the chosen sensor setup, this creates a representation with a sensor-specific degree of detail of environmental factors or higher-level situations. As a driver also perceives the environment with her/his senses and develops a representation of the situation, influences of environmental factors such as traffic density on the driver can be compared and deduced. If a certain driver reaction is linked to an environmental factor, based on the sensor representation of the environment an assistance system could possibly predict driver behaviour. This information would especially be valuable in the abovementioned take-over situations, in which responsibility shifts from the machine to the human. The mathematical description of the environment could be attributed to subjective complexities, identifying scenarios that cause higher workload and situations in which the driver needs assistance. As there are possibly also inter- and intra-individual differences in perceived subjective complexity, an ideal assistance system would adapt individually and specific to different driving situations. This could lead to a better usability, correct allocation of assistance and acceptance of automated driving function.

Sensors such as cameras, radars and lidars collect data that describe an abstraction of their perceived environment and allow interpretation either through humans or computational algorithms. In a simplistic form, this data collection is similar to the cognitive processing for the first perception phase towards building situation awareness (Endsley, 1995). In what terms does environmental complexity differ mathematically (objective complexity) to an individual perceived situation complexity (subjective complexity) and how can this be measured? The second part of this question will be addressed in this paper and a solution will be developed to enable the investigation of the first part of the question in future work.

It is worth defining our interpretation of these two different versions of complexity, explicitly regarding driving environments. *Objective complexity* is the mathematical describable driving situation in which all objects within a predefined area are continuously referenced to an ego-vehicle. This mathematical description includes metrics such as the distances, velocities (relative and absolute) and the time to trajectory intersections. The mathematical composure of the factors can vary and yield different values of objective complexity depending on the interpretation of the mathematical description. In a practical example, the data would be obtained from singular or combinations of sensors, capturing information of environmental objects. This differs from general global descriptions of complexity such as the number of vehicles in the environment (Gold, et al., 2016), in which no references to an ego-vehicle and driver are drawn. The problem with global descriptions, without reference to the driver in the environment, is that the dispersion of vehicles is not evident from the point of view of the driver. When listing the amount of vehicles surrounding the ego-vehicle, no information is given where all these vehicles are (front, behind, lane

etc.). *Subjective complexity* is the perceived complexity of a driving situation from the human's perspective. This includes all stationary and moving objects relevant to the driving task. Abstract cognitive and psychological constructs such as driving situation familiarityaffect this complexity and are not measureable with similar accuracy as the metrics of objective complexity. This is mainly because measurements from designated sensors such as electroencephalography, skin conductance or any other psychophysiological measurement are not unambiguously linked to any of these constructs and quantification of human response is not possible. A scale for the subjective complexity is also arbitrary, relative to the psychological constructs and subject to individual differences.

Previous research has focused on describing environmental factors specifically for the driving environment, such as the time to resume control and the quality of the transition depend on driver-vehicle-environment factors (Gold, et al., 2016). Early work resulted in a classification scheme of driving situations with three million unique situations (von Benda, 1977). This classification scheme was later simplified to incorporate only four major aspects; horizontal course, traffic density, special weather and hazards (Fastenmeier, 1995). Due to the high complexity of factors, different types of models have been introduced to predict driver transition behaviour. The first class of models utilizes mathematical models, e.g. regression models, to extrapolate data based on empirical findings post-hoc and explain correlations in the data (McDonald, et al., 2019; Zhang, et al., 2019). A second class of models provides online prediction based on data obtained through driver and environment monitoring (Nilsson et al., 2015; Braunagel et al., 2017; Lotz & Weissenberger, 2019). However, subjective driver interpretation is missing as input data. The problem with all of these models is that defined factors can interact, e.g. traffic density or driver experience, effects that cannot be investigated in isolation within one study. Therefore, the investigation of differences between objective and subjective complexity, as defined above, has been difficult in the past. The investigation was especially difficult as drivers continuously needed to control the vehicle, always generating a response at steering. This has now changed through automated driving.

Subjective relevance is an important factor to predict individual behaviour. Ohn-Bar and Trivedi (2016) conducted research on the subjective relevance of objects in the driving environment, stating that spatio-temporal reasoning is needed to identify relevance by the driver. Therefore, the context of space and time in the driving situation of any automation level should be regarded when investigating environmental effects on the driver.

Through recent technical advances of automated driving, it is possible for the driver to take their hands off the steering wheel and observe the environment. Automated driving, specifically Level 2 and Level 3, is an ideal enabling technology suited for the investigation of differences between objective and subjective complexity. Therefore, it is possible to gather data on subjectively perceived critical complexity where previously the driver continuously generated responses at the steering and the data were open for interpretation. A distinction of intended interventions was difficult, because drivers constantly had their hands on the steering wheel.

This paper introduces a conceptual advanced driver assistance system. The system is designed to learn situations in which the driver takes back control of the self-driving vehicle, when no request to intervene is issued. The assistance system thereby registers situations based on current objective complexity from the vehicular sensors and associates it with subjective complexity. The moment drivers reclaim control through self-initiated take-over, the objective complexities gathered by vehicle sensors can be identified in which no automated driving is desired. Thereby, the assumption is formed that the drivers consider the environment as subjectively complex. Hence, the automated vehicle can learn when the automation function itself can suggest take-over predictively.

**Subjective Complexity Model**

The proposed subjective complexity model relies on the fact that the vehicle has a driver assistance system capable of simultaneous lateral and longitudinal control, i.e. without the need of having the hands on the steering wheel. Typically, this approach is only possible with advanced Level 2 or Level 3 systems. The objective of the proposed model is to make predictions when the surrounding driving complexity reaches a point in which the driver feels intervention is necessary. Thereby, a relationship between objective and subjective complexity can be investigated. The hypothesis is followed that the driver subjectively decides that complexity of the driving environment is too complex and external vehicular sensory data is recorded at intervention. Other reasons for self-initiated take-over are also possible, e.g. low satisfaction with vehicle control, and intention cannot be differentiated. It is worth noting, that the trust in the automated vehicle is affected by driving experience (Gold, et al., 2015). To show the functionality of the model, sideswipe manoeuvres were recorded with an advanced Level 2 automated truck and divided into a training and test dataset. These sideswipe manoeuvres were limited to vehicles crossing onto the ego-lane from the fast lane (left).

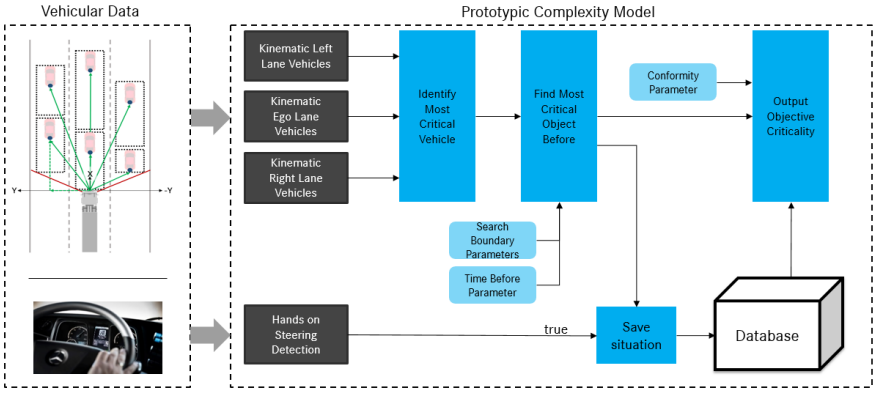



*Figure 8. Workflow of adaptive assistance system. Sensor data are split into regions of interest (left, ego, right). Kinematics of all perceived objects in these regions are calculated upon which the most critical objects is identified (see Figure 2). Output criticality is calculated based on previously recorded data.*

*Concept*

The functionality of the subjective complexity model is presented in Figure 1. The model is split into four major components. First, the perception of vehicles in the periphery of an ego-vehicle are identified, including the calculation of kinematic relationships. Secondly, the most critical object is determined based on the previously calculated kinematics. Thirdly, situations in which the driver intervenes with the vehicle without a take-over warning being displayed, are recorded and saved in a database. Fourthly, the criticality of current driving situations is calculated based on the conformity parameter of current kinematics with saved situation kinematics.

*Sensory perception and kinematics*

Sensory data of the surrounding vehicles are gathered and split into three possible lane positions. This includes the ego-lane as well as the lanes directly to the left and right. The raw data received from the sensors includes the lateral $Dist_y$ and longitudinal $Dist_x$ position as well as the speed of each object relative to the ego vehicle $RelSpd$ and object width $Width_y$. This allows the calculation of lateral $Spd_y$ and longitudinal $Spd_x$ speed of each object. Additionally, a safety corridor is defined through the width of the variable $Buffer$, see Figure 2. In this version of the proposed model, a maximum of six vehicles could be perceived around the ego-vehicle, with a maximum of two objects per lane. It should be noted that different sensor setups can alter the outcome of the system dramatically. By adding different sensors, e.g. cameras for object classification, additional data can offer subsequent critical object identification. Based on available radar data with the current sensor setup, the following kinematic variables were calculated.

$$TT_{cross\_border} = \left( \frac{\left|Dist_y\right| - \left(\frac{1}{2}\left(Width_y\right) + Buffer\right)}{Spd_y} \right) \quad (1)$$

$$TT_{headway} = \frac{Dist_x}{RelSpd} \quad (2)$$

$$TT_{collision} = TT_{headway} - TT_{cross_{border}} \quad (3)$$

$$Dist_{cross\_border} = TT_{cross\_border} * RelSpd \quad (4)$$

In total, ten kinematic variables are taken into account with the available sensors, see Table 1. Every relevant object on any of the three lanes has a separate set of these ten variables. Further variables in following implementation versions could include crossing angles, further crossing times, trajectory predictions.

*Identification of most critical object*

In the case of a self-initiated driver take-over, i.e. no request to intervene, either the complete constellation of the surrounding vehicles or a single object causing the driver to intervene needs to be identified. Here an assumption needs to be formulated, to differentiate between these two options. The proposed model assumes that one object

is the most critical in the environment and it can be defined as the object that would enter the safety corridor first, if all vehicles maintain their trajectory. This assumption corresponds to the smallest $TT_{collision}$, see equation (3), of any of the six surrounding objects. Previous development versions of the adaptive model also incorporated multiple critical objects. However, as there is always one object which is hit prior to all the others, the assumption was made that the driver reacts primarily towards this object. If the constellation of all vehicles were to be recorded, a higher amount of constallations would be possible with less likelihood of reoccuring.

*Recording self-initiated take-over situations*

If a driver intervenes with the automation function controlling the ego-vehicle, the currently most critical object is recorded to a database. Simultaneously, the model identifies where this most critical object was located for a certain amount of time previously to the take-over. The time is an adjustable parameter as well as the size of the search region, defined by a lateral and longitudinal measure. The two sets of ten kinematic variables, current and delayed, are saved with object lane positions resulting in 22 mathematical variables. Every time the driver regains control of the vehicle, the current situation with its delayed prior position is recorded to the database. As the driving environment can vary dramatically based on the type of road or national restrictions, the data and type of driving culture are completely adaptable. Similarly, the driver's interpretation of situations may vary over time and compared to other drivers. As more and more data is recorded the model adapts over time, this enables learning of personalized self-initiated take-over.

*Table 2. Kinematic variables calculated from sensor data.*

| Variable Name | Definition |
|---|---|
| $\boldsymbol{Dist_x}$ | Longitudinal distance from front of ego-vehicle to rear of object. |
| $\boldsymbol{Dist_y}$ | Lateral distance from front of ego-vehicle to rear of object. |
| $\boldsymbol{Spd_x}$ | Longitudinal speed of object relative to the longitudinal axis of the ego-vehicle. |
| $\boldsymbol{Spd_y}$ | Lateral speed of object relative to the lateral axis of the ego-vehicle. |
| $\boldsymbol{EgoSpd_x}$ | Speed of the ego-vehicle along its longitudinal axis. |
| $\boldsymbol{RelSpd}$ | Difference of longitudinal speed between the object and the ego-vehicle |
| $\boldsymbol{TT_{cross\_border}}$ | The time required for the object to cross into the safety corridor. Only considered if the trajectories of the vehicles cross. |
| $\boldsymbol{TT_{headway}}$ | The time required for the ego-vehicle to bridge the longitudinal distance to the object. |
| $\boldsymbol{TT_{collision}}$ | The time required for the ego-vehicle to reach the point where the object crosses into the safety corridor minus the time required to reach that point. This measure considers the time to collision once the safety border is breached. |
| $\boldsymbol{Dist_{cross\_border}}$ | The longitudinal distance the object is from the ego-vehicle once the safety corridor is breached. |

*Continuous Criticality Output*

Upon identifying a most critical object, the model relies on fuzzy logic (Ross, 2010) to compare current situations with previous unforced take-over situations from the abovementioned database. All kinematic variables are taken into account for the prediction method and a majority voting mechanism determines comparability of saved situations with the current driving environment. It is possible to adapt to this mechanism in the future. The model searches through all previous situations, comparing current kinematic variables to the saved situations. As it is highly unlikely that the exact situation appears twice during an self-initiated take-over, a confidence percentage in form of a conformity parameter is introduced. The definition of this confidence percentage has a profound influence on the precision of the model as discussed in the conclusion.

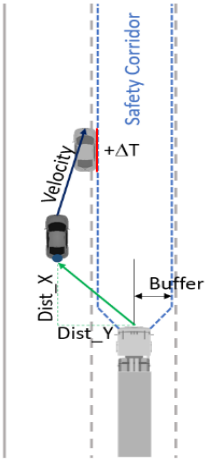



*Figure 9. Overview of distances and variables for kinematic calculations. The vehicle on the left lane requires ΔT time, corresponding to $TT_{cross\_border}$, to cross into the safety corridor on the ego-lane in front of the ego-vehicle.*

*Data Collection*

To present the functionality of the subjective complexity model, a small number of manoeuvres were recorded on a German two-lane federal road with speed restrictions. This dataset is too small to investigate the full potential of the system. However, first indications of the functionality can be examined. The sensors were mounted to a prototype Mercedes-Benz Actros with an Active Drive Assist (Daimler AG, 2019). Over the course of two hours, sideswipe manoeuvres from the left lane towards the ego-lane were recorded. This manoeuvre was an exemplary situation, which our fictive driver was uncomfortable in and chose to take-over. It can be expected that real-world traffic situations in which a driver intervenes with the automation function are seldom and would not deliver adequate data. The dataset was divided into a training and test dataset with proportions of approximately 90% to 10% respectively.

This resulted in a total of 105 training sideswipe manoeuvres, see Figure 3, and 13 test manoeuvres.

**Results**

The model is evaluated based on the self-initiated test manoeuvres that are examined qualitatively. These 13 test manoeuvres are not limited to sideswipes, they consist of take-overs due to a construction site, one sideswipe in a traffic jam at low speeds, five delayed take-overs due to sideswipes and six sideswipes from motorway entry-ramps, i.e. right side. A qualitatively comparison of vehicular signals synchronised with a dashcam video was realized for the model proof of concept. A quantitative analysis was not meaningful, as the data are limited. The complete model was simulated in MATLAB/Simulink, see Figure 4 and Figure 5, which depict the prediction value of the most critical object currently and delayed as well as the hands-on signal when the driver intervened.

*Figure 10. Exemplary sideswipe manoeuvres recorded in the training dataset.*


*Figure 11. Two exemplary self-initiated take-over situations that were not trained in the training set. Predicted sideswipe manoeuvres never reach a confidence greater 80%.*

Figure 5 displays the qualitative comparison of the five sideswipe manoeuvres, which the driver initiated with a varying delay. As shown in the graphs portraying the current similarity prediction, delayed similarity prediction of 0.5 sec and when the take-over was initiated (top to bottom), the snapshot of the actual sideswipe was predicted very accurately (vertical blue line). Overall, in four of the five delayed take-overs, the model correctly identifies a sideswipe manoeuvre with 100% confidence. The third depicted sideswipe take-over in Figure 5 with a delayed response shows a low prediction quality. It can also be seen, that sensor dropout appears quite frequently throughout the drives.

The self-initiated take-overs that were not included in Figure 5 and consisted of the six take-overs from sideswipes at motorway entry-ramps, displayed a poor quality of prediction and are not depicted. Overall take-over prediction value by the model in these other eight situations never reached over 80%. Two of these eight exemplary situations are depicted in Figure 4.

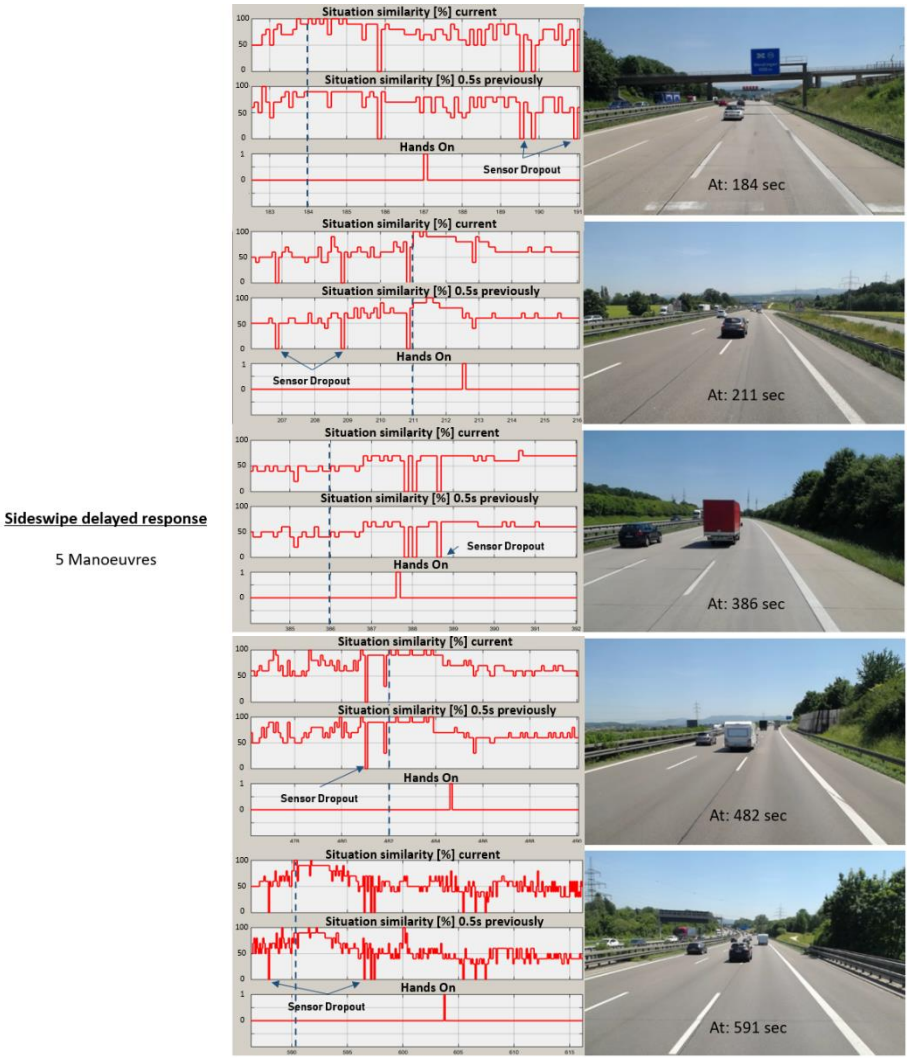



*Figure 12. Five sideswipe manoeuvres with delayed driver response. Trace data depicts confidence values for current and delayed most critical objects in the environment. The hands-on signal generated by the driver is also depicted. The blue line indicates the point in time, to which the video-snapshot corresponds.*

## Conclusion

The introduction of our subjective complexity model is an innovative solution of a learning and adaptive assistance system. By recording driver self-initiated unforced take-overs during automated driving, it is possible to monitor take-overs without interpretation of intent and behaviour. If proven reliable and beneficial, this system can predict preferences in which the driver does not trust the self-driving vehicle or feels the need to manage the driving situation. Trust is an essential component in the human-machine-interaction during automation, as drivers should be able to anticipate

system behaviour. If this is not possible, self-initiated take-overs are likely and the model offers assistance. Through continuous learning of relevant situations in which the driver wishes to control the vehicle, the vehicle itself can suggest take-over predictively. This offers different configurations of predictions for different drivers and roads. The functional layout of the model also allows the adjustment of sensors, where the effect on the predictability of take-over can be tested.

Apart from being an adaptive driver assistance system, the model can function as a carrier technology for the investigation of objective and subjective complexity. Thereby, a solution for the second part of our research question is proposed. One of the main obstacles is that sufficient data are difficult to record for this theoretical comparison. On the brink of introducing automated driving to vehicles, previously the driver continuously held control of the vehicle. This made a differentiation difficult between instances, in which the driver considered the environment to be complex. Self-initiated take-overs are valuable for the interpretation of subjective complexity. These situations show that meaning of the temporal and spatial characteristics of surrounding objects from the drivers' perspective was complex enough to motivate a take-over. It should be mentioned that self-initiated take-overs could also occur due to uncritical situations, e.g. terminating automation. In order to truly investigate the differences of subjective and objective complexities, the first part of the research question, a long-term data collection of individual drivers is required that needs documentation of driver intent.

The results of the prediction accuracy of the model shows satisfactory results. While the sideswipe manoeuvres in the test dataset were identified prior to delayed take-over in four of the five instances, one unlearned situation was not identified, see Figure 5. However, there are several reasons and possibilities to improve prediction and the validation of the model. A filtering of the signal is required for subsequent versions to bypass sensor dropout and smooth the prediction value signals. Another shortcoming in our proof of concept are the high number of false positives. It should be investigated whether these false positives occurred, due to the low distinction between a sideswipe manoeuvre being initiated and vehicle continuing in their lane. Furthermore, the point in time in which the driver initiates take-over can vary dramatically, making it difficult for the system to reference the correct critical object to the situation. Reaction times of a driver must possibly be taken into account. Finally, the system can never abstract the data to new situations. Each situation has to have happened similarly in order for the model to predict the situation in the future. However, based on the introduced conformity parameter, see Figure 1, the model can parameterise to achieve different levels of generalisation.

The model shows that this type of assistance system has promising applications in the driving context as well as research. An applied subject complexity assistance would require larger datasets, a higher variance of critical unrequested take-overs as well as seldom occurrences. The model could also be realized with a machine learning approach, however, the presented solution has the added benefit of clearly showing how and why the system functioned with specific predictions. In the future, a large dataset will be utilized as a basis for a parametrization of all variables as well as the expansion of vehicular sensor for further kinematic description of the environment.

**References**

Braunagel, C., Rosenstiel, W., & Kasneci, E. (2017). Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intelligent Transportation Systems Magazine*, 9, 10-22.

Daimler AG. (2019, May 02). Daimler Global Media Site. Retrieved July 08, 2019, from The new Actros - all new features in detail: https://media.daimler.com/marsMediaSite/ko/en/43216408

Endsley, M.R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors, 37*, 32-64.

Fastenmeier, W. (1995). Autofahrer und Verkehrssituation: neue Wege zur Bewertung von Sicherheit und Zuverlässigkeit moderner Strassenverkehrssysteme. Köln, Germany: TÜV Rheinland.

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in Automation - Before and After the Experience of Take-over Scenarios in a Highly Automated Vehicle. *Procedia Manufacturing, 3*, 3025-3032.

Gold, C., Korber, M., Lechner, D., & Bengler, K. (2016). Taking Over Control From Highly Automated Vehicles in Complex Traffic Situations - The Role of Traffic Density. *Human Factors, 58*, 642-652.

Lotz, A., & Weissenberger, S. (2019). Predicting take-over times of truck drivers in conditional autonomous driving. In: N. Stanton (Ed.), *Advances in Human Aspects of Transportation. AHFE 2018. Advances in Intelligent Systems and Computing, 786*, (pp. 329-338).

Lotz, A., Russwinkel, N., & Wohlfarth, E. (2019). Response Times and Gaze Behavior of Truck Drivers in Time Critical Conditional Automated Driving Take-overs. *Transportation Research Part F, 64*, 532-551. doi:10.1016/j.trf.2019.06.008

McDonald, A., Alambeigi, H., Engström, J., Markkula, G., Vogelpohl, T., Dunne, J., & Yuma, N. (2019). Toward Computational Simulations of Behavior During Automated Driving Takeovers: A Review of the Empirical and Modeling Literatures. *Human Factors, 61*, 642-688.

Nilsson, J., Falcone, P., & Vinter, J. (2015). Safe Transitions From Automated to Manual Driving Using Driver Controllability Estimation. *IEEE Transactions on Intelligent Transportation Systems, 16*, 1806-1816. doi:10.1109/TITS.2014.2376877

Ohn-Bar, E., & Trivedi, M.M. (2017). Are all objects equal? Deep spatio-temporal importance prediction in driving videos.*Pattern Recognition, 64*, 425-436.

Ross, T.J. (2010). *Fuzzy Logic with Engineering Applications*. United Kingdom: John Wiley and Sons, Ltd.

SAE J3016. (2018). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*, June 2018.

Von Benda, H. (1977). Die Skalierung der Gefaehrlichkeit von Verkehrssituationen. I. Teil: Ein Klassifikationsschema für Verkehrssituationen aus Fahrersicht. *FP 7320 im Auftrag der Bundesanstalt für Strassenwesen*. München, Technische Universität München.

Zhang, B., De Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F, 64*, 285-307.

# Information needs regarding the purposeful activation of automated driving functions – an exploratory study

*Simon Danner[1], Matthias Pfromm[2], Reimund Limbacher[2], & Klaus Bengler[1],*
*[1]Chair of Ergonomics, Technical University Munich, Germany*
*[2]AUDI AG, Germany*

**Abstract**

Research mostly focuses on the period of automated driving and the transition back to manual driving, while overlooking the period before the activation of a conditionally automated driving (CAD) function. Attempting to close this gap, factors influencing the intention to use CAD, such as the potential to engage in non-driving related activities (NDRAs), were analysed by performing a focus group discussion involving automated driving experts to anticipate drivers' information needs regarding an activation of CAD. These information needs as well as the drivers' expectations regarding the availability duration of CAD were investigated in an exploratory driving simulator study. For this purpose, participants ($N$ = 15) experienced four scenarios with variable durations of availability regarding the CAD function in combination with NDRTs of different lengths. The information needs anticipated by the focus group were evaluated. Results show that before activating the automation, participants mainly desired to receive information on the availability duration, or otherwise, on the duration until CAD will be available. When CAD was not available, participants wanted to know the detailed reasons. The determined information needs are assumed to assist drivers in purposefully using CAD considering their planned NDRTs.

**Introduction**

One advantage of SAE level 3 driving functions over SAE level 2 functions is that drivers do not have to monitor the system anymore while driving automated (SAE, 2018). Consequently, drivers have the option to engage in non-driving related activities (NDRAs) while automation is active. However, L3-automated systems are not designed to work under all conditions. Therefore, users can only activate the L3-automation when all conditions of its use are met. Moreover, the driver needs to be permanently prepared for a Request to Intervene (RtI) if system limitations are reached. Thus, the driver is in control over the vehicle before and after a period of CAD. Which information do drivers need before activating an automated driving function purposefully? Why do drivers want to activate CAD and which reasons would discourage them from doing so? A focus group interview and an exploratory driving simulator study were conducted and analysed on a descriptive level to receive initial answers to these questions.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

## Theoretical Background

### *Transitions to automation*

Transitions in the context of automated driving are mostly discussed when investigating the transition from automated to manual driving. This might be explained by this type of transitions' criticality (Lu et al., 2016). Lu et al. (2016) state that the activation is trivial as it seems comparable to the activation of ACC. However, the authors also state that activations pose a risk when conducted at the wrong time. For the activation of CAD, specific conditions have to be met and therefore drivers need an appropriate mental model of the system's functions and limitations in order to handle the automation safely (Forster et al., 2019). Mental models are mental representations of real objects or systems and include functionalities and logical relations (Bach, 2000). Forster et al. (2019) have evaluated two different approaches, namely working through an interactive tutorial and reading a manual before driving automated in a simulation and conducting various transitions. Results show that both concepts led to an increased understanding in comparison to a baseline group, which only received generic information about the system. Since mental models are prone to changes over time and learned system limitations can be forgotten when not experienced (Beggiato & Krems, 2013), this approach of educating the driver before usage is not considered sufficient.

### *Expectations and attitudes towards automated driving*

The possibility of conducting NDRAs is an important aspect of people's expectations towards automated driving (Howard & Dai, 2013) and thus it is indicated to investigate which kinds of NDRAs are likely to be conducted while the user is driven automatically. Pfleging et al. (2016) found that people would like to talk to occupants, watch the road, read, text, sleep, watch movies and play games during their extra time while driving automated. Hecht et al. (2019) found that people spend most of the automated drive watching videos on a mounted tablet, watching the surrounding traffic and the landscape or conducting activities on their smartphones. Participants showed a high variance regarding their NDRAs and their average activity duration.

Acceptance is a construct often used to express the willingness to accept new technologies, such as self-driving vehicles (Payre et al., 2014). According to Davis (1989) and his technology acceptance model (TAM), acceptance depends on perceived usefulness and perceived ease of use, which together predict the intention to use new technology. The possibility of conducting NDRAs free of interruptions is associated with perceived usefulness (Naujoks et al., 2017), which on the other hand is correlated with acceptance (Venkatesh & Davis, 2000). Therefore, the possibility of conducting NDRAs uninterrupted could be associated with the intention to use and thus activate CAD.

### *Information needs*

People desire driving task related information during manual driving and information related to transparency, system status and comprehensibility of system actions during CAD (Beggiato et al., 2015). These include information regarding current and next

manoeuvre as well as reasons for missed manoeuvres. Furthermore, time left in the current system status should be presented to the user. These information needs, especially the ones addressing transparency and comprehensibility, can differ between people depending on the individual trust and aim on building the same (Beggiato et al., 2015). Displaying the duration of the automated drive increases acceptance towards the system (Richardson et al., 2018) and improves take-over performance (Wandtner et al., 2018). None of the discussed information relates to a purposeful activation that would enable users to achieve their set goal by using CAD. Moreover, there are no findings in literature on information needs regarding CAD when the automation is not available.

**Research questions**

Purposeful activation of CAD requires the driver to know what purpose he pursues by activating as well as the knowledge if an activation could help him serve this purpose. Consequently, a correct mental model of the system functionality is necessary. When planning to modify the mental model by giving information, it is helpful to know what concepts of automated driving are present in mental models today. Therefore, the first research question is: *What do novices expect regarding the availability of L3-automation?* As it is assumed that these expectations require adjustment, the second research question is: *What kind of information do potential users need before activating the automation?* Furthermore, as conditions for availability are not necessary intuitively understandable, the third research question is: *What reasons for non-availability do participants assume when automation is not available in the simulation and what information do they desire regarding the automation?* Since these questions have not been addressed in research so far, this study aims on finding first answers to build hypotheses on. Moreover, the reasons why participants would or would not use CAD are questioned.

**Method**

*Focus group discussion*

For obtaining first answers to the aforementioned questions, a focus group interview involving five automated driving experts from AUDI AG was conducted. The participants are considered experts for two reasons: firstly, they are involved in the technical development of automated driving functions (either as engineers or as human factor experts), and secondly, they all experienced automated drives with novices using prototype vehicles. The discussion lasted one hour and was recorded using audio equipment. Afterwards, the record was transcripted and analysed. A research associate from TU Munich moderated the discussion using an interview-guideline prepared beforehand. The guideline consisted of four thematic blocks involving questions about their experiences with novices in automated vehicles, the novices' expectations regarding the automation's availability duration, how realistic these expectations are and what kind of information could help decrease the discrepancy between the expectations and actual functionalities at the time such a system is launched. The transcript was analysed with the focus on finding answers to these specific questions. The analysis was conducted following the approach of the qualitative content analysis with focus on deductive category assignment (Mayring,

2015). The categegories were: experiences with novices, novices' expactations, estimations about how realistic thes expectations are, potential information needs.

*Driving simulator study*

A driving simulator study was conducted to evaluate the information that emerged from the focus group discussion. Furthermore, the test persons' expectations regarding L3-automations were examined.

*Simulator and routes*
The study was carried out in a fixed-base driving simulator at AUDI AG. The driving tracks were simulated using the software Virtual Test Drive. For this study, one highway route was used which differed only regarding the availability of the automation, the traffic density or the motorway exit taken by the test persons. In all four drives, the participants started from a motorway lay-by.

*Participants*
Overall, 15 participants took part in this study. The sample consisted of 4 women and 11 men. The mean age was 27.5 years (*SD* = 3.1) and participants stated that they drove 12,214 km (*SD* = 13,009) per year on average. 20% of the participants reported that they have an ACC, 13% a lane assistance and 13% parking assistant in their own car but all of the participants had heard about these systems.

*Procedure*
The participants were informed about the procedure and a written consent was obtained. After filling out a demographic questionnaire, all participants started with a five minutes test drive experiencing manual drives as well as transitions to L3-automation and vice versa. In this way, the participants got to know the notification for availability and the RtI. Afterwards, the test persons completed four consecutive trips, filled out questionnaires and answered semi-standardised interview questions between the rides. The order of the four trips was randomised. The test persons started and ended the trips on a motorway lay-by. The automation was available during three of the four trips. Three trips took about 5 minutes each, while one trip took about 8 minutes. When the automation was not available on the routes, it was due to a missing emergency lane. If test persons nevertheless tried to activate the automation, a pop-up appeared in the instrument cluster saying "automation not available: route section not appropriate". After completing all four trips and qualitative interviews, the participants rated the information needs derived from the focus group discussion.

Before the first trip, participants were instructed to imagine a drive home from work, which they want to use to watch a short video on a tablet mounted below the central information display. They were also instructed to take the next highway exit stopping at the lay-by. Test persons were not allowed to watch the video during manual drive and had to stop the video in case of an RtI. Automation was available after 20 seconds on the highway until about 15 seconds before the next highway exit. The video was 4 minutes long, so the participants had the chance to finish it during the automated drive. This scenario illustrates the ideal situation in which a user is able to conduct an NDRA without interruption.

The instruction before the second trip was nearly the same. The only difference was that the test persons received instructions to not take the next highway exit but the one after that and were told to watch another video. This video took 8 minutes and therefore, the participants could not finish it before the RtI was issued at the first highway exit after 5 minutes. After passing this exit, the automation did not become available again and the participants drove to the next exit manually. This scenario illustrates the case where the user cannot conduct an NDRA without interruption and has no chance to finish it after being interrupted.

The instruction before the third trip was the same as before the first trip but without the instruction to watch a video. The participants were told to drive as they wished – manually or automated. The traffic density was higher in this scenario to create an unpleasant and dull highway scenario without the chance to distract oneself by an NDRA.

The instruction before the fourth trip was the same as before the first trip. The difference in this scenario was that the automation did not become available. Thus, this situation illustrates the case where an automation, which should apparently be available, is not without any notices. The route was the same as during the other trips but without an emergency lane, to examine whether the test persons were able to recognise reasons for non-availability.

*Measures*

Participants rated the information needs that emerged from the focus group discussion on a five-point Likert scale indicating how important and useful a display of this information is considered. To answer the research questions a semi-structured interview of five to ten minutes was conducted after every test drive. The investigator noted the participants' answers.

## Results

*Focus group discussion*

With regard to experiences with novices, the experts reported that people who have never had contact with automated vehicles often overtrust the automation after a short time. Furthermore, they feel disturbed by RtIs, do not understand and – in some cases – do not accept system limitations. The focus group participants stated that novices expect an automation to be available all the time even though they were informed of possible RtIs. When novices were told that an automation only works on motorways and its availability is dependent of further conditions, novices are still surprised when the automation is not available on the motorway for some time. Furthermore, experts reported that people often think they could sleep when the automation is active even though they know they have to act as fallback level. When asking how realistic the experts assess the novices' expectations, they stated the expectations are not realistic or achievable within the next years when the first L3-automations enter the market. They also reported that periods of 30 to 40 minutes of automated driving on motorways are realistic, but interruptions will be most likely. The focus group participants assumed that the discrepancy between the expectations and technical

possibilities come from non-transparent system limitations and thus incorrect mental models.

As a failure to achieve the goal – thus, discrepancies between people's expectations and the outcome of an event – leads to frustration (Ochs et al., 2008), the experts were asked which information could be displayed in the HMI to lower this discrepancy and therefore frustration. Experts stated that a display of the availability duration before and after activation of the automated system would help adapt the expectations to realistic system capabilities and therefore prevent users from frustration. Furthermore, suggestions of NDRAs, which can be conducted within an availability period, are assumed useful. In addition, an overview of all route sections where automation is probably available could be presented to make it easier for the user to organise NDRAs on a trip. Moreover, a display when automation will be available if it is currently not available could prevent frustration especially if users expect the automation to be available without limitation, at least on a motorway. Table 1 shows the potential information needs anticipated in the focus group discussion.

*Table 1. Potential information needs anticipated in the focus group discussion*

| *Anticipated information needs when automation is available* | *Anticipated information needs when automation is not available* |
|---|---|
| Estimated availability duration of the automation before activation | Reasons for non-availability |
| Certainty of availability duration | Duration until automation is available |
| Overview of availability periods on whole route | |
| Suggestions of NDRAs feasible during automated drive | |

*Driving simulator study*

The test persons experienced four test drives in permuted order. However, the rides are referred to as first trip, second trip etc. analogue to the aforementioned descriptions.

To answer the first research question, the test persons were asked how long they would have expected the automation to work. As this question is explicitly important when the participant conducts an NDRA, which is either feasible in the period of automated driving or not, it was asked after trips one and two. These trips represented the ideal and non-ideal situations in which the NDRA is either feasible (first trip) or interrupted due to an RtI (second trip). Seven participants experienced the first test drive before the second. On the first time asked, eight participants answered that they had expected the automation to be available infinitely long and therefore until they leave the highway. Figure 1 shows all answers and their quantities.

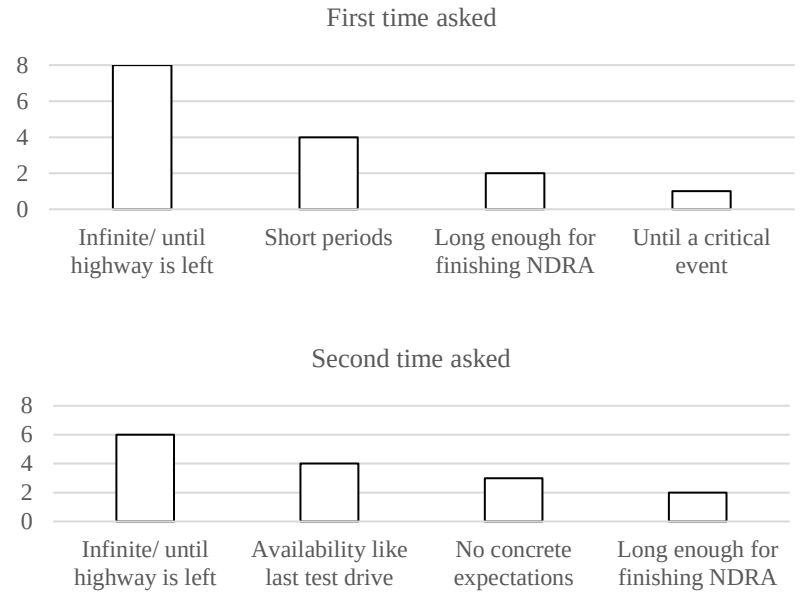



*Figure 1. How long test persons expected the automation to be available.*

To answer the second research question, the test persons were asked what kind of information they wished to be displayed before activating the automation. Ten participants stated they wished for a display of the period or distance the automation would be available. Two test persons stated they did not wish for more information before activating the automation but a display of the automation period after activating. Figure 2 shows all answers to research question two.

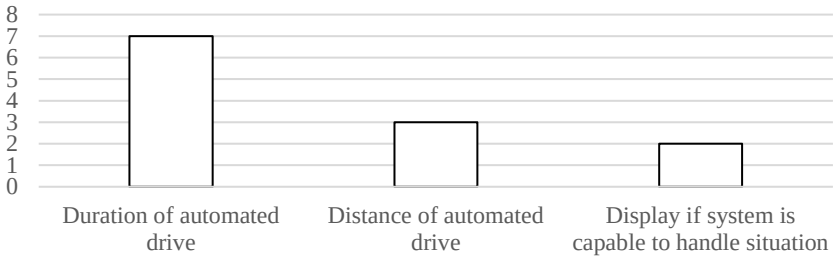



*Figure 2. Information test persons wished to be displayed before activating the automation.*

Research question three was what kind of information test persons desire when automation was not available. Participants answered this question after experiencing test drive four during which the automation did not become available. Eight participants stated they wished to know when the automation would be available, in either time or distance. All answers are shown in figure 3.

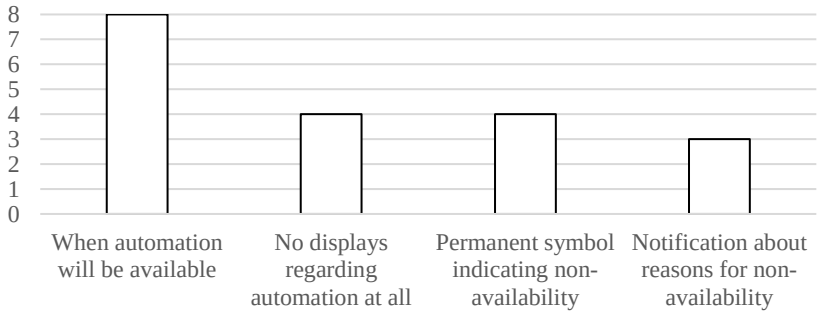



*Figure 3. Information test persons wished for when automation was not available.*

If participants tried to activate when automation was not available a pop-up message appeared. Eleven participants desired more detailed reasons, stating this feedback was not sufficiently understandable. Six participants tried to activate and saw the feedback while nine participants experienced it when the investigator instructed them to try to activate. When asked which reasons for non-availability the test persons assumed, six participants stated they believed the traffic density to be the reason while four thought some technical issues to be responsible for non-availability. None of the participants guessed the right reason, which was a missing emergency lane.

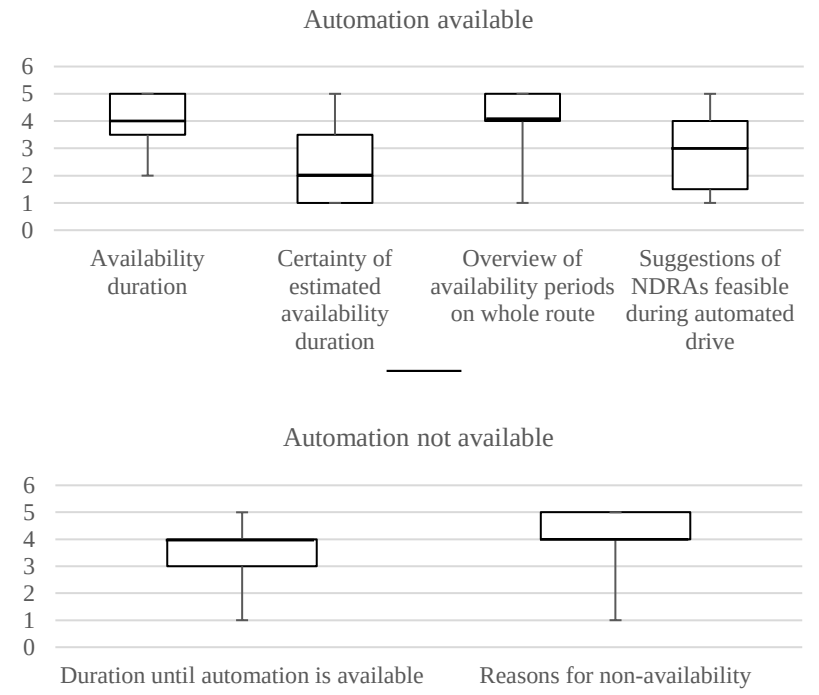



*Figure 4. Participants' ratings of the potential information needs from the focus group.*

Furthermore, it was investigated which reasons people have to activate CAD when available and what reasons would prevent them from doing so. Interview data showed that the main reason for activating is the desire to conduct an NDRA while the main reason for not activating was the desire to drive faster than an automation would. Another factor for the potential activation was driving pleasure with having fun while driving leads to no activation. After all trips the test persons were asked to rate the potential information emerged from the focus group discussion. Figure 4 shows the medians of the ratings.

**Discussion and conclusions**

The present data suggests that potential users of future L3-automations have too high expectations regarding the availability periods of the automation and consequently the NDRAs feasible without interruption. They expect an automation to be available for an infinitely long time within the most apparent limitations, for instance on a highway, and do not expect further limitations leading to RtIs. Interview data showed that test persons mainly desired a display of time or distance the automation will be available for in order to be able to compare the estimated duration of their NDRAs with the duration of automated driving. Some test persons even stated they would not wish to activate CAD if their planned NDRA was not feasible during the automated drive. Furthermore, participants desire a display of the anticipated time until the automation will be available while it is not. This information need was not anticipated in the focus group but would be covered by a display of an overview of all availability periods as it would contain the time or distance between two of the same. Interestingly, when automation is not available, some participants desired an extra symbol indicating non-availability while others explicitly stated they do not want an extra display for non-availability, as this would be redundant, revealing individual differences. Another important result is the desire to know the reasons for non-availability. This may lead to a higher perceived understanding of the system as it does when RtIs come with an explanation (Körber et al., 2018). As participants desire to know why the automation is not available, a display explaining the reasons seems all the more important, as no participant was able to recognise the reason in the simulation by oneself. Investigating why participants would use the automation or not, answers mostly referred to either conducting NDRAs or driving faster than the automation would. This indicates, these two factors mainly influence the decision wheter to activate or not.

