Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference

Varieties of interaction: from User Experience to Neuroergonomics

Edited by
Dick de Waard, Francesco Di Nocera, Denis Coelho, Judy Edworthy, Karel Brookhuis, Fabio Ferlazzo, Thomas Franke, and Antonella Toffetti
ISSN 2333-4959 (online)

Please refer to contributions as follows:
[Authors] (2018), [Title]. D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference (pp. pagenumbers). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Available as open access

Published by HFES

Contents

AUTOMATION

Automated driving: subjective assessment of different strategies to manage drowsiness
Veronika Weinbeer, Julian-Sebastian Bill, Christoph Baur, & Klaus Bengler

Eye movements and verbal communication as indicators for the detection of system failures in a control room task
Carmen Bruder, Carolina Barzantny, & Dirk Schulze Kissing

A method to improve driver's situation awareness in automated driving
Yucheng Yang, Martin Götze, Annika Laqua, Giancarlo Caccia Dominioni, Kyosuke Kawabe, & Klaus Bengler

HMI

Persuasive assistance for safe behaviour in human-robot collaboration
Matthias Hartwig, Vanessa Budde, Alissa Platte, & Sascha Wischniewski

Comparing the effects of space flight and water immersion on sensorimotor performance
Bernhard Weber, Simon Schätzle, & Cornelia Riecke

Analysis of potentials of an HMI-concept concerning conditional automated driving for system-inexperienced vs. system-experienced users
Kassandra Bauerfeind, Amelie Stephan, Franziska Hartwich, Ina Othersen, Sebastian Hinzmann, & Lennart Bendewald

Canary in an operating room: integrated operating room music
Alistair MacDonald & Joseph Schlesinger

SURFACE TRANSPORTATION

Relevant eye-tracking parameters within short cooperative traffic scenarios
Jonas Imbsweiler, Elena Wolf, Katrin Linstedt, Johanna Hess, & Barbara Deml

Modelling driver styles based on driving data
Peter Mörtl, Andreas Festl, Peter Wimmer, Christian Kaiser, & Alexander Stocker

Graded auditory feedback based on headway: an on-road pilot study
Pavlo Bazilinskyy, Jork Stapel, Coert de Koning, Hidde Lingmont, Tjebbe de Lint, Twan van der Sijs, Florian van den Ouden, Frank Anema, & Joost de Winter

AVIATION

User performance for vehicle recognition in three-dimensional point clouds
Patrik Lif, Fredrik Bissmarck, Gustav Tolt, & Per Jonsson

Impulsivity modulates pilot decision making under uncertainty
Julia Behrend, Frédéric Dehais, & Etienne Koechlin
Innovative cockpit touch screen HMI design using Direct Manipulation
Marieke Suijkerbuijk, Wilfred Rouwhorst, Ronald Verhoeven, & Roy Arents

OTHER

Assessment of stress sources and moderators among analysts in a cyber-attack simulation context
Stéphane Deline, Laurent Guillet, Clément Guérin, & Philippe Rauffet
Potential of wearable devices for mental workload detection in different physiological activity conditions
Franziska Schmalfuß, Sebastian Mach, Kim Klüber, Bettina Habelt, Matthias Beggiato, André Körner, & Josef F. Krems
Ocular-based automatic summarization of documents: is re-reading informative about the importance of a sentence?
Orlando Ricciardi, Giovanni Serra, Federica De Falco, Piero Maggi, & Francesco Di Nocera
If Nostradamus were an Ergonomist: a review of ergonomics methods for their ability to predict accidents
Eryn Grant, Paul M. Salmon, & Nicholas J. Stevens

Automated driving: subjective assessment of different strategies to manage drowsiness

Veronika Weinbeer[1,2], Julian-Sebastian Bill[3], Christoph Baur[2], & Klaus Bengler[2]
[1] AUDI AG, [2] Technical University of Munich, [3] Otto-von-Guericke University Magdeburg
Germany

Abstract

It is likely that driver drowsiness will gain in significance as automation increases. However, as long as the automation system is unable to deal with every kind of traffic situation, it will still be necessary to get the driver back into the loop or, for example, to initiate a minimum risk manoeuvre should the transfer of the driving task to the driver fail. This article assumes that drivers are not yet allowed to sleep during an automated drive (AD). To date, it is unknown how the system should react in the case of elevated drowsiness. To evaluate this, participants (N = 31) subjectively assessed various options of a driver-state related strategy and of a system-based strategy before and after a tiring simulated AD. Assessments revealed that reducing the maximum speed was the best rated system-based option and that a targeted use of non-driving related tasks was the driver-state related option that was most widely supported. This article provides initial insights into the acceptance of various strategies for managing drowsiness during an AD from a user perspective. Further research is needed to evaluate the efficacy and safety outcomes for different strategies.