Further research should validate the information needs reported in this study even though the information coming from the experts and from the novices mainly coincide. The focus of this study and thus of the study design was on NDRAs and conducting them free of interruptions and therefore the results may be biased in this direction. Moreover, a naturalistic driving study could lead to further results. There might be more information needs regarding the automation before activating the same. Generally, there is a gap in research concerning the activation of automated driving functions, which should be closed. This work suggests the activation of automated driving by the driver to be an important step, which should not be perceived as trivial, especially as wrong or purposeless activations and consequently not feasible NDRAs may lead to frustration or decreased acceptance and thus decreased usage.

**References**

Bach, N. (2000). Mentale Modelle als Basis von Implementierungsstrategien. *Konzepte für ein erfolgreiches Change Management, Wiesbaden*.

Beggiato, M., & Krems, J.F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: Traffic Psychology and Behaviour*, *18*, 47-57.

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J., Othersen, I., & Petermann-Stock, I. (2015). What would drivers like to know during automated driving? Information needs at different levels of automation. *7. Tagung Fahrerassistenzsysteme*. Munich: TÜV SÜD Akademie GmbH.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

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*, 143.

Hecht, T., Feldhütter, A., Draeger, K., & Bengler, K. (2019). What Do You Do? An Analysis of Non-driving Related Activities During a 60 Minutes Conditionally Automated Highway Drive. In *International Conference on Human Interaction and Emerging Technologies* (pp. 28-34). Cham: Springer.

Howard, D., & Dai, D. (2014). Public perceptions of self-driving cars: The case of Berkeley, California. In *Transportation Research Board 93rd Annual Meeting* (pp. 1-16).

Körber, M., Prasch, L., & Bengler, K. (2018). Why do I have to drive now? Post hoc explanations of takeover requests. *Human Factors*, *60*, 305-323.

Lu, Z., Happee, R., Cabrall, C.D., Kyriakidis, M., & de Winter, J.C. (2016). Human factors of transitions in automated driving: A general framework and literature survey. *Transportation research part F: Traffic Psychology and Behaviour*, *43*, 183-198.

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In *Approaches to qualitative research in mathematics education* (pp. 365-380). Dordrecht: Springer.

Naujoks, F., Forster, Y., Wiedemann, K., & Neukum, A. (2017). Improving usefulness of automated driving by lowering primary task interference through HMI design. *Journal of Advanced Transportation*, *3*, 1-12.

Ochs, M., Pelachaud, C., & Sadek, D. (2008). An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1* (pp. 89-96). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation research part F Traffic Psychology and Behaviour*, *27*, 252-263.

Pfleging, B., Rang, M., & Broy, N. (2016). Investigating user needs for non-driving-related activities during automated driving. In *Proceedings of the 15th*

*international conference on mobile and ubiquitous multimedia* (pp. 91-99). New York: ACM.

Richardson, N., Flohr, L., & Michel, B. (2018). Takeover Requests in Highly Automated Truck Driving: How Do the Amount and Type of Additional Information Influence the Driver–Automation Interaction? *Multimodal Technologies and Interaction, 2*, 68.

SAE Internation (2018). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J3016*.

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*, 186-204.

Wandtner, B., Schömig, N., & Schmidt, G. (2018). Secondary task engagement and disengagement in the context of highly automated driving. *Transportation research part F: Traffic Psychology and Behaviour, 58*, 253-263.

# Driver's Experience and Mode Awareness in between and during Transitions of different Levels of Car Automation

*Paula Lassmann[1], Ina Othersen[2], Matthias Sebastian Fischer[1], Florian Reichelt[1], Marcus Jenke[1], Gregory-Jamie Tüzün[1], Kassandra Bauerfeind[2], Lisa Mührmann[2], & Thomas Maier[1]*
*[1]University of Stuttgart, IKTD, Germany*
*[2]Volkswagen AG, Wolfsburg, Germany*

**Abstract**

Highly automated driving will have a significant impact on our future mobility. When a driver uses a system that comprises different SAE levels (L0, L2 and L3) the Human Machine Interface (HMI) needs to support the mode awareness of the driver at all times. While in L2 the driver has to monitor constantly, in L3 he can spend time on non-driving-related-tasks. The publicly funded project TANGO (Technology for automated driving, optimized to the benefit of the user) enables the design of an "attention and activity assistant" for automated truck driving in L2 and L3. The HMI of the project provides information about the automation level through different modalities: visually (instrument cluster & LED strip), auditory (sounds and voice announcements) and haptically via a tactile seat matrix. By conducting a driving simulation study, the usability of the HMI was investigated. The goal was to determine the ability of the driver to differentiate cognitively three SAE levels with the support of the TANGO HMI.

**Introduction**

The vision of automated driving stands for an increase in road safety and efficiency, a fatigue-free and stress-free driving experience as well as a safe use of built-in information and communication systems while driving. The fact that such a need exists among car drivers has already been sufficiently demonstrated in various studies (cf. Petermann-Stock et al., 2013; Wulf et al., 2012). However, the benefits of automation can also be beneficial for another group of users - professional drivers. They could be supported in their daily work routine by hours of monotonous journeys. The altered human-vehicle interaction has not yet been sufficiently examined in the field of trucks.

Within the project of TANGO, the research project concentrates on SAE Level 2 (L2) and SAE Level 3 (L3) systems (SAE International, 2018) and their transitions. L2 provides the driver with combined support in longitudinal and lateral guidance by means of an automated function. However, the driver must constantly monitor the driving situation and be prepared to intervene immediately. This monitoring function

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

is omitted in L3, where the system takes over the complete vehicle guidance in certain conditions (e. g. traffic jam or on motorways). In these situations, the driver no longer has to be “in the loop" and can therefore potentially turn to activities not related to driving. The driver only has to be ready for manual vehicle guidance within a period of several seconds when the system requests the driver to take over the vehicle guidance.

Especially in these quite similar modes of automation, an adequate awareness of the situation and the system is mandatory in order to avoid errors and to achieve an ideal human machine interaction (Sarter & Woods, 1995; Kolbig & Müller, 2013). Situation awareness is defined as “the perception of the elements in the environment within a span of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1988a, p. 97). It is understood as a dynamic process in which errors but also corrections can take place at all three levels (Endlsey, 1988a, 1995a, 1995b). In the course of automated systems, situation awareness must also be enriched by system awareness. It can be understood as part of situation awareness and thus includes the same processes: the knowledge and understanding of system information and system-relevant environmental information as well as the anticipation of the information (Sarter & Woods, 1995; cf. Othersen, 2016). If this system awareness is incomplete, the probability of mode confusion raises. Thereby, the system reacts differently than expected by the user. The user may behave inappropriately (e. g. monitoring activities in L3) or miss actions (e. g. missing monitoring activities in L2; Bredereke & Lankenau, 2002). Studies could identify the missing supervision and higher attention to side task or longer viewing distances from the road in L2 (cf. Buld et al., 2002). Petermann-Stock (2015) also identified uncertainties regarding the system status and the required action through increased focus on relevant displays. Above all, an over-confidence in low automation levels, where a lack of monitoring with the potential oversight of system errors occurs, should be prevented.

The human machine interaction changes significantly through the use of automation, so that earlier actions are replaced by supervision or withdrawal from the driving task. The aim of efficient HMI should therefore be to provide important information for adequate awareness of the situation and the system as well as to prevent mode confusion as far as possible. The multimodal HMI developed in the TANGO project will be evaluated for the first time in a driving simulator study with professional drivers. The research questions of the study are as follows:

- Do people know which SAE Level they are in?
- Do people know their tasks according to the SAE level?
- How efficient, effective and satisfyingly is the level change supported by the TANGO HMI?
- How do people react to a critical driving situation?

## Methodology

*Experimental setup*

The study took place at the vehicle ergonomics test facility at the research and teaching area Industrial Design Engineering of the University of Stuttgart. This fully variable model of a vehicle interior is based on a static, electrically adjustable seat box with driver and passenger seats. The driving simulation is shown on five monitors (Samsung, 1920 x 1080 pixels, 59") with a 210° field of vision covering, two side mirrors and a rear-view mirror. For simulation, the software SILAB (WIVW, Version 5.0) was in use. The rides took place on a two-lane motorway with an emergency lane. During all automated rides, the vehicle's speed was set to 100 km/h on the right lane. The vehicle was occasionally overtaken by other road users.

The study design included two independent variables: One variable was automation level (L0, L2, L3), respectively transitions, as a within factor. Each participant experienced each automation level and possible transition. However, the order was counterbalanced in between two groups. The second independent variable was the arousal of a critical event (same situation, either in L2 or L3) as a between factor, which occurred at the end of test run two. The difference between the take-over situations before the critical situation was the fact, that in L2 no warning was given, whereas in L3 the system gave a take-over request (TOR) (for an overview of the ride see fig. 2 below).

*User centred HMI*

For promoting the mode awareness, the automation level state was supported by different HMI elements. A schematic layout of the vehicle cockpit is given in fig. 1. In this HMI, the L2-mode was called "Assistance Plus" (colour code: blue) and L3-mode "Autopilot" (green). Orange and red were used for warnings. The HMI included an instrument cluster, a head unit display above the centre console and a detachable tablet placed on a holding to the right. A colour-coded LED strip showed the level colour and was positioned at the bottom edge of the windshield. For (de-)activating L3, two push-buttons (that lit up in green when the L3 was available) on the steering wheel had to be pushed simultaneously. One push-button on the centre console, which lit up in blue, (de-)activated L2. The push-button in the middle of the steering wheel was used for the Sign Detection Task (SDT; see chapter *Data and analysis*). A tactile seat matrix (TSM; Schwalk et al., 2015), was used for tactile feedback during the TOR and after activation of the automation. It consisted of a 4x4-matrix in the backrest and a 3x5-matrix inside the seating area. All processes were supported by voice announcements and icons within the instrument cluster. As soon as a change from L2 to L0 was recommended, the driver had 2.2 s to deactivate Assistance Plus either by pressing the above-mentioned button or by driving related intervention (using the accelerator or brake pedal or oversteering). However, Intervention always resulted in a level change to L0. If the driver neither pressed the button nor intervened, the automation switched off. In L3 the driver had 15 s to react to a level change. If the driver did not react the system did a safe stop.

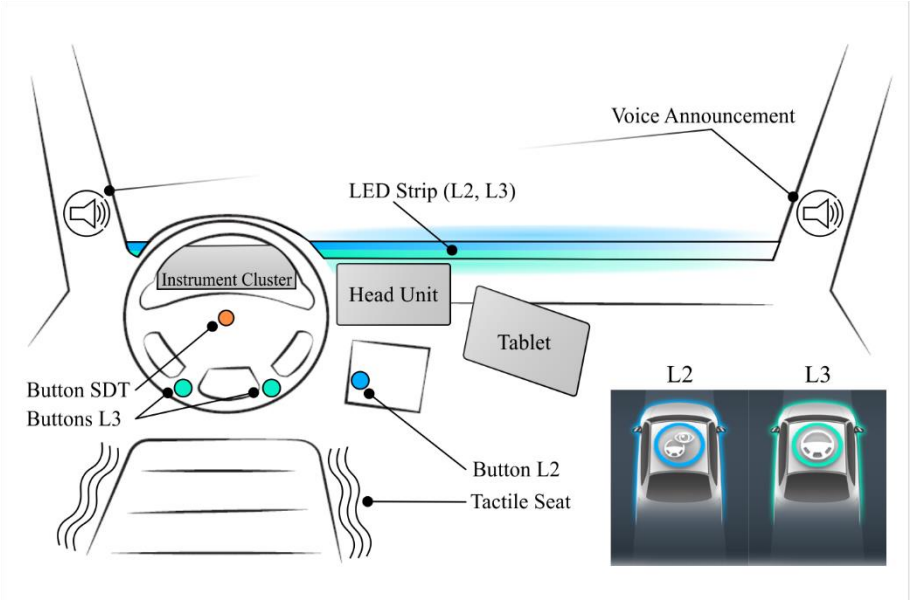



*Figure 1. Simplified representation of the mounted HMI in the driver's cockpit. The icons at the right bottom are displayed in the instrument cluster.*

*Procedure*

At the beginning, participants received an introduction to the research topic, the detailed description of the system functionality and the different tasks they have to perform according to the SAE levels. Afterwards they answered demographic questions and went through an acclimatisation ride in which all SAE levels and all tasks could be experienced.

One test ride consisted of three consecutive runs with automation and four transitions, with a total run time of approximately 90 min. Transitions were announced by the system according to the study setup. After each transition, questions were asked about the transition itself in terms of effort, mode awareness as well as the HMI without pausing the ride (see fig. 2). Participants answered all questions throughout the study verbally, which the investigator documented. In addition, the simulation was paused in the middle of each test drive (system freeze) for answering questions on the mode awareness. At the end of the first drive, the participants took a break during which they had to complete another questionnaire regarding the overall driving experience in this first half of the simulation. Subsequently, the second test drive took place – similar to the first one. However, at the end of the test drive, a critical event (system error) occurred, in which the driver had to take over. This critical event consisted of a traffic jam that suddenly occurred after a hill and therefore could not have been noticed early. In L2 the system did not brake on time. Therefore the drivers needed to initiate a transition and brake themselves to avoid a crash. However, in L3 the system announced a take-over within 15 s. After finishing the second ride, in addition to the same questions that were answered at the end of the first drive, another questionnaire referred to the perception of the critical event.

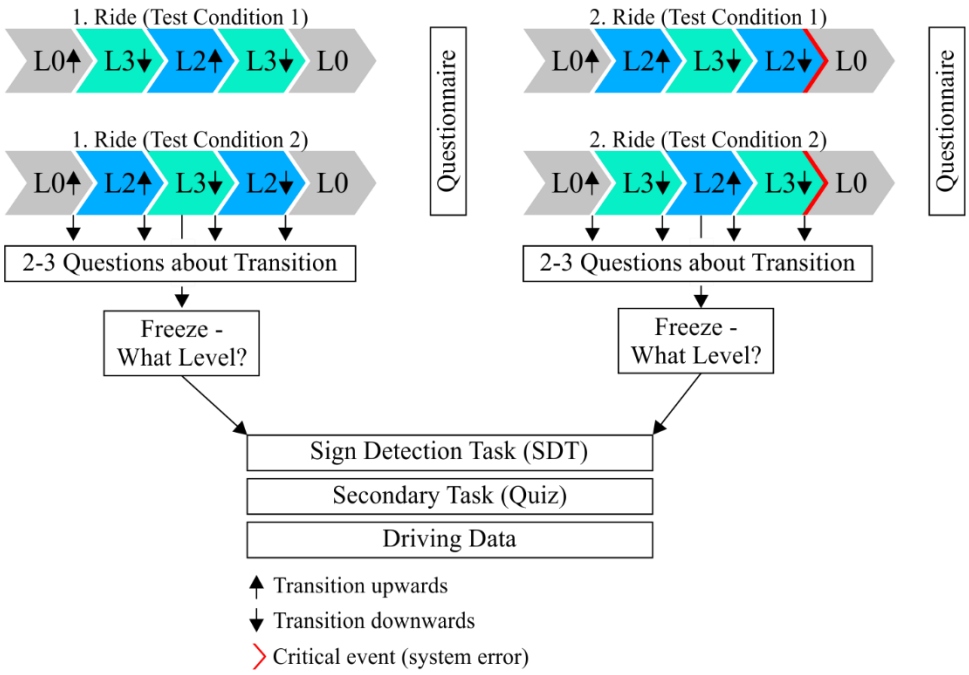



*Figure 2. Schematic study design and procedure (two rides per person with two different test conditions as in between factor).*

*Data and analysis*

A non-driving-related-task as well as a driving-related task assessed the mode awareness. The driving related task was supposed to measure the monitoring performance of the drivers and at the same time, assess the mode awareness. The drivers had to detect speed limit signs, which is called the Sign Detection Task (Lassmann et al. 2019). The design of the SDT is based on detection tasks of driving relevant stimuli that have been used for assessing vigilance and therefore monitoring performance (cf. Greenlee et al. 2017; Heikoop et al. 2017). In L2, the participant was supposed to press the SDT-button located on the steering wheel when a specified speed limit sign (100 km/h) could be seen on the roadside for 3.5 s. If the driver reacted to another road sign, this was considered as error. The SDT requires visual attention and could be compared to monitoring activity in terms of suddenly appearing obstacles in L2 (Lassmann et al. 2019). The hit rate and the response time to the stimulus were recorded. In case of pressing the SDT-button in L3 it was considered as a mode confusion since monitoring is not required in this level. In addition to the SDT, a secondary task (quiz, based on Petermann-Stock et al., 2013) was offered on the tablet to the right. The drivers could decide themselves if they interacted with the quiz (either in the fixed position or hand-held). It consisted of 262 questions in German with four answer options each. The quiz has been proofed as an engaging task for truck drivers in several studies within the TANGO project (e.g. Bieg et al. 2019).

Another objective measure of mode awareness is a variation of the SAGAT (Situation Awareness Global Assessment Technique; Endsley, 1988b) survey method. The SAGAT record the respective situation-specific knowledge of the person via questions concerning perception, understanding and anticipation of the situation after freezing the simulation. In this study, the HMI was hidden or frozen after a transition instead of the entire simulation and questions were asked about the respective automation level and the distribution of responsibilities. Besides that, the Mode Awareness questionnaire from Benecke (2014) was used as subjective data. This includes the areas of perception, understanding and anticipation of the system status. In addition, individual items were asked for the critical event with regard to effort (Subjective Experienced Strain Scale; Eilers et al., 1986), subjective reaction quality and subjective criticality. All questions were implemented using a five-level Likert scale.

For this purpose, mean value differences were calculated using the Wilcoxon rank-sum test as well as variance statistical methods with and without repeated measurements for the factors measurement time and automation level at a significance level of $\alpha = .05$.

*Participants*

The driving simulation was performed with 30 participants (aged 22 to 60 years, $M$ = 41.6, $SD$ = 10.8). The group consisted of twenty truck drivers, three bus drivers and seven other frequent drivers with an average annual kilometrage of 85,500 ± 36,667 km (range 20,000 to 200,000 km). For 2.5 hours of simulated driving and questioning the participants received an incentive of 100 Euro. Due to measurement failures, motions sickness or language problems during the study, eight participants were excluded from the analysis, which leaves 22 subjects.

**Results**

*Mode Awareness*

*Sign Detection Task (SDT)*
Only in L2, people should perform the SDT. However, two participants hit the button continuously, three subjects once, while being in L3. The rest (77.0 %) performed correctly by not hitting the button. In L2 hit rates reached a mean of 77.6 % (16.9) with values ranging from 50 to 100 % with no change over time ($F$[2,42] = 1.900; $p$ = .162; $\eta^2$ = .083). Seven persons had a mean under 70.0 %, whereas the hit rate of three of them increased throughout the study. 15 had a hit rate of over 88 %. The mean of the reaction times of the hits was 1.95 s (.35). No change over time occurred either ($F$[2,42] = 1.042; $p$ = .362; $\eta^2$ = .047).

*Secondary Task – Quiz*
The results showed a shift of attention in higher automation level ($F$[2,20] = 103.33; $p \leq .001$; $\eta^2$ = .91) (see fig. 3). The drivers did less quiz questions during L0 ($M_{L0}$=1.33) than in L2 ($M_{L2}$=34.77) and L3 ($M_{L3}$=56.39). In addition, there was an effect of time (First, Second and Third Time in either L0, L2 or L3) for all three

automation levels ($F$[2,20] = 5.14; $p \leq .05$; $\eta^2 = .34$). All test persons performed less quiz questions over time ($M_{M1}$=36.01; $M_{M2}$=31.02; $M_{M3}$=25.41).

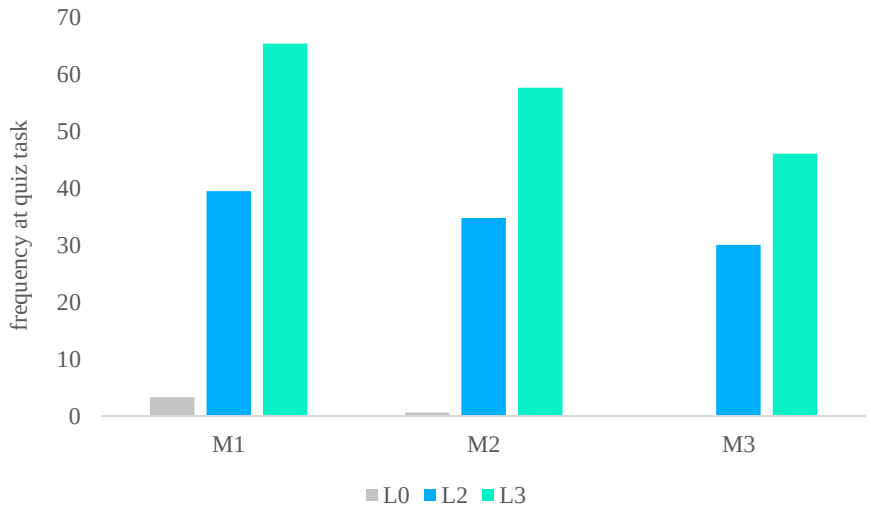



*Figure 3. Performance frequencies on secondary task for the complete drive; Automation level: L0 – L3; time of measurement: M1-M3 (first, second and third Time in either L0, L2 or L3).*

The correlation of the performance frequencies of the quiz and the hit rates of L2 are negatively correlated (Spearman: $r$ = -.488; $p \leq .05$). Looking at the correlations of each measurement time, only a correlation of M3 exists: r = -.566 ($p \leq .01$).

*Subjective Mode Awareness*

The analysis of the subjective evaluation of mode awareness showed that participants in the two groups did not differ in terms of mode awareness ($F$[1,20] = 0.004; $p$ = .952; $\eta^2$ = .000). However, there was an effect on the time of measurement ($F$[1,20] = 10.664; $p \leq .01$; $\eta^2$ = .348). The mode awareness improved during the ride ($M_{before}$ = 4.47; $M_{after}$ = 4.68).

During the freezing situation, 21 of 22 participants were able to reproduce the current automation level they were in, as indicated by the correct labelling of the automation mode (i.e. 'Assistance Plus'/ 'L2'), or the matching colour (i.e. 'blue mode'). Participants in L2 and L3 were equally aware of the system mode during the freezing situation ($Mdn_{L2}$=5.00, $Mdn_{L3}$=4.83, $W$=70, $p$=.209). However, 31 % of the participants in L2 and 22 % in L3 were unsure about the correct tasks they had to perform.

Concerning the system error that had to be detected either by the participants themselves in L2 or by a TOR by the system in L3, no significant differences could be found in the degree of effort ($Mdn_{L2}$=2.17, $Mdn_{L3}$=1.60; $W$=38, $p$=.125), the speed while overtaking ($Mdn_{L2}$=4.90, $Mdn_{L3}$=4.67, $W$=74, $p$=.220) and their confidence with the quality of their reaction ($Mdn_{L2}$=4.83, $Mdn_{L3}$=4.80, $W$=58, $p$=.882). Finally, both

groups evaluated the driving task as well adopted ($Mdn_{L2}$=4.83, $Mdn_{L3}$=4.90, $W$=64, $p$=.689), and quickly surveyed ($Mdn_{L2}$=4.83, $Mdn_{L3}$=4.80, $W$=58, $p$=.882).

*Transition and Critical TOR*

In the following section, only the transitions into and between automation levels (see fig. 4) are addressed. The change into L0 (critically and uncritically) is focused in fig. 5. The analysis of the transitions shows that a few participants had problems changing levels. 22.73 % did not make the safe transition from L0 to L2 at the first attempt. They switched at least once to L2 and back to L0 due to another button, pedal or steering wheel operation. 18.19 % had the same problems transitioning from L0 to L3. The same number of participants did not change directly from L3 to L2. They changed into L0 before reaching the right level. There were no problems when changing from L2 to L3.

The transition times of three participants were considered as outliners (3σ) and therefore excluded from statistical analysis. Each participant changed from L0 to L2 and L3 once. They changed twice from L2 to L3 and L3 to L2 due to the study setup, without an effect of time of measurement (first or second time; L2 to L3: $t(18) = 0.498$; $p = .624$; L3 to L2: $t(18) = 1.895$; $p = .074$). Therefore the mean of both values was used for the following analysis. Reaction times differed in terms of transitions ($F[2,253] = 8.480$; $p = .001$; $\eta^2 = .320$). The transition from L3 to L2 ($M_{L3toL2} = 6.90$) took more time than the changes from L0 to L2 ($M_{L0toL2} = 4.28$; $p = .005$) and L2 to L3 ($M_{L2toL3} = 4.12$; $p \leq .001$).

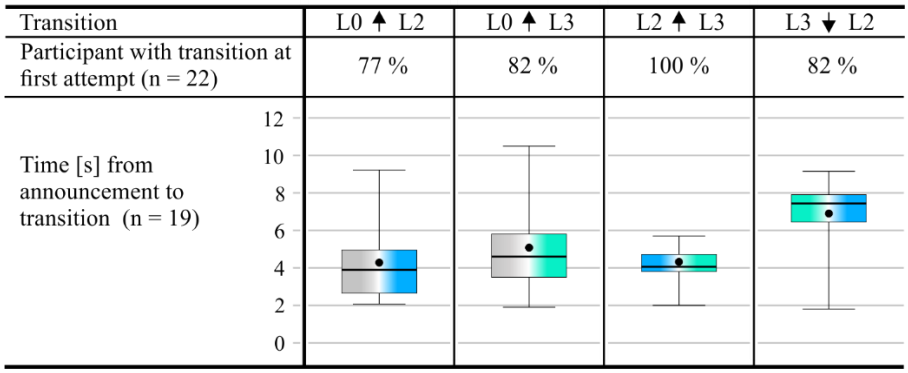
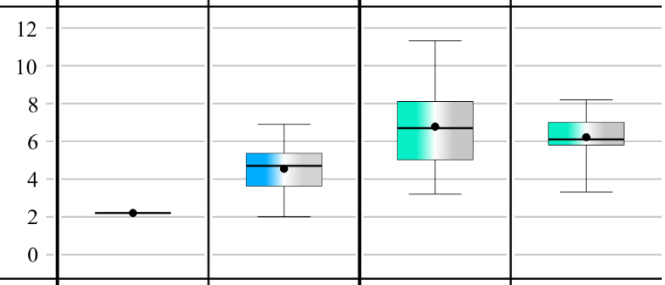



*Figure 4. Transition times to automation and within automation modes (initiated via button press).*

The following results are displayed in fig. 5. Due to small in between group sizes (9 and 10 participants), the data was not analysed statistically. After the request to change from L2 to L0, the automation was switched off after 2.2 s. No driver reacted via intervention or button press within this time. At this point, no statement can be made about the drivers' handling of the situation. In the critical event, all drivers acted within 6.90 s with a mean of 4.54 s (onset: fastest transition minus 2 s reaction time)

by using the pedal or steering wheel intervention during the critical event. None of the participants collided with another vehicle.

The uncritical transition from L3 to L0 was made within 3.20 and 11.32 s with a mean of 6.77 s, whereas in the critical situation, the reaction time decreased slightly to a mean of 6.22 s (range from 3.31 to 8.2 s). Without a critical event, 50 % used the buttons to change level. All other participants used pedal or steering wheel intervention. The buttons were used less in the case of the critical event (33 %). None of the participants provoked a safe stop (15 s after the announcement).

| Transition | | L2 ↓ L0 | L2 ↓ L0 critical event | L3 ↓ L0 | L3 ↓ L0 critical event |
|---|---|---|---|---|---|
| Number of participants in group | | 9 | 10 | 10 | 9 |
| Time [s] from announcement to completed transition | |  | | | |
| Collision during critical event | | N/A | 0 % | N/A | 0 % |
| Transition initiation | Button(s) | 0 % | 0 % | 41.67 % | 30 % |
| | Interventions | 0 % | 100 % | 58.33 % | 70 % |
| | No reaction * | 100 % | 0 % | 0 % | 0 % |



* *L2: no reaction within 2.2 s*
*L3: no rection within 15.0 s (safe stop)*

*Figure 5. Transition times for changing into L0 in an uncritical and critical situation (between factor).*

**Discussion**

The study identified the influences of the various automation levels on mode awareness. According to the SDT, two subjects had a continuous task confusion. Seven people failed the monitoring performance, since they would not have detected important signs and reacted to them within 3.5 s in over 30 % of the cases in L2. There might be three factors which had influence on the SDT performance: distraction, task confusion due to insufficient knowledge about the tasks or mode confusion. The quiz was a visual distracting non-driving related task, which was available throughout the levels. Visual distraction leads to a bad performance for detecting obstacles (Lorenz et al. 2015) as well as in the SDT (Lassmann et al. 2019). This thesis is supported by the negative correlation of the SDT hit rates and the performance frequency of the quiz. Subjects who were rather involved in the quiz, missed road signs. Regarding the factors task and mode confusion, the objective data gathered during the freezing situation might help to explain the results. Overall 21 of 22 participants were able to

correctly reproduce the current automation level they were in, but 31 % of the participants in L2 and 22 % in L3 were unsure about the respective tasks. These results could actually be an indicator of mode confusion, which refers to a discrepancy of the participants' belief about which aspects of vehicle performance are controlled by themselves and which are controlled by the automation at a particular instance (Cummings & Ryan, 2014).

In terms of the transition, participants did fairly well. After having tried the transition once during the acclimatisation ride, most succeeded in transitioning at the first attempt within a few seconds. Most problems that occurred were due to the fact that people either still pressed a pedal, pressed the button for too long or did not trust the trajectory of the simulation. For most people this happened only once during a whole test ride. In addition, subjects did not change levels faster while doing it the second time which speaks for good usability at the first place. In summary, according to the results, the HMI supported the user during transitions well. Nevertheless, a quote of 100 % transitions at first attempt would be desirable.

For the changes from L3 into a lower level, people took more time, which is in line with the findings of Gold et al. (2013): the longer the possible time frame for take-over, the longer the take-over takes. Even during the critical situation, take-over times did not change much, which supports the thesis, that people were rather trustful of the system. A timeframe of 2.2 s for the transition from L2 to L0 was not enough for drivers to react to the change and the readiness of the driver for take-over was not checked. This shows the danger of a L2-system: undertaking a transition from L2 to L0 without an explicit driver interaction, monitoring or a fallback action might lead to a situation of an unsupervised car in motion. For this reason a driver monitoring will be implemented in the TANGO system to check the driver's readiness. However, all drivers became aware of the critical situation in L2 and reacted in time to prevent an accident, which leads to the assumption that drivers were aware of the monitoring task and also the mode. In terms of take-over from automation to manual driving, the intervention seems to be more intuitive for drivers than pressing a button.

In summary, the results of the study seem divers. According to the findings of Lee and See (2004), the misbelief about the vehicle's operation is a result of overtrust or undertrust in the automated system. In this context, it could be assumed that participants in L2 had overtrust in the automated system, as they incorrectly thought that they could fill in the quiz, despite being supposed to watch the traffic. Overtrust can lead to misuse of the automated system, where the driver applies the automation to a roadway environment that is outside of the automation's operational scenarios. In the critical take-over situation, however, the drivers with L2 were able to update their situational awareness quickly enough so that there were no problems with out-of-the-loop performance. On the contrary, participants in L3 had possibly distrust in the system, as they thought that they had to watch the traffic, despite the automated system taking over this task completely. Participants believed that the automation performance was less than it actually was, which leads to a disuse of the automated system and thus removing the possible benefits of the automation (Lee & See, 2004).

## Conclusion

This study on mode awareness with regard to different automation levels was able to show that the test persons could subjectively indicate the correct automation level, but made mistakes in indicating the tasks which they had to perform. This corresponds to the objective performance in the secondary task. On the one hand these results show, that the HMI succeeded to convey the information of different automation modes that were obvious to the driver. On the other hand the results could actually be an indicator for mode confusion which refers to a discrepancy between how the participants believed the vehicle to be operating and how the vehicle was actually operating during L2 and L3 – e.g. that monitoring the system could be achieved while performing a visual non-driving related activity. Therefore the tasks during automation should be emphasised more clearly – either by instruction or by the system - and internalised by the driver. However, in conclusion the TANGO HMI supports the driver well, especially in regard to transitions, but can be improved regarding assistance of mode and task awareness.

## References

Benecke, S. (2014). *Mode Awareness im Fahrkontext – ein Vergleich zweier Bedienkonzepte für die teilautomatisierte Fahrt.* Unveröffentlichte Bachelorarbeit. Lüneburg: Leuphana Universität Lüneburg.

Bieg, H.J., Daniilidou, C., Michel, B., & Sprung, A. (2020). Task load of professional drivers during level 2 and 3 automated driving. In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference* (pp. 41-52). Available from https://hfes-europe.org

Bredereke, J. & Lankenau, A. (2002). A Rigorous View of Mode Confusion. In S. Anderson, S. Bologna & M. Felici (Hrsg.), *Computer Safety, Reliability and Security* (S. 19–31). Berlin Heidelberg: Springer-Verlag.

Buld, S., Krüger, H.-P., Hoffmann, S., Kaussner, A., Tietze, H. & Totzke, I. (2002). *Wirkungen von Assistenz und Automation auf Fahrerzustand und Fahrsicherheit. Veröffentlichter Abschlussbericht Projekt EMPHASIS: Effort-Management und Performance-Handling in sicherheitsrelevanten Situationen* (Förderkennzeichen: 19 S 9812 7). Würzburg. Retrieved 03.12.2015 from http://www.psychologie.uni-wuerzburg.de/izvw/texte/2002_buld_krueger_Wirkungen_von_Assistenz_und_Automation.pdf

Cummings, M.L., & Ryan, J. (2014). Point of view: who is in charge? The promises and pitfalls of driverless cars. *TR News*, 292.

Eilers, K., Nachreiner, F., & Hänecke, K. (1986). Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. *Zeitschrift für Arbeitswissenschaft, 40*, 215-224.

Endsley, M.R. (1988a). Design and evaluation for situational awareness enhancement. *Proceedings of the Human Factors Society 32rd Annual Meeting* (pp. 97-101). HFES, Santa Monica.

Endsley, M.R. (1988b). Situation Awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference* (NAECON) (pp. 789-795). New York: IEEE.

Endsley, M.R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors, 37*, 65–84.

Endsley, M.R. (1995b). Torward a Theory of Situation Awareness in Dynamic Systems. *Human Factors, 37*, 32–64.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 1938-1942). Sage CA: Los Angeles, CA: SAGE Publications.

Greenlee, E.T., DeLucia, P.R., & Newton, D.C. (2017). Driver Vigilance in Automated Vehicles: Investigating Hazard Detection Performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1369-1369). Sage CA: Los Angeles, CA: SAGE Publications.

Heikoop, D.D., de Winter, J.C., van Arem, B., & Stanton, N.A. (2017). Effects of platooning on signal-detection performance, workload, and stress: A driving simulator study. *Applied Ergonomics, 60*, 116-127.

Kolbig, M. & Müller, S. (2013). Mode Awareness im Fahrkontext: Eine theoretische Betrachtung. In E. Brandenburg, L. Doria, A. Gross, T. Günzler & H. Smieszek (Hrsg.), 10. *Berliner Werkstatt Mensch-Maschine-Systeme. Grundlagen und Anwendungen der Mensch-Maschine-Interaktion* (pp. 1–8). Berlin: Universitätsverlag der Technischen Universität Berlin.

Lassmann, P., Fischer, M.S., H.-J., Jenke, M., Reichelt, F., Tüzün, G.-J., & Maier, T. (2019). Keeping the balance between overload and underload during partly automated driving: relevant secondary tasks. Submitted to *17. ATZ-Fachtagung.* Springer-Verlag.

Lee, J.D., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50-80.

Lorenz, L. & Hergeth, S. (2015). Einfluss der Nebenaufgabe auf die Überwachungsleistung beim teilautomatisierten Fahren. In *VDIWissensforum GmbH (Hrsg.), 8. VDI Tagung: Der Fahrer im 21. Jahrhundert* (S. 159–172). Düsseldorf: VDI Verlag GmbH.

Othersen, I. (2016). *Vom Fahrer zum Denker und Teilzeitlenker - Einflussfaktoren und Gestaltungsmerkmale nutzerorientierter Interaktionskonzepte für die Überwachungsaufgabe des Fahrers im teilautomatisierten Modus.* Wiesbaden: Springer Fachmedien.

Petermann-Stock, I. (2015). *Automation und Transition im Kraftfahrzeug. Nutzerzentrierte Gestaltung von Übergabe- und Übernahmesituationen innerhalb eines mehrstufigen Automationsansatzes.*

Petermann-Stock, I., Hackenberg, L., Muhr, T. & Mergl, C. (2013). Wie lange braucht der Fahrer? Eine Analyse zu Übernahmezeiten aus verschiedenen Nebentätigkeiten während einer hochautomatisierten Staufahrt. Paper präsentiert auf der *6. Tagung Fahrerassistenz*, November 2013, München.

SAE International (2018). Taxonomy and definitions for terms related to on-road automated motor vehicles (J 3016 Aufl.).

Sarter, N. & Woods, D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors, 37*, 5-19.

Schwalk, M., Kalogerakis, N., & Maier, T. (2015). Driver support by a vibrotactile seat matrix–Recognition, adequacy and workload of tactile patterns in take-over scenarios during automated driving. *Procedia Manufacturing*, *3*, 2466-2473.

Wulf, F., Rimini-Doering, M., Arnon, M., & Gauterin, F. (2012). Approaches of user-centered interaction develop-ment for highly automated vehicles in traffic-jam scenarios. In SAE-China and Fista (Hg.), Proceedings of the *FISTA 2012 World Automotive Congress*. Heidelberg: Springer.

# Workload evaluation of effects of a lane keeping assistance system with physiological and performance measures

*Yu-Jeng Kuo, Corinna Seidler, Bernhard Schick, & Dirk Nissing*
*[1]Kempten University of Applied Science, [2]Rhein-Waal University of Applied Science*
*Germany*

**Abstract**

The present study investigated the mental workload associated with driving a vehicle equipped with Lane Keeping Assistance System (LKAS). Specifically, an experiment was carried out with 16 participants driving with LKAS in four real-world scenarios. Effects on mental workload were evaluated with psychophysiological measures such as heart rate and skin conductance response (SCR). The driving performance, which is also a measure of evaluating mental workload, was assessed by measures such as steering reversal rate, variation of lateral position and steering effort. The result suggested that LKAS has reduced physical workload in the steering task. However, the lane keeping performance was not improved. Moreover, the NASA-TLX showed that participants perceived higher mental workload while driving with LKAS. This effect was mirrored in the SCR. The objective data showed that LKAS was associated with higher steering reversal rate, which might explain the reason of participants perceiving higher mental workload. Overall, it was suggested that the mental workload was higher with the tested LKAS.

**Introduction**

The development of the Advanced Driver Assistance Systems (ADAS) has advanced a lot since the late 90s. From the passive Anti-lock Braking System (ABS) in 1987 (Bosch), to the introduction of Adaptive Cruise Control (ACC) in 1999, various driving tasks in modern cars have been gradually delegated to automated control system (Bengler et al., 2014). Few years after the introduction of ACC, the Lane Keeping Assistance System (LKAS) was introduced by Honda in 2004 (Ishida & Gayko, 2004). In contrast to longitudinal motion managed by ACC, the LKAS is designed specifically for lateral control. The idea behind LKAS is simple: to support staying in a lane. The system constantly measures the distance to the lane marking via one or more camera, and applies steering torque to keep vehicle from leaving the lane.

Discussions regarding the effect of vehicle automation on human can be found widely in the literatures. Stanton and Marsden (1996) stated a number of arguments favouring automation of the driver role, such as the automation could improve well-being, improve road safety, and enhance product sales. Ultimately, it may relieve the driver of excessive and complex driving activities. Brookhuis et al. (2001), however, pointed out that the ADAS may introduce these benefits, but the consequences (e.g. increasing

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

complexities of the cockpit, decreasing alertness and attention from the driving task, and negative effect on skills) should also be identified. Although in the context of Intelligent Vehicle-Highway System, Hancock and Parasuraman (1992) also commented that such assistive function might intend to mitigate the mechanical effort from the driver, yet it could also hypothetically increase driver's cognitive workload if monitoring the system is required.

When a driver assistant system functions as expected, it has been reported in a review paper that the averaged self-reported workload (0% = minimum, 100 = maximum) decreased from 43.5% in manual driving to 38.6% in ACC driving (De Winter et al., 2014). Furthermore, some evidences even suggested that lateral support relieves mental workload to a greater extent than ACC (Young et al. 2002; Carsten et al. 2012). In contrast, if automation does not behave as one anticipates, it could result in increasing driver's mental workload. For instance, Banks and Stanton (2015) showed in a field study that participants reported higher subjective mental workload and lower trust when driving with automated vehicle (with longitudinal, lateral support and auto-overtake system) in comparison to manual driving. The results indicated that the unexpected lane changes and unsafe auto-overtake offerings were possibly part of system's weaknesses.

As argued by Sarter et al., (1997), when a new automation is introduced into a system, new coordination demands between human and machine often come along. Moreover, it is particularly difficult for human to coordinate activities, when the intentions of machine agents are not clear. This observation is similar to the findings in our previous pilot study, in which the participants subjectively reported overall higher workload levels while driving with LKAS than driving without it. The paper concluded that the unexpected system failure, inconsistent feedback and lack of transparency were the main reasons of having this outcome (Schick et al., 2019).

Similar to our pilot study, the primary purpose of this study is also to investigate the mental workload associated with LKAS. However, it differs in two ways. Firstly, only drivers who have had experience with LKAS were selected as participants. The experience with automation, as suggested in Stapel et al., 2017, is a prerequisite of reducing perceived workload. Secondly, the objective data (i.e. driving performance) were presented, which should reveal the driving behaviour when driving with LKAS. In total, four distinct real-world scenarios were designed, which consisted of various curviness of the route and driving velocity. The participants were asked to drive through all scenarios two times (with and without LKAS). The mental workload was assessed with objective and subjective measures.

Based on the work mentioned above, two hypothesis have been formulated:

H1: Drivers' perceived mental workload would be higher with LKAS than without it.

H2: Lane keeping performance would be better when driving with LKAS.

**Experimental Design**

To elicit different levels of workload with LKAS, four driving scenarios were designed based on the combination of cruising velocity and road geometry. The cruising speed was either at 120 km/h (Low-speed, L) or 160 km/h (High-speed, H), whereas the road geometry was either curvy (c) or straight (s). Hence, the four scenarios were abbreviated as Lc (Low-speed-curvy), Ls (Low-speed-straight), Hc (High-speed-curvy) and Hs (High-speed-straight).

The scenario Lc was a 7 km rural road (B19, Waltenhofen – Oberdorf) that consisted of a number of minor curvy sections. The scenario Ls and Hs were each 5 km straight motorway sections (A980, Waltenhofen – Dreieck Allgäu). Essentially, these two scenarios shared the same motorway, but in opposite direction. Finally, the scenario Hc was a 10 km motorway (A7, Dreieck Allgäu – Oy-Mittelberg) which consisted of two high radius curves (each with a radius of approximately one km). The order of the scenarios was predefined, as driving through all scenarios in a randomized order would have taken too much time travelling between each scenario.

For one complete lap, the participant first started with scenario Lc (go and back), followed by a single scenario Ls (go), then through scenario Hc (go and back), and finally finish in scenario Hs (back). Unless explicit speed limit encountered, the driver tried to maintain the speed at 120 km/h in scenario Lc and Ls, and at 160 km/h in scenario Hc and Hs. It took in total about 25 minutes to finish one lap. In order to investigate the effect of LKAS in different scenarios, the participant had to drive through all scenarios two times (laps) i.e. with and without LKAS. The order of introducing LKAS was counterbalanced. The routes are illustrated in Figure 1.

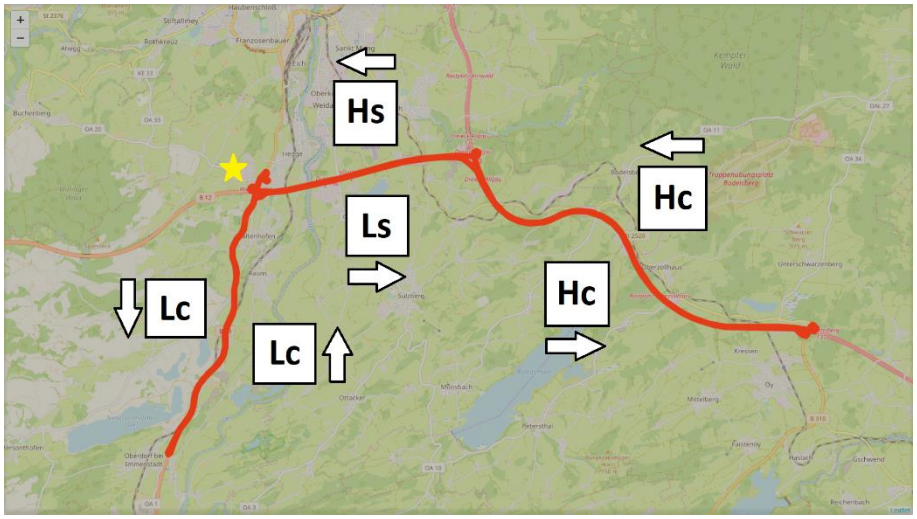



*Figure 13. Four real-world scenarios (Lc: Low-speed-curvy, Ls: Low-speed-straight, Hc: High-speed-curvy, Hs: High-speed-straight). The yellow star indicates the starting and ending of one complete lap. (Figure adapted from openstreetmap.org)*

*Participants*

In total, 21 volunteers between 19 and 65-year of age participated in the experiment (M: 32.6; S.D. 13.5). They had participated in a previous pilot study. All participants possessed a driving licence for at least three years, their self-reported average annual driving mileage was 16333 km (S.D. = 5365 km). The participants signed an informed consent form before taking part in the experiment. Due to adverse weather, traffic conditions and technical issues, the data of five participants were discarded.

*Objective measures*

To assess mental workload, the skin conductance response (SCR) was taken as an indicator of the activity of sweat glands. In the literature, both the SCR and its tonic counterpart (skin conductance level, SCL) have been used for measuring mental workload (Gris et al., 2012; Zangróniz et al., 2017). In this experiment, the count of SCR per kilometre was taken as a workload indicator. In addition, the heart rate (HR) and heart rate variability (HRV) were also taken as dependent variables. It has been shown that HR and HRV are sensitive to evaluate operators' effort (Aasman et al., 1987) and mental workload (De Waard & Brookhuis, 1991; Wilson & Eggemeier, 1991). For HR, these measures in the time-domain were included:

- Inter-beat-interval (IBI)
- Root-mean-square of successive R-R interval differences (RMSSD)
- Standard deviation of N-N intervals (SDNN)
- Percentage of successive R-R intervals that differ by more than 50 ms (pNN50)

To describe driver's performance and behaviour, the standard variation of lateral position (SDLP) and the steering reversal rate (SRR) were used. Before computing SDLP, as suggested by Östlund et al. (2005), the distance-to-line was filtered with second order Butterworth 0.1 Hz high-pass filter to ignore the variation within 10 seconds of observation window. In addition, the data that were 5 seconds before and after any lane-crossing events were excluded. The SRR was defined as the number of times per minute that the direction of steering movement was reversed through a small finite angle (3-degree). Finally, the steering effort was included to quantify the level of physical effort required to perform the steering task. It was calculated as the product of steering angle (degrees) and steering torque (Nm).

The LKAS tested in this study was equipped in a premium class vehicle. The vehicle parameters were assessed with a data acquisition system (DEWE2-A4, Dewetron) installed in the rear trunk. The physiological data were recorded with a wireless wearable system (BioNomadix, BIOPAC Systems Inc.).

*Subjective measures*

The NASA-TLX (Hart & Staveland, 1988) was used to measure subjective mental workload. A 21-point scale was used to map workload level from 0% to 100% for six subscales (mental demand, physical demand, temporal demand, effort, performance and frustration). The result (Raw-TLX, R-TLX) was obtained by averaging the ratings across subscales for four conditions (120/160 km/h x with/without LKAS).

*Protocol*

The experiment was conducted in late April until early May 2018 in the Allgäu region, Germany. Upon arrival, each participant was briefed about the routes and the goal of the study. Starting from the research centre, the driver used the first 8 km to become familiar with the test vehicle before the starting point (the yellow star in Figure 1). One research staff member sat on the passenger seat to operate the measurement devices. After finishing the first lap, the participant parked the car in a parking lot nearby and filled the questionnaire (NASA-TLX) before starting the second lap. The LKAS was then switched either on or off here. For safety reasons, the driver had their hands on the steering wheel all the time. In case of an unexpected system failure occurred, the driver should perform counter steering or any other measures to correct the vehicle's trajectory. It took about one hour for each participant to finish one complete test run (two laps). A summary of the each scenario is listed in Table 1.

*Table 3. Experiment design for one complete lap (Lc-Lc-Ls-Hc-Hc-Hs). Each participant had to drive two laps: with and without LKAS.*

| Scenarios | LKAS | Velocity | Length (km) | Route |
|---|---|---|---|---|
| Lc | with / without | | 7 | Curvy |
| Lc | | 120 km/h | 7 | Curvy |
| Ls | | | 5 | Straight |
| Hc | | | 10 | Curvy |
| Hc | | 160 km/h | 10 | Curvy |
| Hs | | | 5 | Straight |

**Results**

For data analysis, the objective data were submitted to 2 (LKAS: ON, OFF) x 4 (scenarios: Lc, Ls, Hc, Hs) analysis of variance (ANOVA) with repeated measures. Greenhouse-Geisser corrections were applied in case where the data failed to pass Mauchly-Test. Post-hoc test with pairwise comparisons were corrected by Bonferroni corrections. The p-value for significance test was 0.05.

*Physiological measures*

In terms of HR, neither a main nor an interaction effect of LKAS was found. In contrast, a main effect of scenarios was found significant, $F(3, 45) = 7.931$, $p < .005$. The post-hoc pairwise comparison showed that the HR in scenario Ls (75.10 bpm) was significantly lower than the curvy scenarios (Lc = 76.67 bpm, Hc = 76.61 bpm), both $p < .01$. In other words, HR was in general higher in the winding route than on the straight motorway. For other dependent variables, only a significant main effect of scenarios on IBI was found, $F(3, 45) = 7.65$, $p < .005$.

Apart from the effect of LKAS and scenarios, the learning effect between the HR with groups (between-subject effect) and number of trials (within-subject effect) was investigated. The data were submitted to a two-way mixed ANOVA. With respect to HR, a main effect of trial numbers was observed, $F(2.1, 29.4) = 5.35$, $p=.01$. In

contrast, the difference between groups was not significant. This result suggests that the group, which experienced LKAS in the first lap, showed a lower HR in the second lap. In contrast, the HR of another group (without LKAS in the first lap) remained at a similar level in the second lap where LKAS was switched on (Figure 2).

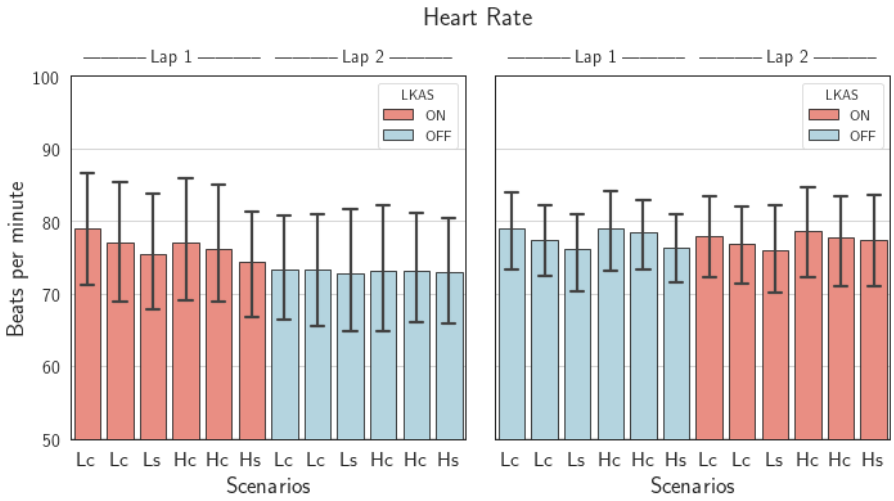



*Figure 14. The heart rate over scenarios in a chronological order (left to right). The group on the left started with LKAS, while the group on the right started without LKAS. (Lc: Low-speed-curvy, Ls: Low-speed-straight, Hc: High-speed-curvy, Hs: High-speed-straight)*

For SCR, a main effect of LKAS on the average count per kilometre was found, $F(1, 15) = 4.62$, $p = .048$. However, no difference was found with scenarios as well as their interactions. The result of average SCR / km over the scenarios is shown in Figure 3.

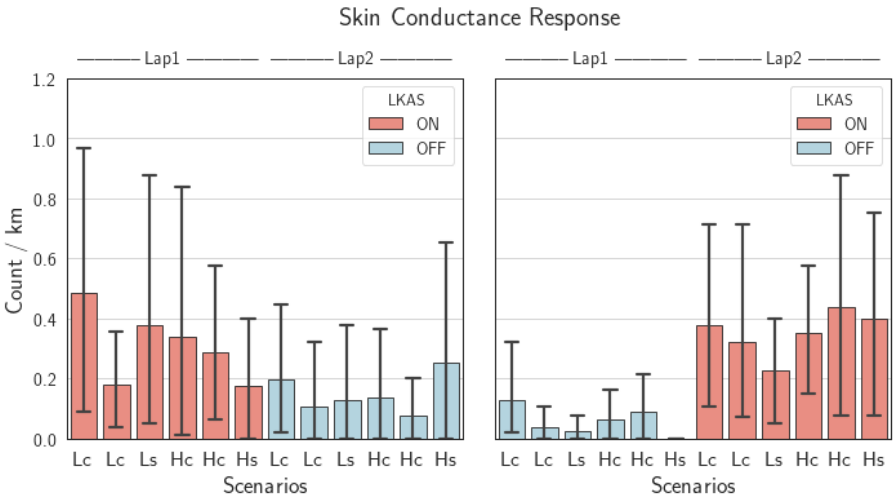



*Figure 15. Averaged SCR / km over all scenarios in a chorological order. (Lc: Low-speed-curvy, Ls: Low-speed-straight, Hc: High-speed-curvy, Hs: High-speed-straight)*

*Performance measures*

The analysis of SRR revealed a significant main effect of LKAS ($F(1, 15) = 55.24$, $p < .005$) as well as of four scenarios ($F(3, 45) = 679.1$, $p < .005$). An interaction effect was also found between LKAS and scenarios ($F(1.7, 26.1) = 9.07$, $p = .002$). A pairwise t-test showed that the SRR was always higher when driving with LKAS in the curvy scenarios (in Lc, $t = 4.767$, $p<.005$; and Hc, $t = 5.475$, $p < .005$), whereas the difference was not significant in straight scenarios (Ls and Hs). On the other hand, the SRR in curvy scenarios (Lc vs. Hc) was significantly different from each other irrespective of LKAS (all $p < .005$), while no difference between straight scenarios (Ls vs. Hs) was found. This result is illustrated in Figure 4a.

In terms of steering effort, ANOVA showed that significant main effects of LKAS ($F(1, 15) = 242.3$; $P < .005$) and scenarios ($F(1.74, 26.1) = 635.6$, $p < .005$) were found. In addition, the interaction effect between two factors ($F(1.45, 21.8) = 165.0$, $p < .005$) was also significant. In contrast to SRR, the steering effort in every scenario was greater when driving without LKAS ($p < .05$), except of scenario Ls. However, when driving with LKAS, no significant difference was found between curvy scenarios (Lc vs. Hc), as well as between straight scenarios (Ls vs. Hs). Overall, the steering effort in the curvy scenarios (Lc and Hc) was significantly greater than straight scenarios (Ls and Hs). This observation is illustrated in Figure 4b.

For SDLP, no difference was found between scenarios, and between LKAS. The post-hoc paired-sample t-test indicated that the SDLP was only significantly higher in scenario Hc ($M = .120$ m) than scenario Lc ($M = .107$ m) when driving without LKAS. The rest comparisons were all not different from each other.

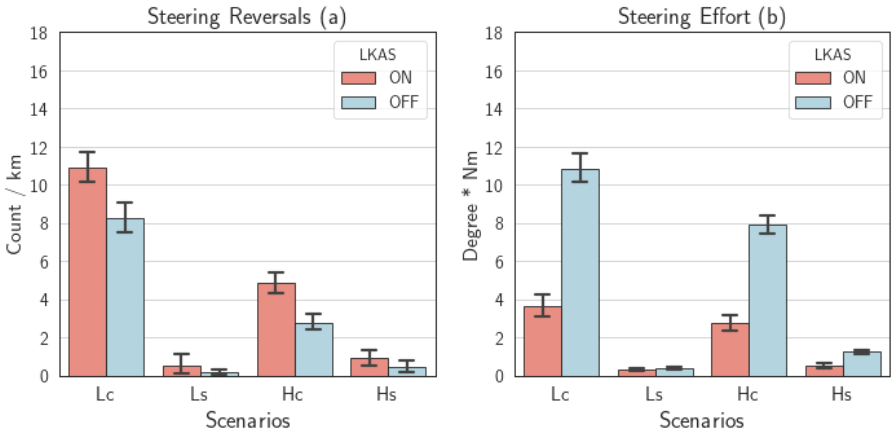



*Figure 16. Driving performance measures. a) Steering Reversals b) Steering Effort (Lc: Low-speed-curvy, Ls: Low-speed-straight, Hc: High-speed-curvy, Hs: High-speed-straight)*

*Subjective measures*

A two-way (LKAS x velocity) ANOVA was performed on the results of R-TLX. It was observed that the LKAS had a main effect on the subjective rating of mental workload ($F(1, 60) = 6.17$, $p = .016$). In contrast, no difference in mental workload was found between two velocity settings. There was also no interaction effect. The result of NASA-TLX is presented in Figure 5.

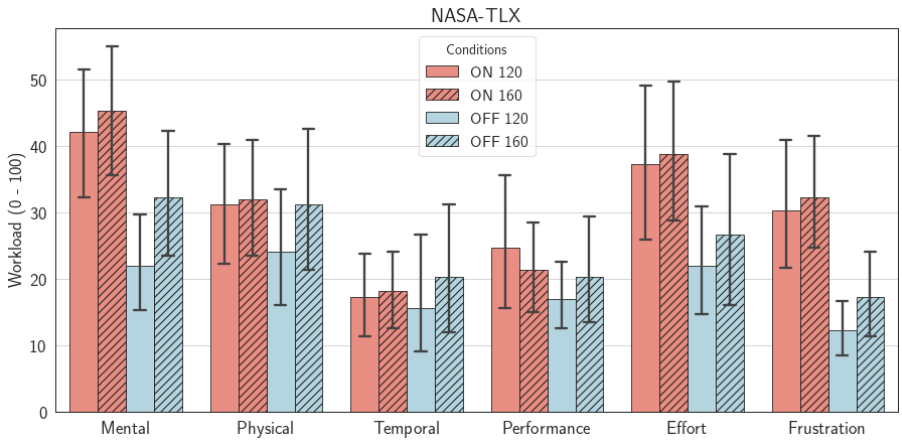



*Figure 17. These subjective ratings were collected after the each lap. Depends the order of introducing LKAS, the participant would answer in each lap for either LKAS ON or OFF, in both velocity conditions (120 and 160 km/h). For analysis, the R-TLX was obtained by averaging the score of subjective workload over six subscales.*

**Discussion**

The result of R-TLX shows that the participants rated their mental workload overall higher when the LKAS was switched on. This result supports H1 that the LKAS increases drivers' perceived workload, which is also in line with our pilot study (Seidler & Schick, 2018). However, the perceived mental workload was not different in low and high-speed scenario. This different result might be due to a small sample size, or driver's experience with automation (Stapel et al., 2017). Out of six subscales from the NASA-TLX, it can be seen in Figure 5 that the difference between two LKAS conditions (ON vs. OFF) was particularly huge in the subscale mental demand, effort and frustration. The reasons may be explained by the objective data.

The analysis of SRR (Figure 4a) reveals that, in the curvy scenarios, drivers performed more counter-steering to correct the trajectory while driving with LKAS. This suggests that constantly correcting LKAS's output might be annoying and disturbing, which results in frustration and mental effort, regardless of the steering effort under the same LKAS setting was actually lower (Figure 4b). This is an interesting result, as reducing physical workload is one of the goals of ADAS (Tanaka et al., 2000). However, this contradictory observation shows that the drivers might prefer applying more steering torque (manual driving) than having an assistance system that reduces physical workload but requires more mental effort.

In contrast to performance measures, the physiological data only partially supports the hypothesis H1 that driving with LKAS induces mental workload. On one hand, the SCR in Figure 3 demonstrats that the LKAS introduced a significant effect on the average SCR/km in different laps. On the other hand, the HR did not show a statistical difference between LKAS conditions (ON vs. OFF). Instead, it is only found that the HR was higher in the curvy scenarios (Lc, Hc) than the straight scenario (Ls). This result is however expected, because the task of keeping a vehicle between lanes depends highly on a psychomotor eye-hand coordination of the driver (De Waard, 1996). Although the observation in Figure 2 could be another evidence that LKAS induces mental workload (as decreasing HR over time was not found in both groups), it is known that higher HR does not necessarily correlate with increasing mental workload, since HR is also sensitive to physical workload e.g. as a result of steering reversals (Jahn et al., 2005).

In terms of driving performance (H2), the objective data reveals that LKAS did not improve the lane keeping performance. Even though the SDLP was significantly higher in the curvy scenario (Hc) than other three scenarios when driving without LKAS, it is difficult to conclude that the driver experienced more mental workload here, since the result of HR did not support this observation. Moreover, it is still an open question whether the measure SDLP could truly reflect mental workload in a field study, despite the fact that data during overtaking and lane changing events were excluded. As Östlund et al. (2005) points out, the width of the route and observation window may heavily influence the reliability of this measure.

Finally, it is realized that certain driving performance measures e.g. SRR and SDLP, though may be helpful interpreting the driving behaviour, are not ideal for examining mental workload associated with LKAS. The argument is that LKAS's performance (whether it applies enough torque or counter steers at the right time) is often associated with curves, in which the lateral position/control is heavily influenced by the system itself. This means that even if the performance measure can truly mirror the variation of mental workload, the interpretation would also not be easy. In this case, subjective measure (NASA-TLX) is a relatively robust way to assess mental workload.

**Conclusion**

Overall, the steering effort has shown that the LKAS has reduced physical workload significantly, particularly in the curvy scenarios. However, the reduced physical effort did not result in a better lane keeping performance as no difference of SDLP was observed. Moreover, the result of NASA-TLX shows that drivers experienced higher mental demand and frustration while interacting with LKAS. This could be explained by the frequent steering reversals required to correct the driving trajectory while driving with LKAS, as shown in the SRR. This increasing physical activity possibly led to an increase in HR, which results in difficulties in assessing changes in mental workload from this psychophysiological measure. However, the difference in count of SCR suggests that the driver could be annoyed or surprised by the LKAS behaviour. Therefore, the results from this study suggests that mental workload is higher when driving with this tested LKAS.

**References**

Aasman, J., Mulder, G., & Mulder, L.J.M. (1987). Operator Effort and the Measurement of Heart-Rate Variability. *Human Factors, 29*, 161–170.

Banks, V.A., & Stanton, N.A. (2016) Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53*, 389-395

Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., & Winner, H. (2014). Three Decades of Driver Assistance Systems: Review and Future Perspectives. *IEEE Intelligent Transportation Systems Magazine, 6*, 6–22.

Brookhuis, K.A., De Waard, D., & Janssen, W.H. (2001). Behavioural impacts of Advanced Driver Assistance Systems - an overview. *European Journal of Transport and Infrastructure Research. 1*. 245–253

Carsten, O.M.J., Lai, F.C.H., Barnard, Y., Jamson, A.H., & Merat, N. (2012) Control Task Substitution in semiautomated driving: Does it matter what aspects are automated? *Human Factors, 54*, 747–761

De Waard, D., & Brookhuis, K.A. (1991). Assessing driver status: A demonstration experiment on the road. *Accident Analysis & Prevention, 23*, 297–307.

De Waard, D. (1996). *The measurement of drivers' mental workload*. PhD thesis, University of Groningen. Groningen, the Netherlands: The Traffic Research Centre.

De Winter, J.C.F., Happee, R., Martens, M.H., & Stanton, N.A. (2014) Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation Research Part F: Traffic Psychology and Behaviour, 27*, 196–217

Hancock, P.A., & Parasuraman, R. (1992). Human factors and safety in the design of intelligent vehicle-highway systems (IVHS). *Journal of Safety Research, 23*, 181–198.

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.

Ishida, S., & Gayko, J.E. (2004) Development, evaluation and introduction of a lane keeping assistance system. *IEEE Intelligent Vehicles Symposium.* IEEE

Jahn, G., Oehme, A., Krems, J.F., & Gelau, C. (2005). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*, 255–275

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., Horst, D., Juch, S., Mattes, S., & Foehl, U. (2005). Driving performance assessment methods and metrics. Deliverable 2.2.5, AIDE Integrated Project, IST-1-507674-IP.

Sarter, N.B., Woods, D.D., & Billings, C.E. (1997) Automation surprises. In G. Salvendy (Eds.), *Handbook of Human Factors & Ergonomics, second edition* (pp. 1926-1943). Wiley, New York

Schick B., Seidler C., Aydogdu S., & Kuo Y.J. (2019) Driving experience vs. mental stress with automated lateral guidance from the customer's point of view. In P. Pfeffer (Eds.) *9th International Munich Chassis Symposium 2018*. (pp. 27-44). Wiesbaden: Springer Verlag

Seidler, C., & Schick, B. (2018). Stress and workload when using the lane keeping assistant: Driving experience with advanced driver assistance systems. *27th Aachen Colloquium - Automobile and Engine Technology*. Aachen. Germany

Stanton, N.A., & Marsden P. (1996) From fly-by-wire to drive-by-wire: Safety implications of automation in vehicles. *Safety Science, 24*, 35-49

Stapel, J., Mullakkal Babu, F.A., & Happee, R. (2017). Driver Behavior and Workload in an On-road Automated Vehicle. In *Proceedings Road Safety & Simulation International Conference 2017*.

Tanaka, J., Ishida, S., Kawagoe H., & Kondo, S. (2002). Workload of using a driver assistance system. *ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.00TH8493).* IEEE

Wilson, G.F., & Eggemeier, F.T. (1991). Physiological measures of workload in multi-task environments. In D. Damos (Eds). *Multiple Task Performance*. (pp. 329-359) London: Taylor & Francis.

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

Zangróniz, R., Martínez-Rodrigo, A., Pastor, J.M., López, M.T., & Fernández-Caballero, A. (2017). Electrodermal Activity Sensor for Classification of Calm/Distress Condition. *Sensors (Basel, Switzerland), 17*, 2324.

# Evaluation of physiological responses due to car sickness with a zero-inflated regression approach

*Rebecca Pham Xuan[1], Adrian Brietzke[1], & Stefanie Marker[2]*
*[1]Volkswagen AG, [2] Technical University Berlin*
*Germany*

**Abstract**

Motion sickness as a reaction to passive movement is a serious issue in various forms of transportation like cars. The goal of the study is to identify physiological changes that can be measured as a response to motion sickness in a real driving environment. The observed features were heart rate, pulse, respiration, skin temperature and electrodermal activity. Forty volunteers were passengers in a car while watching a movie. Meanwhile the car moved in a half-automated stop-&go-scenario, which represented the motion sickness stimulus. A remarkable part of the recorded data had to be neglected due to a high level of signal noise caused by the car environment. The minutely recorded subjective sickness feedback had a zero-inflated poisson distribution. Therefore a zero-inflated regression model was used to identify the relevance of each of the aforementioned features. The model shows that electrodermal activity and pulse were the most relevant features indicating an increase in motion sickness. The observation of physiological parameters in the car environment is a promising method to objectively determine motion sickness.

**Introduction**

The issue of motion sickness (also called kinetosis) has a long history and occurs in all cultures, ages, and genders. Being out of the loop regarding the driving task bears a higher risk of getting motion sickness (Diels, Bos, Hottelart, & Reilhac, 2016). With the ongoing development of fully automatic cars the risk of having more passengers experiencing motion sickness gets more attention. Passengers should be able to enjoy the given opportunities to fill the spare time i.e. with reading in automated cars. Currently, the process of evaluating countermeasures against motion sickness requires the subjective passenger's feedback. The development of countermeasures that ought to reduce motion sickness illustrates the deficiency of objective motion sickness detection. Approaches vary from enhancing situation awareness (Yusof, 2019) to display concepts (Diels & Bos, 2015). In order to evaluate those countermeasures and objectively estimate the passengers' state in terms of motion sickness, more work is needed. The aim is to have objective feedback through physiological measurement in the future. This study provides preparatory work regarding the opportunities coming from the relationship between physiology and self-rated motion sickness.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

A short overview on some relevant research and findings, done so far, is given here. The idea of measuring physiological parameter to obtain objective motion sickness levels is decades old. Thereby only those features will be considered, which can be collected without making the customer (passenger in the car) feel less naturalistic or be restricted in any way (for example due to head-worn tracking systems). Some research groups focused on single items while others looked at multiple physiological features. In the following, some results are described. A rise in heart rate for motion-sick participants was found by several studies; however, some of those changes were only weak and not significant (Yates & Miller, 1996; Yates et al. 1998; Graybiel & Lackner, 1980). A significant change in heart rate could be found in the beginning of the trial by Cowing (1985), whereas Holmes & Griffin (2001) found significant changes when strong nausea occurred. It has been observed that the respiration frequency, as a further physiological feature, rises shortly before and while vomiting due to motion sickness (Yates et al., 1998). Deep breathing can be used to combat motion sickness (Jokerst et al., 1999). Nobel (2010) found that motion sickness leads to a dysfunction in the autonomous thermoregulation. His result supports the findings that body temperature is not a good indicator for motion sickness (Scott, 1988; Graybiel & Lackner, 1980). On the other hand, it could be shown that skin conductance is a robust and reliable predictor for motion sickness. The electrical skin potential rises when motion sickness increases (Crampton, 1955; Bertin, 2005; Meusel, 2014).Yates & Miller (1996) indicate that skin colour could play an important role when detecting motion sickness using physiological data, since pallor changes with sickness and is seldom a response to other stressors. Since pallor seems to proceed the onset of nausea (Crampton, 1955), it has a high potential of being a good indicator of motion sickness (Scott, 1988; Holmes et al., 2002).

In short, some features show potential, but most features are not cause-specific: the change of a single feature cannot be traced back to motion sickness with certainty. Therefore, finding a pattern of multiple physiological changes is required. Such a pattern could detect or predict motion sickness more robustly and would not be as vulnerable to unexpected physiological behaviour of the individual. The aim of this investigation was to develop an objective rating method allowing the evaluation of countermeasures without using self-rated indicators by the help of multiple features. An approach in a real driving scenario is presented along with first results.

## Method

### *Ethical Approval*

Participants read and signed an informed consent prior to participation. Any participants with one of the following conditions were excluded from the trial: cardiovascular weakness, hypertension, hypotension, epilepsy, balance disorder, pregnancy, other health impairments or of age younger than 18 years. For a conducted trial the participants received a voucher (value €20) as compensation regardless of the trial duration. All participants reported normal or corrected-to-normal visual acuity. The experiment was approved by the Ethics Committee of the Brandenburg University of Technology Cottbus-Senftenberg. To prevent participants from harm, those with a high risk of getting severe motion sickness (high susceptibility) were

excluded from the trials. The derivation of the participants' susceptibility is explained in Table 1.

*Participants*

Forty volunteers (20 women, mean age = 37.9 years, SD = 11.4 ranging from 21-57 years) which were employees of the Volkswagen AG participated. They are not involved in motion sickness research and participated during their private time. The recruiting process contained an assessment of the participants' susceptibility. By using the Motion Sickness Susceptibility Questionnaire – Short (MSSQ) (Golding, 1998) susceptible (n = 23) and non-susceptible (n = 17) participants for the trial were chosen. Therefore categories were defined using the MSSQ-Score (final score) and the item regarding the experienced motion sickness over the last 10 years in cars (interim score). The categorization can be found in Table 1.

*Table 1. Susceptibility Categories*

| Category | MSSQ-Score | Interim Score | Accounted as |
|---|---|---|---|
| A | Final score = 0 | 0 or 1 | Non-Susceptible |
| B | Final score > 0 | 1 | |
| C | Final score > 6 and < 11 | 2 or 3 | Susceptible |
| D | Finale score >11 | 2 or 3 | |
| E | Final score > 20 | 4 | highly susceptible |

Interim code: never felt sick '1', rarely felt sick '2', sometimes felt sick '3', frequently felt sick '4'

*Materials and Set up*

The motion sickness stimulation during the trial was a stop-&go-scenario.Two cars drove behind each other and the participant sat in the rear car in the front passenger seat. A vehicle acceleration profile was created before the trial and replayed for the vehicle in front, while the participants' car followed with adaptive cruise control. This should assure a constant motion sickness provocation in all trials for all participants. A trained security driver was in the driver seat and the experimenter in the rear passenger seat.

The lead car was a VW Passat, while the rear car was an Audi A8 D5. During the experiment the participant had a display (Nanovision MIMO UM-1010S, 10.1" USB Multi-Touchscreen Display) fixed to the leg. They were asked to give feedback every minute about their motion sickness status on a seven-point scale ranging from zero - "no symptoms" to six - "unbearable" in German language. The scale, illustrated in Figure 1, was located on the bottom of the touch screen and the participants gave feedback by tapping on the screen.

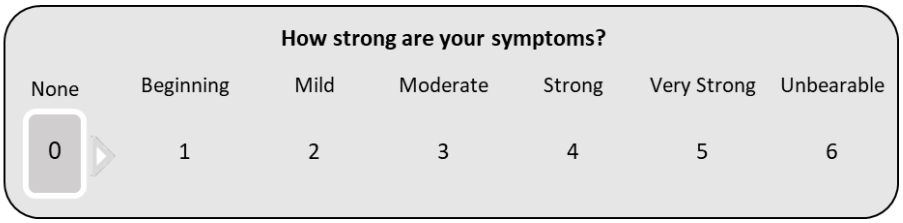



*Figure 1. Illustration of the Questionnaire*

Kinetosis appears more often if passengers are involved in tasks in which their eyes are off the street. Therefore participants were instructed to keep their eyes on the monitor during the whole drive. To ensure that participants would be watching the monitor, they had to count either jelly fish or clown fish (randomized over the trials) in a coral reefs film. The film was screened throughout the entire time on the upper part of the display above the questionnaire.

The study was conducted in November and December 2018. All participants were able to get acclimated for several minutes after getting into the car, coming from the cold temperatures outside (approximately 5°C). The car temperature was set to constant 23°C, which is supposed to be the optimal temperature for measuring electrodermal activity (Boucsein, 2012).

*Procedure*

Each participant completed two trials to increase reliability of the data which were organized on different days. After giving informed consent, participants were seated in the car. During the time given for acclimation, the sensors were attached to the participants. The first part of the experiment was a seven-minute session in the standing car, therein the recorded data was used to create a baseline. The baseline measures were followed by the actual trial, where participants would experience the stop-&go-driving scenario for a maximum of 20 minutes or until an abort criterion was reached. During both sections, the baseline and the drive, participants had the visual counting task. There was always only one participant at a time. After the trial, the vouchers were handed over,participants were provided refreshment and asked to stay at the location until the symptoms fully disappeared.

*Physiological measurements*

The physiological data acquisition was carried out by the use of a ProComp Infinity encoder with ProComp Infinity Sensors and recording from the BioGraph Infinity Software (Thought Technology Ltd, 2019). Electrocardiac activity (ECG) and blood volume pulse (BVP) were recorded at 2048 Hz. Electrodermal activity (EDA), temperature and respiration were measured at a sampling rate of 256 Hz. The respiration sensor was placed in a stretch belt and placed around the chest. Skin conductivity was measured by placing sensors on the pointer and ring finger of the non-dominant hand, while ECG was recorded using wrist straps. The BVP sensor as well as the temperature sensor were placed on the middle finger of the non-dominant hand. Furthermore, a second measurement of temperature and pulse was derived from

the inner ear by using the device Cosinuss° One (Cosinuss°, 2019). The accuracy from the temperature in the inner ear is a constant offset to the body core temperature but dynamic changes can be recorded precise enough for most medical applications. Pulse oximetry in the external auditory canal is comparable to pulse oximetry on the finger, while it is more robust towards motion artefacts. (Kreuzer, 2009) The approach in measuring the features twice was realized to improve overall data quality. The dynamic environment could cause a low signal-to-noise-ratio which therefore requires a backup system.

*Data Analysis*

On average the time series of the 70 trials per physiological parameter containing over 2100 observations in total were used for the data analysis. Each of the parameters was statistically and visually screened for outliers and noise. Initial analysis for the heart rate signal included cascading high- and lowpass filtering, afterwards QRS complexes were detected using wavelet analysis. Downsampling processes were done for the blood volume pulse on the finger (finger pulse) as well as the temperature data. The finger pulse and the respiration signal were waveform data, wherein a peak was considered a beat or a breath respectively. Electrodermal activity was divided into tonic and phasic movement with a 0.5 Hz highpass filter. From the phasic component skin conductivity reactions (SCR) were extracted. SCRs were identified as responses with an amplitude of SCRs/min $\geq 0.03$ µS. Rejection rate was set to 10 %, meaning that amplitudes SCRs/min $< 0.003$ µS were rejected. Almost all of the features were normalized using the baseline measurements and were averaged per minute. Only the SCRs were not normalized, since its appearance itself is an indicator for motion sickness (Golding, 1992).
Since several physiological features were derived from the participants and a human body rarely shows any independent physiological changes, it has to be assured that no information is used in the analysis multiple times (multicollinearity). Multicollinearity describes the case when information is redundant in a set of variables and the redundancy gets apparent in a combination of several variables. Physiological reactions of humans are mostly dependent which increases the chance of multicollinearity in the data. To test that no harmful multicollineariy was present firstly pairwise correlation was calculated. Before calculating the correlations, the features need to be centred and scaled which led to lead to a mean of zero and standard deviation of one for all of the features. If the pairwise correlation shows high coefficents (Pearson's $r < 0.7$) this is considered as indicator for severe multicollinearity. In addition, the variance inflation factor (VIF) was calculated. The VIF is a predictor of whether variables have a strong relationship to one or more variables. The calculation of VIF was necessary since multicollinearity can also appear, if pairwise correlations are low. A conservative threshold indicating harmful multicollinearity is VIF = 4 (Slinker& Glantz, 1985).
In accordance to the rating distribution, a zeroinflated poisson regression model was computed. For the model all ratings of 4 were transformed to 3, because the amount of reported 4s was too little. Furthermore, only cases where recording of all features was successful, could be considered, resulting in 895 observations. The model consists of two separate processes: one considers the count part of the model. The count model examined how ratings evolve, if the participant experiences motion

sickness at some point (susceptible to the provocation). The second process contains a logistic regression considers those participants which are unsceptible to the stop&go scenario and reported only zeroes. The results of the zero-inflation model coefficients, shows the odds of reporting no motion sickness symptoms (Atkins, 2007). To verify the model, the combined probability of no symptoms (Rating = 0) were calculated and compared to the actual appearance of no symptoms.
A 5 % significance level was accepted in all tests, data analysis and statistical calculation were carried out using Matlab 2016b and R 3.5.3.

**Results**

Correlation between Blood Volume Pulse on the finger (finger pulse) and inner ear was high (r = 0.78), therefore the ear pulse was not further used. Furthermore, skin temperature was not used, since the measurement showed high fluctuation which has most likely been caused by the airconditioning fan of the car, instead, the temperature derived from the inner ear was used. The measurement of the heart rate showed a low signal-to-noise-ratio, possibly due to the unsteady environment of the movement and electrical components in the car led to many artefacts. Therefore the heart rate was excluded from further analysis. The remaining features were the finger pulse, inner ear temperature, respiration rate and skin conductivity components (tonic and SCRs). Table 2 lists the features wherein all but the SCR-Peak were normalized by the baseline (substraction of baseline mean from each data point).

*Table 2. Normalized features used along variance inflation factor*

| **Measurement** | **Derived Feature** | | **Mean** | **SD** | **VIF** |
|---|---|---|---|---|---|
| Blood Volume Pulse | Peak | [Counts per minute] | 0.83 | 3.79 | 1.02 |
| Temperature | Mean Temperature | [K] | 0.58 | 0.88 | 1.06 |
| Respiration | Peak | [Counts per minute] | 0.35 | 2.97 | 1.06 |
| Skin Conductivity | Mean Tonic Level | [µS] | 0.26 | 0.48 | 1.27 |
| | SCR-Peak | [Counts per minute] | 2.22 | 1.79 | 1.37 |

The listed measurements in Table 2. were used for the further analysis and have pairwise correlations r < 0.7. Each of the features has a VIF < 4, therefore none of them indicated harmful multicollinearity. For those remaining components, the correlations to the ratings are plotted in Figure 2. The displayed boxplots bring out the partial relationships between the dependent variable and the indented regressors. The negative SCR-correlations appeared, when particpiants showed a relieve in symptoms but SCRs still occurred. In the tonic part of the EDA a positive tendency can be observed. The correlation of respiration to rating has a negative tendency, while the residual parameters BVP and temperature have mainly positive correlations.

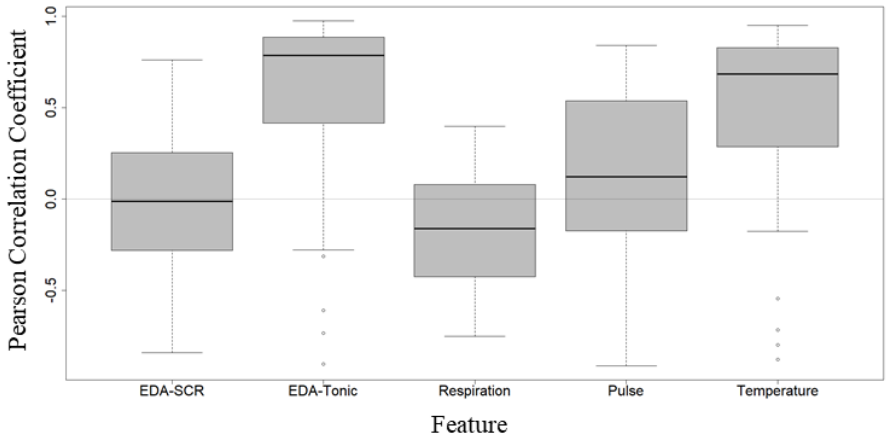



*Figure 2. Boxplot of correlations of physiological features to sickness rating*

The given ratings plotted in a histogram (Figure 3) indicate that the distribution is not Gaussian, but tends to a Poisson distribution which was also found by Reason (1967) for a motion sickness rating. In total 1293 ratings were given during the provocation wherein 718 were '0 – no symptoms' and 21 ratings were '4 – strong symptoms'. Testing a zeroinflation with the Score-Test from van den Broek (1995) reveals that the data have a zeroinflation ($\chi^2$ = 195,99, df = 1, p < .001).

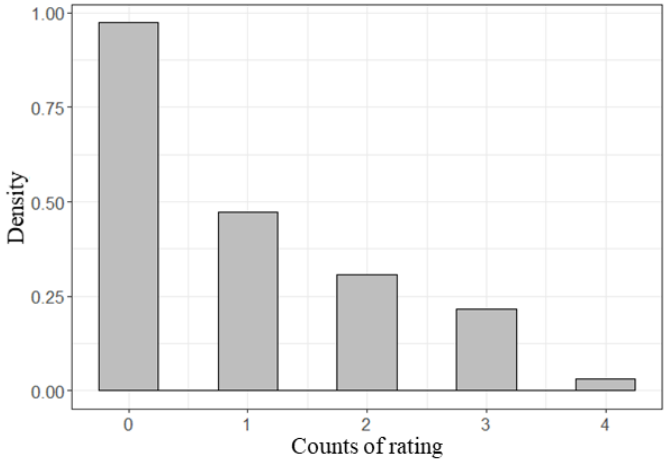



*Figure 3. Histogram of the rating during the drivings*

The reported mean sickness development over all subjects are plotted in Figure 4. The '+' at rating 4 represents the break-off criterion of which 19 occured in total due to subjects reporting the level of 4. Two times participants reported a motion sickness level of 4 very early which was assumed a mistake until the rating was repeated. When a rating of 4 occurred, the remaining minutes were filled with 4 to enable the plot.

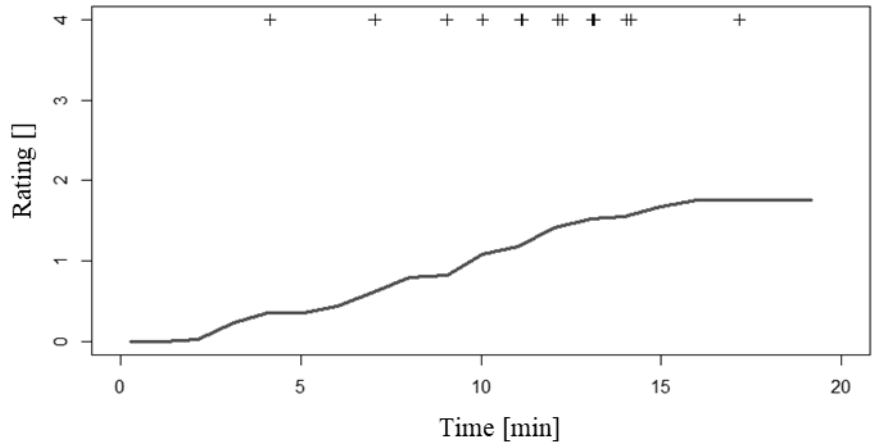



*Figure 4. Mean reported motion sickness development*

The resulting model is shown in Table 3. The unprocessed results of the modelling lead to numbers which are calculated using log link. Therefore the estimated slopes (Est) of the coefficients are on a log scale and shown along with their exponentiated values (Exp(Est)) to ease interpretation. The estimate of any coefficient in the count model describes how the rating changes if the respective coefficient changes one unit. Generally one outcome of the log link function is a non-linear relationship of the predictor variables with the result (Beaujean & Morgan,2016). The percentage of the rating change can be calculated using Equation (1).

$$\text{Percentage of Rate-Change} = 100\text{x}[\exp(b_0)\ \text{x}\ \exp(b_1\ \text{x}\ \Delta\text{EDA, SCR})\ \text{x}\ \exp(b_2\ \text{x}\ \Delta\text{EDA, tonic})\ \exp(b_3\ \text{x}\ \Delta\text{Respiration})\ \text{x}\ \exp(b_4\ \text{x}\ \Delta\text{Temperature})\ \text{x}\ \exp(b_5\ \text{x}\ \Delta\text{BVP})] \quad (1)$$

Therein $b_0$ represents the intercept while $b_{1\text{-}5}$ are the regression coefficients and $\Delta$ are the changes in the respective predictors. The distance of the result to 1 can be interpreted as the increase or decrease of the percentage (Atkins et al., 2013). For a better understanding the influence of Blood Volume Pulse change shall be described as an example. The residual parameters are kept at an average level (as measured during the baseline condition). The coefficient calculated by the model for the influence on rating due to the change of BVP is 1.37 (exp(Est)), the BVP changes in the range of its standard deviation (1 unit, since the data were centred and scaled). The influence on rating can be calculated using Equation (1):

$$\text{Percentage of Rate-Change} = 100\ \text{x}\ \exp(0.01)\ \text{x}\ \exp(0.31*1) = 137.94$$

Meaning that there is approximately 38% of increase in motion sickness rating, when the BVP changes one peak/minute.

*Table 3. Zeroinflated poisson model*

| Coefficient | Est | SE | exp(Est) | z-Value | Exp(95% CI) Lower | Exp(95% CI) Upper | p | |
|---|---|---|---|---|---|---|---|---|
| Count model coefficients | | | | | | | | |
| Intercept | 0.01 | 0.06 | 1.01 | 0.19 | 0.91 | 1.12 | 0.85 | |
| EDA, SCR | -0.12 | 0.05 | 0.89 | -2.19 | 0.80 | 0.99 | 0.03 | * |
| EDA, tonic | -0.00 | 0.06 | 1 | -0.05 | 0.89 | 1.12 | 0.96 | |
| Respiration | -0.04 | 0.03 | 0.96 | -1.21 | 0.9 | 1.03 | 0.23 | |
| Temperature | -0.30 | 0.07 | 0.74 | -4.60 | 0.68 | 0.80 | <0.001 | *** |
| BVP | 0.31 | 0.04 | 1.37 | 8.41 | 1.28 | 1.45 | <0.001 | *** |
| Zero-inflation model coefficients | | | | | | | | |
| Intercept | -2.42 | 0.48 | 0.09 | -5.07 | 0.02 | 0.24 | <0.001 | *** |
| EDA, SCR | 0.71 | 0.19 | 2.04 | 3.82 | 1.42 | 2.94 | <0.001 | *** |
| EDA, tonic | -3.68 | 0.60 | 0.03 | -6.17 | 0.01 | 0.08 | <0.001 | *** |
| Respiration | -0.33 | 0.15 | 0.72 | -2.17 | 0.51 | 0.95 | 0.03 | * |
| Temperature | -2.66 | 0.62 | 0.07 | -4.27 | 0.01 | 0.28 | <0.001 | *** |
| BVP | 0.33 | 0.17 | 1.39 | 1.93 | 0.97 | 2.06 | 0.05 | |

Note. Est: Unstandardized coefficient (log link), SE: Standard error, exp(Est): 95% CI confidence interval: Exponentiated regression coefficient. Log Likelihood: -978.5 (df = 12)

The number of correctly and incorrectly predicted observations can be found in the confusion matrix (Table 4) along with the derived sensitivity (proportion of positive cases correctly predicted).

*Table 4. Confusion matrix*

| **Observed** / **Predicted** | **0** | **1** | **2** | **3/4** | **Total** |
|---|---|---|---|---|---|
| 0 | 460 | 140 | 91 | 38 | 729 (81.45%) |
| 1 | 57 | 30 | 28 | 38 | 153 (17.10%) |
| 2 | 0 | 2 | 7 | 3 | 12 (1.34%) |
| 3 | 0 | 0 | 0 | 1 | 1 (0.11%) |
| 4 | 0 | 0 | 0 | 0 | 0 |
| Total | 517 (57.77%) | 172 (19.22%) | 126 (14.08%) | 80 (8.94%) | 895 |
| Sensitivity | 88.97% | 17.44% | 5.55% | 1.25% | |

## Discussion

The conditions of the real driving experiment introduced confounding factors that cause notable noise as influence, which negatively affects the signals (i.e. temperature, influence of sun, driving conditions, car movements or technical artifacts). These factors as well as internal biological variations have an impact on the variance of the data (Scholz, 2006) and prevent a full use of all of the measurements.