Motivation

Driver drowsiness plays an important role in vehicle safety because an increase of drowsiness is often associated with a decline in driver performance (e.g., Sagberg, 1999). So far, the study results have provided a mixed picture. Some researchers found no influence of drowsiness or automation duration on take-over performance (Feldhütter et al., 2017; Schömig et al., 2015; Jarosch et al., 2017), whereas others found a negative influence of drowsiness on the lateral acceleration during the transition (Goncalves et al., 2016) and on the time until situation awareness was reached after the transition (Vogelpohl et al., 2017). Despite these partially contradictory results, this study assumes that drivers will not be allowed to sleep during an AD as it was found that drowsiness, which Johns (1998) describes as "a transitional state between wakefulness and sleep", can already negatively influence take-over performance and the subsequent driving performance. Hence, strategies are needed to manage driver drowsiness during an AD.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Key elements of a strategy in the context of drowsiness and automated driving

Various definitions of the term "strategy" exist. Drucker (2006, p. 352) described strategic decisions as follows: "They involve either finding out what the situation is, or changing it, either finding out what the resources are or what they should be." Rumelt (2013, p. 2) described the key elements of strategic working as *"discovering the critical factors in a situation and designing a way of coordinating and focusing actions to deal with those factors"*. Based on those strategy definitions, the following concept presents the derived key elements of various strategies for dealing with drowsiness during an AD (see figure 1).

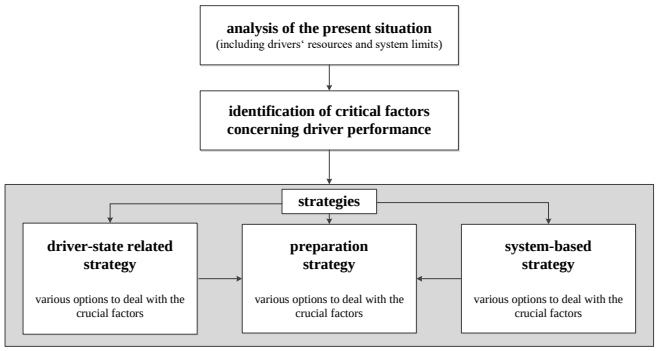


Figure 1. AD and drowsiness: Elements of a strategy and relation between different strategies

Analysis of the present situation

In order to assess the current situation, knowledge of the system state and of drivers' drowsiness states is needed. Hence, a driver monitoring system (DMS) is needed for assessing driver's drowsiness state. The technical system is understood as an Automated Driving System (ADS) according to the SAE (2016) and supplemented by a DMS. The system must be able to detect system limits, to initiate a request to intervene (RtI) and to return the driving task. For example, a motorway exit or a stationary object in front of the ego vehicle may represent system limits (Bahram et al., 2015). Further, it is assumed that driver drowsiness will also represent a system limit as long as drivers are not allowed to sleep during an AD. Reaching a system limit leads to a RtI. This article does not consider any further sensor or hardware failures.

Identification of critical factors

Two types of critical driver reactions might occur when drowsy drivers need to take over control from an automated system. On the one hand, drowsy drivers might need more time for a sufficient understanding of the current situation as found in a driving simulator study (Vogelpohl et al., 2017). On the other hand, drowsy drivers might also react in a startled or surprised way in the event of an unexpected take-over situation or RtI. Effects of being startled or surprised have already been observed in the field of aviation (Martin et al., 2012). In addition, it has been assumed that the

consideration of startle effects are of great importance, especially when the automation mode changes unexpectedly (Jacobson, 2010).