The rating data are derived from susceptible and non-susceptible participants. The non-susceptible participants are a source contributing only zeros to the rating, therefore the distribution of the rating results in a zeroinflation. The use of an zero-inflated model is therefore appropriate. Each observed feature of the model is within the confidence interval. The models' overall validity is therefore considered to be given. Interpretation of the slopes in Poisson models (which become multiplicative models) has to be done very carefully, it is described in more detail by Atkins et al. (2013). Generally, it is shown that Skin Conductivity Responses, temperature and Blood Volume Pulse have a significant explanation range regarding the rating. The negative relationship between sweat (SCR) and motion sickness is surprising. A calculation according to Equation (1) results in a decrease of the rating when the SCR rises 1 unit. It was expected that sweat activity rises along with a development of motion sickness. The findings, as in several studies, of a higher amount of perspiration as one of the characteristics of a motion-sick group compared to a non-motion-sick group, could not be found here (Crampton, 1955; Scott, 1988; Golding, 1992; Bertin et al., 2005). Temperatures seems to reduce as motion sickness rises. An increasement of temperature in a thermoneutral environment was also described by Nobel (2010). Contrarily in preceding studies temperature was behaving variable (Jarvis & Uyede, 1985) or did not change significantly (Drylie, 1987). The significant effect found is therefore surprising. Further the model indicates that a rise of BVP leads to a rise of a motion sickness rating. This is in agreement with findings from literature (Crampton, 1955; Dahlman, 2009). The output of the model dealing with the zeros would be interesting regarding the onset of motion sickness symptoms. This would require, the threshold of 0 - "no symptoms" to 1 - "beginning" symptoms was similar understood by all of the participants. Correct categorization of the participants' motion sickness into the scale was assumed but due to subjective judgement it cannot be assured, especially when "beginning" symptoms were reported.

## Conclusion

The presented study examined the relationship between physiological data and reported motion sickness. Participants were situated in a stop&go-scenario, while being involved in a non-driving related task, which caused them to have their eyes off the street. This scenario was sufficient to provoke motion sickness over time: Out of the 40 participants 7 participants had severe motion sickness, while 27 participants had at least mild or a higher degree of motion sickness symptoms. The recordings were done in a real-driving scenario, where the challenge of transfering and reproducing results from laboratory environments in real-driving experiments became apparent. Physiological features were used to perform regression analysis in order to analyse the associations between a reported motion sickness level and physiological reactions. The distribution of the rating led to a zero-inflated poisson model.

The generated model revealed that sweat (SCR), temperature and Blood Volume Pulse changes significantly with the rise of motion sickness. Reversely these results indicate that sweat and blood volume pulse are good indicators for motion sickness. The model had a good sensitivity considering the predicition of 'no symptoms' (~89%). Ratings indicating the appearance of motion sickness (Rating > 0) were in average predicted lower than the observation. The significant features along with narrow confidence intervals substantiate that motion sickness expresses itself in physiological changes, which can be recorded during a real-driving scenario. This is considered a promising basis when continuing the work towards objective detection of motion sickness.

**Future Strategy**

The model can be adjusted in two possible ways. One will be to change the general model. The zero-inflated poisson model considers the rating as an numeric value, while the numbers 0-4 represent the categories of having "no symptoms" to "strong symptoms". Therefore an zeroinflated ordered probit regression model will be calculated, which does not assume the numbers 0-4 to be equidistant but still represents an ordered scale. Alternatively, a binary model will be computed, wherein ratings of 0 and 1 are grouped as "no symptoms" and ratings greater than 1 as "symptoms present". This will allow to overcome the uncertainty of the onset of reported motion sickness. Comparison of the models will allow to choose the best fit.

After choosing the best model the independent variables could be varied. According to literature motion sickness is influenced by several factors, for example personality, sex, age, exposed time to stimulus (Brietzke et al., 2017; Dahlman, 2009) or theoretically derived susceptibility via a questionnaire (MSSQ by Golding, 1998). Therefore including such parameters into the model should influence the outcome and informative value of any model. It is expected, in example, that the results from a model including data of self-assessed susceptible passengers are more precise in the outcome. The adjustments should confirm if the grouping factors significantly influence of the participants' rating of motion sickness. This allows conclusions, whether the model can be built more accurately if certain groups are considered. Practically this includes assertions on how motion sickness is connected to physiology in people with a certain profile and which indicators are important. Adjusting the models towards the actual susceptibility (i.e. choosing people with a rating higher than 2 – "mild symptoms") would probably lead to the most reliable results. By taking the temporal development of the physiology with regard to the onset of motion sickness into account it could be feasible to recognize motion sickness even before the passenger is totally aware of it. In an additional step, it will be tested to what extent the accuracy of prediction can be enhanced using the aforementioned factors. In general the research question regarding the potential of objective motion sickness detection in cars is currently referred based on literature that mostly addresses the laboratory context. These results need to be proven relevant and feasible for implementation and application in the car. The presented work is one method towards transporting laboratory findings into the car. The approach of using multiple features in a mathematical model will leed to helpful results in the progress of evaluating the importance of physiology for objective detection of motion sickness in cars.

**References**

Atkins, D.C., & Gallop, R.J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*, 726-735.

Atkins, D.C., Baldwin, S.A., Zheng, C., Gallop, R.J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors, 27*, 166-177.

Beaujean, A.A., & Morgan, G.B. (2016). Tutorial on Using Regression Models with Count Outcomes using R. *Practical Assessment, Research & Evaluation, 21(2),*1-19.

Bertin, R.J.V., Collet, C., Espié, S., & Graf, W. (2005). Objective measurement of simulator sickness and the role of visual-vestibular conflict situations. In *Driving Simulation Conference North America* (pp. 280-293). Orlando, USA. University of Central Florida

Braithwaite, J.J., Watson, D.G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology, 49*, 1017-1034.

Brietzke, A., Klamroth, A., Dettmann, A., Bullinger, A.C. (2017). Motion sickness in cars: Influencing human factors as an outlook towards highly automated driving. Poster presented at Human Factors and Ergonomics Society Europe Chapter Annual Conference, Rome. Available from https://www.hfes-europe.org/wp-content/uploads/2017/10/Brietzke2017poster.pdf

Boucsein, W. (2012). *Electrodermal activity*. New York, USA. Springer Science & Business Media.

Chen, C.-L., Li, P-C., Chuang, C-C., Lung, C-W., & Tang, J-S. (2016). Comparison of motion sickness-induced cardiorespiratory responses between susceptible and non-susceptible subjects and the factors associated with symptom severity. In 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (pp. 216–221). New Jersey, USA. Institute of Electrical and Electronics Engineers.

Cosinuss° (2019). Retrieved November 01, 2019, from: https://www.cosinuss.com/products/one/

Cowings, P.S., Suter, S., Toscano, W.B., Kamiya, J., & Naifeh, K. (1986). General autonomic components of motion sickness. *Psychophysiology, 23*, 542-551.

Crampton, G.H. (1955). Studies of motion sickness: XVII. Physiological changes accompanying sickness in man. *Journal of Applied Physiology, 7*, 501-507.

Dahlman, J. (2009). *Psychophysiological and performance aspects on motion sickness*. PhD thesis, Linköping University. Linköping, Sweden. Department of Clinical and Experimental Medicine.

Diels, C., & Bos, J.E. (2015). Design guidelines to minimise self-driving carsickness. In *7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. New York, NY, USA. Association for Computing Machinery

Diels, C., Bos, J.E., Hottelart, K., & Reilhac, P. (2016). Motion sickness in automated vehicles: the elephant in the room. In *Road Vehicle Automation 3*, (pp. 121-129). Cham, Springer

Drylie, M.E. (1987). An Analysis of Physiological Data Related to Motion Sickness for Use in a Real-Time Motion Sickness Indicator (No. AFIT/GE/ENG/87D-16). Masterthesis, Air University, Ohio, USA. Air Force Institute of Technology Wrigth-Patterson Air Force Base.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Great Britain. Sage publications.

Golding, J. (1992). Phasic skin conductance activity and motion sickness. *Aviation, space, and environmental medicine, 63*, 165-171.

Golding, J.F. (1998). Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Research Bulletin, 47*, pp. 507-516.

Graybiel, A., & Lackner, J.R. (1980). Evaluation of the relationship between motion sickness symptomatology and blood pressure, heart rate, and body temperature. *Aviation, space, and environmental medicine, 51*, 211-214.

Holmes, S.R., & Griffin, M.J. (2001). Correlation between heart rate and the severity of motion sickness caused by optokinetic stimulation. *Journal of Psychophysiology, 15*, 35-42.

Jarvis, N.R., & Uyeda Jr, C.T. (1985). *An Analysis of Potential Predictive Parameters of Motion Sickness Using a Computerized Biophysical Data Acquisition System* (No. AFIT/GSO/ENG/85D-1). Masterthesis, Air University, Ohio, USA. Air Force Institute of Technology Wrigth-Patterson Air Force Base.

Jokerst, M.D., Gatto, M., Fazio, R., Stern, R.M., & Koch, K.L. (1999). Slow deep breathing prevents the development of tachygastria and symptoms of motion sickness. *Aviation, space, and environmental medicine, 70*, 1189–1192.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement, 20*, 141-151.

Kreuzer, J. (2009). *Alltagstaugliche Sensorik: Kontinuierliches Monitoring von Körperkerntemperatur und Sauerstoffsättigung*. PhD thesis, Technische Universität München, Germany. Fakultät für Ekeltrotechnik und Informationstechnik.

Kushki, A., Fairley, J., Merja, S., King, G., & Chau, T. (2011). Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological measurement, 32*, 1529-1539.

Meusel, C.R. (2014) *Exploring mental effort and nausea via eda within scenario-based tasks.* Master thesis, Iowa State University, Iowa

Nobel, G. (2010). *Effects of Motion Sickness on Human Thermoregulatory Mechanisms*. PhD thesis, Royal Institute of Technology. Stockholm, Sweden. Department of Environmental Physiology.

Ohsuga, M., Kamakura, Y., Inoue, Y., Noguchi, Y., Shimada, K., & Mishiro, M. (2011). Estimation of driver's arousal state using multi-dimensional physiological indices. In *International Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 176-185). Berlin, Germany: Springer.

Reason, J.T. (1967). *Relationships between motion after-effects, motion sickness susceptibility and "receptivity"*. PhD thesis, University of Leicester. Leicester, United Kingdom.

Scholz, M. (2006). *Approaches to analyse and interpret biological profile data*. PhD thesis, Universität Potsdam. Potsdam, Germany.

Scott, M.F. (1988). *A study of motion sickness: mathematical modeling and data analysis* (No. AFIT/GEO/ENG/88D-4). Masterthesis, Air University, Ohio, USA. Air Force Institute of Technology Wrigth-Patterson Air Force Base.

Slinker, B.K., & Glantz, S.A. (1985). Multiple regression for physiological data analysis: the problem of multicollinearity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 249*, R1-R12.

Stevens, J.P. (2012). *Applied multivariate statistics for the social sciences*. New York, USA. Routledge.

Thought Technology Ltd (2019). Retrieved November 01, 2019, from: http://thoughttechnology.com/index.php/procomp-infiniti-333.html

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics, 51*, 738-743.

Yates, B.J., Miller, A.D.(Eds.). (1996). *Vestibular autonomic regulation*. CRC Press.

Yates, B.J., Miller, A.D., & Lucot, J.B. (1998). Physiological basis and pharmacology of motion sickness: an update. *Brain research bulletin, 47*, 395-406.

Yusof, N.B.M. (2019). *Comfort in autonomous car: mitigating motion sickness by enhancing situation awareness through haptic displays*. PhD thesis, Technische Universiteit Eindhoven. Eindhoven, The Netherlands. Department of Industrial Design.

# Interpersonal trust to enhance cyber crisis management

*Florent Bollon[1,2], Anne-Lise Marchand[3], Nicolas Maille[1], Colin Blättler[3], Laurent Chaudron[4], & Jean-Marc Salotti[2]*
*[1]ONERA (French Aerospace Lab), [2]IMS UMR CNRS 5218,*
*[3]CREA (French Air Force Academy Research Centre),*
*[4]Theorik-Lab*
*France*

**Abstract**

In the field of cyber-security, software performance optimization is a major focus of research to better prevent cyber threats. However, once threats are detected, they have to be managed by a human operator or more often by human operators' joint actions. The purpose of this study is to show that in these collaborative situations, the interpersonal trust level between these actors shapes their handling of the threat. Forty-five participants performed, with twenty-eight different fictive teammates, a collaborative counting task that included aleatory phases of jamming. Each fictive teammate was described through two adjectives selected to induce a predefined level of interpersonal trust (low or high). The subject and his collaborator worked on different systems with different objects to count and different jamming phases. Nevertheless, each participant had the possibility of supervising his teammate's work by checking out his task and modifying his answers (number of targets and jamming events reported) if required. The subject was responsible for validating the team's final result. The experimental data show that, in this type of collaborative task, the interpersonal trust level has indeed an influence on the supervision strategy used and the team performance.

**Introduction**

In order to prevent the increase in the number of cyber-attacks, States are setting up cyber operations centers (C2Cyb). The operators of these C2Cybs, who monitor the state of systems and the information flows, are collectively responsible for detecting, correlating and analyzing the various indicators that can *make sense* of a cyber crisis (Boin, Busuioc, & Groenleer, 2014). These indicators, which are difficult to perceive but that predict perturbations in the system, are called *weak signals* (Saritas & Smith, 2011) and are discrete, ephemeral, distributed and difficult to interpret.

In a complex and highly interconnected cyberspace, the collection, detection, analysis and comprehension of *weak signals* requires aggregating information from various actors, both human and material, engaged in monitoring the global system. The amount and complexity of the information available in cyberspace makes it impossible for a single operator to compile all the information in a limited amount of time. The heterogeneous nature of the signals also increases the uncertainty of operators about how to interpret them. As a result, decisions made by the C2Cyb team

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

leader are based on information that is usually unverifiable and transmitted by his/her teammate. This information can sometimes contradict the leader's information. A question therefore arises: how does the team leader in C2Cyb consider this contradictory information when making decisions in a situation of uncertainty?

The decision-making strategies studied in psychology and economics are sometimes based on theories that adopt probabilistic visions. In particular, the dual-process theory presupposes the existence of two distinct rationality processes (De Neys, 2006; Evans, 2003; Evans, 2011; Kahneman & Frederick, 2007) used in optimizing decision-making. According to this theory, two systems, called system 1 and system 2, coexist. System 1 is a fast, intuitive system that does not require the use of working memory (Evans, 2011). System 2 is used for tasks requiring thoughtful decision-making, and, by extension, a calculation of the probabilities of possible futures generated by the decision. System 2 is slower than system 1 and requires greater cognitive resources and task-specific access to working memory (Evans, 2011). Thus, when a person uses system 2, s/he performs a conditional probability calculation in order to make the best decision.

In the work underlying this theory, the probability distributions of the different options are usually clearly identifiable by the participant, assisting decision-making (Kahneman & Tversky, 1979). However, due to the abundance of information in cyberspace, no probability distribution seems to be applicable by the operator to analyze the veracity and the impact of *weak signals*. In fact, when a team leader has to make a decision, he can only do it based on his own information (the *weak signals* directly perceived) and the information transmitted by his teammates without being able to check it or to compare it with a probability distribution. In these cases, other mechanisms that facilitate decision-making should therefore come in play. Among these mechanisms, trust is often described as a uncertainty reducer (Meyerson, Weick, & Kramer, 1996) that facilitates decision-making (Bell, 1982). This article proposes to study in environments with high uncertainty, what the role of trust is in the leader's decision-making when he cannot verify the data transmitted by his teammate and when these data are different from his own.

Posten and Mussweiler (2019) established a trust predictability function, i.e. trust would allow us to anticipate the possibilities by calculating their probabilities of occurrence. This is what Gambetta indicated (1988: p. 217) when he defined trust as "a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to monitor it) and in a context in which it affects his own action". Gambetta's definition and, more generally, the research conducted in economics (Williamson, 1993) and sociology (Coleman, 1990) link the phenomenon of trust to the notion of probabilistic evaluation and are thus in accordance with the dual-process theory approach. Trust can be considered as the calculation of the perceived cost-benefit (Williamson, 1993) of a relationship. In this calculative approach, "Trust emerges when the trustor perceives that the trustee intends to perform an action that is beneficial" (Rousseau, Sitkin, Burt & Camerer, 1998, p.399). Indeed, trust can only occur in relationships that bring rewards to both

parties (Lewicki & Bunker, 1995) and can be summarized, from an economic perspective, by a probability calculation (Williamson, 1993).

This notion of probability calculation is the link between the literature on trust and the literature on decision-making. In theory, the decision corresponds to "a choice or a set of choices drawn from the available alternatives" (Bellman & Zadeh, 1970). Like trust, decision-making is the choice of the alternative that subjectively presents the best cost/benefit ratio. In this approach, decision-making is no more than the result of a probabilistic assessment of the consequences of different choices (Lowenstein, 2003). In the decision-making process, the trust mechanism could therefore be seen as a readjustment of the probabilities perceived by an operator of the possible futures generated by different options, the option chosen by the operator being the option with the best cost/benefit ratio. This interpretation is consistent with Lewis and Weigert's (1985, p.969) definition of trust when they describe it as "to trust is to live as if certain rationally possible futures will not occur". In teams operating in uncertain environments such as cyberspace where operators cannot assign probabilities about future events generated by a decision (Duncan, 1972), trust may therefore facilitate decision-making. In cases where the leader cannot verify *in situ* the information transmitted by his teammate, and therefore assign a probability as to the accuracy of this information, the level of trust could be a determining factor in decision-making, in particular by facilitating acceptance by the leader of the information transmitted by his teammate. When the level of trust between a leader and his teammate is high, the information provided by the teammate should be perceived by the leader as probably more accurate than when the level of trust is low.

Hypothesis 1: For a team leader, a high level of trust in his teammate leads to a greater acceptance of the unreliable information that the teammate transmits.

In C2Cyb, *weak signals* reported by a teammate are often unverifiable by the leader. This impossibility of verifying the information means that it is impossible for the leader to assign an effective probability to these *weak signals*. When the leader cannot rely on actual probabilities, he has to assign a subjective probability (Kahneman & Tversky, 1972) to these *weak signals*. To do this, he can only rely on his own information, particularly the evaluation of the *weak signals* that he has himself received. He can therefore compare the *weak signals* he has perceived directly with those communicated to him; if all these *weak signals* correspond, they will be considered *consistent*. In this case, the leader should perceive the information transmitted by his teammate as probably more reliable than in the case of *non-consistent signals*.

According to the dual-process theory, in the case of *weak consistent signals*, decision-making is fast and intuitive (system 1). In the case of *non-consistent* signals, because of the necessary probability calculation, the response is slower (system 2) (Hypothesis 2). In this case, when the level of trust between team members is low, if the leader has not perceived any evidence of an attack "directly", he may judge as unlikely the elements in favour of an attack that are provided by the teammate. In other words, a leader will be more inclined to accept the contradiction if he trusts his teammate (hypothesis 3).

Hypothesis 2: *Consistent* signals are processed more quickly by the leader than *non-consistent* signals

Hypothesis 3: The level of trust has an indirect effect on decision-making by modulating the *consistency* consideration

**Material and procedure**

*Method*

To test these hypotheses, it is necessary to create an experimental context similar to that faced by cyber leaders. This environment must offer the participant (here, a team leader) a main task and a supervision task on which can be grafted one or more weak signals directly perceived by the leader or transmitted by a teammate. Despite the "*weak*" character, these signals must be sufficiently detectable. The leader has to make a decision based on these *weak signals* that he cannot verify in situations where he has a variable level of trust in his teammate and where these signals are not always *consistent*.

The chosen task fulfils these conditions: it offers the participant a main task of counting aircraft on a photograph with the possibility of checking a similar task with a teammate. The teammate is fictional and only presented by a predefined and controlled level of trust (Bollon, Maille, Marchand, & Blättler, 2019). During this task, "jamming" (see Figure 1) constituting the *weak signals* may occur. The participant has to indicate the number of jamming events without being able to check the number indicated by his teammate. This consideration of the teammate's data corresponds to a "blind" decision. It is this decision that is analyzed in this study and not the decisions related to the main task that can be checked on the teammate's side.

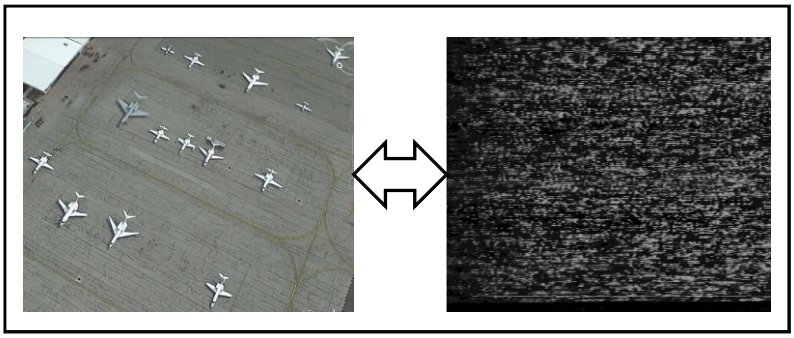


*Figure 1. The picture on the left is an example of a photograph used in the experiment; the picture on the right is the jamming that can occur at any time. In the event of jamming, the image on the right appears for one second before disappearing.*

In order to test the impact of trust on acceptance of the information transmitted (hypothesis 1), it is necessary to induce different levels of trust in the participant, to check this induction and to test, for each level, the percentage of information transmitted by the teammate and accepted by the participant. The trust-level induction is an independent variable (IV) with two controlled levels (low and high) that will be

called “trust-levels” in the following section of this article. The percentage of information transmitted by the teammate and accepted by the leader (in %) is a dependent variable (DV) collected during the experiment that will be called “decision” in the following section of this article.

In order to test the impact of consistency on the choice of decision system (system 1 or system 2) (hypothesis 2), it is necessary to induce *consistent* and *non-consistent* signals and to compare the time taken by participants to validate a decision according to these signals. The *consistent* or *non-consistent* nature of the signals is an IV which will be called “consistency” in the following section of this article. The *consistency* distribution is controlled by the occurrence of the *weak signals* transmitted. The time taken by participants to validate a decision (in ms) is a DV, called “ time “, collected during the experiment.

In order to determine the impact of trust on decision-making during consistent and/or non-consistent events (hypothesis 3), the two IVs explained above as well as the DV “decision” are used.

*Participants*

45 people (46.6% women and 53.3% men) with an average age of 22.7 years (SD: 1.09%) participated in this study. All participants were second-year engineering students. No participants had any health problems; all had normal or corrected vision.

*Protocol*

Before the start of the experiment a briefing was carried out, and all participants completed an informed consent sheet. Following this, the participants carried out a 5-minute training session before starting the experiment. The experiment was divided into 28 trials, each with 4 phases. For each trial the participant worked with a different teammate (computer simulated behaviour). 14 trials were performed with a trustworthy teammate (high trust) and 14 trials with a non-trustworthy teammate (low trust). In order to avoid an order effect, teammates’ profiles were randomly drawn. All participants therefore worked with all teammate profiles but in a different order.

Participants performed the experiment in groups in computer rooms that did not allow them to see what was happening on the other participants’ screens. At each trial, the participant thought s/he was doing the task in collaboration with one of the other participants in the room, although in reality all the teammates were fictitious. Each participant performed the task on an ordinary desktop computer using the keyboard and mouse. The screens of all participants were similar in terms of resolution and brightness.

The task was carried out in 4 phases. The first phase of each trial was designed to introduce to the participant the characteristics of his new teammate who was more or less trustworthy (IV “trust level”). The first display showed a pair of words characterizing this teammate (see Figure 2 “1”). This pair of words allowed the participant to induce a level of trust in his teammate, either low (thanks to rather negative elements: unreliable, disloyal, etc.) or high (thanks to rather rewarding

elements: professional, organized, etc.) (Bollon et al. 2019). These word pairs were obtained by following the protocol described by Bollon et al. (2019) which uses social psychology methods to identify social representations of trust in given social groups. In order to ensure that the participant had taken the teammate's characteristics into account, he was asked, on a second display, to find these two characteristics among 8 distractors (Bollon et al. 2019).

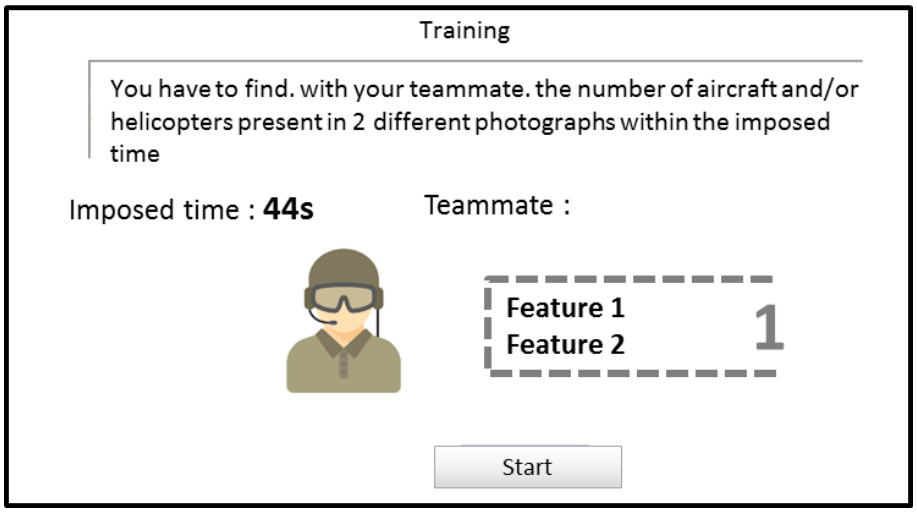



*Figure 2 . First display of phase 1. On this display the participant was informed of the instructions (similar throughout the experiment), the time allocated to the task (the time differed depending on the photograph) and the characteristics of his teammate (noted "1" on the image above). These characteristics induced a low or high level of trust in the participant.*

The second phase corresponded to the completion of the aircraft counting and jamming counting tasks. The participant had a control display that allowed him to see the countdown of the remaining time as well as the sum of the aircraft counted in the two photographs. This screen contained 4 buttons that allow the participant to (see Figure 3):

1. Display the image on which s/he had to count the aircraft and jamming events
2. Display his teammate's image in order to check the count made by his teammate if necessary
3. Modify the total score, if the participant considered that the number of aircraft counted in the two photographs was not correct
4. Complete this task and move on to the next phase. This button was only active after the participant had validated the number of aircraft present in his photograph.

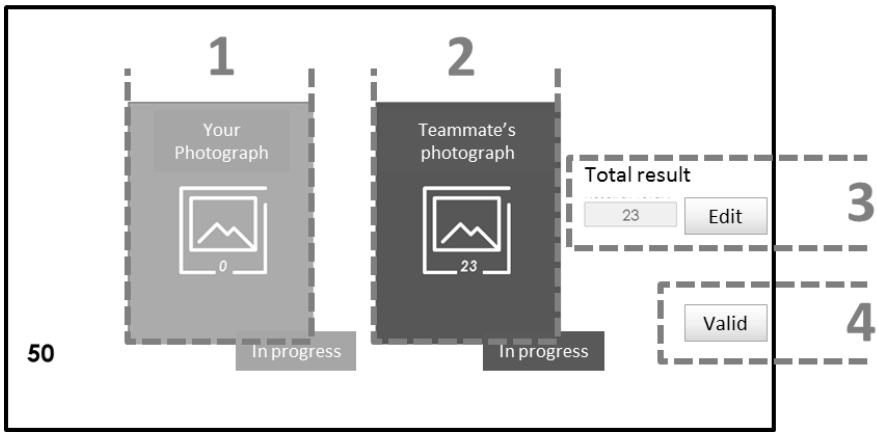



*Figure 3. Main display of phase 2. On this display the participant can see the time remaining as well as the number of aircraft counted by his teammate. With the help of different buttons, the participant can access his own image (button "1"), access the image of his teammate (button "2"), modify the total score (button "3") or complete phase 2 (interlocutor "4")*

On the display allowing him to perform his own counting task, the participant found his photograph, the remaining time (see Figure 4 "1") as well as 5 buttons that allowed him to:

- Increment or decrement the count by the number of aircraft (see Figure 4 "2"),
- Increase the number of jamming events detected (see Figure 4 "3"),
- Validate his count of the number of aircraft (see Figure 4 "4")
- Return to the control display (see Figure 4 "5").

The teammate's display was exactly the same as the participant's one. However, on the teammate's screen the buttons were not clickable (except for the button used to come back to the control screen). On the teammate's display, the photograph was different from the one presented on the participant's screen and s/he had to do the aircraft and jamming counting tasks on this other photograph. Moreover, on the teammate's screen, it was impossible for the participant to see the jamming (jamming events were never displayed on the teammate's screen). In this experiment, the participant was not aware that it was impossible for him to see the jamming events occurring on the teammate's screen.

The participant had to count the aircraft in his photograph and validate the team's total result before the end of the time limit. If this was not the case, the trial was failed and an additional trial with a teammate of the same level of trust was added at the end of the session. The validation of the total score allowed the participants to move on to the next phase.

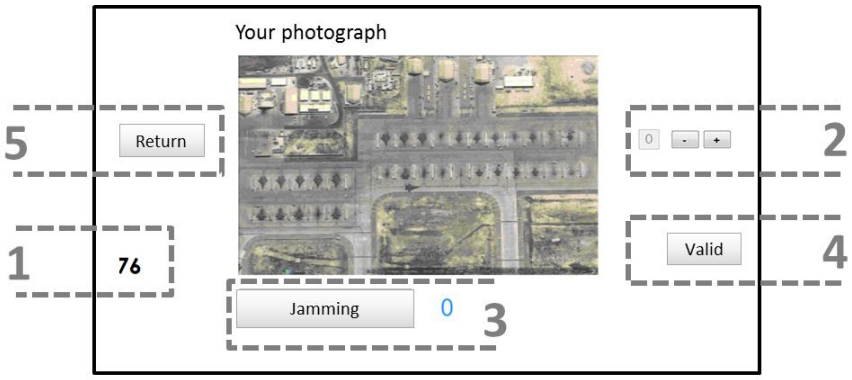



*Figure 4. Display used by the participant in phase 2 to perform his counting task. On this display the participant can see the remaining time ("1"). Using different buttons, the participant can count the aircraft (button "2"), increment the interference counter (button "3"), validate his aircraft count (button "4") or return to the main display (button "5") (see Figure 3).*

The third phase was devoted to validation by the participant of the jamming events detected on the two photographs. The display showed the number of jamming events detected by the teammate (see Figure 5 "1") and the number of jamming events detected by the participant (see Figure 5 "2"). Because no real jamming was displayed on the teammate's screen, the participant could not see these jamming events and therefore could not assess the validity of the information transmitted by his teammate. Next to each of these numbers, there were 3 buttons to validate or invalidate the jamming (none, 1 or more). The participant had to make a decision on the number of jamming events to validate on the teammate's photograph (see Figure 5 "3") as well as the number of jamming events to validate on his own photograph. Once this was done, the participant could move on to phase 4.

The different DVs used to test the 3 hypotheses were collected in phase 3. The "decision" DV used to test hypotheses 1 and 3 corresponds to the percentage of jamming events transmitted by the teammate and not validated by the participant (in %). The "time" DV used to test hypothesis 2 corresponds to the time taken by the participant to validate this third phase (in milliseconds).

In order to control the IV "consistency", in this experiment, the jamming events presented to the participant were linked to the jamming events transmitted by the fictitious teammate in order to obtain the following 4 cases:

- No jamming was presented to the participant and the number of jamming events detected by the teammate was 0 (25% of cases)
- 1 or 2 jamming events were presented to the participant and the number of jamming events detected by the teammate was 1 or 2 (25% of cases)
- No jamming was presented to the participant but the number of jamming events detected by the teammate was 1 or 2 (25% of cases)
- 1 or 2 jamming events were presented to the participant but the number of jamming events detected by the teammate was 0 (25% of cases)

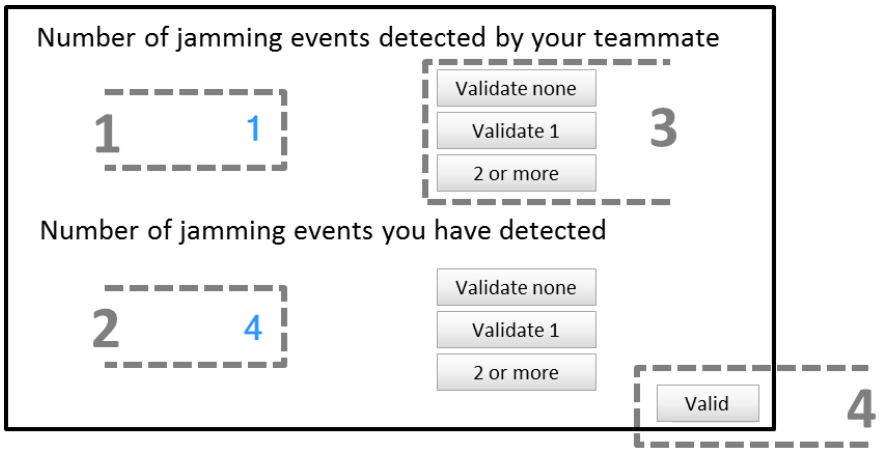



*Figure 5. Display used by the participant in phase 3. On this display the participant can see the number of jamming events detected by his teammate ("1") and the number of jamming events he had himself indicated ("2"). The participant had to make a decision on the number of jamming events to be validated for the participant (button "3") and for himself before he could complete phase 3 by clicking on the validation button (button "4").*

The first two cases were the so-called *consistent* cases and the other two *non-consistent* cases.

Finally, Phase 4 was the subjective assessment of the participant's level of trust in the results (number of aircraft) reported by his teammate. The purpose of this evaluation on non-segmented scales was to verify that the experimental trust induction equipment was working well and that the participant was working with teammates whom he perceived as trustworthy and others as less trustworthy (Bollon et al., 2019). As a high level trust induction should lead to a higher subjective evaluation by the participant of his teammate's performance than a low level trust induction (Dirks & Ferrin, 2001), the smooth operation of the experimental protocol should therefore lead the participant to assign a high evaluation to teammates in whom he had high trust and a lower one to teammates in whom he had less trust.

**Results**

Data from the 45 participants were included in the analysis. Before analysing the results required for hypothesis testing, the verification of the induction of trust in the experimental protocol was performed. The subjective evaluation data of the results transmitted by the teammate (recovered in Phase 4) show that when the trust level was high ($M$ = 5.68, $SD$ = 2.33) the subjective evaluation of the teammate's performance seem to be higher than when the trust level was low ($M$ = 5.34, $SD$ = 2.44). In order to validate these results, a one-way repeated measure ANOVA, with the IV "trust level" as a factor, has been carried out. The significant results ($F(1,44) = 4.11$, $p$ = .04) validated the presence of two levels of trust (high and low).

In order to test the hypothesis that a high level of trust between team members leads to greater acceptance by the leader of the information transmitted by his teammate (hypothesis 1), the DV "decision" and the IV "trust level" were used. For each participant, the data obtained were averaged, for each level of trust. The data indicate (see Figure 6) that between the low trust level ($M$ = 23.4%, $SD$ = 35.7%) and the high trust level ($M$ = 21%, $SD$ = 33%) the performances are relatively similar. A one-way repeated measure ANOVA, with the IV "trust level" as a factor, has been carried out. The insignificant results ($F(1,44) = 1.1$, $p = .30$) do not support hypothesis 1. In other words, trust does not seem to have a direct effect on the validation of the results reported by the teammate.

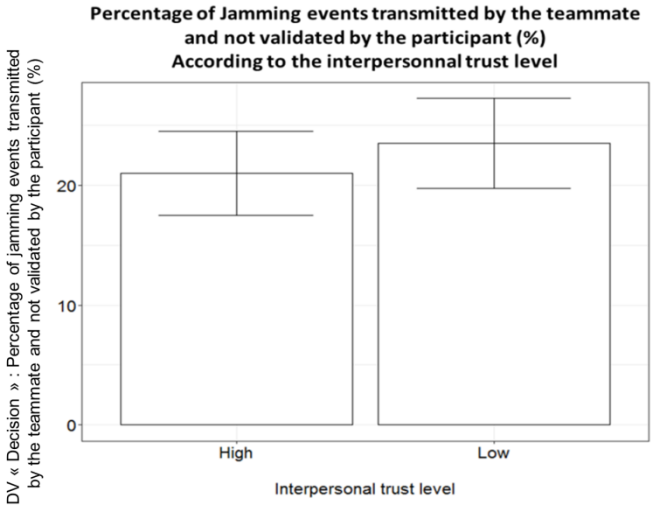



*Figure 6 . Percentages of jamming events transmitted by the teammate and not validated by the participant according to the trust level*

In order to test the hypothesis that consistent signals are processed faster by the leader than non-consistent signals (hypothesis 2), the "time" DV and the "consistency" IV were used. For each participant, the data obtained were averaged, for each level of consistency. The results show that when the jamming events were *consistent* ($M$ = 3750.5 ms, $SD$ = 1192.4 ms) the participants seem to validate phase 3 more quickly than when the jamming events were *non-consistent* ($M$ = 4272.7 ms, $SD$ = 1515.5 ms). In order to validate these results, a one-way repeated measure ANOVA, with the IV "consistency" as a factor, has been carried out. The significant results ($F(1,44)$ = 13.37, $p$ <.001) validated hypothesis 2. It would seem that the participants had a different perception of the consistency of the signals.

In order to test the hypothesis that the trust level has an indirect effect on decision making through the modulation of the *consistency* consideration (hypothesis 3), the DV "decision", the IV "trust level" and the IV "consistency" were used. For each participant, the data obtained were averaged, for each trust level, according to their *consistency*. The data show (see Figure 7) that when the information transmitted by the teammate was in line with the event perceived as the most likely (consistent case)

the leader seems to validate the information transmitted by his teammate, irrespective of whether the teammate was associated with a high (*M* = 21.3%, *SD* = 30.8%) or low (*M* = 18.5%, SD = 32%) trust level. However, when the information transmitted by the teammate supported an event perceived as unlikely (non-consistent cases), when the trust level was high (*M* = 20.6%, *SD* = 35.4%), the leader seem to validate the information transmitted by his teammate more easily than when the trust level was low (*M* = 28.4%, *SD* = 38.7%). In order to validate these results, a two-way repeated measure ANOVA, with the IV "trust level" and the IV "consistency" as a factor, has been carried out. The results of the ANOVA showed an interaction effect (F(1,132) = 4.86, p = .02 (eta-squared =.068))). A post hoc analysis performed with a Tukey HSD test indicated a significant difference in trust levels for *non-consistent* trials (*p* = .02) and no difference for *consistent* trials (*p* = .40).

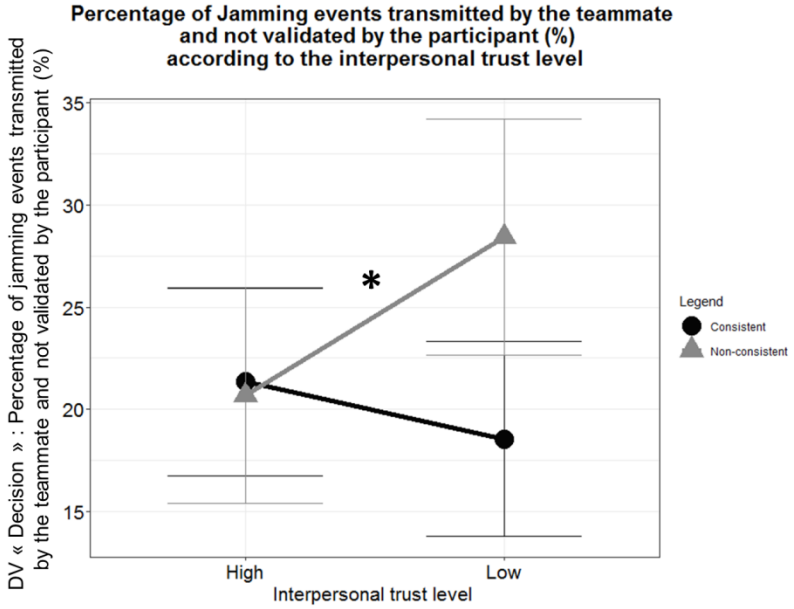



*Figure 7. Percentages of interference transmitted by the teammate and not validated by the participant according to the trust level and the consistent or non-consistent character of the tests*

**Discussion**

This study has investigated the relationship between interpersonal trust and decision-making in uncertain environments. On the basis of the dual-process theory (De Neys, 2006; Evans, 2003; Evans, 2011; Kahneman & Frederick, 2007), it is expected that decision-making can be supported either by a rapid and intuitive mechanism (system 1) that requires few resources or by a slower mechanism (system 2) (Evans 2011) involving an assessment of probabilities in relation to the possible situations, risks and benefits of certain alternatives. Applied in a micro-world resulting from cyber crisis management, the experiment aimed to better understand the impact of trust between operators and the consistency of the information exchanged on the decision-making

mechanism (through the time taken to complete the task), but also on the decision itself (validation of the partner's response).

The results show that trust does not directly impact decision-making when it is made on unverifiable elements (hypothesis 1). This result seems to contradict existing models that link trust and decision-making (Kim, Ferrin, & Rao, 2008). However, the current literature studies trust in collaborative tasks where participants can at least access the teammate's work to assess it (Bollon et al., 2019; Dirks, 1999), while the protocol presented here proposes a "blind" decision. It seems necessary to further study this type of situation and its impact on trust. On the other hand, the *consistency* of the information exchanged directly modifies the time taken to take the decision (hypothesis 2). In other words, the *consistency* of the elements exchanged between operators appears to be the primary criterion that determines the mechanism underlying the decision-making process. Once system 1 or 2 has been chosen, trust comes into the decision to the extent that the system 2 leader agrees more with the teammate's result when he or she has trust even if the information given is *non-consistent*. Once system 1 or 2 has been chosen, trust becomes an important factor in the decision-making. In fact, the leader in system 2 accepts the teammate's result to a greater extent when he trusts him even if the information given is *non-consistent*. (Hypothesis 3) (*see* Figure 8).

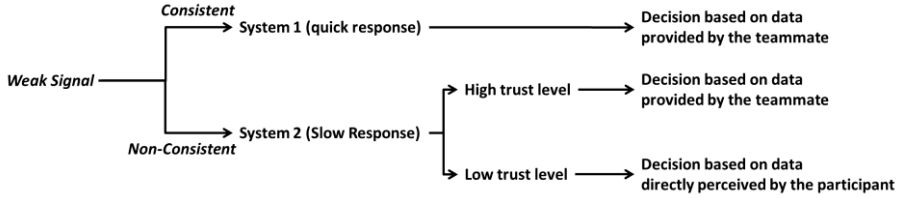



*Figure 8 . When the weak signals directly received by the leader and the weak signals transmitted by the teammate are perceived as consistent, decision-making is fast and intuitive (system 1) and is independent of the trust level. However, in the case of weak signals perceived as non-consistent, decision-making is slower (system 2) and involves the trust level. When the trust level is high, the leader's decision-making is in line with the information provided by the teammate and when the trust level is low, the decision making is in line with the information he has himself perceived.*

Thus, the study shows that the result of the decision, in terms of the acceptance or non-acceptance of the teammate's information, is linked both to the consistency of the information transmitted and to the level of trust between the operators. When the information received is consistent with the teammate's observations, decision-making is intuitive and not linked to the level of trust between operators and all information is accepted by the leader. On the other hand, when the information is non-consistent and the leader uses system 2 to make his decision, then the level of trust in the teammate who gave him the information can change the decision; the more trust the leader has in his teammate, the more inclined he is to accept his information, whether the latter confirms or invalidates his observations. The level of trust appears therefore to have a significant impact on the probability that the leader will associate with the information received, which the literature has suggested since Gambetta's (1988) work.

The direct impact of consistency is significant in the implementation of C2Cyb. Indeed, it is important in these safety-critical operations to better understand what can impact the way decisions are made. This can make it possible to adapt operator training by making them aware of the effect of consistency on their decision-making (rapid decision versus rational decision). These results can also contribute to a better understanding of how information is presented on the interfaces in order to help in better decision-making.

In terms of trust, the experiment shows that in the context of a decision made by assessing the risks or costs associated with each choice, trust in the source of the information changes the decision. This result is also important from an applicative point of view because it shows that some weak signals sent back to the decision-maker could be taken into account differently in the decision depending on the relationship between the people. Trust between people therefore changes the trust placed in the data itself. It will therefore be important for socio-technical systems such as C2s to take this dimension into account to optimize its effect on the functioning of the system.

One of the methodological contributions of this study is that we have confirmed experimentally the implementation of different decision-making mechanisms according to consistency, in accordance with the dual process theory. In other words, this *micro-world* may affect decision-making in either system 1 or system 2. However, the protocol used does not make it possible to check whether the time delay observed as a function of consistency corresponds to a probability calculation. A future study should make it possible to test this probability calculation by detailing how the decision-making process is carried out. It could use this micro-world to better understand the cognitive mechanisms really at work in each strategy.

This study considered two factors, consistency and trust, which combine to modulate the decision-making mechanism and decision content in collaborative activities. It would now be appropriate to investigate how these results are related to the interaction between human operators or whether they are more general. Are the mechanism and decision similar if the operator acts in cooperation with an automated system or artificial intelligence?

**References**

Bell, D. (1982). Regret in Decision Making Under Uncertainty. *Operations Research, 30*, 961–81.

Bellman, R.E., & Zadeh, L.A. (1970). Decision-Making in a Fuzzy Environment. *Management Science, 17*(4), B141–B164.

Boin, A., Busuioc, M., & Groenleer, M. (2014). Building European Union capacity to manage transboundary crises: Network or lead-agency model? *Regulation & Governance, 8*(4).

Bollon, F., Maille, N., Marchand, A., & Blättler, C. (2019). Cyber-attaques : Organiser la confiance. *Dixième colloque de Psychologie Ergonomique*, (pp. 243–250). Arpege Science Publishing.

Coleman, J.S. (1990). *Foundations of Social Theory*. Harvard University Press.

De Neys, W. (2006). Dual Processing in Reasoning: Two Systems but One Reasoner. *Psychological Science, 17*, 428–433.

Dirks, K.T. (1999). The effects of interpersonal trust on work group performance. *The Journal of Applied Psychology, 84*, 445–455.

Dirks, K.T., & Ferrin, D.L. (2001). The Role of Trust in Organizational Settings. *Organization Science, 12*, 450–467.

Duncan, R. (1972). The characteristics of organizational environments and perceived environmental uncertainty. *Administrative Science Quarterly, 17*, 313–327.

Evans, J.S.B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*, 454–459.

Evans, J.S.B. (2011). Dual-Process Theories of Reasoning: Contemporary Issues and Developmental Applications. *Developmental Review, 31*, 86–102.

Gambetta, D.G. (1988). Can we trust trust? In D.G. Gambetta (Ed.), *Trust* (pp. 213-237). New York: Basil Blackwell.

Kahneman, D., & Frederick, S. (2007). Frames and brains: elicitation and control of response tendencies. *Trends in Cognitive Sciences, 11*, 45–46.

Kahneman, D., & Tversky, A. (1972). Subjective probability: judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*, 263.

Kim, D.J., Ferrin, D.L., & Rao, H.R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems, 44*, 544–564.

Lewis, J.D., & Weigert, B. (1985). "Trust as a Social Reality." *Social Forces, 63*, 967–985.

Lewicki, R.J., & Bunker, B.B. (1995). Trust in relationships. *Administrative Science Quarterly, 5*, 583-601

Lowenstein, G. (2003). *The Role of Affect in Decision Making*. Handbook of Affective Sciences, ( pp. 619–642).

Meyerson, D., Weick, K.E., & Kramer, R.M. (1996). Swift trust and temporary groups. In *Trust in organizations: Frontiers of theory and research* (pp. 166–195).

Posten, A.C., & T Mussweiler. T. (2019) Egocentric Foundations of Trust. *Journal of Experimental Social Psychology, 84*, Article 103820. https://doi.org/10.1016/j.jesp.2019.103820

Rousseau, D.M., Sitkin, S.B., Burt, R. S., & Camerer, C. (1998). Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review, 23*, 393–404.

Saritas, O., & Smith, J. (2011). The big picture-trends, drivers, wild cards, discontinuities and weak signals. *Futures, 43*, 292–312.

Williamson, O.E. (1993). Calculativeness, Trust, and Economic Organization. The Journal of Law and Economics, 36 (Part 1, Part 2), pp. 453–486.

# Identification of behaviour indicators for fault diagnosis strategies

*Katrin Linstedt & Barbara Deml*
*Karlsruhe Institute of Technology*
*Germany*

**Abstract**

In manufacturing, the increasing automation leads to a rising demand for professionals fulfilling non-routine tasks like fault diagnosis of complex systems. Low reoccurrence rates of faults and working conditions, like shift work, hinder learning and make measures for knowledge support especially attractive. Additional information can be offered during the diagnosis process but the needs of the operators vary. One way to estimate the useful amount of information could be to recognize if the operator uses an associative, experience-based or an elaborate, structure-based strategy. In an attempt to identify reliable criteria to distinguish these strategies, we asked 40 participants to operate a waste water treatment simulation and confronted them with six fault scenarios. All participants received intensive training on the start-up and operation of the simulation and practiced the fault diagnosis and documentation beforehand. Through gaze behaviour analysis, a strong preference for attention focussing emerged for participants with an associative approach. Additionally, significant differences between both strategic approaches were found for Need for Cognition and prior technical knowledge.

## Introduction

With the rise of cyber-physical production systems, the transformation of the workplace of human operators is proceeding (Müller, 2019). One core demand on humans in these systems is troubleshooting, or fault diagnosis. Fault diagnosis includes the detection and localisation of faults and is the prerequisite for an efficient and effective repair and a sustainable maintenance of the system (DIN EN 13306:2018-02). Typical characteristics of fault diagnosis tasks are time pressure and a low reoccurance rate of faults. At the same time the systems are characterized by a lack of transparency which makes symptoms and their cause hard to detect. An unambiguous relation between symptom and cause is rare, more often the maintenance personnel is dealing with networks of reciprocal influence and estimate probabilities for various fault causes (Bergmann et al., 1997; Rothe & Timpe 1997). In conclusion, the cognitive demands for fault diagnosis on maintenance personnel are high.

To reduce the demands of fault diagnosis, various measures can be imagined. Fault diagnosis is a knowledge-intensive task requiring declarative knowledge of the system

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

as well as procedural knowledge of the interaction with the system and the diagnosis itself. As will be seen, different fault diagnosis strategies relate to different knowledge requirements and thus are proposed as essential indicators to inform the choice of a measure. Since recognition of different strategies is challenging, the study presented here aims at analysing behaviour correlates, specifically of gaze behaviour, to facilitate strategy recognition. To this end, two classes of strategies shall be contrasted in the following.

From a cognitive perspective, that task of diagnosis is often described in terms of reasoning and problem solving (e.g. Reed & Johnson, 1993; Schaafstal, 1993; Schmidt et al., 1990). An intensively discussed approach to describe the process of reasoning are dual-process theories. The underlying idea is the existence of two different processing types (Type I and Type II) while the specific characteristics vary between authors (e.g. Evans & Stanovich, 2013; Kahneman, 2012; Smith & DeCoster, 2000). Evans and Stanovich (2013) describe defining features of both types: Type I processes do not require working memory capacity and are autonomous, Type II processes require working memory capacity and use cognitive decoupling or mental simulation. Typical correlates of Type I processes are high speeds, parallel processing, automatic and associative thinking and experience-based decision making. Type II processes are rather slow, processing takes place in a serial, rule-based manner while thinking is more abstract and controlled. Intuitive answers are created quickly and with little effort but can be misleading, especially when reasoners lack experience. Through the intervention of reflective Type II reasoning, these intuitions can be corrected. While the insufficiency of Type I answers has been studied widely, dual-process theorists also stress the adaptivity of these answers (e.g. Kahneman, 2012). With regard to preconditions for different processing types, higher prior knowledge and experience (Smith & DeCoster, 2000) is expected to promote the use of Type I reasoning while thinking dispositions like Need for Cognition (NFC, Cacioppo & Petty, 1982; Stanovich et al., 2011) are expected to promote Type II reasoning (but see also Pennycook et al., 2017).

Critics of the dual process approach take issue with the notion of two qualitatively distinct systems and pursue a unified theoretical approach for intuitive and deliberate judgement (e.g. Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011). The latter proposed a framework which states that both types of judgement are rule-based, and even can use the same rules, but vary in their difficulty of application. The theory states that rule selection depends on individual memory constrains and processing potential, the task itself and the ecological rationality of the rule. The speed and accuracy of the rule execution are controlled by individual differences of cognitive capacities (Kruglanski & Gigerenzer, 2011).

Rouse (1983) examined human problem-solving during system failures more specifically and contrasts context-specific pattern recognition with context-free search strategies. In his model of human problem solving, decisions are preferably based on state information, assessed by pattern-recognition, while structure information is included if this fails. Rouse (1983) builds on work of Rasmussen (1978) who distinguishes between symptomatic and topographic strategies. Important aspects of symptomatic strategies are the comparison with known abnormal system states; the interaction with the system is guided by previously experienced faults. A topographic

search implies comparisons against a norm planned system performance which is led by the structure of the system. The use of available information can be rather uneconomic. Due to the difference in necessary prior knowledge, topographic strategies are expected to be applied when encountering unknown situations. Ham and Yoon (2007) analysed existing literature regarding the potential of principle vs. procedural knowledge to improve fault diagnosis performance and distinguish between forward reasoning "along the direction of the causalities of the circuit" (p.280), which poses higher demands, and backward reasoning. Reed and Johnson (1983) observed various expert strategies for fault diagnosis including what they termed heuristic path following. The core aspect is the focus of attention on relevant parts of the material to reduce the search space. This is in line with work by Van Meeuwen et al. (2014) who extracted three visual problem solving from the literature, namely attention focusing (i.e. focusing on relevant information in the current situation), perceptual chunking (i.e. combining elements to reduce necessary effort and ignore details) and means-end analysis (i.e. starting from the goal working backwards). They could show differences in the eye movements between novices, intermediates and experts in the number of fixation, fixation duration, number of transitions and time to first fixation in accordance to their hypotheses. In specific, experts showed more perceptual chunking and followed less a means-end strategy. Also, they reduced the amount of information more strongly than other groups.

Taken together, behaviour during fault diagnosis can be classified roughly into two classes: (1) a more associative, experience-based approach which is based on information reduction and includes pattern-recognition, and (2) a more elaborate, structured approach which is based on information exploitation. While no clear predictions regarding the fault diagnosis success can be made, cognitive and knowledge demands are expected to vary between theses approaches and influence strategy choice.

In the following, an empirical study will be presented which confronted participants with a fault diagnosis task to elicit the application of individual strategies and analyse behaviour correlates. After outlining the design and method of the study, detailed hypotheses will be introduced and tested. Finally, conclusions will be drawn and discussed as to which behaviour correlates are associated with either the associative, experienced-based approach or the elaborate, structured approach to fault diagnosis.

## The study

### *Design*

The aim of this study was to investigate behavioural correlates of fault diagnosis strategies, especially in gaze behaviour. To this end, the process control simulation WaTr Sim (waste water treatment simulation, Urbas & Heinath, 2007) was employed. In the first part of the study, all participants underwent a training for the start-up and operation of WaTr Sim as well as the procedure of fault diagnosis and reporting. In the second part of the study, participants were entrusted with the task of operating the simulation during nine simulated production weeks and asked to report and diagnose all faults that might occur during this time. The behaviour of the simulation was controlled by nine scenarios of which six contained faults. The order of the fault scenarios was randomized except of the final one. Behavioural data was gathered

throughout all nine productions weeks via eye tracking, screen and interaction recording as well as subjective questionnaires. In this contribution, the focus lays on the final production week, the analysis follows a between group approach.

*Participants*

The present study included 40 volunteers of which ten had to be excluded because of technical issues (n=4), insufficient training performance (n=1) and failure to detect the fault during the last production week (n=5). Participant acquisition took place in the university's environment. The remaining sample consisted of 19 men and 11 women with an average age of M = 27.2 (SD = 8.6). Twelve participants practised a profession, 17 were students, one was unemployed. Most participants (n=20) had no prior knowledge on the task of fault diagnosis while six had high to very high prior knowledge (M = 2.1, SD = 3.5, 9-point Likert scale). Additionally, prior knowledge in related technical fields was assessed via a 9-point Likert scale (1 = none, 9 = excellent). The results show moderate technical knowledge (M = 4.1, SD = 1.7). All participants had no prior knowledge of the simulation WaTr Sim before the study and were compensated at the end of the study in the amount of €20.

*Materials*

*WaTr Sim*
WaTr Sim (Urbas & Heinath, 2007) simulates a waste water treatment facility with waste water feeding in via truck deliveries and multiple stages of processing taking place until fresh water and a purified gas is produced. Altogether six stages can be distinguished: delivery, homogenisation, separation, an intermediate product repository, gas scrubbing, and a final product repository (see Figure 1, from top-left to right). While the first four stages and the sixth stage included automatic functions for information acquisition and analysis (cf. Parasuraman, Sheridan & Wickens, 2000), mainly via an alarm function based on tank level thresholds, the fifth stage is fully automated when quality of production and valves settings of the previous stages are within the normative range.

Operators are responsible for the start-up of the facility and a safe and efficient production, which maximises the amount of fresh water and purified gas and minimises the amount of waste produced. The interface allows, inter alia, for adjustments of set points of valves and heating systems and offers detailed views of component groups, information on current alarms and a trend visualisation for the final product. Fig. 1 shows the main control interface. Each run of the simulation consists of one production week with a predefined length measured in simulation steps. Each step lasts 2000ms.

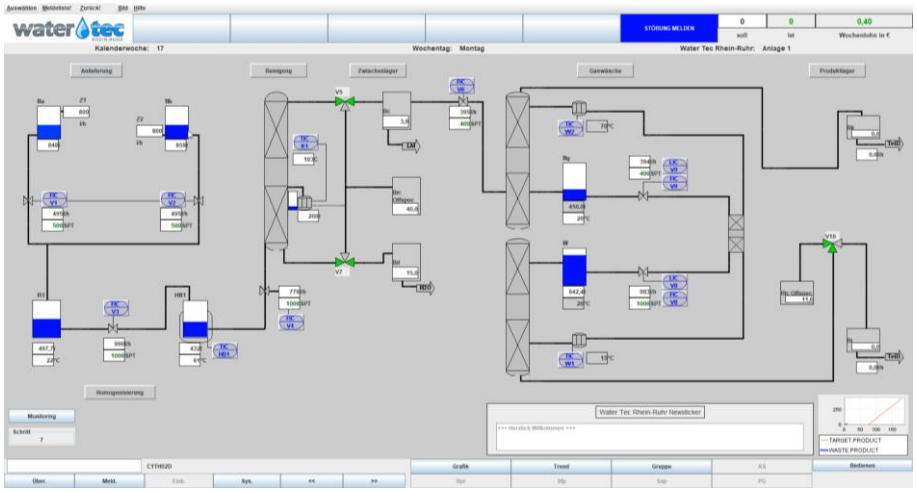


*Figure 1. Screenshot of WaTr Sim with valves (e.g. V1, V6), heaters (e.g. H1, W1) and tanks (e.g. Ba, Bc).*

*Scenarios*

The operation of the simulation was predetermined by nine scenarios: three control scenarios and six fault scenarios. The scenarios defined all set points at the first simulation step and lasted for either four or six minutes. Faults included fully and partially defective units and were visible through component observation, system alarms and/or a news ticker. For example, in the last scenario the heating unit of the gas scrubber fails, the output only reaches a temperature of 50°C instead of 70°C.

*Training*

The training for operating the simulation WaTr Sim followed the principles of instruction (Merrill, 2002) and was guided by a handbook presented on a 10.8" tablet. All participants were trained to execute a specific start-up procedure; they gained knowledge on all components and their functionality and practiced the interaction with the interface and the fault report. The training was led by the experimenter who followed standardized instructions for the interaction with the participants. It concluded in two knowledge tests, one written test on declarative knowledge regarding the facility and one practical test on start-up, operation and fault diagnosis of the facility. Passing these tests was a prerequisite for participating in the second part of the study. Altogether, the training lasted about 60min.

*Data Acquisition*

*Eye Tracking*

The experiment took place at the institute's lab rooms with illumination held constant. The simulation was presented on a 24" LCD screen at a resolution of 1920x1080pi. Eye movements were recorded using an EyeLink 1000 Plus desktop eye tracker in head-free mode at a sample rate of 1000Hz (accuracy: 0.25-0.5°, spatial resolution: 0.05). Parsing of eye data followed default thresholds. Participants were calibrated with a 9-point-calibration which was checked before every production week with a drift assessment and repeated if the deviation was 1° visual angle or higher.

*Questionnaires*
The study included multiple questionnaires, inter alia to assess demographic data, and prior knowledge, and a German version of the short scale on Need for Cognition (NFC, Beißert et al., 2014).

*Fault report*
Participants were instructed to report each fault after detection via a button implemented in the simulation before they began searching for the cause. Description of the fault was done after the production week had finished.

*Think-aloud interview*
After the last production week, the screen recording of this week was replayed for the participants and an interview following the think aloud method was conducted and recorded. During the interview, participants were encouraged to report on their actions and thoughts with questions from an unstructured interview guideline (e.g. *"What are you doing at this moment?"* or *"Please describe your thoughts in more detail."*).

*Data analysis*

Statistical analysis was conducted with R (R Core Team, 2018) and a significance level of $\alpha$=.05. For directional hypotheses, one-tailed tests were used. The data was tested on deviation from normal distribution with the Shapiro Wilk test for each group. In case of a detected deviation, Wilcoxon rank sum tests were employed instead of t-tests for independent samples. Because of unequal group sizes, the effect size was calculated with Hedge's correction.

For the analysis of eye tracking data, the screen was divided into multiple areas of interest (AOI) including the processes of delivery, gas scrubber and final repository as well as separate components and information sources. As the size of the areas varied, parameters like number of fixations ($n_{fix}$) and fixation duration ($t_{fix}$) were normed on the size of the current AOI. Eye movement data was included for a 30s time window before the fault report via button press.

Recordings from the interviews were transcribed and, based on a guideline with category descriptions and examples, categorized into two classes of strategies: (1) an associative, experienced-based approach which is based on information reduction and (2) an elaborate, structured approach which is based on information exploitation. To ensure reliability, a third of the material was categorized by two raters. The agreement of the raters was acceptable with Cohen's $\kappa = 0.61$. In a second step, participants were assigned to two groups (associative vs. elaborate) depending on the ratio of statements in each category.

Accuracy of diagnosis was evaluated on a scale from 0 to 3 with a grading scheme including the ratio of the number of correctly vs. incorrectly identified symptoms and the correctly identified cause of the fault.

*Hypotheses*

Building on the insight of existing research, multiple hypotheses were deduced (Table 1).

*Table 1. Overview over hypotheses*

| | *hypotheses* | | *assessed behaviour indicators* |
|---|---|---|---|
| Participants with an associative approach show… | …lower or higher NFC (H1)… | …than participants with an elaborate approach. | sum NFC scale |
| | …lower or higher prior technical knowledge (H2)… | | sum prior technical knowledge |
| | …more attention focussing (H3)… | | $n_{fix}$ on delivery lower<br>$t_{fix}$ on delivery lower<br>$t_{fix}$ on tank Bk lower |
| | …more backward reasoning (means-end) (H4)… | | more saccades to the left (sum)<br>$n_{fix}$ on final repository higher<br>$t_{fix}$ on final repository higher |
| | …more perceptual chunking (H5)… | | lower number of components fixated |
| | … no difference in fault diagnosis performance (H6)… | | accuracy of diagnosis equal |

## Results

The strategy classification resulted in two unequally sized groups, 13 participants followed an elaborate approach while 17 followed an associative approach.

In Table 2, results for all dependent variables are summarized. In accordance with H1, there is a significant difference between groups on NFC (t=3.948, df=16.7, p=.001, 95% CI [-9.2, -2.8]). Figure 2 visualises the result. Participants with a more associative approach showed a higher NFC than participants with a more elaborate approach. The effect is large ($g_{Hedge's}$=-1.5). H2 can be accepted as well with participants with an associative approach showing higher prior technical knowledge than participants with an elaborate approach (W=44, p=.006, 95% CI [-2.9, -0.4], see figure 3). The effect is large ($g_{Hedge's}$=-1.8). Additionally, the results show strong support for H3, but only limited support for H4 and no sppurt for H5. There was no significant difference between groups regarding the diagnosis performance (W=131, p=.250, 95% CI [-2.0, 0.0], the effect was small ($g_{Hedge's}$=-0.5). Figure 4 visualises the data. The implications will be discussed in the following chapter.

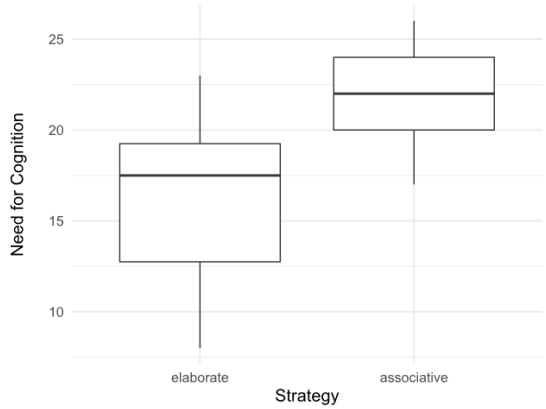



*Figure 2. Box-Whiskers-Plot for Need for Cognition.*

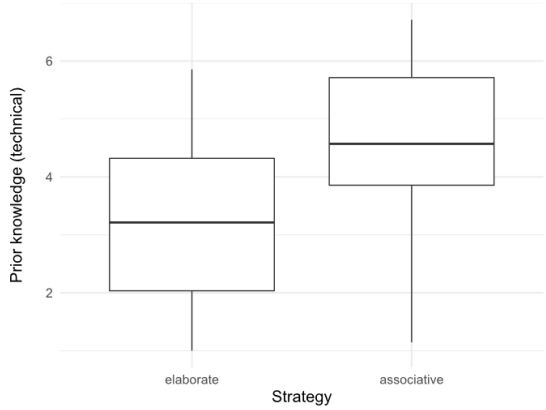



*Figure 3. Box-Whiskers-Plot for prior technical knowledge.*

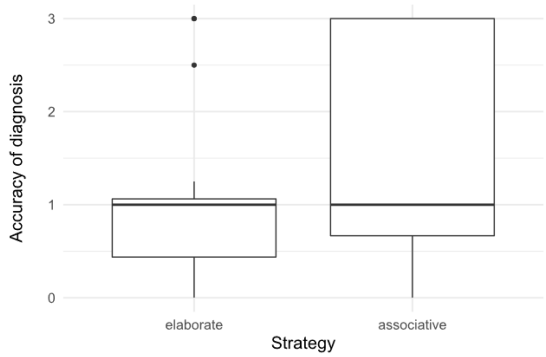



*Figure 4. Box-Whiskers-Plot for accuracy of diagnosis.*

*Table 2. Overview over results*

| *hypothesis* | *dependent variable* | *t (df) / W* | *p* | *95% CI* | $g_{Hedge's}$ |
|---|---|---|---|---|---|
| H1 | sum NFC scale | t=3.948, df=16.7 | .001 | -9.2, -2.8 | -1.5 |
| H2 | sum prior technical knowledge | W=44 | .006 | -2.9, -0.4 | -1.8 |
| H3 | $n_{fix}$ on delivery | W=78.5 | .002 | 0.4, ∞ | 1.7 |
| | $t_{fix}$ on delivery | W=85 | <.001 | 204.3, ∞ | 2.2 |
| | $t_{fix}$ on tank Bk | t=2.100, df=4.4 | .049 | 5.6, ∞ | 1.4 |
| H4 | Sum saccades to the left | t=-0.663, df=26.7 | .744 | -4.6, ∞ | -0.2 |
| | $n_{fix}$ on final repository | t=1.479, df=22.5 | .076 | -0.1 ∞ | 0.5 |
| | $t_{fix}$ on final repository | W=130 | .045 | 5.2, ∞ | 0.6 |
| H5 | Number of components fixated | t=0.640, df=25.6 | .264 | -1.8, ∞ | 0.2 |
| H6 | Accuracy of diagnosis | W=131 | .250 | -2.0 , 0.0 | -0.5 |

**Discussion and conclusion**

The aim of this study was to investigate behaviour correlates of fault diagnosis strategies. Based on a review of existing theory and research, two classes of strategies have been defined: an associative, experienced-based approach and an elaborate, structured approach. Participants were split into these two groups based on a content analysis of verbal reports.

The results show large and significant differences between participants from both groups before the study, supporting the claim that strategy choice is influenced by individual differences of prior knowledge and motivation (e.g. Stanovich et al., 2011; Kruglanksi & Gigerenzer, 2011). It should be noted that all participants had no experience with the operation of WaTr Sim before the study and were exposed to the same scenarios – the knowledge gain during the study was thus dependent on the individual learning performance.

With regard to attention focussing, the results strongly support the hypothesis, that an associative approach includes higher attention focussing. Participants with an elaborate approach spend more time fixating components of the first step of the process. Also, they fixate this step more often. During the final scenario, only parts of the gas scrubber and the final repository showed symptoms of the faults. Such behaviour can be understood as a more thorough use of information with the gaze being diverted from the more obviously affected components. This is also true for the tank Bk which is part of the final repository – in past scenarios, analysis of the tank's

behaviour was not necessary for the fault diagnosis. Therefore, participants with an associative approach were not expected to spend attention on this component as experience taught them it is not necessary. The results agree with this expectation as participants with an elaborate approach spend more time fixating tank Bk.

Backward and forward reasoning have been mentioned by various researches to describe diagnosis strategies, e.g. the topographic search described by Rasmussen (1978) which includes searching systematically through the system and which can be classified as elaborate approach. The results show that participants with an associative approach spend more time on the goal state of the system but there is only a marginal difference in the number of fixations on the goal state and no difference in the number of gaze switches to the left vs. to the right. Taken together a preference for means-end analysis seems to exists within the associative approach but the direction of the reasoning stays unclear.

As chunking includes grouping of elements, the expectation was to find participants with an associative approach fixate less components but instead choosing representative components for different parts of the process. This expectation was disappointed. Possible reasons included insufficient training on the system as chunking is especially seen within experts (van Meeuwen et al., 2014).

Various authors stress the claim that success of strategies depends on the task at hand and the performing individual, therefore a superiority of one class of strategies was not expected and also not found. Accordingly, Figure 4 shows equal medians in both groups, but a striking difference in the variance of the data. To understand this result better, analysis of supplementary data will be necessary.

In conclusion, participants differed meaningfully in their attention focussing according to their strategic approach. Individual differences of motivation and prior knowledge seem to play an important role for strategy choice. To understand this relationship better, more insights on strategy development over time and specific use of knowledge are necessary. Nevertheless, the distinction between an associative and an elaborated approach has been proven useful and behaviour indicators emerged.

**References**

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2014). Eine deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition. Die Need for Cognition Kurzskala (NFC-K). *GESIS-Working Papers, 2014|32*.

Bergmann, B., Wiedemann, J., & Zehrt, P. (1997). Konzipierung und Erprobung eines multiplen Störungsdignosetrainings. In K. Sonntag and N. Schaper (Eds.), *Störungsmanagement und Diagnosekompetenz* (pp. 235-254). Zürich: vdf Hochschulverlag.

Cacioppo, J.T., & Petty, R.E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116-131.

DIN EN 13306:2018-02. *Instandhaltung – Begriffe der Instandhaltung; Dreisprachige Fassung EN 13306:2017*. Beuth, Berlin.

Evans, J.St.B.T., & Stanovich, K.E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science, 8*, 223-241.

Ham, D.-H., & Yoon, W.C. (2007). The training effects of principle knowledge on fault diagnosis performance. *Human Factors and Ergonomics in Manufacturing, 17*, 263-282.

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*, 500-533.

Kruglanski, A.W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*, 97-109.

Merrill, M.D. (2002). First principles of instruction. *Educational Technology, Research and Development, 50*, 43-59.

Müller, R. (2019). Cognitive challenges of changeability. Adjustment to system changes and transfer of knowledge in modular chemical plants. *Cognition, Technology & Work, 21*, 113-131.

Pennycook, G., Ross, R.M., Koehler, D.J., & Fugelsang, J.A. (2017). Dunning-Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*, 1774-1784.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rasmussen, J. (1978). *Notes on diagnostic strategies in process plant environment*. Risø-M-1983.

Reed, N.E., & Johnson, P.E. (1993). Analysis of expert reasoning in hardware diagnosis. *International Journal of Man-Machine Studies, 2*, 251-280.

Rothe, H.J., & Timpe, K.P. (1997). Wissensanforderungen bei der Störungsdiagnose an CNC-Werkzeugmaschinen. In K. Sonntag, and N. Schaper (Eds.) *Störungsmanagement und Diagnosekompetenz* (pp. 137-154). Zürich: vdf Hochschulverlag.

Rouse, W.B. (1983). Models of human problem solving: Detection, diagnosis, and compensation for system failures. *Automatica, 19*, 613-625.

Schaafstal, A. (1993). Knowledge and strategies in diagnostic skill. *Ergonomics, 36*, 1305-1316.

Schmidt, H.G., Norman, G.R., & Boshuizen, H.P. (1990). A cognitive perspective on medical expertise: Theory and implication. *Academic Medicine, 65*, 611-621.

Smith, E.R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review, 4*, 108-131.

Stanovich, K.E., West, R.F., & Toplak, M.E. (2011). The complexity of developmental predictions from dual process models. *Developmental Review, 31*, 103-118.

Urbas, L., & Heinath, M. (2007). *AWASim Handbuch*. Technische Universität Dresden.

Van Meeuwen, L.W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P.A., De Bock, J.J.P.R., & Van Merriënboer, J.J.G. (2014). Identification of effective visual problem-solving strategies in a complex visual domain. *Learning and Instruction, 32*, 10-21.

# Investigating the effects of passive exoskeletons and familiarization protocols on arms-elevated tasks

*Aurélie Moyon[1,2], Jean-François Petiot[1], & Emilie Poirson[1]*
*[1]Ecole Centrale de Nantes, Nantes*
*[2]Europe Technologies, Carquefou*
*France*

**Abstract**

Exoskeletons present interesting qualities for high demanding physical tasks, but their integration in companies is still a challenge. This study aims to evaluate the effects of exoskeletons on the completion of arm-elevated tasks. Three categories of dependent variables are studied in a lab experiment: physical measurements (cardiac cost), performance indexes (quality and duration) and perceived benefits (reported by subjects on quantitative scales). The independent variables of the experiment are the presence (or not) of the exoskeleton, and the media used for the familiarization process of the subject before the use of the exoskeleton. Two levels of familiarization are proposed to the subjects: brochure of the exoskeleton manufacturer, and live tutorial demonstration by a skilled experimenter. A laboratory study (n=36 participants) involving two arms elevated tasks was specifically designed to simulate industrial work situations. Results show that the use of the exoskeleton reduces cardiac cost, global and local perceived effort, number of errors, and increases task performance. Concerning the familiarization process, the live tutorial demo provides higher task performances and users acceptance, lower global and local perceived effort and the number of errors. These results confirm that user acceptance and integration of exoskeletons in companies require dedicated training supports.

**Introduction**

Passive exoskeletons started to enter the market of New Assistive Technologies (NAT) in various industries where handling tasks are still involving human control and know-how. This growing interest forces companies to relate the claimed effectiveness of occupational exoskeletons as a solution that could release muscle activity and task-related strain. Even if functional effects have been established in reducing muscular demand (Huysamen et al., 2018; Theurel & Desbrosses, 2019) these exoskeletons are still facing ergonomics barriers such as discomfort (de Looze et al., 2016), movements limitations, low usability and acceptance of end-users. (Graham et al., 2009).This is why previous studies suggest a more holistic approach (Bosch et al., 2016) to investigate dimensions of usability, moreover on realistic work settings (Baltrusch et al., 2018). Recent studies suggest focusing on the actual use, to better understand expected and potential unexpected effects (Kim et al., 2018). This is why the evaluation of Human Exoskeleton Interaction (HEI) should focus on

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Usability. Last years, Europe Technologies has been training future users and product managers to the use of exoskeleton, in order to enhance potential adoption. However, no evidence has been found on the effectiveness of a specific familiarization protocol on user's acceptance and on task-related performance. Consequently, the main purpose of the current study is to validate the claimed positive effects of the exoskeleton prototype, as well as the effectiveness of a familiarization protocol on objective performance, perceived benefits and user acceptance. A second aim is to highlight specifications of human-exoskeleton interaction to guide further product development and familiarization program. The remainder of the paper is organized as follows. The second section presents the material and method and the description of the experiment. Results are presented in third section. The concluding section provides implications and perspectives for further work.

## Materials and methods

### *Participants and ethics approval*

36 healthy participants (50% male, 50% female) with no current injuries / musculoskeletal disorders volunteered and gave written consent before the experiment according to the tenets of the Declaration of Helsinki. Current health status was evaluated using the Nordic questionnaire (Descatha et al., 2007). Their age span from 20 to 65 years old with a range of height between 163 to 175cm. Participants had never been trained to use exoskeleton nor performing tasks.

### *Occupational exoskeleton*

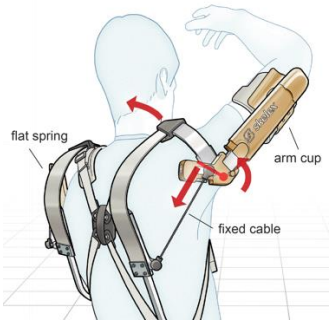



*Figure 1. Product architecture and the mechanical principle of operation of the tested exoskeleton. Flat springs in the back apply a progressive strength upwards.*

The exoskeleton used is a wearable passive system provided by our partner *SkelEx* (SkelEx, Rotterdam, The Netherlands). It was co-developed with this partner from various field studies and user's feedbacks (Moyon et al., 2018). As shown in figure 1, its design is based on a backpack style with two flat springs in the back that can store kinetic energy when lowering the arms. Reversely, the spring strength is then applied upwards and help reducing upper body strain while performing arm-elevated tasks. This constitutes the first independent variable of our experiment with the two conditions (Exo/No Exo). Two versions of the prototype called Exo A and ExoB have

been tested for a secondary design purpose, so differences won't be discussed here. All variables were tested for both versions, results are merged into an Exo condition.

*Familiarization protocol*

In our observations of the spreading to exoskeletons in industry, we noticed that companies are starting to buy exoskeletons without considering the familiarization phase and potential fail of acceptance for occupational use. In order to protect future users, the French Institute of normalization is working on an agreement and a potential future norm about Human-Exoskeleton Interaction ergonomics. Europe Technologies takes actively part in this project, by sharing field insights. A global acceptance program has been designed to foster better integration of exoskeletons in companies. A key element of this program is a familiarization protocol (labelled F2), designed to optimize user's performance and acceptance. It is based on our previous expertise to give users the best level of knowledge and practice in the shortest amount of time (to match real-time constraints). To do so, this protocol F2 is composed of the following steps: Demystification, Technics, Potential, Limits, Donning/Adjusting/doffing, Free experience (without industrial constraints). It aimed at providing certification of a level 4 based on a 1-7 scale of knowledge/practice (appendix). Level 4 means that participants are aware of basic technical, safety and usability principles, and know how to don/doff quickly the exoskeleton. In the following experiment, F2 is performed by a skilled experimenter and materialized by a written script. Another familiarization protocol, F1, corresponds simply to the manufacturer's brochure, materialized by a paper brochure. The two familiarization protocols (F1 or F2) were administered to the participants before the execution of the task. This constitutes the second independent variable of our experiment. Between tasks, participants could adjust the exoskeleton again if needed. They could read the brochure F1 or ask the experimenter to repeat an item in tutorial F2. But the experimenter couldn't take any additional initiative, to not distort the results.

*Testing equipment*

The heart rate was measured in real-time during the tasks. We used a heart rate computer POLAR RS800CX and its dedicated professional software POLAR Trainer 5. This system is composed of an emitter attachable on a thoracic belt. The data transfer was realized from the emitter to the software by an infrared USB adapter. For precision task performance, user lines were obtained by an interactive whiteboard SMART Board 800. This system projects and records automatically produced pixels. 1 pixel = 1mm. All tasks were camera recorded to help further interpretation of results.

*Design of experiments*

For a secondary product design purpose, all participants tested two versions of the exoskeleton prototype called A and B, so as the NoExo condition. Concerning the familiarization protocol, given that protocol F2 is more informative than F1, it was irrelevant for the same participant to test protocol F1 after F2. For this reason, the only possible orders for the test were F1->F1, F1->F2 or F2->F2. To limit the number of experiments (two tests with two exoskeletons A and B), a balanced incomplete block

design was defined, presented in table 1. Six blocks were considered, with six participants in each block.

***Table 1***. *Experimental design for the two variables Exoskeleton and Familiarization protocol with two conditions (NoExo/Exo) and (F1/F2). The rows correspond to the first combination tested by the participants, the column to the second (for example, 6 participants tested first ExoB with protocol F1 (BF1), then ExoA with protocol F1 (AF1).*

| | AF1 | AF2 | BF1 | BF2 |
|---|---|---|---|---|
| AF1 | | | 6 | 6 |
| AF2 | | | | 6 |
| BF1 | 6 | 6 | | |
| BF2 | | 6 | | |

**Previous analysis of industrial tasks**

Assembling tasks involve arm-elevated postures that could be assisted by an exoskeleton. The manufacturer *SkelEx* (SkelEx, Rotterdam, Netherlands) provided the model that was designed specifically to assist the strain related to this posture. Constraints of the real work situation such as average duration of steps, the weight of the tool, precision standards have been integrated into the lab experiments. Experiments took place between January and May 2019 on the site of LS2N laboratory, Nantes.

*Lab tests*

From an analysis of the previous industrial tasks, a controlled laboratory experiment was built in order to not disturb the manufacturing process of the industrial. These tasks in a laboratory have furthermore the following advantages:

- To measure more easily the effects of the exoskeleton and the familiarization protocol on user performance, perceived benefits, and acceptance with a reproducible procedure.
- To involve more participants, with a larger diversity of profiles

The idea was to create a simple laboratory protocol that could easily evaluate the potential of exoskeletons for repetitive and precision tasks.

*Repetitive task (R)*

According to real constraints observed previously, a repetitive task was designed to reproduce arm-elevated posture (Figure 2). A board with eight lines of industrial nuts was placed vertically on the wall. The size and height of the board were adjusted so that any participants could reach at least 7/8 lines with a tool of 6kg. Setting movements were paced at 20 actions/min using a metronome. Participants had to set as many nuts as they can. They stopped when they experienced fatigue or high

discomfort or failed pace three times in total. Errors were observed: nuts should be correctly set, we tolerate a space of 5 millimetres corresponding to nut thickness. Data collected were: total time, time per line, number of nuts correctly set, number of errors/line. Four dimensions questionnaire including the following items: perceived exertion, fatigue, comfort, quality, performance, task-related usability assessment: perceived utility, easiness of use and move with the exoskeleton.

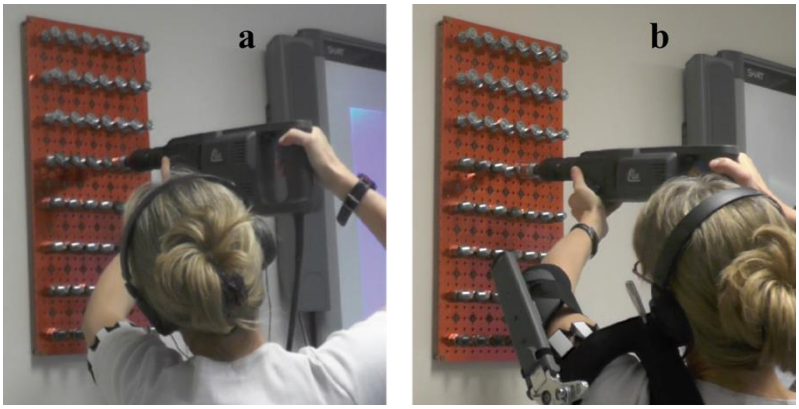



*Figure 2. (a) A participant without the exoskeleton performing the repetitive task R and with the exoskeleton (b).*

*Precision task (P)*

This task aimed at testing the potential benefits of wearing the exoskeleton (less perceived effort and fatigue, respect of quality and natural moves) while performing repetitive and accurate movements, as observed in the real work situation. A background of lines was projected on the wall by an interactive whiteboard system (Figure 3). The test consisted of redrawing the same signs with an interactive pen with maximum accuracy. Seven lines of ten signs each are displayed on the background. Participants started by the line at their eye-level and moved progressively upward to an overhead position. They had to stand behind a line placed at 40cm from the wall but could move parallel to the wall. Distance from the wall was visually controlled so that arms elevated posture targeted by assistance would be respected. The test ended when participants experienced fatigue, discomfort or traced all signs. Movements were paced at 4second/sign using a voice recorded metronome. Data collected were: traced signs, time per line, number of completed signs, and number of errors/line.

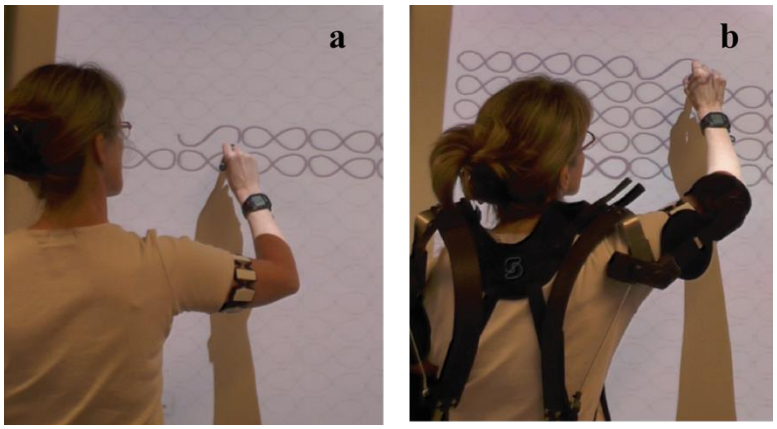



*Figure 3. (c) A participant performing the precision task P without exoskeleton (a) and with the exoskeleton (b).*

**Objective measurements**

*Familiarization performance of donning/adjusting*

Familiarization performance was measured by chronometer for doffing/donning procedure after the participant had experienced the brochure (F1) or the tutorial (F2). Measurements were organized as follows: 5 min to read the manufacturer's brochure or to listen to the tutorial performed by a skilled experimenter, 3 min of testing alone, finally, the participant was challenged to install it and control adjustments. The recording was stopped above 3 minutes. This is the duration limit evaluated previously as a standard because operators have to be very quick at doffing/donning in a real situation in order to be flexible on other tasks.

*Global physical workload*

This work situation has been previously targeted by an internal ergonomic study. Laboratory tasks were designed to approach real perceived effort with similar postures and duration constraints. The condition Exo/NoExo was measured on both tasks R and P, always in the same order and separated by a break while they seated. A reference heartbeat (HR) was recorded while seating 5min before performing the task. Activity blocks were analyzed with the conditions Exo/NoExo. The measurements were separated by a 10 minutes break while operators remained seated. According to Meunier protocol (Meunier, 2014), in order to compare two different conditions of the activity (NoExo, Exo), we calculated the Absolute Cardiac Cost (ACC) according to the duration of the activity. ACC is the difference between the average heart rate (Ha) and the Reference Heart rate (Hr) and it is expressed in beat per minute (bpm). ACC*duration is expressed in heart rate (h) according to the duration of the task (in min). It represents the number of pulses 'consumed' during the task. The definition of the Absolute Cardiac Cost is represented in Figure. 4.

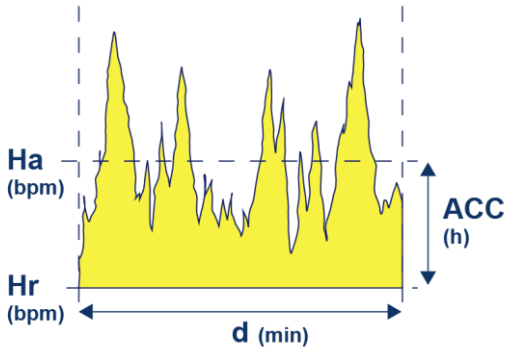



*Figure 4. ACC*d is the difference between Reference Heart rate (Hr) and Average Heart rate (Ha) expressed in beat/min multiplied by task duration (min).*

*Tasks performance*

On the repetitive task R, the number of settings was observed and the duration recorded by chronometer. A speaker connected to a digital metronome indicated the rhythm to respect. The performance of precision task P was measured by chronometer and counting the numbers of symbols.

*Subjective measurements*

A four dimensions questionnaire (Cognitive, Occupational, Physical and Affective) built from a previous study (Moyon et al.) recorded user's subjective effects of exoskeleton on tasks. The perceived musculoskeletal strain was evaluated with Borg Scale (CR-10) (Hill et al., 1992). We recorded on Likert scales (0-10) factors such as Easiness of learning, Evolution of perceived musculoskeletal effort, Perceived Usability for industrial constraints, Physical Comfort, Intention to use daily and Acceptance after use.

**Data analysis**

To investigate significant differences in user performance, perceived benefits and acceptance between Exoskeleton, differences in means were analyzed by comparisons of NoExo (without exoskeleton)/Exo (with exoskeleton) using an ANOVA (mixed linear model, that considers the subject as a random effect and the factor "Exoskeleton" as a fixed effect) and a one-tail one-sample T-test was applied to determine a significative threshold for Exo condition subjective results according to the variables. Also, the effectiveness of the familiarization protocol (F1/F2 conditions), was analyzed for the same variables and for Exo condition only, by a two-samples two-sided T-test, which calculate the difference of means between the six groups. The statistical significance was set to $p<0.05$ (*) and $p<0.001$ (**). Statistical analyses were performed using XLSTAT 2019. For each dependent variable, the results for the different conditions are reported as means (with their standard errors) in original units.

**Results**

*Study of exoskeleton effects on Global physical workload*

The evolution of Absolute Cardiac Cost (ACC) with task duration (ACC*d) is expressed in number of heart rate (h). The results are shown in figure 5. For both tasks, the lowest values of ACC*d are found while wearing the exoskeleton (Exo). Without the exoskeleton (NoExo), ACC*d is increased by 32 h ± 2.9for the task R and by 27.1 h ± 5.9 for the task P.

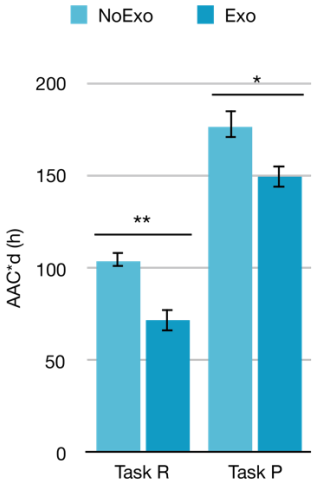



*Figure 5. Evolution of ACC*d (h) for Task R and task P with (Exo) and without (NoExo) exoskeleton*

Despite the weight and physical constraints produced by springs, the exoskeleton seems to reduce the cardiac cost for all tasks.

*Study of exoskeleton effects on tasks performance*

Hypothesis: Performance is better when the participant is wearing the exoskeleton. For task R, the highest number of valid actions (45.5±1, $p<0.0001$) and the lowest number of errors (4.4±0.3, $p<0.0001$) is found when wearing the exoskeleton. A similar effect is found for task P: highest number of valid signs (49.6±0.9, $p<0.0001$) and lowest average number of errors (5.1±0.3, $p<0.0001$) were found when wearing the exoskeleton. We conclude that for all tasks, Human-Exoskeleton performance is better than NoExo condition with a higher number of actions and a lower number of errors.

**Subjective measures**

*Physical aspects: evolution of perceived musculoskeletal strain*

Hypothesis: perceived exertion could be reduced while wearing the exoskeleton. Global exertion for tasks R and P has been evaluated respectively with a mean of 6.99/10 ± 0.21 and 6.45/10 ± 0.25 for NoExo condition and 4.22/10 ± 0,14 and 3.61/10 ± 1.16 for Exo condition. These results are represented by dotted lines in Figure 6. Results indicate that globally the strain is lower when wearing the exoskeleton, with a significant ($p<0.0001$) reduction of global strain respectively of 3.06/10 and 3.12/10 for task R and task P.