Strategies to deal with drowsiness

In order to derive strategies for managing driver drowsiness in the context of automated driving, a fundamental understanding of the underlying mechanisms is necessary. As a result, the four-process model developed by Johns (1998) is presented. This model consists of a total "wake" and a total "sleep" drive. Both types of drive inhibit each other. The wake drive consists of a primary and secondary wake drive. It is assumed that, in most cases, the secondary wake drive will determine whether the driver will fall asleep. A person's ability to avoid falling asleep may be strongly influenced by emotional and cognitive inputs (Saper et al., 2005) and by motivational aspects (Rowley, 2006). However, the secondary wake drive can change within seconds (Johns, 2000). Overall, during automated driving, the secondary wake drive can determine whether drivers will fall asleep, depending on human behaviour and the type of input. The options of a driver-state related, a system-based and a preparation strategy are presented below.

Driver-state related strategy

A driver-state related strategy is used to minimise drivers' drowsiness. Various drowsiness countermeasures during manual driving were intensively studied under certain conditions (e.g., Oron-Gilad et al., 2008; Davidsson, 2012; Gaspar et al., 2017). Nevertheless, the possibilities for minimising drowsiness during a less automated drive are limited. However, during an AD, more motivating tasks can be offered, which help drivers to avoid falling asleep or at least extend the period in which drivers' drowsiness state is acceptable. This consideration is supported by a study showing that the nature of non-driving related tasks may significantly influence participants' drowsiness level (Jarosch et al., 2017). In addition, drowsiness did not further increase when a non-driving related task (quiz) was executed, whereas high levels of drowsiness were observed when participants did not have to execute any motivating non-driving related task (Schömig et al., 2015). However, the reactivation potential of various non-driving related tasks has not yet been sufficiently investigated. In addition, a driver-state related strategy should not be condescending to drivers by limiting them to a few specific non-driving related tasks during a longer AD. Further research is thus needed in order to investigate the reactivation potential of various non-driving related tasks when drivers are already experiencing drowsiness. This raises the question of how often and at which drowsiness level a reactivation would be useful and accepted by the users. Furthermore, it needs to be taken into account that measures against sleepiness are no longer effective at higher drowsiness levels, as they are no longer executed by drivers (Hargutt, 2002, p.196). Thus, the drowsiness management concept allows a single exceedance of a critical drowsiness level (DLx) accompanied by the offer of the driver-state related strategy. If this strategy fails and DLx is exceeded on one more occasion, driver's drowsiness level is considered a system limit (see figure 2).

System-based strategy

In contrast to the driver-state related strategy, which was intended to impact the driver's drowsiness level, the system-based strategy is aimed at ensuring vehicle safety. If there is any uncertainty about whether a driver may safely retake control, the system can try to reach a service station in order to give the driver the chance to recover. In addition, the system might not perform lane changes any longer in order to be prepared, should a minimal risk condition need to be reached. The ways of achieving a minimal risk condition may differ, depending on the type of system failure (SAE, 2016):

> It may entail automatically bringing the vehicle to a stop within its current travel path, or it may entail a more extensive maneuver designed to remove the vehicle from an active lane of traffic and/or to automatically return the vehicle to a dispatching facility. (p. 9)

A speed reduction could increase the time available for a take over and decrease the intensity of the deceleration if it is necessary to stop the vehicle. The adjustment of speed under consideration of a constant deceleration as a strategy was calculated and illustrated by Bahram et al. (2015). In addition, the system can return the driving task to the driver in order to avoid a further increase in driver drowsiness during the AD. Consequently, drivers would be responsible for driving the vehicle safely after the transition. However, such a sudden RtI might be unexpected and could lead to startled or surprised reactions. A preparation strategy might be appropriate to reduce unwanted driver reactions.

Preparation strategy

One preparation strategy aims at reducing surprise factors and at reactivating the driver as well as possible within a short period of time. Therefore, suitable driver-state related and system-based options are performed simultaneously. This strategy is executed if the system limit is drowsiness and no other system limit (e.g., sensor failure) exists. For instance, drivers can obtain specific information on the current situation (such as speed limits) and they can also be asked to check the mirrors in order to obtain a sufficient overview of the situation before taking over control. Furthermore, additional system-based options should be performed to enhance overall safety.

The findings and considerations reported were grouped into the following drowsiness management concept (see figure 2). In this concept Part A represents the technical system. Part B shows the developed state machine.

A drowsiness management concept in the context of automated driving