Perceived local strain shows lower scores when wearing the exoskeleton and an effect of transfer towards other parts of the body has shown in figure 8 (both tasks merged). Indeed, participants perceived a mean reduction of strain on upper parts of the body, on Shoulders (2.32/10; ± 0.15, $p<0.0001$), on Arms (2.93/10 ± 0.12, $p<0.0001$), Elbow/forearms (0.06/10 ± 0.16, $p<0.0001$), neck (1.41/10 ± 0.14, $p<0.0001$), in the Upper and lower back (0.79/10 ±0.09, $p<0.0001$ and 0.46/10 ±0.1, $p<0.0001$) and on legs (0.17 ±0.06, $p<0.0001$). Also, the perceived strain has been transferred to other parts of the body, with a small mean increased of 0.4± 0.16, $p= 0.002$ in the Elbow/Forearm part.

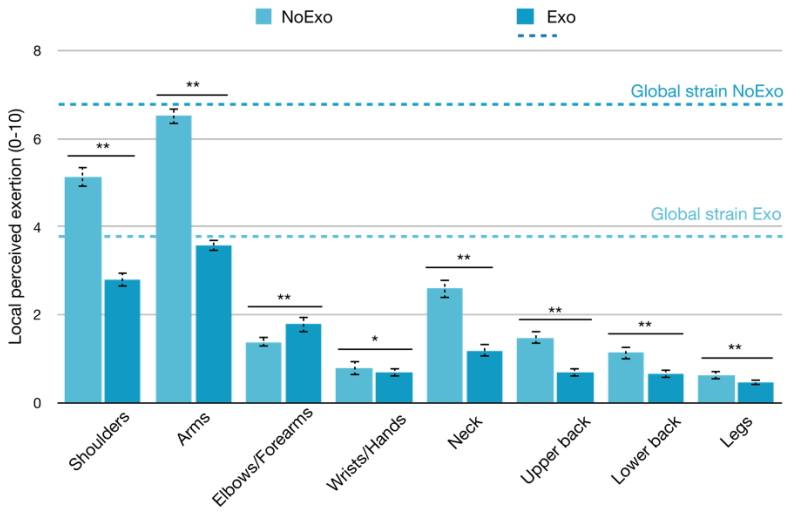



*Figure 6. Evolution of global and local perceived effort for specific parts of the body without (NoExo) and with Exoskeleton (Exo) for all tasks. A global effort is represented by the lines.*

We can conclude than the evolution of perceived exertion could be reduced globally while wearing the exoskeleton (Exo). However, we observed a transfer effect of local strains with a very small local decrease on Wrist/Hand and and a non-expected increase on Elbow/Forearm.

*Cognitive and Occupational aspects*

Regarding Affective aspects, no participant found that wearing the device was devalorizing. To check if the exoskeleton is suitable to perform simulated tasks constraints, we observe the evolution of focus demand, perceived quality and performance while wearing the exoskeleton.

*Questions (Likert scale 0-10):*
*With the exoskeleton, I can perform the task with the same quality (strongly disagree- totally agree)*
*With the exoskeleton, I feel (much less effective-much more effective)*

*Two reverse questions:*
*To use the exoskeleton involves an extra focus demand (strongly disagree- totally agree)*
*To master the exoskeleton involves an effort (marginal – extremely important)*

For all results except the two last inverse sentences (Effort to master and Extra focus demand), results <5 are interpreted as a negative effect and results >6 are interpreted as a positive effect. A score between 5 and 6 corresponds to indecision or average effect. The effort to master and Extra focus demand, results <5 are interpreted as a positive average effect and results >4 are interpreted as a positive effect. A one-tail one-sample T-test was applied to determine a significative threshold according to the variable. Significant results are shown in Table 3. For both tasks in average regarding cognitive aspects, Perceived performance was positively significant with the exoskeleton (mean = 7.19, lower mark interval: 6.88, $p<0.0001$), participants reported that wearing the exoskeleton didn't involve important supplementary focus demand (mean = 4.21, upper mark interval: 4.64, $p=0.001$) or involved an important effort to master (mean=4.07, upper mark interval: 4.39,p= $p<0.0001$). Also, they could perform the same quality standards (mean=7.17, lower mark interval: 6.84, $p<0.0001$). All differences in means between tasks were not significant ($p>.05$). We can conclude than the use of exoskeleton on the simulated industrial tasks does not disturb the respect of quality standards, perceived performance and doesn't imply extra mental load concerning focus demand.

**Effects of familiarization protocol (F1/F2)**

**Objective results**

*Donning performance*

Hypothesis: lowest donning duration performed with F2 protocol. Results showed a significant decrease of donning performance (adjustments included) with the lowest duration of 93.97±26.47s for F2 vs 171.97 ±26.36s for F1 as shown in figure 7. Donning performance is expressed in seconds, the full line indicates the limit duration expected by partners, the dotted line represents the maximum duration users have to reach to pass level 4 of familiarization on our internal scale (HEFL: Human Exoskeleton Familiarization levels). Otherwise, user certification is not delivered.

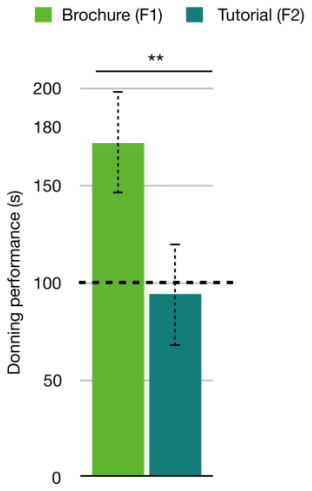



*Figure 7. Evolution of donning/adjusting performance (s) according to familiarization protocol F1 or F2.*

F2 has a positive effect on donning performance. All participants who experienced F2 reached a duration lower than 100s. The manufacturer's brochure F1 is much less efficient and not enough to reach the certification level (100s).

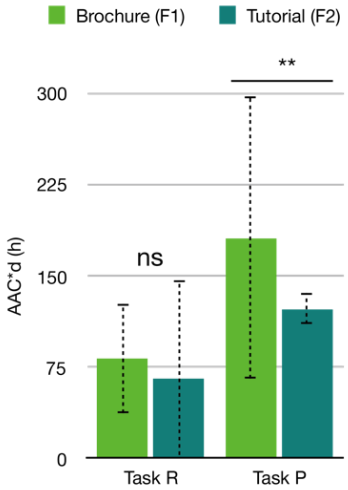



*Figure 8. Evolution of mean CCA*d (h) for Task R and task P according to familiarization protocol F1 or F2.*

*Global physical workload*

Hypothesis: F2 allows to have a lower physical strain by optimizing installation, adjustment, and use. If experiencing F2, ACC*d is reduced by 64.28 h ±80.3 for the task P with p=0.008. The decrease for task R is not significant, as shown in figure 8.

These results showed a higher reduction of global strain while experiencing F2 protocol. We can conclude that F2 had a positive effect on global strain for both tasks.

*Effectiveness on task performance*

Hypothesis: Performance is better when a participant has been familiarized with expert tutorial (F2). The evolution of the number of actions and error for the repetitive task R with familiarization protocol (F1 or F2) is shown in figure 9. The highest number of valid actions (49.22±7.62, $p<0.0001$) and the lowest number of errors (3.52±1.61, $p<0.0001$) were found when experiencing the expert tutorial F2.

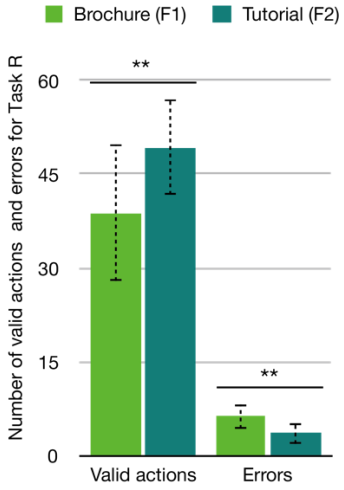



*Figure 9. Task performance indicators and errors for repetitive task R according to familiarization protocol (F1 or F2). Brackets indicate significant differences between F1 (manufacturer's brochure) and F2 (expert tutorial) condition.*

Results for the precision task P with familiarization protocol (F1 or F2) are shown in figure 10. The highest number of valid signs (52.92±7.93, $p<0.0001$) and the lowest average number of errors (3.81±2.55, $p<0.0001$) were found when wearing the exoskeleton.

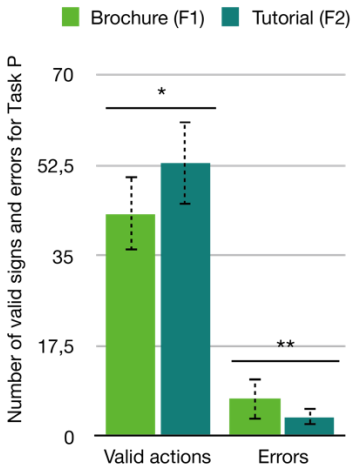



*Figure 10. Task performance indicators and errors for precision task P according to familiarization protocol (F1 or F2).*

We conclude that for all tasks, F2 has given a better Human-Exoskeleton performance than manufacturer's brochure F1with a higher number of actions and a lower number of errors.

**Perceptive results**

*Physical, Cognitive and Occupational aspects*

Hypothesis: F1 protocol produces lower perceived benefits, usability and acceptance score than F2 protocol. The effectiveness of familiarization protocol (F1/F2) on user's perception is verified by two-samples two-sided T-test to compare the means of these two groups. The results are presented in Table 5. Higher scores given on Likert scale (0-10) have been found when participants experienced F2 familiarization protocol. The most significant differences were found in this order for easiness of learning (donning and adjusting) with an increased score of +4.15/10, comfort (+3.36/10), easiness to move with (+2.95/10), focus demand (+2.63/10). They are shown in bold in Table 2 with all variables.

*Questions (Likert scale 0-10):*
*To learn how to don/adjust the exoskeleton is easy (totally agree-strongly disagree)*
*To master the exoskeleton is easy (totally agree-strongly disagree)*
*For the tasks, the exoskeleton benefits are (marginal- extremely important)*
*To learn how to move with the exoskeleton is easy (totally agree-strongly disagree)*
*The exoskeleton is comfortable (extremely uncomfortable- extremely comfortable)*
*With the exoskeleton, I need extra focus demand (strongly disagree-totally agree)*
*With the exoskeleton, I feel (much less effective-much more effective)*
*I can perform the task with the same quality (totally agree-strongly disagree)*

*User acceptance*

Hypothesis: the user's acceptance score is higher when experiencing F2. Acceptance is scored through a three-dimensional question: Q1: 'My global satisfaction for the exoskeleton is (extremely low- extremely high), Q2: 'If needed, I would use the exoskeleton (Never-Everyday), Q3: I would recommend the exoskeleton to a colleague (Not at all- absolutely). The validity of three questions toward a global Acceptance dimension is verified by alpha's Cronbach >0.80.

*Table 2. Descriptive statistics (mean, standard deviation, difference, p-value) and comparison of familiarization protocol (F1 or F2) on perceived benefits and acceptance dimensions (*p < .05).*

| Dimension | Brochure (F1) | Tutorial (F2) | Difference, p value |
|---|---|---|---|
| Easiness of learning (donning and adjust) | 4.38 (2.49) | 8.18 (1.29) | 4.15, <0.0001 |
| Perceived support | 6.06 (2.22) | 7.29 (2.02) | 1.24, 0.001 |
| Master demand | 5.11 (2.33) | 3.03 (1.92) | 2.09, <0.0001 |
| Perceived global strain | 5.29 (1.57) | 3.81 (1.37) | 1.48, <0.0001 |
| Easiness to use | 5.81(2.28) | 7.86(2.15) | 2.04, <0.0001 |
| Easiness to move with | 4.91 (2.62) | 7.86 (1.92) | 2.95, <0.0001 |
| Comfort | 4.53 (2.10) | 7.88 (2.23) | 3.36, <0.0001 |
| Focus demand | 5.52 (3.03) | 2.88 (2.67) | 2.63, <0.0001 |
| Performance | 6.28 (2.31) | 8.11 (1.77) | 1.83, <0.0001 |
| Respect of quality | 6.34 (2.52) | 8 (1.86) | 1.65, <0.0001 |
| Acceptance | 6.42 (1.99) | 8.25 (1.70) | 1.82, <0.0001 |

We conclude that for all aspects presented (Cognitive, Occupational and Physical), F2 protocol has given a better perceived performance, benefits and user acceptance than the manufacturer's brochure (F1). Human-Exoskeleton performance could be significative influenced by the familiarization experience that includes different type of knowledge and practice.

**Discussion and Conclusion**

Firstly, some interesting contributions to Human-Exoskeleton Interaction on simulated industrial tasks have been found. Significant positive effects have shown a reduction in Global physical workload and perceived strain, an increase in task performance, in relation to positive effects on subjective benefits as perceived performance, the respect of quality standards and the lack of extra focus demand. These positive effects on physical, cognitive and occupational aspects are strategic to ensure occupational exoskeleton adoption in industries. Also, if the expected reduction of perceived strain is significant in targeted muscles (shoulder, arms), some muscular strain increased while wearing exoskeleton and highlights the possible influence of load transfer that should be investigated. A further study could aim at simulating muscle activation of the Human-Exoskeleton system to better understand this effect. Secondly, a key finding of this study is a significant positive effect of an expert familiarization protocol on perceived benefits, usability and user acceptance. These results suggest that the use of exoskeleton is not intuitive. A familiarization experience that includes specific knowledge and practice could help optimize Human-Exoskeleton performance and user acceptance, that could eventually lead to a quicker adoption in companies. It is not easy to study the familiarization process as it is related to time. And long experiments would not be appropriated as they would involve participants to endure high strains. The suggested laboratory protocol is easily repeatable and allows the test of familiarization dimensions using a short duration of physio pathogenic activity. Further work could deal with the influence of panel diversity that has not been taking into account in this study. Also, differences of effects on all variables could be investigated, to bring manufacturer interesting feedbacks on the effect of claimed design improvements from Exo A to B prototypes.

**References**

Baltrusch, S.J., Van Dieën, J.H., Van Bennekom, C.A.M., & Houdijk, H. (2018). The effect of a passive trunk exoskeleton on functional performance in healthy individuals. *Applied Ergonomics, 72*, 94-106. https://doi.org/10.1016/j.apergo.2018.04.007

Bosch, T., van Eck, J., Knitel, K., & de Looze, M. (2016). The effects of a passive exoskeleton on muscle activity, discomfort and endurance time in forward bending work. *Applied Ergonomics, 54*, 212-217. https://doi.org/10.1016/j.apergo.2015.12.003

De Looze, M.P., Bosch, T., Krause, F., Stadler, K.S., & O'Sullivan, L.W. (2016). Exoskeletons for industrial application and their potential effects on physical work load. *Ergonomics, 59*, 671-681. https://doi.org/10.1080/00140139.2015.1081988

Descatha, A., Roquelaure, Y., Chastang, J.-F., Evanoff, B., Melchior, M., Mariot, C., Ha, C., Imbernon, E., Goldberg, M., & Leclerc, A. (2007). Validity of Nordic-style questionnaires in the surveillance of upper-limb work-related musculoskeletal disorders. *Scandinavian Journal of Work, Environment & Health, 33*, 58-65.

Graham, R.B., Agnew, M.J., & Stevenson, J.M. (2009). Effectiveness of an on-body lifting aid at reducing low back physical demands during an automotive assembly task : Assessment of EMG response and user acceptability. *Applied Ergonomics, 40*, 936-942. https://doi.org/10.1016/j.apergo.2009.01.006

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklade, A.L., & Christ, R.E. (1992). Comparison of Four Subjective Workload Rating Scales. *Human Factors, 34*, 429-439. https://doi.org/10.1177/001872089203400405

Huysamen, K., Bosch, T., De Looze, M., Stadler, K.S., Graf, E., & O'Sullivan, L. W. (2018). Evaluation of a passive exoskeleton for static upper limb activities. *Applied Ergonomics, 70*, 148-155. https://doi.org/10.1016/j.apergo.2018.02.009

Kim, S., Nussbaum, M.A., Mokhlespour Esfahani, M.I., Alemi, M.M., Alabdulkarim, S., & Rashedi, E. (2018). Assessing the influence of a passive, upper extremity exoskeletal vest for tasks requiring arm elevation : Part I – "Expected" effects on discomfort, shoulder muscle activity, and work task performance. *Applied Ergonomics, 70*, 315-322. https://doi.org/10.1016/j.apergo.2018.02.025

Meunier, P. (2014). *Cardiofréquencemétrie pratique en milieu de travail : Une approche objective de la pénibilité professionnelle*. Éd. Docis.

Moyon, A., Petiot, J.-F., & Poirson, E. (2019). Development of an Acceptance Model for Occupational Exoskeletons and Application for a Passive Upper Limb Device. *IISE Transactions on Occupational Ergonomics and Human Factors, 7*:3-4, 291-301, https://doi.org/10.1080/24725838.2019.1662516

Moyon, A., Poirson, E., & Petiot, J.-F. (2018). Experimental study of the physical impact of a passive exoskeleton on manual sanding operations. *Procedia CIRP, ELSEVIER, 70* (pp. 284-289). https://doi.org/10.1016/j.procir.2018.04.028

SkelEx (Rotterdam, Netherland.). https://www.skelex.com/ consulted 30.05.2019

Theurel, J., & Desbrosses, K. (2019). Occupational Exoskeletons : Overview of Their Benefits and Limitations in Preventing Work-Related Musculoskeletal Disorders. *IISE Transactions on Occupational Ergonomics and Human Factors, 1-17.* https://doi.org/10.1080/24725838.2019.1638331

**Appendix**

Human-Exoskeleton Familiarization Levels. According to our field expertise, a certified user should reach at least level 4.

**HEFL**

| | Human-Exoskeleton Familiarization levels | Description |
|---|---|---|
| 0 | No previous experience with the exoskleleton. It is an unkown system. | The worker has no previous experience with the exoskeleton neither knowledge of the system. His perception can be biased by' science fiction effect' (movies, advertising...). |
| 1 | First visual contact. Basic knowledge. | The worker has no experience with the exoskeleton but a basic knowledge of general principles. His perception can be biased by' science fiction effect' (movies, advertising...) |
| 2 | Knowledge of technological basics and utility (case studies). | The worker has basic knowledge of general and technical principles. Limits and application have been explained. A test of donning/adjusting may have been tried. |
| 3 | Knowledge of technological basics and utility (case stuies). Donning/adjudting/doffing are mastered. | The worker knows basic, mechnical and adjustment principles. He knows how to donn/adjust/doff the exoskeleton within 100 seconds with safety checks, including experience of limitations. |
| 4 | Knowledge of risks. Subjective validation of the exoskeleton on simple simulated tasks. | The worker has all previous knowledge and experience and he has felt potential assistance and limitations with a tailored protocol (simulation of case study tasks) in the range of assistance. |
| 5 | Subjective validation of the assistive device with an intermediate test out of the work situation and in real situation. | The worker has all previous knowledge and experience and he has felt potential assistance and limitations with simulation protocol of minimum 15 minutes. He has positively evaluated the device (acceptance score >7/10). |
| 6 | Subjective validation of the assistive device with an long test out in real situation. | The worker has all previous knowledge and experience and used the exoskeleton at least for 48h (2h/day). He has positively evaluated the device (acceptance score >7/10). |
| 7 | Subjective validation of the assistive device with an long test out in real situation. | The worker has a good knowledge on use principle and a minimum experience of 3 weeks (4h/day max). His subjective evaluation is globally positive and effects on eventual changes in organisation/process/habits have been adaopted. |

# Why is circular suturing so difficult?

*Chloe Topolski[1,2], Cédric Dumas[2,3], Jerome Rigaud[4], & Caroline G.L. Cao[1,2,3]*
*[1]Wright State University,[2]IMT Atlantique Bretagne-Pays de la Loire*
*[3]Laboratoire des Sciences du Numérique de Nantes*
*[4]Centre Hospitalier Universitaire de Nantes*
*[1]United States of America, [2,3,4]France*

**Abstract**

Suturing is a basic surgical skill that requires much training to achieve competency. Circular suturing is even more challenging, especially in minimally invasive surgery. In a radical prostatectomy procedure, circular suturing is performed to reconnect the bladder and urethra after the prostate has been removed. Task analysis of linear suturing and circular suturing, in laparoscopic and robot-assisted surgery, was performed and validated. Results revealed that circular suturing involves more motoric and perceptual constraints than linear suturing, requiring depth perception for proper alignment of two differently sized circular structures. Robotic surgical systems such as the da Vinci Surgical System can reduce some of these constraints by providing a stereoscopic view of the circular structures and increasing the manipulability of the needle and tissue, compared to the laparoscopic approach. These findings have implications for the design of training and assessment, as well as assistive tools to enhance the performance of circular suturing.

**Background**

In surgery, suturing is performed to close incisions or gaps in the anatomy when diseased tissue has been removed. Suturing is one of the most difficult basic technical skills in surgery (Ghazi & Joseph, 2018). It requires hand-eye coordination, dexterity and precision to place evenly spaced stitches with equal tension to achieve good approximation of tissue (Secin et al., 2006). In minimally invasive surgery such as laparoscopic surgery, intracorporeal suturing is even more difficult due to the limited degrees of freedom in manipulation and constrained space (Cao et al., 1996). In laparoscopic surgery, 4 or 5 small incisions are made in the abdomen into which the laparoscopic instruments are inserted. The tools are long and thin in order to fit into small incisions while still reaching the desired points inside the body. The insertion point creates a fulcrum effect which forces the surgeons to move their hands in the opposite direction they want the end-effector of the tool to move. This skill is non-intuitive and complicates the procedure for surgeons. The surgical site, provided by an endoscope which is also inserted into the abdomen through an incision, is displayed on monitors around the operating room for the surgical team. This 2D view of the surgical field makes it difficult to manoeuvre within a 3D space. Overall, these constraints can complicate many surgical tasks, especially intracorporeal suturing.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In certain cases, suturing may be required around circular anatomical structures. For example, in urology, after a radical prostatectomy (complete removal of the prostate) is performed to reduce the risk of cancer or to mitigate the spread of cancerous cells, the urethra and bladder neck are joined together with sutures in a process called the urethrovesical anastomosis. This anastomosis involves circular suturing and is considered to be the most difficult part of the entire operation (Ghazi & Joseph, 2018).

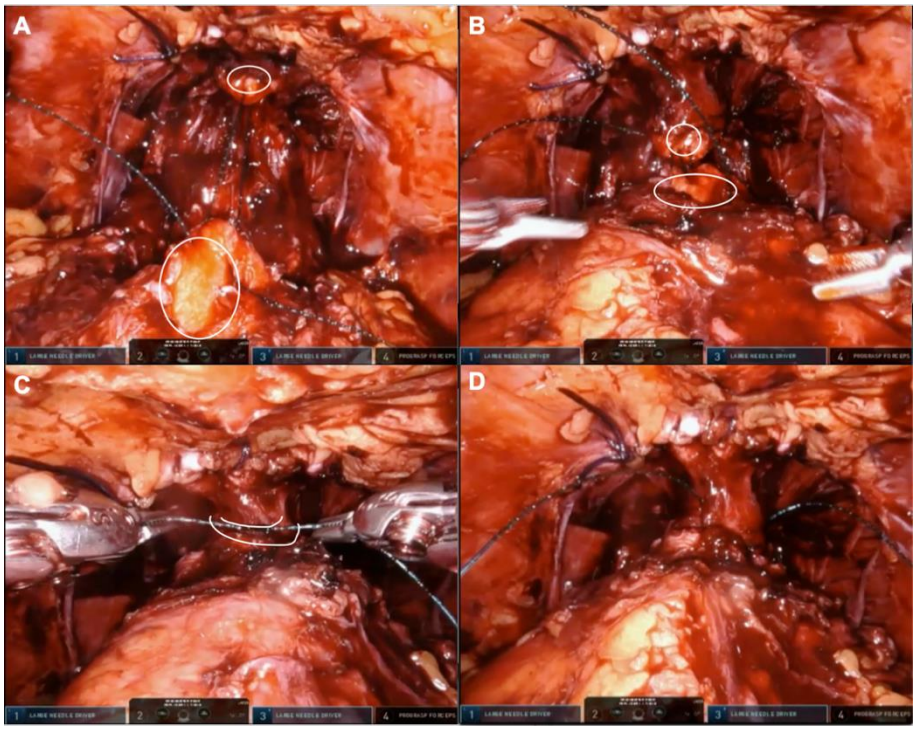



*Figure 1. Illustration of four different stages of circular suturing in an urethrovesical anastomosis. As the surgeon progresses, the urethra (indicated by small white circle in A) and bladder neck (indicated by large white circle in A) are brought closer together (part B) and joined (part C) and secured (part D), thus completing the anastomosis.*

The urethrovesical anastomosis involves the joining of the ends of two tubular structures –the urethra and the bladder (see Figure 1). This means that the surgeon must suture around the outside circumference of both tubes to ensure the tissues are securely connected while still allowing fluid to pass through the lumen of the tubes. As this method differs from the more common linear suture where the stitches are made across a straight line, a drastically different technique is needed. The intricacies of these different tasks are outlined in many surgical texts but are not explained in detail. Novice surgeons have to rely on guided training with expert surgeons in order to fully grasp the concepts and methods of circular suturing that make it so

challenging. Not only is the task difficult to learn, it is also difficult to teach to novice surgeons, especially in the minimally invasive approach.

Surprisingly, the robotic surgical system da Vinci (Intuitive Surgical, Inc.; Figure 2) that had been struggling to demonstrate value in laparoscopic surgery provided the solution to this difficult urological procedure. In fact, the use of the da Vinci Surgical System in urological procedures increased from 8% in 2004 to 67% in 2010 and is now used in more than 70% of prostatectomy procedures (Voilette et al., 2015).

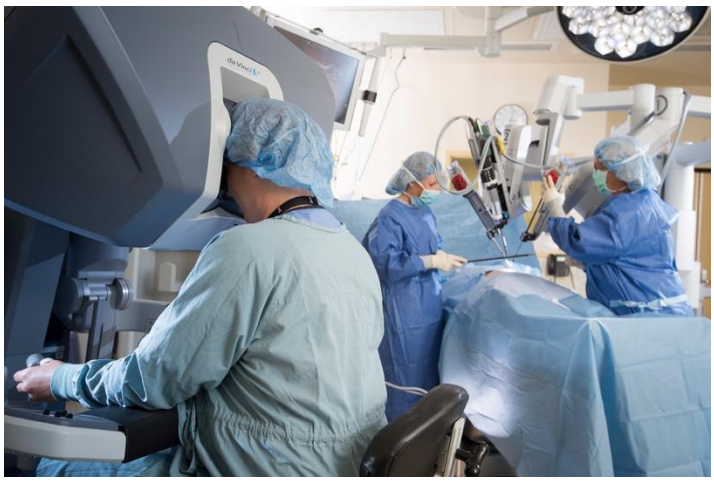


*Figure 2. The da Vinci Surgical System includes a control console where the surgeon is seated (left) and surgical instrument dock that is positioned over the patient (right). Image from: https://www.franciscanhealth.org/health-care-services/robotic-assisted-surgery-334*

The robotic surgical system, da Vinci Surgical System, provides the surgeon with a stereoscopic view of the surgical field while being positioned in an ergonomic seat. The joysticks and pedals included on the control console allow the surgeon to control all of the tools connected to the surgical instrument dock quickly and easily.

Additionally, the joysticks allow the surgeon to control more intricate movements of the surgical instruments such as graspers and scissors. With the da Vinci, these tools have more degrees of freedom than traditional laparoscopic tools (Figure 3). The wrist-like joints on the da Vinci-compatible tools allow the surgeon to more easily manipulate tissue or other medical equipment such as sutures (Chellali et al., 2014). Nevertheless, the task of suturing, and in particular, circular suturing, in the minimally invasive environment remains challenging.

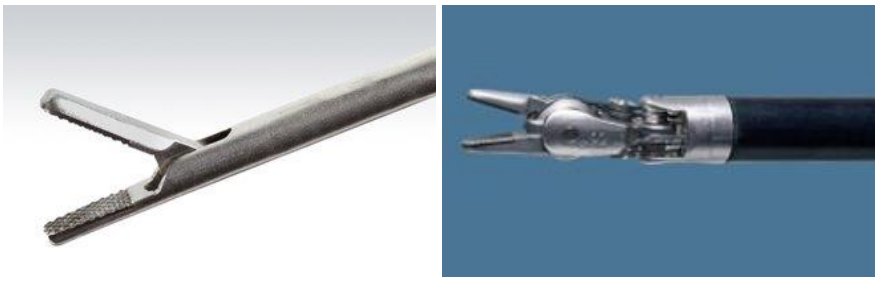


*Figure 3. Laparoscopic needle drivers (left) and da Vinci laparoscopic needle drivers (right) are similar in design, but the da Vinci tools have significantly more degrees of freedom with their included wrist-like joints. Images from: Microlap® Needle Drivers. ConMed https://www.conmed.com/products/laparoscopic-robotic-and-open-surgery/instruments/low-impact-laparoscopic-instruments/low-impact-needle-drivers (left) Endowrist® MEGA™ Needle Driver. Intuitive Surgical https://www.intuitivesurgical.com/test-drive/pages/endowrist-instruments.php (right)*

Nevertheless, the robotic system has not been able to completely nullify the difficulties inherent to the urethrovesical anastomosis, such as bimanual dexterity in instrument manipulation (Chen et al, 2018). While the da Vinci® has no doubt improved many aspects of minimally invasive surgery (Ballantyne, 2002), the urethrovesical anastomosis still proves to be a challenging task for many surgeons.

This study is the first step towards an understanding of the requirements and constraints in circular suturing for the purpose of surgical skills training, as well as for developing an objective assessment metric for circular suturing performance. Ultimately, an assistive tool may be developed to make explicit the requirements to augment the performance of novice and expert surgeons alike.

## Materials and methods

### *Data collection*

To gather initial information about circular suturing tasks, ten surgical texts and manuals were consulted and reviewed to learn the basic steps necessary to complete a urethrovesical anastomosis procedure (Croce & Olmi, 2000, Davis, 2016, Ghazi & Joseph, 2018, Hudgens, 2015, Johnson & Cadeddu, 2019, Joseph, 2008, Lierse, 1987, Secin et al. 2006, Sundaram et al., 2010, Yuh & Gin, 2018). Observation and recording of five robot-assisted radical prostatectomy surgeries procedures were completed at the Centre Hospitalier Universitaire de Nantes, supplemented by 12 videos of the same surgery found online from other hospitals and training programs. The live procedures ranged from 1.5 hours to 6 hours in duration. The online videos were a mix of laparoscopic or robot-assisted radical prostatectomies; each video averaged around two hours long. Surgeon consent was obtained for the operating room observation portion of the process. Visual recordings of the live observations were taken from the da Vinci® intraoperative camera; no patient data or audio were included in the recordings.

Four expert surgeons were interviewed. All surgeons consented to being video recorded as they were interviewed. The interview consisted of three main portions: review of a pre-selected video, a structured interview, and reviewing the hierarchical task analysis diagrams. First, the surgeons were asked to observe a video of an expert completing an urethrovesical anastomosis and make comments throughout the video relating to technique and procedure (Mollo & Falzon, 2004). Next, the interviewer asked questions about certain aspects of the procedure and the surgeon's past experiences with the procedure. Finally, the surgeons were asked to review the four task analyses and verify the content and sequence of steps.

*Data analysis*

A task analysis was performed following the procedure in Cao et al. (1996) and four hierarchical task analysis (HTA) diagrams (linear and circular suturing, and laparoscopic and robotic suturing) were constructed to match the techniques observed in the operating rooms. All HTA were validated by four expert surgeons.

A cognitive task analysis was performed by interviewing four expert surgeons at the Centre Hospitalier Universitaire de Nantes in Nantes, France. The transcripts of each of the interviews was synthesized to extract common themes based on the language used. This information was organized and classified to supplement the HTA. By doing this, it became easier to address the specific differences in each of the tasks and which steps of the tasks were more difficult overall.

**Results**

*Hierarchical Task Analysis*

Figures 4-7 show the hierarchical decomposition of the four suturing tasks: laparoscopic linear suturing, laparoscopic circular suturing, robotic linear suturing, and robotic circular suturing. Comparing linear and circular suturing, the first sublevel of the task decomposition was similar; this sublevel contained six to seven steps. The only difference was between circular and linear suturing where two steps were needed to penetrate the tissue since there are two distinct structures to pass the needle through. Distinct differences appeared in the second sublevel of the task decomposition. Circular suturing was more complex than linear suturing, requiring more sub-tasks that were not necessary for the linear suture.

When comparing the robotic approach with the laparoscopic approach, the task decomposition showed that in many of the second-level subtasks, the robotic approach was less constrained than the laparoscopic approach. In the robotic approach, it was not necessary for the needle to be set as meticulously as in laparoscopy since the robot wrist motions can adapt easily to different angles. While there were notable differences in the content of the subtasks, the procedure ultimately remained very much the same.

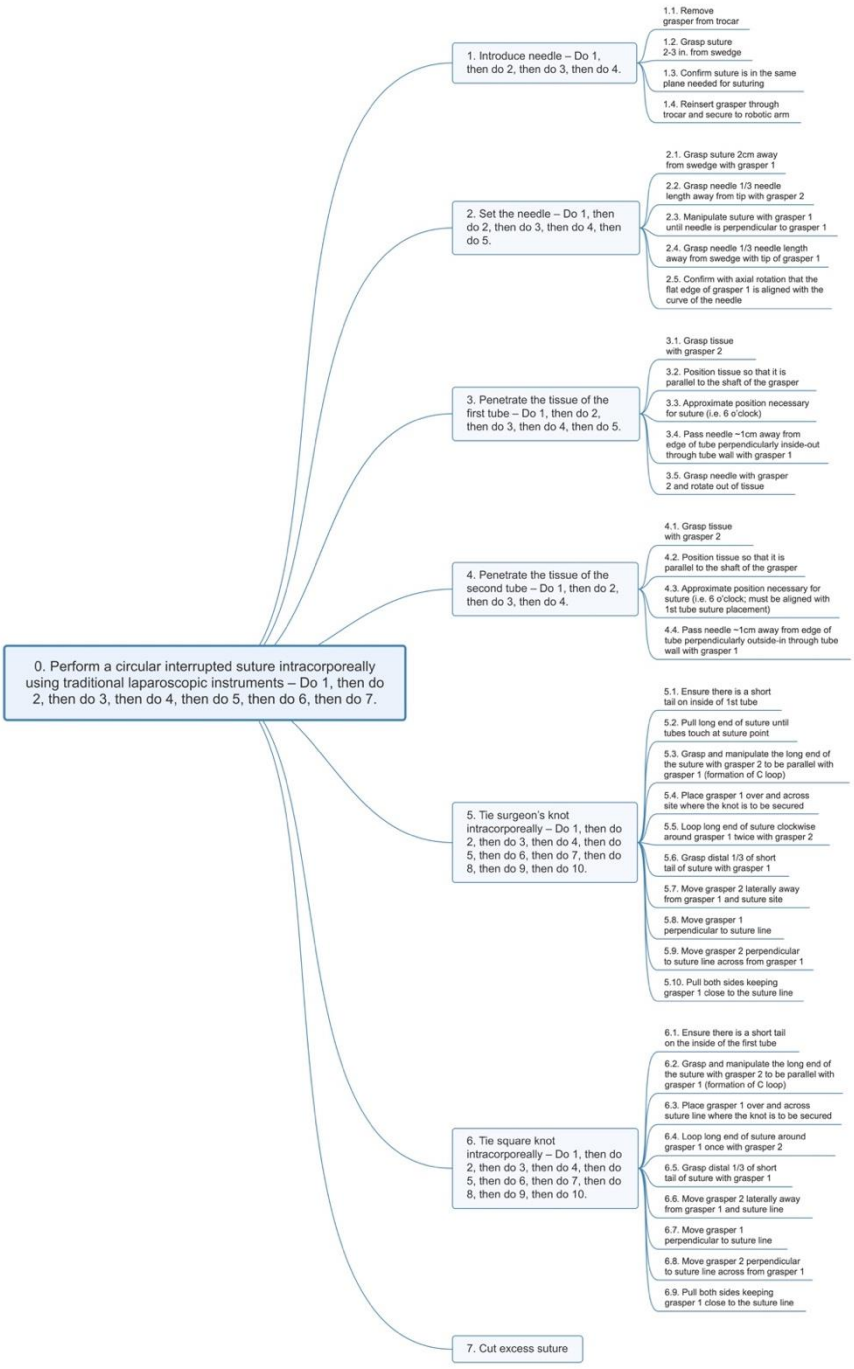



*Figure 4: HTA of a circular suturing task using the laparoscopic approach. There are seven first-level subtasks and 37 second-level subtasks included in the diagram, all of which are necessary to perform a circular suture using this approach.*

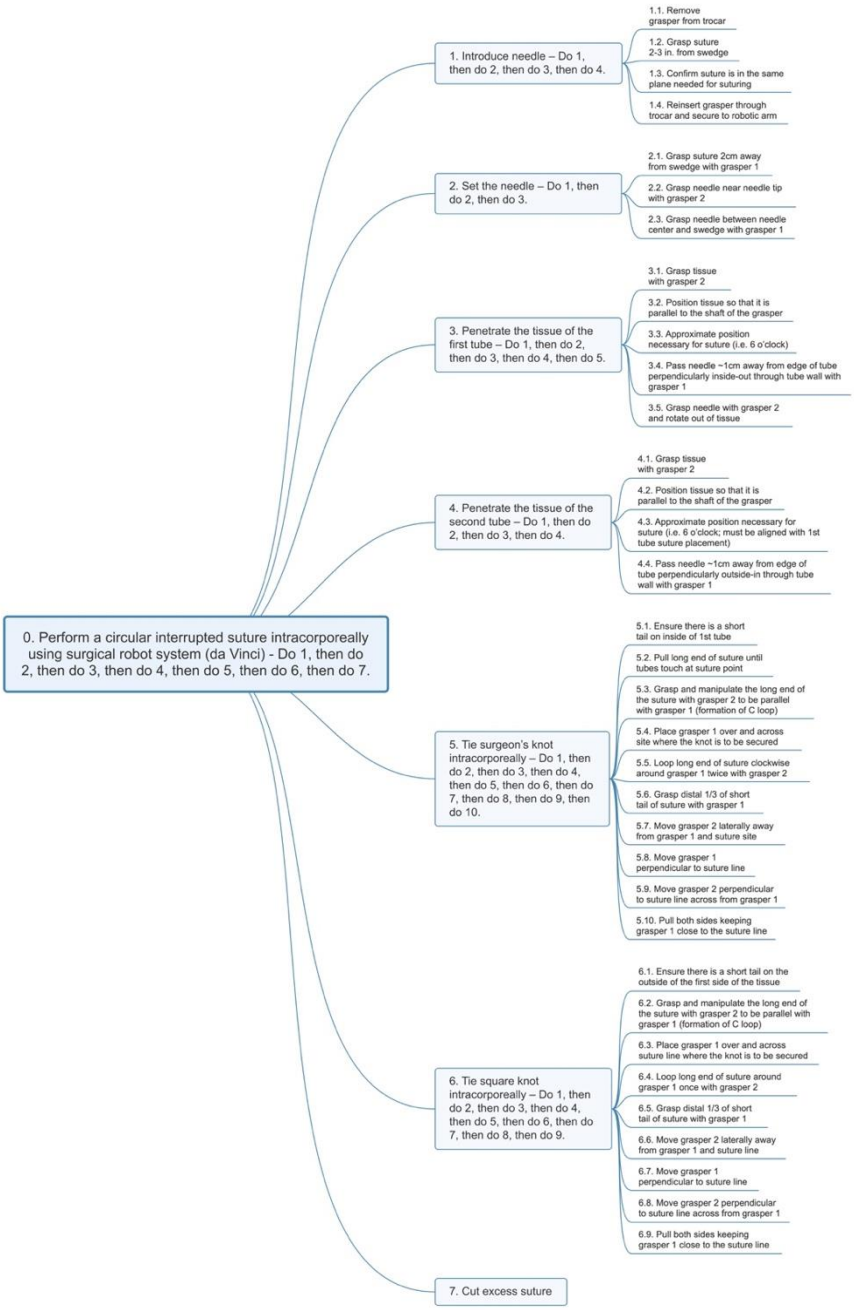



*Figure 5: HTA of a circular suturing task using the robot-assisted surgical approach. The second level not only has 12 fewer subtasks than the laparoscopic approach, but the tasks are also simpler and less exigent.*

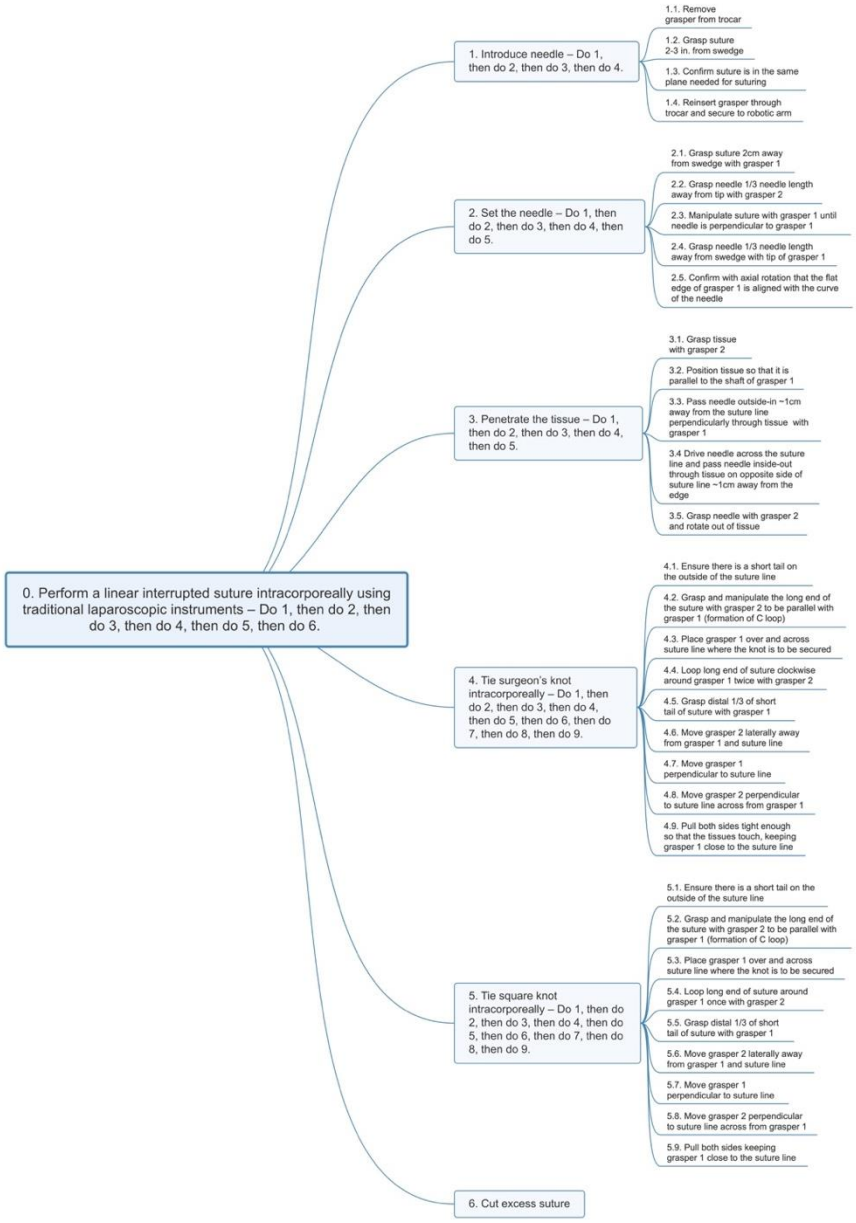



*Figure 6: HTA of a linear suturing task using the laparoscopic approach. There are 6 first-level subtasks and 32 second-level subtasks necessary in order to complete a linear suture using this approach.*

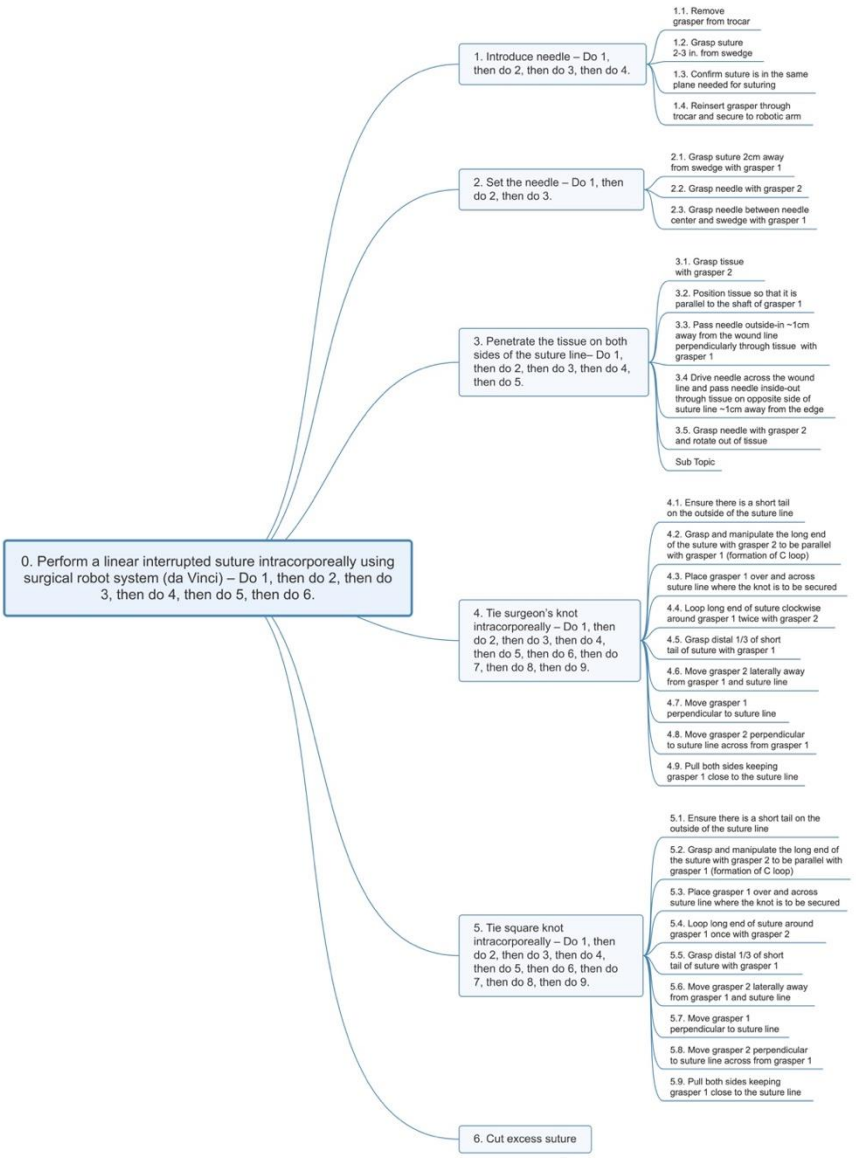



*Figure 7: HTA of a linear suturing task using the robot-assisted surgical approach. This approach has 2 fewer subtasks than the laparoscopic approach and is lower in complexity in the "set the needle" task.*

## *Cognitive Task Analysis*

Tables 1-3 summarize the results of the cognitive task analysis. Task requirements and constraints were abstracted from the interviews and classified into two levels of abstraction: execution (skills) and planning. The execution or skill of the surgeon was further broken down into two more levels: motor movements and perception. Table 1 reveals the additional degrees of freedom that the robotic system afforded in manipulating tissue and orienting the needle. Table 2 reveals additional requirements for the circular suturing task, such as the changing orientation of the needle for each stitch, which align with the capabilities of the robotic system in Table 1. Finally, the need to visualize and plan extensively in circular suturing compared to linear suturing is summarized in Table 3. Notably, the placement of the stitches in circular suturing required mental imagery in planning, and constant adjustments during execution.

*Table 1. Comparing laparoscopic and robot-assisted suturing techniques.*

| *Laparoscopic* | *Robot-assisted* |
|---|---|
| Few degrees of freedom – one axis of rotation | More degrees of freedom – wrist motion extremely helpful for needle orientation |
| Better for linear sutures, circular sutures become more difficult with changing angles of insertion | Can easily adapt to linear or circular sutures |
| Orientation of needle in grasping tool critical | Orientation of needle in grasping tool not as important |
| 2D view of surgical field lacking depth for circular suturing | High-definition and stereoscopic view of surgical field good for circular suturing |

*Table 2. Comparing the execution tasks (motor movements) of linear and circular suturing.*

| *Linear* | *Circular* |
|---|---|
| Angle of insertion remains consistent | Angle of insertion changing |
| Alignment of needle the same for each stitch | Alignment and orientation of needle has to be varied precisely |
| Easy alignment, no concern with twisting or stretching | Different size circumference of openings complicates alignment |
| Can most often use dominant hand to do majority of suture | Required to use left and right hand with same amount of dexterity |

*Table 3. Comparing the planning tasks (perceptions) of linear and circular suturing.*

| *Linear* | *Circular* |
|---|---|
| Visualizing placement of suture based on last stitch/set measurement (i.e. 0.5 cm) is very simple | Placement of suture depends on size/shape of tissue and relative difference of size of openings |
| Only have to use one needle | Using and monitoring two needles |
| Can easily anticipate where needle emerges from tissue; mostly driving toward camera | Difficult to see where needle will emerge especially when driving needle away from camera |
| Can often be completed with one grasper, no alternative for manipulation around suture site | Passing and manoeuvring the needle with both left and right graspers – must decide when to switch and how |

**Discussion**

From the hierarchical task analysis alone, it is not clear why circular suturing is more difficult than linear suturing. Even though there are differences in the number of subtasks at the second level of task decomposition, the differences seem minor as the suturing tasks follow the same technique of needle insertion-needle pull through-suture pull through-repeat needle insertion. Similarly, whether the suturing is performed laparoscopically or with the robotic system, the steps and subtasks are similar, further confirming that these different approaches follow the same technique in performing a suturing task.

While the execution steps used in linear and circular suturing are essentially the same, the cognitive task analysis revealed marked differences at the execution and planning levels. As linear suturing involves working in one plane, the angle of needle insertion remains consistent for all stitches. In circular suturing, however, the angle of insertion changes with each subsequent stitch. This varying angle of the needle must vary with the tangent of the curve around the circular structure.

Additionally, in urethrovesical anastomosis, the two structures being sutured together have different circumference which complicates the alignment process. Linear sutures which often bring two pieces of tissue together in the same plane are easy to align without any stretching or twisting. In circular suturing, the surgeon must also be able to use both the left and right tools with the same amount of dexterity. A linear suture can often be completed entirely with one hand, while both hands are need to achieve multiple angles of the needle in circular suturing.

Not only is circular suturing more difficult in terms of motor control, but perceptual constraints also play a major role in how a circular suture is completed. In linear suturing, visualizing where the needle should be placed next, based on the position of the previous suture, is relatively easy. However, in the anastomosis task, the positioning of the structures, as well as the difference in size of the structures, makes it more difficult to determine where the next stitch should be placed. Circular suturing

most often involves using two needles and keeping track of these needles and sutures can become confusing. Additionally, visualizing these two needles around the circumference of the bladder neck can be difficult. As the surgeon has to drive the needle through the back of the bladder neck, away from the camera, to a point occluded by tissue, where the needle exits the tissue is often a matter of guessing.

The planning process throughout all of these steps also changes between linear and circular suturing. For example, the spacing of stitches in a linear suture can be pre-determined based on the length of the suture, such as 5 mm. For a circular suture, the spacing is different on each of the two structures to be joined, due to their size difference. The corresponding stitches on the bladder neck and the urethra must align to ensure an even and tight closure. The passing and manipulating of the needle also require more planning and adjustments in a circular suture. While a linear suture can be conducted simply with one grasper, a circular suture requires the surgeon to decide when to switch directions, when to switch needles, and when to switch hands and grasps to maintain the optimal physical control over the process.

Clearly, many of these requirements are being addressed by the increased degrees of freedom in the surgical robot. Laparoscopic tools are very rigid compared to the robotic end-effectors; the wrist motion of the robotic tool allows for easier needle manipulation that is crucial in circular suturing. Laparoscopic instruments are adequate in linear suturing where the suture is only being applied across a single plane of tissue. However, in circular suturing where the plane of action is constantly changing, the wrist motion of the robotic tools allows the surgeon much more freedom. The setting of the needle in robot-assisted surgery is not as strict as it is in laparoscopic surgery because the wrist motion allows for rotation in different directions rather than just the one axis of rotation that the laparoscopic tools offer. Presumably, the increased degrees of freedom allow for more dexterity, hence usability (Chellali et al, 2014). Another major benefit of the robotic system is the stereoscopic view provided in the operational console. This stereoscopic view is useful in visualizing the circular structures. Laparoscopic screens display the surgical site in 2D only, not allowing the surgeon to have accurate depth perception within the surgical field. It is also possible that surgeons' situation awareness may be limited by the 2D view. However, we did not examine this dimension of the problem.

*Considerations for future work*

What is not included in this analysis is the timeline of each approach for the suturing task. A separate timeline analysis, in combination with the task analysis, would more precisely reveal which subtask is time-consuming or which subtask is more difficult.

Current teaching materials for minimally invasive linear suturing may be adequate for teaching the order of steps when adapted for circular suturing. However, it is clear that there are additional perceptual and motoric requirements that need to be included in the training instructions. More explicit instructions can be developed for training, as well as for evaluation of performance in circular suturing.

## Conclusion

In both laparoscopic and robot-assisted minimally invasive surgery, circular suturing is considered a challenging task to teach and to learn. The joining of the bladder and urethra after a radical prostatectomy procedure is just one example of this type of task. In this study, analysis of four different intracorporeal suturing approaches was conducted through observations of live surgeries, interviews, and video review with expert surgeons. The results of this analysis revealed that circular suturing requires depth perception and proper alignment of two differently sized circular structures, as well as additional motoric manipulations of needle and tissue. Utilizing robotic techniques can mitigate some of these constraints by providing a stereoscopic view of the surgical field as well as increasing the manipulability of both the needle and tissue. The ability to use mental imagery during the planning phase seems to be an important factor in the success of the task. These findings will inform future design of training and assessment methods, and assistive technologies for surgical performance.

## Acknowledgements

This project was made possible through funding from the FAME Cluster of the NExT Project in Pays de la Loire. The authors are grateful to Drs. Karam, De Vergie, and Nedelec of the CHU Nantes for their domain expertise.

## References

Ballantyne, G.H. (2018). Robotic surgery, telerobotic surgery, telepresence, and telementoring. *Surgical Endoscopy, 16*, 1389-1402.

Cao, C.G.L., MacKenzie, C.L., & Payandeh, S. (1996). Task and motion analyses in endoscopic surgery. In *ASME Dynamic Systems and Controls Division proceedings, (Fifth Annual Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems),Volume 58*, (pp. 583-590). Atlanta, Georgia USA.

Chellali, A., Schwaitzberg, S.D., Jones, D.B., Romanelli, J., Miller, A., Rattner, D., Roberts, K.E., Cao, C.G.L. (2014). Toward scar-free surgery: an analysis of the increasing complexity from laparoscopic surgery to NOTES. *Surgical Endoscopy 28*, 3119–3133.

Chen, J., Oh, P.J., Cheng, N., Shah, A., Montez, J., Jarc, A., Guo, L., Gill, I.S., & Hung, A.J. (2018). Use of Automated Performance Metrics to Measure Surgeon Performance during Robotic Vesicourethral Anastomosis and Methodical Development of a Training Tutorial. *Journal of Urology 200*, 895-902.

Croce, E. & Olmi, S. (2000). Intracorporeal Knot-Tying and Suturing Techniques in Laparoscopic Surgery: Technical Details. *JSLS : Journal of the Society of Laparoendoscopic Surgeons 4*, 17-22.

Davis, J.W. (2016). *Robot-assisted radical prostatectomy* New York, New York: Springer Science + Business Media.

Gill, I.S. & Hung, A.J. (2018). Use of Automated Performance Metrics to Measure Surgeon Performance during Robotic Vesicourethral Anastomosis and Methodical Development of a Training Tutorial. *Journal of Urology, 200*, 895-902.

Ghazi, A. & Joseph, J.V. (2018). The Urethrovesical Anastomosis in H. John and P. Wiklund (Eds.) *Robotic Urology* (pp. 375-389) Cham: Springer International Publishing.

Hudgens, J.L. (2015). Systematic Method in J.L. Hudgens and R.P. Pasic (Eds.) *Fundamentals of Geometric Laparoscopy and Suturing* (pp. 25-38) Tuttlingen, Germany: Endo Press.

Johnson, B.A. & Cadeddu, J.A. (2019). Radical Prostatectomy in *Robotic-Assisted Minimally Invasive Surgery* (pp. 239-247) Cham: Springer International Publishing.

Joseph, J. (2008). Vesicourethral Anastomosis in H. John, & P. Wiklund (Eds.) *Robotic Urology* (pp. 109-116). Berlin: Springer.

Lierse, W. (1987). Male Urethra in W. Lierse (Eds.) *Applied Anatomy of the Pelvis* (pp. 171-174) Berlin: Springer.

Mollo, V., Falzon, P. (2004). Auto- and allo-confrontation as tools for reflective activities. *Applied Ergonomics, 35*, 531-540.

Secin, F.P., Karanikolas, N., Touijer, K.,& Guillonneau, B. (2006). Laparoscopic Radical Prostatectomy: Techniques and Complications in S. Naito, Y. Hirao, and T. Terachi (Eds.) *Endourological Management of Urogenital Carcinoma* (pp. 129-145) Tokyo: Springer-Verlag.

Sundaram, C.P., Gjertson, C.K., & Koch, M.O. (2010). Laparoscopic and Robotic-Assisted Laparoscopic Radical Prostatectomy in S. Tsuda and O.Y. Kudsi (Eds.) *Essential Urologic Laparoscopy* (pp. 301-331) Totowa, New Jersey USA: Humana Press.

Violette, P.D., Mikhail, D., Pond, G.R., Paulter, S.E. (2015). Independent predictors of prolonged operative time during robotic-assisted radical prostatectomy. *Journal of Robotic Surgery 9*, 117-123.

Yuh, B., Gin, G. (2018). Robot-Assisted Radical Prostatectomy in Y. Fong, Y. Woo, W.J. Hyung, C. Lau, and V.E. Strong (Eds.) *The SAGES Atlas of Robotic Surgery* (pp. 113-125) Cham: Springer International Publishing.

# An extended version of the Dynamic Safety Model to analyse the performance of a medical emergency team

*Thierry Morineau[1], Cécile Isabelle Bernard[1], & Seamus Thierry[2]*
*[1]University of Bretagne Sud, Vannes*
*[2]Scorff Hospital, Lorient*
*France*

**Abstract**

The Dynamic Safety Model (DSM, Rasmussen, 1997) constitutes an original approach to safety issues. The model posits that adverse events are caused mainly by pressures coming from work constraints that lead operators' activity to migrate towards unacceptable limits of performance. In particular, Ramussen calls attention to the economic and workload pressures exerted on activity, insidiously pushing operators to tolerate risky behaviours as long as no critical event occurs. Recently, Morineau and Flach (2019) proposed to extend the DSM in order to integrate this model fully into the Cognitive Work Analysis (CWA) framework. More precisely, they suggested that the work domain analysis, the first stage of CWA, be replaced by the DSM. This use of the DSM would enable the analysis of intentional work systems involving loose coupling between work domain and organization. From this perspective, we present an analysis of the activity of a medical team confronted with a medical adverse event simulated in an emergency room.

**Introduction**

Research in cognitive systems engineering has developed a formative approach to analyse work systems. The basic assumption of this approach is that operators' behaviours are mainly shaped by work constraints, in the same way as animals in an ecosystem must adapt their behaviours to environmental features. At the methodological level, Cognitive Work Analysis (CWA) is the framework commonly used to describe behaviour-shaping constraints (Rasmussen, 1986; Vicente, 1999). It involves five embedded stages of analysis, namely work domain analysis (WDA) describing the constraints arising from the objects (domain) on which the work is performed; control task analysis describing constraints produced by the requirements to perceive and act on the work domain features; strategy analysis describing how performing tasks can be embedded in specific strategies, notably to manage the workload; organizational analysis focusing on workload allocation between human and/or artificial agents; and competencies analysis focusing on the individual inner constraints required to perform tasks.

Numerous studies have shown the relevance of CWA to apprehend work systems (e.g. nuclear plant, aviation, anaesthesia). Some studies indicate that this approach fits

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

particularly well with the analysis of causal work systems in which the work domain constraints directly drive the other embedded work constraints: tasks, strategies, team organization, and competencies (Hajdukiewicz et al., 1999; Wong et al., 1998). In a causal system, a tight coupling exists between work domain and work organization. In an intentional work system, outcomes particularly depend on how work is organized by operators through ad hoc decisions on priorities and adaptive processes to cope with the workload. In intentional systems, relationships between the work domain and the organization are mainly loosely coupled. Hence, in an intentional work system, a major issue for operators is to ensure efficient management of work constraints by coordinating the work requirements. This coordination will ensure that the organization's activity stays synchronized with the requirements imposed by the work domain state.

To analyse loosely coupled work systems, Morineau and Flach (2019) have proposed a new version of CWA, named "heuristic Cognitive Work Analysis" (hCWA). The specificity of this method resides in replacing the first stage of work domain analysis with a more modest, but heuristic, modelling of work constraints based on an extended version of the Rasmussen's Dynamic Safety Model (DSM). After outlining the DSM, we introduce hCWA and a first application on observations collected during medical emergency scenarios simulated in a high-fidelity simulation setting.

**The Dynamic Safety Model**

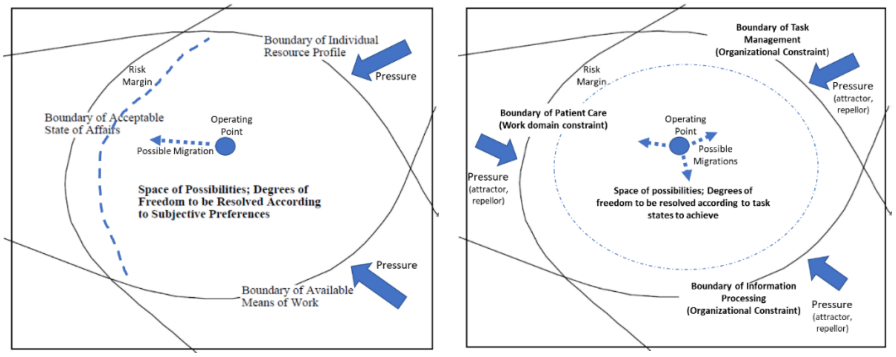
The level of coupling of a work system with its work environment constitutes the cornerstone that led Rasmussen (1990) to propose the DSM as a framework to analyse performance and safety issues. The tighter this coupling is, the more dependency relationships exist between events occurring in the work system.

At the lower level of human-machine interaction, tight links exist between operators and the work environment. Operators rely on the deterministic sequence of behaviours that the machine induces by its sequence of operations. At this level of granularity, sequential task analysis methods can be used to describe how users' behaviours more or less follow expected sequences, considering that deviations potentially represent less efficiency, error, or accident.

At the higher level of socio-technical systems, tight coupling can also be prominent if the work organization is based on a traditional way of working, whereby work processes are strictly decomposed as a set of sequences of discrete states to be reached. In this context, traditional accident analysis based on causal trees can be used to determine at which step operators violated some expectations, which led to the final accident or failure.

However, in modern work systems, automation or high demands of flexibility in activity provide operators with more degrees of freedom. At a high level of automation, the operators' job is to supervise automatic systems. When high flexibility in the work process is requested, the operators need to find ad hoc solutions. These degrees of freedom lead operators to use frequent decision making and implement adaptive strategies. Hence, the basic issue for operators is how to resolve numerous degrees of freedom in a space of possibilities bounded by a set of work

constraints that need to be complied with. In this space of possibilities, operators' activity can be modelled as an operating point with a trajectory in a workspace bounded by work constraints. This is the core of the DSM proposed by Rasmussen (Figure 1).




*Figures 1 and 2: A synthesis of the Dynamic Safety model inspired from Rasmussen (1990 & 1997, left side) and its extended version used in hCWA and applied to healthcare systems inspired from Morineau and Flach (2019, right side)*

Rasmussen has proposed different versions of the DSM. In his 1990 paper, Rasmussen presented the constraining boundaries as respectively referring to the "state of affairs" that corresponds to the state of the work domain, the "available means of work" (e.g. equipment), and the "individual resource profile", which is composed of operators' physiological and psychological capacities. These work constraints can produce pressures on operators that can potentially lead their trajectory to cross a boundary, leading to an accident or a problem. Based on these generic constraints, Rasmussen (1990) suggested to model activity respectively by identifying the space of possibility specific to the analysed work system, the subjective criteria used by operators to make decisions in order to solve degrees of freedom in their trajectories, the strategies used to synchronize with the work constraints, and the team organization aspects and the competencies needed to move the operating point within the workspace.

In comparison with this first generic approach to the workspace that potentially allowed it to be used as a sketch for analysis in relation with CWA, Rasmussen went further in the specification of the DSM in his 1995 and 1997 papers. The constraining boundaries were specified as referring to acceptable performance (firstly named "acceptable state of affairs"), economic cost, and individual workload. Organizations seek to limit the economic cost of activity and operators, their level of workload. The combination of these two pressures may critically and insidiously lead the operating point to migrate towards a single safety margin located near the acceptable performance boundary. To reduce this risk, proposals in safety science can be deployed to increase the safety margin by augmenting the work system reliability, increasing the operators' awareness of this risk, for instance through a safety management culture, or by making the boundary more visible, for instance with the ecological interface design (Vicente & Rasmussen, 1992).

In Cook and Rasmussen (2005), the DSM was used to interpret safety issues in hospitals. Through this modelling, the authors returned to the basic issue of model emergence by considering what happens if an adaptive modern work system uses tight coupling.

**Heuristic Cognitive Work Analysis (hCWA)**

Heuristic Cognitive Work Analysis is a methodological framework that is based on the first approach to the DSM proposed by Rasmussen (1990). The DSM is viewed as having a heuristic value for CWA. In hCWA, the first CWA stage of work domain analysis through an abstraction hierarchy is replaced with the DSM template. Rather than expanding the description of the work domain through an abstraction hierarchy that is particularly well-adapted for causal work systems, hCWA proposes to focus on the dynamics of activity triggered by the necessity to coordinate multiple conflicting constraints arising from the work domain and the organization, with multiple degrees of freedom to resolve in order to find the best adapted trajectory in the space of possibilities.

Figure 2 shows the extended version of the DSM used in hCWA that is specifically applied to medical work systems. Previous research has identified the following three constraints as specific to healthcare systems (Morineau et al., 2017):

- *Patient Care* is the work domain constraint for a healthcare system. Potentially, a patient can evolve towards a deteriorating state, thus putting pressure on the medical team;
- *Task Management* is an organizational constraint. It involves the manipulation of drugs and equipment during care delivery. These elements induce the performance of specific tasks to prepare, control, restore, or store them. To manage drugs and equipment is a peripheral activity for professionals who have been educated and trained mainly to deliver care. However, if these tasks are not performed well, they will produce risky pressure on activity;
- *Information Processing* represents the cost involved in processing information that is exchanged between operators and/or with information systems. Research in distributed and situated cognition has shown that much information processing and storage is performed in the work environment rather than exclusively in individuals' minds (Hollan et al., 2000). Similar to task management, information processing involves resources used to adapt to the work domain constraint (care delivery), but it can also represent a supplementary constraint for the cognitive workload, requiring communicating, reading and writing digital or paper documents. Difficulties or noise in information processing can drastically weaken operators' activity.

All these constraints can function as attractors or repellors for operators; concretely, they may lead to avoidance (repellors) or attraction (attractors) in the course of activity. Contrary to the original version of the DSM, all the work domain constraint can exert pressure on trajectories in the space of possibilities; thus, several forms of migration towards risky margins can occur inside the workspace. Facing these three basic work constraints, operators need to use their inner resources to manage the trajectory of their operating point in the workspace.

In hCWA, the DSM identifies the problem operators need to solve. Modalities to solve this problem can be found in the next analysis stages of CWA, namely control task, strategies, work organization, and competencies analyses.

**Analysis of simulated medical emergency events with hCWA**

We analysed two episodes of care delivery simulated in a high-fidelity simulation room: handoff and bed lowering to facilitate cardiac massage. The patient was represented as a realistic and interactive mannequin, including physiological parameters accessible on a monitor. Participants were professional nurses and nursing aids in the context of training sessions (more details can be found in Morineau & Flach, 2019).

*Episode #1: Handoff and patient monitoring*

This first episode occurs at the beginning of the session, when nurse N1 performs the handoff with nurse N2 and nursing aid NA1. Four main events can be identified that describe the trajectory of the operating point in the workspace according to the attracting or repelling forces induced by the constraining boundaries: Patient Care, Task Management and Information Processing (Figure 3):

1. Handoff at the entrance of the bedroom: Information processing attraction.
2. Call from the patient who is stressed: Patient care attraction;
3. The caregivers continue the handoff around the patient's bed: Patient care attraction despite the necessity to perform the handoff;
4. Nurse1 interrupts Nurse2 who was explaining stressful details of the next analysis to the already stressed patient: Patient care as repellor. Nurse2 must avoid to speak in front of the patient.

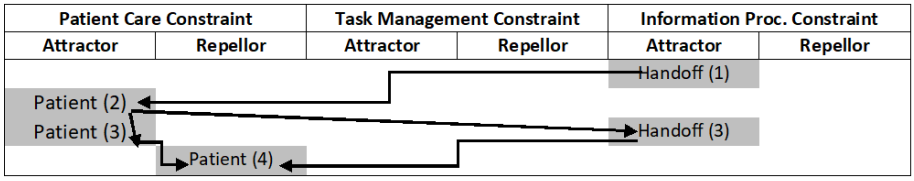



*Figure 3. Trajectory of the operating point during Episode #1 regarding the pressures exerted by the three boundaries: Patient Care, Task Management, Information Processing.*

*Episode #2: Lowering the bed and performing cardiac massage*

This second episode covers the difficulties experienced by NA1 to lower the top part of the patient's bed with the remote in order to facilitate the current cardiac massage. Ten main events can be identified (Figure 4):

1. Nursing Aid1 wants to lower the bed. First, she wrongly raises the bed, then rapidly succeeds, but fails to lower the top part: Attraction from task management.

2. Nurse2 says: ‘You must lower the bed’ and pushes directly on the top part of the mattress, without any success.
3. NA1 takes the remote but fails to lower the top part of the bed.
4. NA1 assists her teammate to place the massage board under the patient.
5. NA1 tries to lower the top part of the bed again, but instead produces a new lowering of the entire bed.
6. NA1 asks N2: ‘Can you lower the bed, please!’
7. N2 lowers the bed while regulating the oxygen flow: Both patient care and task management exert an attraction.
8. While she is massaging, NA1 asks N2 ‘Again, please’.
9. N2 says: ‘It is at the max.’: Bed management is considered as a repellor by the nurse.
10. While waiting for defibrillation, NA1 succeeds in lowering the bed: a waiting stage in patient care is used for bed management.

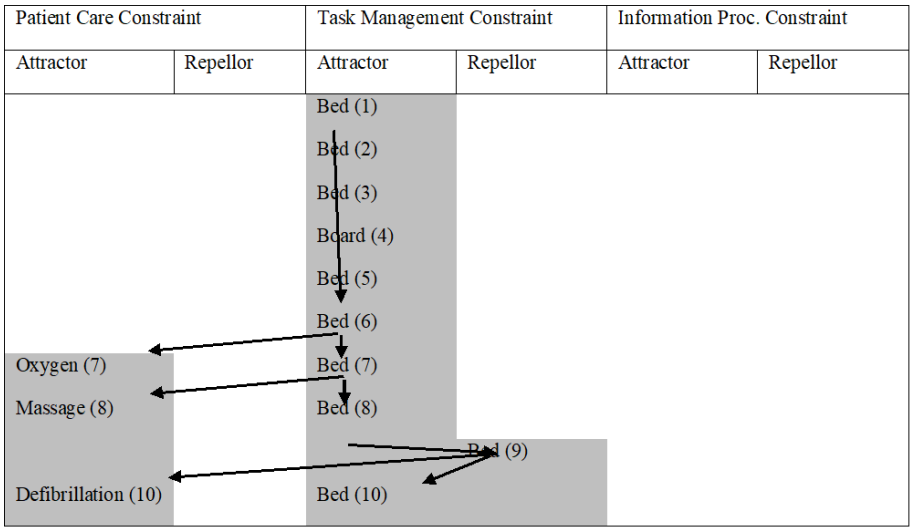



*Figure 4. Trajectory of the operating point during Episode #2 regarding the pressures exerted by the three boundaries: Patient Care, Task Management, Information Processing.*

In these two episodes, the caregivers engage resources in terms of task control, strategies, work organization, and competencies.

**Task control**

Task control refers to an adaptive process based on control loops (e.g., regulation, exploration, anticipation) to coordinate the work constraints. In this context, the normative descriptions of tasks through instructions and procedures can be considered as landmarks to assist implementing these control loops and to avoid violating the risk margins.

*Episode 1:* The handoff represents an anticipatory process. This process is interrupted and modified by the stressed patient, which triggers regulations among caregivers by

responding to the patient's questions and filtering the information communicated to the patient, when N2 stops N1 in her description of details concerning the clinical examination that will be endured by the patient.

*Episode 2:* A global loop of exploration is engaged by NA1 to work out how to lower the bed with the remote during the highly critical moment of cardiac massage. Failures lead to a set of regulations inside this exploration loop, with the support of N2.

**Strategies**

Strategies to perform tasks in loosely coupled work systems involve balancing priorities and values in order to manage the workload. Selecting between possibilities of multitasking or sequential task performance must be made rapidly.

*Episode 1:* First, a sequential activity begins during handoff, then the necessity to manage the patient's interruptions leads to a multitasking configuration of work beside the patient's bed.

*Episode 2:* This episode is markedly interrupted, which leads to multitasking through time-sharing, when NA1 stops care delivery and tries to lower the bed and when N2 tries to lower the bed to assist NA1, or through parallel activity, when N2 lowers the bed and regulates the oxygen flow.

**Work organization**

Work organization concerns allocation and redistribution of the workload among teammates. It also refers to the spatial and temporal organization of the work environment with the purpose of reducing the workload.

*Episode 1:* This episode addresses the issue of where and how the handoff must be performed. Integrating the patient into the handoff becomes problematic.

*Episode 2:* This episode deals with the need to engage the maximum of human resources on patient care, instead of being occupied in trying to lower the bed. Ergonomic solutions to simplify this action upon the patient's bed could be proposed.

**Competencies**

Expertise allows operators to decrease the workload involved in their adaptive processes to work constraints, notably through changes in their level of cognitive control; these changes can be based on knowledge (mental model), rules (heuristics), or skills (Rasmussen, 1986).

*Episode 1:* Handoff mainly involves knowledge-based behaviours through the communication of a mental model of the situation to the next team. This level is particularly sensitive to interruptions that can lead to omissions in the handoff content.

*Episode 2:* Performing cardiac massage is a motor skill that demands considerable physical effort and requires caregivers to adopt a specific posture in order to perform a successful massage. Using the bed remote involves a rule-based control of

behaviour: users need to know how to use the device. If the functioning rules are complex, operators will forget them, as occurred in this episode.

## Conclusion

In Cognitive Systems Engineering, some concerns about the possibility of applying CWA on intentional low coupled work systems have been raised (Wong et al., 1998). Low coupled work systems are governed mainly by constraints emerging from the ways operators organize their work and find solutions to resolve the multiple degrees of freedom that they must deal with.

By considering the constraints arising from both the work domain and the work organization, hCWA proposes an alternative to the traditional CWA that is focused on the work domain constraints. It could deal with the ergonomic issues posed by intentional work systems. hCWA is a heuristic analysis framework since it searches for the basic work requirements that structure the work system and fundamentally shape the operators' behaviours. Rather than describing exhaustively the work domain properties, as the abstraction hierarchy technique does, hCWA points out the consequences of the conflicting interactions between the basic work constraints. These interactions must be dynamically solved by operators in the course of their activity. Such an analysis could be put in relation with the notion of 'elementary structure' developed by the anthropologist Claude Levi-Strauss (1945) or the notion of 'simplexity' proposed by the physiologist Alain Berthoz (2009).

## References

Berthoz, A. (2009). *La Simplexité*. Paris: Odile Jacob.

Cook, R., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *BMJ Quality & Safety*, *14*, 130-134.

Hajdukiewicz, J.R., Burns, C.M., Vicente, K.J., & Eggleston, R.G. (1999). Work domain analysis for intentional systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 43, No. 3, pp. 333-337). Sage CA: Los Angeles, CA: SAGE Publications.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*, 174-196.

Lévi-Strauss, C. (1945). L'analyse structurale en linguistique et en anthropologie. *Word*, *1*, 33-53.

Morineau, T., & Flach, J.M. (2019). The heuristic version of Cognitive Work Analysis: A first application to medical emergency situations. *Applied Ergonomics*, *79*, 98-106.

Morineau, T., Flach, J.M., Le Courtois, M., & Chapelain, P. (2017). An extended version of the Rasmussen's Dynamic Safety Model for measuring multitasking behaviors during medical emergency. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (Vol. 6, No. 1, pp. 238-243). Sage India: New Delhi, India: SAGE Publications.

Rasmussen, J. (1986). *Information Processing, and Human-Machine Interaction: An Approach to Cognitive Engineering*, New York: Elsevier Science.

Rasmussen, J. (1990). The role of error in organizing behaviour. *Ergonomics, 33,* 1185-1199.

Rasmussen, J. (1995). Risk Management and the Concept of Human Error. *Joho Chishiki Gakkaishi, 5,* 39-70.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science, 27,* 183-213.

Vicente, K.J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based Work*, Mahwah, NJ: Lawrence Erlbaum Associates.

Vicente, K.J., & Rasmussen, J. (1992). Ecological Interface Design: Theoretical Foundations, *IEEE Transactions on Systems, Man, and Cybernetics, 22,* 589-606.

Wong, W.B., Sallis, P.J., & O'Hare, D. (1998). The ecological approach to interface design: Applying the abstraction hierarchy to intentional domains. In *Proceedings 1998 Australasian Computer Human Interaction Conference. OzCHI'98* (Cat. No. 98EX234) (pp. 144-151). IEEE.

# The making of Museum works as Smart Things

*Hamid Bessaa, Florent Levillain, & Charles Tijus*
*Laboratoire Cognitions Humaine et Artificielle (CHArt), University Paris 8, France*

**Abstract**

Most important purpose of understanding Human Behaviour in Complex Systems is the making of personalized Human-Artificial dialogs for task-oriented co-operation. Among complex systems are teams of Museum' works that cooperate to build the museum visitors experience (VX), as user experience (UX), to enhance the learner experience (LX). Until now, museums' artworks were passive things people cannot interact with. The "CULTE" project is to offer visitors the possibility to dialogue with connected artworks displayed in the Museum through I.O.T. Thus, as connected objects, Museums' artworks become Smart Things by enriching the visitor experience through trans-media dialogs. We report the rationale for our approach: a problem-solving based approach that is used for designing a smart personalized dialoguing system integrating (i) the context of Museum's complex system, (ii) an ontology of the "what's about" and (iii) the three necessary dialogs components that are the Pragmatic, meta-cognitive and, - as the core of the dialog -, the cognitive components. For the purpose of modelling, from less to more situated, the COGNITION component is embedded in the METACOGNITION component that is in turn embedded in the PRAGMATIC/SEMANTIC component.

**Introduction**

As User Experience, quoted UX, a concept introduced by Don Norman in the 90th to cover all aspects of the experience the person is having with the system (Norman, 2013), Visitor Experience, quoted VX, refers all aspects of the experience the person has with the artwork (Dubois et al., 2011).

As a consequence of technological innovations, VX increases because museums are expanding their system of communication with visitors: before, during and after the visit. Inside and outside the museum walls, visitors can get much more information with the artworks that are connected objects (IOT) and have richer personalized experience. However, if museums deliver this additional information by taking into account the visitor interest, they do it in a way that this is the museum that is talking to the visitor (*when and what*). The visitor is not talking to the museum and there is no dialogue between a visitor and a given artwork.

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

CULTE[1] (*Cultural Urban Learning Transmedia Experience*) is a research project[2] funded by the French National Agency for Research (ANR) about an innovative transmedia pervasive Game which anchors the visitor experience with an in-situ application for Smartphone and an online post-visit platform beyond the museum's wall, in a continuum of visit. The game is also connecting the visitor with others museum's digital tools, which contribute to enhance its experience.

One of the most challenging dimensions of the visitor experience that will make people witnessing an innovative visitor's experience (VX) would be the possibility of dialoguing with any of the artwork connected objects of the Museum as being what we might define as Smart things. This both fundamental and applied research is in the line of research about dialogs with digital media (Bosser et al., 2007; Vandi & Djebbari, 2011; Astic, I., 2014; de los Rios, 2015; Holken et al., 2017).

In this article, we first define what smart things are, what they are made of and then how to design the dynamic interactive dialog of interaction of Smart Things with their users. This new kind of an interaction should be based on a dialogue that is embedded in the dynamic of the visitor route, taking dynamically into account their emerging interests while the process of dialoguing with artworks is evolving. To do so, we are developing the Verbal Interaction with Smart Things model (VIST) which is a general framework of interaction mode that can be used for any subcategory of Smart Things, although the use case reported here is the one of connected artworks in museums.

**What are Smart Things?**

First, Smart Things are Things which means that they are bearer of properties: "*A thing is always something that has such and such properties, always something that is constituted in such and such a way. This something is the bearer of the properties; the something, as it were, that underlies the qualities*." (Heidegger, 2017). A set of properties from which a typology was made: surface, structural, functional, procedural and behaviour properties (Cordier & Tijus, 2001). A typology that can be used for the design of intelligent, companionable objects, such as those designed by Chen et al. (2015) for the Smart Classroom.

In addition, Smart Things are objects that are connected (IOT) and dedicated for making people daily life simpler. "*Because Smart Things are taking decision for people and, for doing so, have to be adapted to their users, they are made of cognitive technologies that are technologies that include knowledge about human and about human cognition in order to process the data users are providing when interacting with Smart Things*" (Tijus, Rougeaux, & Barcenilla, 2016). In short, "*take the idea of*

[1] *This work was performed within the Project CULTE supported by French state funds managed by the ANR, under reference ANR-13-CORD-018-01.*

[2] *CULTE Partners are MQB (Musée du Quai Branly) - Jacques Chirac, Paris, France which is a well-known ten years old museum dedicated to the meeting ground of worldwide past cultures, CEDRIC Laboratory, Centre National des Arts et Métiers, Paris, France, a Game Design research laboratory; LUTIN, Cité des Sciences et de l'Industrie, Paris, France, a usability dedicated research laboratory for digital tools and MAZEDIA, Nantes, France, a multimedia agency, leader in France in the design of multimedia devices for museums.*

*a human-centred approach to technology and run with it* " (Norman, 2014). Based on "affordance", - that is to say the direct coupling of Action to Perception which is what the interface displays as actionable objects for command that seems to match the user's goal (Gibson, 1986; Norman, 2009) -, as well as object's usability based on categorization, reasoning and problem solving (Poitrenaud, Richard & Tijus, 2005, Tijus et al., 2014).

*What are things made of?*

First of all, things as objects have surface features (*colour, texture, size, shape...*). Although of objective evidence based on instruments to measure these visible properties (*spectroscope for colour wavelength, etc.*), these surface properties can match a user's mental representation positively providing affordance or negatively providing false alarm kinds of errors. Thus, for usage, surface properties can be more or less useful.

Things are made of structural properties: their parts and relations between parts and whole that determines in turn functional properties and procedural properties. Thus, things can be used as agent to act on another objects (procedure), realizing some functions that will transform this patient object. Functional properties (*what for*) as well as procedural properties (*how*) being properties attributed by knowledge or inference. Notice that automatic systems are things in which parts are acting on each other to realize some complex functions. This working machinery have behaviour property. These relations have to be used when dialoguing with users; particularly when things have to be Smart.

Relations do exist between these types of properties (Zibetti & Tijus, 2005). On one side, relations exist between structural, functional and procedural properties. On the other side, relations exist between surface and behavioural properties. Both can be used for inducing the adequate functions and procedures, then to trigger action, for instance for the "putting into place of affordances": indicating the where, when, how and on what to act. In opposite, no relation at all will decline accessibility, usability and learnability. Thus, our theory is that smartness comes from smart relations among properties: the relations that increase the guidance of the interaction with the smart thing.

*What are Smart things made of?*

Smart things can be either physical objects *(a robot*) or virtual entities (*an avatar*). In addition, there is smartness: the properties of automatic systems with autonomy, decision-making and adaptation behavioural robotic properties: "*the smart thing can trigger functions and apply procedure to be autonomous, to take decision and to be adapted while having a given appearance and a given behaviour at will. It follows that smartness is the set of relations between "functional - procedural" properties and "surface – behaviour" properties*" (Tijus, Rougeaux & Barcenilla, 2016).

Notice that interaction with smart things can be engaged and sustained mainly by appearance and behaviour. Thus, the design of a smart dialog systems, - as part of a whole Smart Thing-, might be based on appearance and behaviour (Levillain &

Zibetti, 2017). We argue that these properties, their relations, and the underlying logical arguments should be used for the design of smart things dialogs.

**Interacting with museum artworks as Smart Things**

With content based on the typology of properties of a given Smart Thing, this new kind design of verbal interaction should be based on a dialogue that is embedded in the dynamic of the visitor route, taking dynamically into account their emerging interests while the process of dialoguing with artworks is evolving.

Our approach is based on problem solving of explanation (Tijus, Ganet & Brézillon, 2006) in order to design dialog-based intelligent tutoring systems (e.g., D'Mello & Graesser, 2013). Although there are dimensions of dialogue such as emotion, empathy and sympathy, our proposal is about the three necessary components of a dialog: The Pragmatic dimension, the metacognitive dimension and the cognitive dimension.

More precisely, the core of the dialog is the cognitive dimension: the knowledge transmission from the Smart Artwork to the visitor according to her interest. For the purpose of modelling, from less to more situated, there is the COGNITION component that is embedded in the METACOGNITION component that is in turn embedded in the PRAGMATIC component (figure 1).

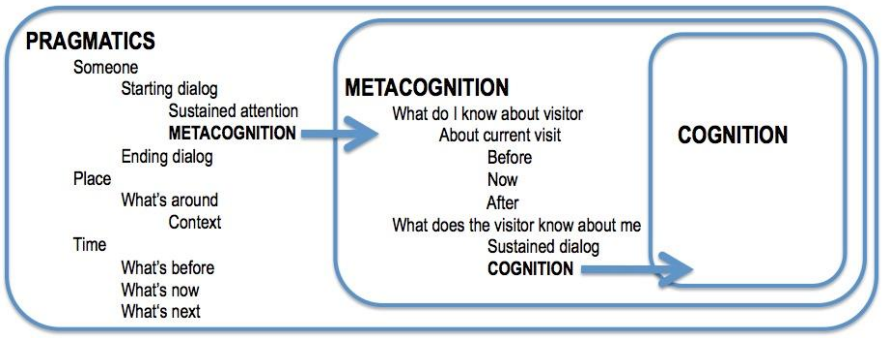



*Figure 1. For a dialog-based intelligent tutoring system, the COGNITION component (knowledge to be delivered through dialog), which is the core of the dialog, is embedded in the METACOGNITION component (meta-knowledge about the purpose of the dialogue and its context), which is in turn embedded of the pragmatics of dialoguing (needs of an interested person, a start and end of the dialog, in a place and at a time for doing so).*

In this brief paper, we shall first introduce the necessary dimensions of an epistemic dialog, which is a dialogue for knowledge transmission and the ontology of knowledge about objects that is to be transmitted, as well as examples of dialogs made by smart artworks in museums.

**What is a dialogue for a smart thing?**

People come to museums to meet things: to see them, to learn about them and to discover new domains of knowledge. Notwithstanding the fact that many works of art are sculptures in human form, it would be "smart" that people can discuss with the

artworks in the museum, as one of the possibilities of interaction. Such an epistemic dialogue would be more than profitable: it might enhance the visitor experience (VX).

Artworks in museums are already connected in a such way visitors can get supplementary information through interaction with some Smartphone (e.g., de los Rios et al., 2015). For instance, thanks to CULTE project, partners developed a transmedia editorial platform, which makes it possible for any museum to develop its own transmedia pervasive devices. Now, partners are going to extend the devices inter-operability and the space and time relationship between the visitor and the museum by adding an off-site mobile application. In that direction, Smart museum Artworks might be capable of discussion with the visitor; as well as being the trigger of the discussion with the visitor than as being triggered by the visitor for discussion. Because till now, much of interactions with museum artworks are determined only by the possibilities offered to the visitor (*ask for [that] by doing [this]*), such an interactive behaviour would be far from what exists.

Such smart things must be based on cognitive technologies that are technologies that include knowledge about human and about human cognition for cognitive processing in order to process the data that visitors are providing when interacting with them. Cognitive computing makes it possible the set of inferences on which dialogue can be built. For the online building of an epistemic dialogue with the purpose of knowledge transmission, the model needs the three embedded components as in Figure 1.

As display in Table 1, although not mandatory, the PRAGMATICS and METACOGNITION components [C-] shall be used to manage the epistemic dialogue. Many different sentences that match these components content can be used. For instance, when by image recognition “a particular person is a possible target for dialogue” [C-1.1], saying “Hello” [C-1.1.1], “Are you interested by me” [C-1.1.1.1], “I think you are because you are a pupil coming in this museum with your class and your professor” [C-2.1], “You have already seen other similar Artworks” [C-2.2.1], “but now you are facing something different” [C-2.2.2], and “I’m the last artwork in your visit” [C-2.2.3]. “So, you already know the country where I come from” [C-2.3], “what do you want to know about me? I have so much to say!” [C-2.3.1], “First of all...” [C-2.3.2], “... as other artworks in this room” [C-1.2.1], “such as the one in your back” [C-1.2.2], “we are talking for long” [C-1.3.1], “ and you already see so many things” [C-1.3.2], “the museum is going to close” [C-1.3.2], “maybe we shall say goodbye” [C-1.1.2].

The tree of categories of the PRAGMATICS and METACOGNITION components can be used to build adapted sentences, as well as to interpret the sentences produced by the visitor. The cognitive computing refers here as the categorization process of affecting visitors’ sentences to the pragmatic and metacognitive categories of human dialog. Note that these categories can be used to question the visitor when interpretation fails. Thus, the tree of categories of the PRAGMATICS and METACOGNITION components can be used to build adapted sentences.

*Table 1. the tree of categories of the PRAGMATICS and METACOGNITION components can be used to build adapted sentences*

| |
|---|
| *C-1. - The PRAGMATICS components are the know-how about the dialogue process.* |
| *C-1.1. - Get [Someone] for dialoguing* |
| *C-1.1.1 - Have a [Starting dialog]* |
| *C-1.1.1.1 - Beware of and control [Sustained attention]* |
| *C-1.1.1.2 - Use the [METACOGNITION] component* |
| *C-1.1.2. - Have an [Ending dialog]* |
| *C-1.2. - Use Information about [Place]* |
| *C-1.2.1 - About [What's around]* |
| *C-1.2.2 - About [Context]* |
| *C-1.3. - Beware and control [Time]* |
| *C-1.3.1 - Use Information about [What's before]* |
| *C-1.3.2 - Use Information about [What's now]* |
| *C-1.3.3 - Use Information about [What's next]* |
| *C-2. - The METACOGNITION components is the knowledge about the dialogue content* |
| *C- 2.1. - Use Information about [What do I know about visitor]* |
| *C-2.2. - Use Information about [its current visit]* |
| *C-2.2.1 - About [Before]* |
| *C-2.2.2 - About [Now]* |
| *C-2.2.3 - About [After]* |
| *C-2.3. - Use Information about [What does the visitor know about me]* |
| *C-2.3.1 - Beware of and control [Sustained dialog]* |
| *C-2.3.2 - Use the [COGNITION] component* |

The following discussion is extracted from the dialog an artwork of the MQB (*Musée du Quai Branly*) is having a visitor. The name of the museum artwork is "Ashura". The related categories of the PRAGMATICS and METACOGNITION components are provided. " *Hello*! " [C-1.1.1], "*I am impressed with the idea of sharing with you*" [C-2.3.1], "*will you talk to me*" [C-2.3.1] "*about Fertility*?" [C-2.3.2], "*During your initiation, you learn that you should not trust appearances*" [C-2.2.1]. Then is "COGNITION" [C-2.3.2]. "*But according to you*" [C-1.1.1.1], "*do I have a link with the costume Gourgecha to my right?*" [C-1.2.1].

Thus, the epistemic talks entail the METACOGNITIVE component that entails the PRAGMATIC components. In the next section, we introduce the ontology of what could be known about a thing that can behave smartly when discussing about itself.

**What a smart thing can tell about itself ?**

There are basic questions about knowledge of things, such as "*Who, what, when, where, why, how*". However, they are not organized in a hierarchy of categories. To do so, we first consider that a thing is a bearer of properties (Heidegger, 1967) and these properties are the components of the COGNITION MODULE. There are extrinsic properties [C-3.1] and intrinsic properties [C-3.2].

Extrinsic properties do not belong to the thing. Thus, Place (*Where*) [C-3.1.1] and Times (*When*) [C-3.1.2] are extrinsic properties that provide the space and time context of the thing. This contextual knowledge (e.g., *where and when the thing was built*) provides relational spatial and temporal properties with other things (*are from the same/different country, were made at the same/different time*). Other extrinsic properties are causal properties [C-3.1.3]: what are the causes of the thing (e.g., *the author, the contingences...*).

In opposite, intrinsic properties are own real properties of the thing. Among intrinsic properties, there are surface properties [C-3.2.1] that are related to perception (e.g., *colour, texture, shape...*) and structural properties [C-3.2.2] that are related to physics: substance (*made of*) and materials (*the parts that composes the thing and how these parts are nested to form a given structure*). There are also cognitive attributed properties [C-3.2.3]: functional, procedural and behavioural properties that are linked to the usage of the thing and rely on structural properties. Finally, there are semantic properties [C-3.2.4] as the thing's name, or other analogical or metaphorical attributes of the thing.

The followings are sentences for a Mask artwork named Ashura Mask: "*I am an Ashura Mask*" [C-3.2.4]. "*My teeth are made of bone fragment*" [C-3.2.2]. "*I come from the oasis of the Algerian Sahara*" [C-3.1.1] "*in which there were happy masquerades in order to celebrate the Ashura festival*" [C-3.1.3], "*on the 10th day of the first month of the Muslim calendar*" [C-3.1.2]. "*It was at the time of an ancient agrarian fertility rite that has survived in some areas since Islamized*" [C-3.1.2]. "*I am of the types of Ashura masks that are called Zalouciou mask*" [C-3.2.4] because Zalouciou means "*Acolyte*" or "*companion*" [C-3.2.4]. They were made and worn by young unmarried men who accompanied a man dressed in his nocturnal wanderings in Gourgecha [C-3.1.3].

## Conclusion

Smart things that are connected objects (IOT) are dedicated for making simpler people's daily life. They are of help for decision-making and problem solving. A number of objects are resources for teaching and learning. As smart things, they will have dialogue competencies and capabilities. Based on a categorization theory, we propose a model and an ontology to design the on-line building of an epistemic dialog. Although done for Museum's objects, the model, its components and the properties that define categories could be of use for designing a large number of types of smart things dialogs (inside a car, with a group of Smart things

## References

Astic, I. (2014). *Rapport sur l'équilibrage des jeux pervasifs transmédias dans les musées.* [Rapport de recherche] CEDRIC-14-3229, CEDRIC Lab/CNAM. 2014. ⟨hal-01126559⟩

Bosser, A. G., Levieux, G., Sehaba, K., Buendia, A., Corruble, V., & De Fondaumière, G. (2007). Dialogs taking into account experience, emotions and personality. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts* (pp. 9-12). ACM.

Chen, C.L.D., Chang, Y.H., Chien, Y.T., Tijus, C. & Chang, C.Y. (2015). Incorporating a smart classroom 2.0 Speech-Driven PowerPoint System (SDPPT) into university teaching". *Smart Learning Environments, 2(1),* 1-11.

Cordier F., & Tijus, C. (2001). Object properties: A typology. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 20*, 445-472

D'Mello, S., & Graesser, A. (2013). Design of dialog-based intelligent tutoring systems to simulate human-to-human tutoring. *Where Humans Meet Machines* (pp. 233-269). New York: Springer.

De los Rios, S., Cabrera-Umpierrez, M. F., Arredondo, M. T., Paramo, M., Tijus, C., Djebbari, E., ... & Santoro, R. (2015). Living Lab Concept Validation Experiment to Experience COOLTURA in the Cité Des Science et de L'Industrie. *Proceedings of the International Conference on Universal Access in Human-Computer Interaction* (pp. 41-52). Springer, Cham.

Dubois, E., Bortolaso, Bach, C., Duranthon, F., & Alanquer-Maumont, A. (2011). Design and evaluation of mixed interactive museographic exhibits. *International Journal of Arts and Technology, 4,* 408-441.

Gibson, J.J. (1986). *The Ecological Approach to Visual Perception*, Hillsadle: LEA.

Heidegger, M. (1967). *What is a Thing?* Indiana: Regenery/Gateway.

Holken, H., Tijus, C., Bessaa, H., Rougeaux, M., Levillain, F., Parmentier, & M. (2017). Personalized Dialog with Artworks: New Visitor Experiences with IoT access for Museums and Historical Sites. *NEM Summit 2017Conference – "Smart Content by Smart Creators".* Museo Reina Sofía. Madrid : 29/30 November 2017.

Levillain, F., & Zibetti, E. (2017). Behavioral artifacts: The rise of the evocative machines. *Journal of Human Robot Interaction, 6(1),* 4-24.

Norman, D. (2014). *Things that make us smart: Defending human attributes in the age of the machine*. Diversion Books.

Norman, D.A. (2009). *The design of future things*. New York: Basic Books / Perseus Book Group.

Norman, D.A. (2013). *The design of everyday things: Revised and expanded edition*. New York: Basic books.

Poitrenaud, S., Richard, J.F., & Tijus, C. (2005). Properties, categories, and categorisation. *Thinking & reasoning, 11(2),* 151-208.

Tijus, C., Barcenilla, J., Rougeaux, M., & Jouen, F. (2014). Open innovation and prospective ergonomics for smart clothes. In Marcelo Soares & Francisco Rebelo (Eds.): *Advances in Ergonomics in Design, Usability & Special Populations: Part III,* (pp. 583-591), AHFE Conference.

Tijus, C., Ganet, L., & Brézillon, P. (2006). Neuf motifs de révision des textes procéduraux: l'apport de la catégorisation et des graphes contextuels à l'explication du savoir-faire. *Langages, 4,* 86-97.

Tijus, C., Rougeaux, M., & Barcenilla, J. (2016). The Making of Smart Things. *Journal of Science and Innovation, 6,* 41-45.

Vandi, C., & Djebbari, E. (2011). How to create new services between library resources, museum exhibitions and virtual collections. *Library Hi Tech News, 28*(2), 15-19.

Zibetti, E. & Tijus, C. (2005). Understanding actions: Contextual dimensions and heuristics. *International and Interdisciplinary Conference on Modeling and Using Context* (pp. 542-555). Berlin: Springer.

# I don’t care what the robot does! Trust in automation when working with a heavy-load robot

*Franziska Legler, Dorothea Langer, Frank Dittrich, & Angelika C. Bullinger*
*Ergonomics and innovation management, Chemnitz University of Technology*
*Germany*

**Abstract**

There are many reasons for the implementation of human-robot collaboration (HRC). HRC enables flexibility of increasingly complex production sites. In contrast to this, the economic aim of process efficiency is threatened by workers’ fear and mistrust in collaborative robots. Fenceless heavy-load collaborative robots have associated risks and so under- or overtrust in automation may result in injuries. An experiment with 25 participants and a heavy-load industrial robot was conducted in a pseudo real-world test environment. Interaction level and robot trajectory were used as within-subject independent variables. Additionally, temporal position of first-failure was varied between participants. Emotional experience and trust were dependent variables. Interaction level, robot trajectory and position of the first-failure did not reveal practical relevant effects on fear or trust. While participants showed short-term responses to first-failure events, following scenarios were not influenced by first-failure regarding emotional experience or trust. Overall, negative emotions were poorly detected and trust in automation was high. These results are in line with findings in the literature regarding overtrust in automation.

## Introduction

Reasons for the implementation of human-robot collaboration (HRC) are diverse. HRC offers new possibilities in the design of ergonomic workplaces. It is also expected that HRC enables the flexibility of increasingly complex production facilities (Oubari et al., 2018). Process efficiency is assumed to increase based on the combination of robots’ repetitive accuracy and workers’ ability to solve ill-defined problems (ISO/TS 15066, 2016). On the other hand, fenceless heavy-load robots increase the risk of injury. Misconduct or technical problems may result in physical contact between workers and robots. Heavy-load robots have carrying capacities up to 500 kg and above. While moving, these robots are capable of exerting forces far beyond the maximum permissible limits associated with the biomechanical threshold of different body parts (see ISO/TS 15066, 2016). Therefore, ISO/TS 15066 (ISO/TS 15066, 2016) specifies strict safety regulations for HRC within shared workspaces. However, little is known about the effectiveness of regulations on the perception of safety by workers and their resulting behaviour. Workers could have concerns simply because of the physical appearance of robots or specific movements. Mental strain, such as negative emotions and mistrust, are likely to occur in these situations (e.g. Arai et al., 2010).

In D. de Waard, A. Toffetti, L. Pietrantoni, T. Franke, J-F. Petiot, C. Dumas, A. Botzer, L. Onnasch, I. Milleville, and F. Mars (2020). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Emotions are characterised by a specific feeling as well as observable physiological and behavioural reactions (Schmidt-Atzert, 1996). Apart from other emotions, fear as a specific negative emotion is important in the context of HRC. Various studies have shown that direct cooperation with a robot results in increased feelings of fear in the workplace (Brending et al., 2016). Fear is also called state anxiety and is defined as "transitory emotion characterised by physiological arousal and consci-ously perceived feelings of apprehension, dread, and tension" (Endler & Kocovski, 2001, p. 2). Behavioural reactions of fear entail bending forward and running for cover to escape from danger (Grèzes et al., 2007). Fearful movements are character-ised by high dynamics (McColl & Nejat, 2014). These sudden movements can lead to physical contact between robots and workers, which may result in worker injury. It is therefore important to study fear in the context of HRC with heavy-load robots.

HRC is only efficient if humans and robots work together to combine their particular strengths. Therefore, another important concept associated with HRC is trust in automation (TiA). TiA is defined as "…the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" (Madsen & Gregor, 2000, n.p.). High trust reduces cognitive complexity in the face of highly automated systems and mistrust leads to rejection of automation (Lee & See, 2004). Consequently, one could infer that high TiA is associated with an efficient robot collaboration. In contrast, it has been shown that overtrust can also cause critical outcomes. Overtrust characterises inappropriate trust calibration that exceeds the capabilities of the automated system. This inappropriate trust may lead to overreliance and misuse of the system (Lee & See, 2004). Reduced situation awareness as a consequence of overtrust (Hancock et al., 2011) may again result in physical harm of workers in the event of system automation failure. Hancock and colleagues (2011) conclude that an appropriate level of trust that neither includes under- or overtrust is necessary for a safe and efficient interaction of humans and robots. Unfortunately, there has been no clear definition or specification of this appropriate level to date.
The relationship between fear and trust in automation is insufficiently researched (Stokes et al., 2010). Lee and See (2004) cautiously summarise that emotional reactions seem to be a critical contributor of trust. Both constructs should therefore be researched in context of HRC.

Various factors influence trust and emotional experience in HRC and are important for ensuring safe and efficient collaboration. Some of the characteristics of robot motion, such as speed, distance to robot (Arai et al., 2010; Desai et al., 2013) and unexpected movements (Desai et al., 2013; Dragan et al., 2015), were found to be important factors influencing people's fear. Nevertheless, theoretical concepts of concrete robot motion trajectories are rare. Dragan et al. (2015) suggest a distinction between predictability and legibility of trajectory. Both are defined by human inferences in collaborations. "Predictable motion is functional motion that matches what the collaborator would expect, given the known goal. (…) Legible motion is functional motion that enables the collaborator to quickly and confidently infer the goal" (Dragan et al., 2015, p. 51). As a result, predictable motion requires know-ledge of the robot's target position, while legible motion enables the user to infer the goal directly from robot's first movements even if the target position is unknown. Legible

robot paths were preferred by users and resulted in higher trust than predict-able trajectories (Dragan et al., 2015). Therefore, legibility also shows the potential to reduce negative emotions. To date, trajectories have only been examined with lightweight robots and the transferability of results to heavy-load robots is unknown.

In general, people expect automated systems to perform well and as a result, complacency is often observed when interacting with an automated system (Parasuraman & Manzey, 2010). It was even possible to transfer the so-called *positivity bias* found in social psychology to interactions with automated systems. These studies have shown that people expect good performance prior to interaction, even without any detailed information of the system (Dzindolet et al., 2003). All automation systems still have their limitations and it is widely known that failures of automated systems affect TiA (e.g. Parasuraman & Manzey, 2010). The concept of first automation failure is most important in this area of research and it is also referred to as *first-failure effect* (Wickens & Xu, 2002). Firstly, a reduction of trust level after occurence of the first failure of a seemingly perfect automation is postulated. Secondly, trust only slowly recovers and often remains on a lower, probably more appropriate, level of trust (Lee & Moray, 1994). One reason for mixed results in first-failure literature is attributed to prior information about system reliability. Wickens and Xu (2002) conclude that without this prior information, a reduction of trust is likely to occur. The first-failure effect was particularly observed in driver-vehicle interaction with automated systems. Strong reduction in trust was found when no information about potential system limitations was given prior to usage (Beggiato & Krems, 2013).
Effects of failures were also observed in human-robot interaction and according to literature in this context, even showed effects of the temporal position of failures. An early automation failure in interactions caused a greater reduction of real-time trust than a late event (Desai et al., 2013). Trust decreases even if system failures do not directly contribute to system performance loss (Muir & Morey, 1996). In most literature, failures are simulated as software conditioned automation breakdown. People miss the occurrence of automation breakdown due to overreliance and reduced situation awareness (see Hancock et al., 2011), resulting in performance loss. To date, no research examining the effects of first automation failure in HRC with heavy-load robots is known to the authors – a critical research gap. Most robot control systems work with point-to-point movements, where trajec-tories between points are not completely pre-programmed. System malfunctions can therefore result in varied robot paths. Given the fact of fenceless interaction and reduced situation awareness while working with automated systems, the risk of physical contact between robot and worker increases. While working fenceless with heavy machines, robot hardware malfunctions can also cause harm or at least result in fear and reduced trust from near-misses without physical consequences.

Another increasingly important novel factor for heavy-load robots is interaction level. Bdiwi, Pfeifer and Sterzing (2017) introduced a classification of four HRC-levels of fenceless collaboration, structured by the specification of the shared task.

- HRC-level 1: No shared task (e.g. because of limited space)
- HRC-level 2: Shared task, no physical interaction (e.g. robot as simple “third arm” without movement in the shared workspace)

- HRC-level 3: Shared task, "handing-over task" (e.g. robot hands over an object or robot reacts to motion of the humans' hand; still no physical contact during robot movement)
- HRC-level 4: Shared task, physical interaction (human forces are applied directly on the robot)

To date, HRC-level 4 with heavy-load robots is poorly viable due to technical inadequacy and safety reasons. Therefore, comparison of HRC-levels 1 and 3 is desirable as they represent the most different and technical feasible levels of interaction. HRC-level 1 is often realised with a physical barrier between the robot and worker (e.g. an assembly table), resulting in some distance between them. In contrast, realization of HRC-level 3 necessarily results in a low distance to the robot. Furthermore, the robot is moving while workers are within the collaboration zone. As distance is an important predictor of trust (see Arai et al., 2010) and robots can produce high forces, it is probable that direct interaction with a moving robot in HRC-level 3 causes higher fear and less trust than HRC-level 1.

Three research questions arise that should give further insight in HRC:

*Research Question 1:* What effect has HRC-level on fear and trust in automation with heavy-load robots?

*Research Question 2:* What effect has robot trajectory on fear and trust in automation with heavy-load robots?

*Research Question 3:* What effect has first-failure on fear and trust in automation with heavy-load robots?

To study research question 1 to 3, an experiment that varied interaction level, robot trajectory and temporal position of first-failure was designed. The experiment took place in a novel pseudo real-world test environment realised at Fraunhofer IWU Chemnitz.

## Method

### *Test environment*

An industrial KUKA robot (Quantec prime KR 180), classified as heavy-load robot, was used. The subjects' task was modelled after a real workplace from the automotive industry. The demo-task consisted of assembling eight hook-and-pile tapes on a front axle carrier. A flexible layout equipped with zone-based robot control (Bdiwi, Krusche & Putz, 2017) was implemented to create two different interaction levels (see Figure 1). In both interaction levels, participants remained at an equal distance from the robot outside of the collaboration zone while the robot moved with a speed of 1000 mm/s.

- HRC-level 1 was realised by placing an assembly table in a robot cell that acted as physical barrier. Therefore, the subjects had no physical contact or interaction with the robot. A certified gripper for front axle carriers was unavail-able as the

specific workplace with HRC does not currently exist in the real-world (the task is done with a handling device instead). To overcome this limitation, the robot never put the component down onto the table. The robot only simulated placement of the component on the table as well as its storage. To enable the assembly task, one front axle carrier was lying on the table and another was fixed to the robot flange all the time (see Figure 1 left and middle).

- In contrast, HRC-level 3 was realised by direct assembly at the front axle carrier, located at the robot flange. Additionally, subjects were able to adjust the assembling height through camera-based gesture control (Bdiwi, Pfeifer & Sterzing, 2017) for better ergonomics. A vision sensor tracked the palm of the subject's hand and the robot arm reacted to upward or downward hand movements accordingly. Thus, subjects were able to control the robot directly at a minimum distance but without physical interaction. It should be mentioned that gesture control showed some unintended problems during a few of the trials.

*Figure 1. Real scenario (left), comparison of HRC-level 1 (middle) and HRC-level 3 (right). Pictures taken from participants' view in a virtual environment. In HRC-level 1, the robot moved to a waiting position after simulated placing of front axle carriers on the table (see middle).*

Both interaction levels contained the same two explorative robot trajectories (see Figure 2). Based on Dragan and colleagues (2015), we defined a legible trajectory "from side" (below head-level, robot arm stretched to side) and a predictable trajectory "from above" (above head-level, robot arm angled, downward movement to the assembly position).

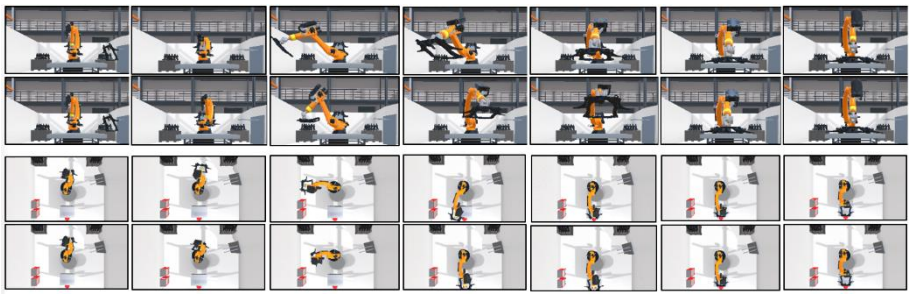


*Figure 2. Comparison of trajectories "from side" (row 1 and 3) and "from above" (row 2 and 4) over time in front and aerial view. After storage of the front axle carrier at the right position of the pictures, the robot returned 270° for the admission of the next component.*

The robot system was capable of simulating a system failure. Because of safety requirements, sudden unexpected or abrupt movements of the robot were not included.

Instead, failure was implemented as the opening of a compressed air valve, resulting in an abrupt and loud noise to simulate hardware technical failure. For participant safety, failure occurred at the beginning of the subjects' assembly task, when the robot had already stopped moving.

Due to safety requirements, participants were objectively located outside of the robot cell at all times whilst the robot was moving with high speeds of 1000 mm/s. To still maintain realistic perception, the cell was visually enlarged by boundary lines on the ground and partition walls.

*Sample*

Twenty-five subjects participated in the experiment. Participants' mean age was 30.2 years. Fifteen participants were male, ten were female. The sample was characterised by a medium to high affinity for technology. Two thirds of the participants had previously interacted with an industrial robot. One third of all participants worked at the time or had worked in the production industry before. Participants received financial compensation for their participation.

*Experimental design*

A 2 (interaction level) x 2 (robot trajectory) x 2 (temporal position of failure) mixed-design was applied. Interaction level (HRC-level 1 vs. 3) and trajectory ("from above" vs. "from side") were within-subject factors. Position of system failure ("early" - after part 1 vs. "late" – after part 2 of the experiment) was conducted as a between-subject factor. All participants completed two parts of experiment that were determined by the interaction level and randomised across participants. Each interaction level started with a baseline assessment as the zero reference. Participants practiced the assembly task but without movement of the robot (in accordance with Bortot et al., 2013). After each baseline, the trajectory was varied in a randomised order within each interaction level. The experimental design resulted in seven scenarios that occurred in partly randomised order within test blocks 1 to 7 (see Table 2 for two exemplary orders).

*Table 2. Exemplary experimental variations*

| Test block (temporal position) | Subject A | Subject B |
|---|---|---|
| 1 | Baseline HRC level 1 | Baseline HRC level 3 |
| 2 | **HRC level 1: from side** | **HRC level 3: from side** |
| 3 | **HRC level 1: from above** | **HRC level 3: from above** |
| 4 | Baseline HRC level 3 | **Early Failure Scenario (HRC 3)** |
| 5 | **HRC level 3: from above** | Baseline HRC level 1 |
| 6 | **HRC level 3: from side** | **HRC level 1: from side** |
| 7 | **Late Failure Scenario (HRC 3)** | **HRC level 1: from above** |

*Measurements*

**Control variables (pre-survey).** Demographic information such as sex, age, as well as experience with industrial robots and production work, was captured in a pre-survey. Additionally, trait anxiety (STAI-T; Spielberger, 1989; $\alpha$ = .80) and Affinity for Technology Interaction (ATI; Franke et al., 2018; $\alpha$ = .92) were assessed.

**Outcome measures (post-scenario).** Outcome measures were assessed after each of the seven scenarios/test blocks. Mean Cronbach's alphas over all seven test blocks are given in brackets. The STAI-S (Spielberger, 1989; α = .90) was assessed to measure state-anxiety. TiA was measured via German translation (Pöhler et al., 2016) of the Jian-Scale. Pöhler and colleagues suggest using two distinct scales; trust and mistrust. Exploratory factor analysis revealed superiority of a two-factor model in all seven test blocks (varimax rotation; see Table 1). Reliability for the factor trust (6 items) revealed an $\alpha$ = .90 and for the factor mistrust (5 items) $\alpha$ = .80.

*Table 1. Exemplary fit-indices of two competing factor-models modelling Jian-Scale for baseline 1*

| | *RMSA* | *TLI* | *RMSEA* | *BIC* | $\chi^2$ | *df* | $\chi^2/df$ |
|---|---|---|---|---|---|---|---|
| 1-factor model | .11 | 0.728 | .21 | -70.67 | 70.96 | 25 | 2.84 |
| 2-factors model | .06 | 0.921 | .15 | -69.68 | 39.76 | 25 | 1.59 |

*Note.* RMSA = root mean square of the residuals; TLI = Tucker Lewis Index; BIC = Bayesian information criterion.

*Procedure*

In advance, subjects were informed about the procedure of the experiment through participant information. After the subjects were welcomed, an informed consent was signed and they filled in a pre-survey on a touchscreen tablet. Subsequently, subjects watched two videos as a cover story (enlargement of existing workplace through HRC). Video1 showed the real workplace with handling device and Video 2 showed an exemplary robot movement in our test environment to lower tenseness of participants. Afterwards, subjects were instructed about the assembly task. Participants were told that they were only allowed to leave their start position, and consequently enter the collaboration zone, if the robot stopped moving. They learned the gesture control of the robot for HRC Level 3. Following this, subjects went through seven test blocks (2 baselines, 5 experimental conditions), each lasting around two minutes and containing three assembly cycles. The test blocks were followed by short post-scenario surveys to measure outcomes via touchscreen tablet.

*Data Analysis*

Statistic Software R (R Core Team, 2018) was used for data analysis. Due to small group sample sizes and non-symmetric distribution of data, nonparametric data analysis was applied. If not specified otherwise, the Wilcoxon Signed-Rank Test was used. The nonparametric effect size, *r*, was calculated according to Tomczak and Tomczak (2014).

For simplification of results, failure scenarios of HRC-Level 1 and HRC-Level 3 were combined as these did not reveal significant differences between scenarios in outcome measures. For group comparisons, relative values of outcomes were calculated by subtraction of participants' first baseline assessment to control for basic differences of groups.

## Results

*Overall results*

State-anxiety was low across conditions. The highest difference in means was between scenario "HRC-level 1/from side" ($M = 27.49$) and "HRC-level 3/from side" ($M = 32.69$). The value of the failure scenario was in between ($M = 29.94$). Accordingly, Friedman's Rank Sum Test showed a significant effect across five experimental conditions (baselines left out; $\chi^2 = 12.54$, $p = .014$).

The baseline (BL) of HRC-level 1 showed higher state-anxiety in comparison to both experimental conditions of HRC-level 1 ($.010 \leq p \leq .074$), resulting in small effect sizes ($.253 \leq r \leq .364$). There was no significant difference in baseline (BL) of HRC-level 3 compared to experimental conditions of HRC-level 3.

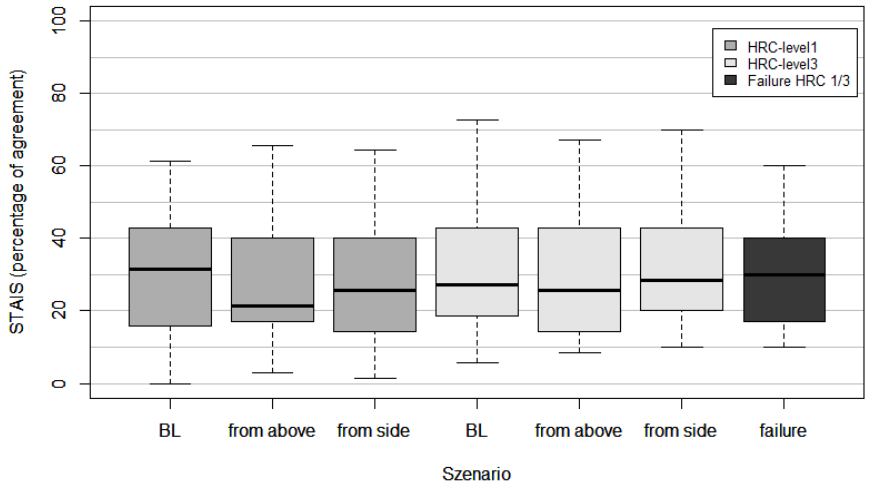



*Figure 3. State-anxiety across scenarios (BL = baseline).*

Trust was high across all scenarios with means ranging between $M = 5.11$ (scenario "BL HRC-level 2") and $M = 5.90$ (scenario "HRC-level 1/from side"; see Figure 4). Accordingly, results for mistrust (same scale range as trust) showed low means across scenarios ranging between $M = 2.33$ and $M = 3.02$. For both trust and mistrust, significant differences to baseline occurred only in HRC-level 1 with medium effect sizes ($.502 \leq r \leq .531$). Friedman's Test showed a significant effect across five experimental conditions for both trust and mistrust (baselines left out; $\chi^2 = 19.49/17.35$, $p < .001$).

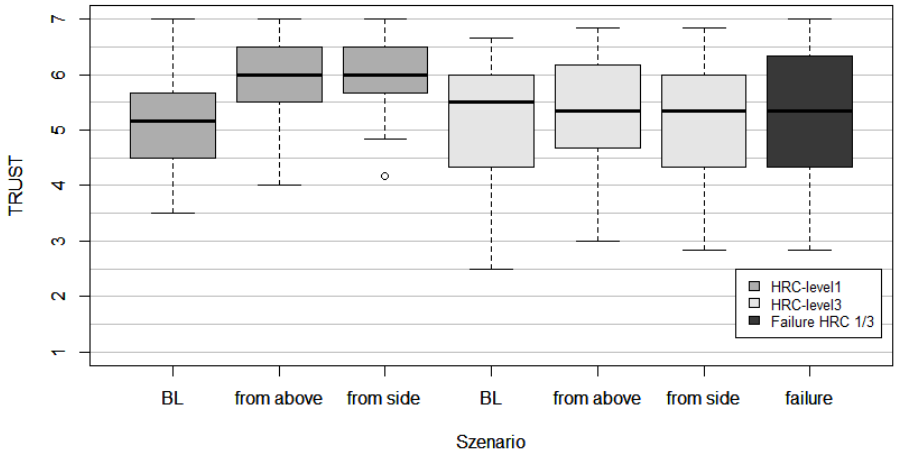



*Figure 4. Trust across scenarios (BL = baseline).*

*Effect of HRC-Level*

Mean state-anxiety was marginally higher in HRC-level 3 than in HRC-level 1 (see Figure 3). Paired comparisons of HRC-levels did not show significant differences in trajectory "from above" ($Z$ = -1.05, $p$ = .294, $r$ = .148, $|M_{diff}|$ = 2.17). In contrast, for trajectory "from side", HRC-level 3 resulted in significantly higher state-anxiety with medium size of effect ($Z$ = -2.60, $p$ = .009, $r$ = .367, $|M_{diff}|$ = 5.20).

Figure 4 shows differences of means for HRC-levels regarding trust. As so, for trajectory "from above", trust was significant lower in HRC-level 3 than in HRC-level 1 ($Z$= -2.45, $p$ = .014, $r$ = .347, $|M_{diff}|$ = 0.59). Similar results were found for trajectory "from side" ($Z$= -3.35, $p$ < .001, $r$ = .474, $|M_{diff}|$ = 0.75), and adequately for mistrust, where mistrust was higher in HRC-level 3 than in HRC-level 1.

*Effect of robot trajectory*

Mean state-anxiety was marginally higher for trajectory "from side" in comparison to "from above" (see Figure 3). Paired comparisons of trajectories did not show significant differences in HRC-level 1 ($Z$ = -0.46, $p$ = .648, $r$ = .065, $|M_{diff}|$ = 0.40) or HRC-level 3 ($Z$ = -1.38, $p$ = .167, $r$ = .195, $|M_{diff}|$ = 2.63).

Figure 4 shows no differences in trust between trajectories. Accordingly, paired comparisons of trajectories did not show significant differences in HRC-level 1 ($Z$ = -1.01, $p$ = .313, $r$ = .143, $|M_{diff}|$ = 0.07) or HRC-level 3 ($Z$ = -0.62, $p$ = .532, $r$ = .088, $|M_{diff}|$ = 0.08). Similar results were found for mistrust.

*Interaction of HRC-level and robot trajectory*

Figure 5 shows interaction plots of HRC-level and robot trajectory for state-anxiety and trust. The interaction plot for mistrust is very similar to the interaction plot of trust. Visually, an interaction effect for state-anxiety is probable while there is no effect for trust. Trajectory seems to be irrelevant in HRC-level 1. In contrast, trajectory

"from side" compared to "from above" seems to result in higher state-anxiety in HRC-level 3. The general linear model shows no significant interaction effect of HRC-level and robot trajectory regarding state-anxiety ($F = 1.65$, $p = .211$).

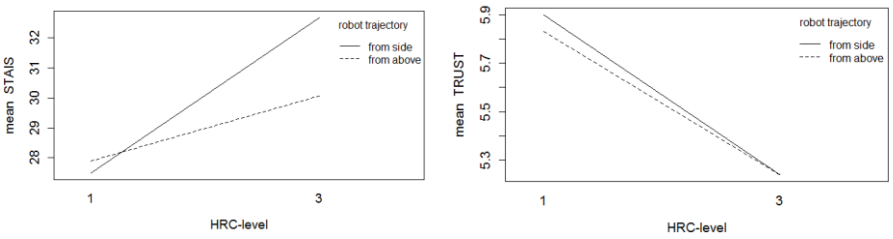



*Figure 5. Interaction plots for state-anxiety (left) and trust (right).*

*Effect of first failure*

To analyse effects of first-failure, a dataset sorted chronologically (in comparison to a dataset sorted by scenarios) was used. The dataset was divided into groups experiencing "early failure" (after first part) and "late failure" (at the end of the experiment). After automation failure occurence, two trends are predicted:

- increased state-anxiety and decreased trust for following test blocks within group "early failure" and
- increased state-anxiety and decreased trust in group "early failure" in comparison to according test blocks in group "late failure".

Figure 6 shows the results for state-anxiety for both temporal positions of failure. Mean state-anxiety was not increased after failure occurrence in group "early failure". Also, state-anxiety in test blocks following failure scenario in group "early failure", was not higher than according test blocks in the group "late failure". Overall, the trend in Figure 6 suggests a decrease of state-anxiety with the time of interaction. Only late system failures resulted in an increase of state-anxiety compared to the test block preceding the failure scenario. Still, state-anxiety in the failure scenario was lower than in baseline 1 for both groups.

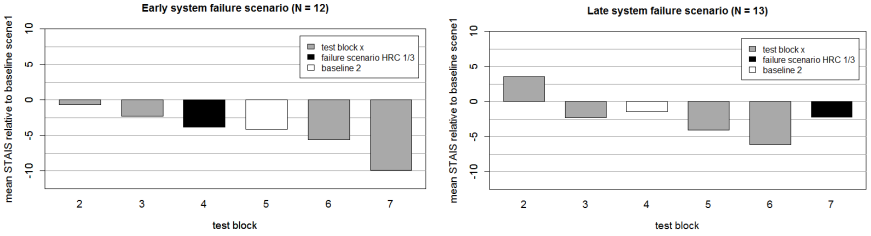



*Figure 6. Effect of first automation failure on state-anxiety for two temporal positions of failure. Means are relative values to test block 1 of participants.*

Figure 7 shows results of trust for both temporal positions of failure. Mean trust is not reduced after failure occurrence. In contrast, mean trust following automation failure is slightly reduced in comparison to group "late failure". Overall, Figure 7 also suggests a slight increase of trust over time of interaction. Early failure only slightly

reduced trust while late failure resulted in a strong decrease of trust compared to the previous test block.

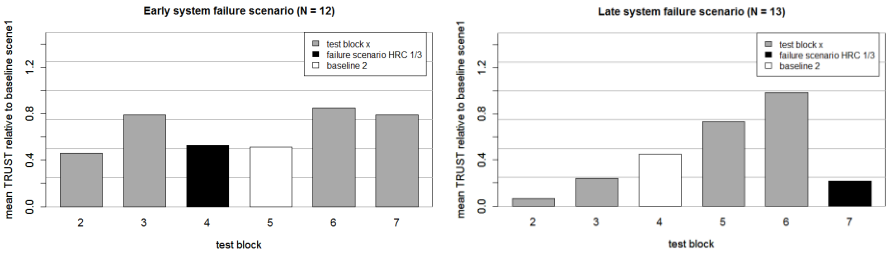



*Figure 7. Effect of first automation failure on trust for two temporal positions of failure.*

## Discussion

Overall, the effect of interaction level was found to be inconsistent with expectations for state-anxiety, and consistent for trust. It is also possible that technical problems with gesture control in the test environment could have influenced the effect of HRC-level on trust and mistrust. Questionnaire items of trust and mistrust included statements about system functionality. Here, gesture control malfunctions may have confounded ratings of trust and mistrust in HRC-level 3. Although the effect of HRC-level on trust was significant, the practical implication of the absolute differences in means is questionable.

Results for robot trajectory are not in line with the findings of Dragan and colleagues (2015). It is possible that our definitions of legible and predictable trajectory differ from these researchers. Another possible explanation is general transferability. At similar speeds, lightweight robots span smaller distances than heavy-load robots. This leads to reduced time for mental anticipation and processing of lightweight robot trajectories. It can be concluded that difference in legibility and predictability may have less relevance, for both state-anxiety and trust, in the case of heavy-load robots.

With regards to the effects of the temporal position of first-failure on state-anxiety and trust, the fact of small group sample sizes should be considered. Therefore, random effects may have caused the differences in results between both groups. In general, relative deviations from baseline 1, shown in Figure 6 and 7, are small. State-anxiety levels remained low and trust levels remained high, which further supports research on positivity bias and overreliance on automation.

### *Limitations of experimental design*

The experimental design allowed systematic variation of different independent variables. As expected, limitations with regard to the transferability of experimental results for real-world industrial settings, exist. Firstly, the assembly task was designed without time constraints. It is probable that pressure due to time constraints would influence emotional experiences of the subjects (e.g. Cœugnet et al., 2013). Additionally, the scenario-based design impedes the subjective experience of workflow and this may lower emotional attachment and presence in the situation. Each scenario/test block lasted only about two minutes, and due to post-scenario surveys, subjects may have been aware of experimental variations and expected some

manipulation. Experimental manipulations should therefore be better researched without scenario pausing to create assembly flow.
As previously mentioned, the missing certified gripper required that front axle carriers could not be placed on the assembly table in HRC-level 1. This could again result in the perception of an artificial situation. Unintentional background noises occurring in the research factory of Fraunhofer IWU may have influenced subjects' perception of the intended experimental system failure. This could have lowered the effect of system failure. Specific emotions other than fear should also be examined in further research. Participants were located outside of the production cell when the robot was moving. Results should be confirmed with subjects remaining inside the collaboration zone. Finally, participants of the study were young and had an affinity for technology. Effects of age on emotional experience and trust when working with heavy-load robots could not be assessed with this sample.

**Conclusion**

The aim of this paper was to study the effects of HRC-level, robot trajectory and temporal position of first-failure on emotional experience and trust, while working with a heavy-load industrial robot. An experiment in a pseudo real-world test environment was therefore designed. Inconsistent effects of HRC-level were found for state-anxiety. In contrast, effects due to trust were in line with expectations. Trust was lower in HRC-level 3, characterised by direct interaction between the human and the robot. Unfortunately, this effect may have been confounded by technical system functionality. Therefore, the effects of HRC-level remain unclear. The present study was not able to transfer results regarding effects of trajectory (Dragan et al., 2015) to heavy-load robots. No differences regarding state-anxiety and trust were found between a novel designed legible ("from side") and predictable ("from above") robot path. First insights for transferability of first-failure effect (Wickens & Xu, 2002) to HRC with heavy-load robots were found. Although some of the observed effects were significant and resulted in medium effect sizes, the observed absolute differences in means between scenarios or test blocks were rather small. In accordance with Schäfer and Schwarz (2019), we concentrate on observed absolute deviations. It is clear that robot movements and their determinants like HRC-level and system failure are important factors for consideration in workplace design, especially for anxious individuals. Still, the present study did not show practical relevant effects on emotional experience and trust in automation.

It was found that state-anxiety decreases, and trust increases over time of interaction. In combination with the overall low levels of state-anxiety and high levels of trust, these results are in line with the literature regarding overtrust effects in automation (see e.g. Lee & See, 2004; Dzindolet et al., 2003). Overreliance and overtrust can result in injuries while working with heavy-load robots. Consequently, research on the strategies to maintain situation awareness and sensitisation for limitations of automation is important to reduce overtrust-effects and to ensure workplace safety. Additionally, effects of reduced situation awareness and overtrust on process efficiency should be examined.

**Acknowledgements**

This research took place within the scope of project "3DIMiR" (project number 03ZZ0459D) supported by German Federal Ministry of Education and Research. The authors acknowledge this financial support. We thank Mohamad Bdiwi, Lena Winkler and Shuxiao Hou from Fraunhofer IWU for the realisation of the test environment. We also thank Charl Jacobs for proofreading of this paper.

**References**

Arai, T., Kato, R., & Fujita, M. (2010). Assessment of operator stress induced by robot collaboration in assembly. *CIRP Annals, 59*, 5-8. doi:10.1016/j.cirp.2010.03.043

Bdiwi, M., Krusche, S., & Putz, M. (2017). Zone-Based Robot Control for Safe and Efficient Interaction between Human and Industrial Robots. Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17), 83-84. doi:10.1145/3029798.3038413

Bdiwi, M., Pfeifer, M., & Sterzing, A. (2017). A new strategy for ensuring human safety during various levels of interaction with industrial robots. *CIRP Annals, 66*, 453-456. doi:10.1016/j.cirp.2017.04.009

Beggiato, M., & Krems, J.F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour, 18*, 47–57. doi:10.1016/j.trf.2012.12.006.

Brending, S., Khan, A.M., Lawo, M., Müller, M., & Zeising, P. (2016). Reducing anxiety while interacting with industrial robots. In Proceedings of the 2016 ACM International Symposium on Wearable Computers (ISWC '16), 54-55. New York, NY, USA: ACM. 54-55. doi:10.1145/2971763.2971780

Bortot, D., Born, M., & Bengler, K. (2013). Directly or on detours? How should industrial robots approximate humans? In Proceedings of 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13), 89-90. Piscataway, USA: IEEE Press. doi:10.1109/HRI.2013.6483515

Cœugnet, S., Naveteur, J., Antoine, P., & Anceaux, F. (2013). Time pressure and driving: Work, emotions and risks. *Transportation Research Part F: Traffic Psychology and Behaviour, 20*, 39–51. doi:10.1016/j.trf.2013.05.002

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In H. Kuzuoka, V. Evers, M. Imai, J. Forlizzi (Eds.), Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13), 251-258. Piscataway, USA: IEEE Press.

Dragan, A. C., Bauman, S., Forlizzi, J., & Srinivasa. S. S. (2015). Effects of Robot Motion on Human-Robot Collaboration. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15). 51-58. New York, NY, USA: ACM. doi:10.1145/2696454.2696473

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies, 58*, 697–718. doi:10.1016/S1071-5819(03)00038-7

Endler, N.S., & Kocovski, N.L. (2001). State and trait anxiety revisited. *Journal Of Anxiety Disorders, 15*, 231–245. doi: 10.1016/S0887-6185(01)00060-3

Franke, T., Attig, C., & Wessel, D. (2018). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction, 35*, 456-467. doi:10.1080/10447318.2018.1456150

Grèzes, J., Pichon, S., & de Gelder, B. (2007). Perceiving fear in dynamic body expressions. *Neuroimage, 35*, 959–967. doi:10.1016/j.neuroimage.2006.11.03

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*, 517–527. doi:10.1177/0018720811417254

ISO/TS 15066:2016-02. Roboter und Robotikgeräte – Kollaborierende Roboter. Ausgabedatum: 2016-02. Beuth-Verlag, Berlin.

Lee, J.D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153–184. doi:10.1006/ijhc.1994.1007

Lee, J.D., & See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46*, 50–80. doi: 10.1518/hfes.46.1.50.30392

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. In Proceedings of the 11th Australian Conference on Information Systems, Gladston, Australia, 6-8. doi:10.1.1.93.3874

McColl, D., & Nejat, G. (2014). Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics, 6*, 261–280. doi:10.1007/s12369-013-0226-7

Muir, B., & Moray, N. (1996) Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*, 429-460. doi:10.1080/00140139608964474

Oubari, A., Pischke, D., Jenny, M., Meißner, A., & Trübswetter, A. (2018). Mensch-Roboter-Kollaboration in der Produktion: Motivation und Einstellungen von Entscheidungsträgern in produzierenden Unternehmen. *ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, 113*, 560–564. doi:10.3139/104.111971

Parasuraman, R., & Manzey, D.H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*, 381–410. doi:10.1177/0018720810376055

Pöhler, G., Heine, T., & Deml, B. (2016). Itemanalyse und Faktorstruktur eines Fragebogens zur Messung von Vertrauen im Umgang mit automatischen Systemen. *Zeitschrift Für Arbeitswissenschaft, 70*, 151–160. doi:10.1007/s41449-016-0024-9

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers In Psychology, 10*(813), 1-13. doi:10.3389/fpsyg.2019.00813

Schmidt-Atzert, L. (1996). Lehrbuch der Emotionspsychologie. Stuttgart: Kohlhammer.

Spielberger, C.D. (1989). *State-Trait Anxiety Inventory: Bibliography* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010). Accounting for the human in cyberspace: Effects of mood on trust in automation. 2010 International Symposium on Collaborative Technologies and Systems, Chicago, IL, 180-187. USA: IEEE Press. doi: 10.1109/CTS.2010.

Tomczak, M.T., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *21*, 19–25.

Wickens, C.D., Xu, X. (2002). Automation trust, reliability and attention (Tech. Rep.AHFD-0214/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab.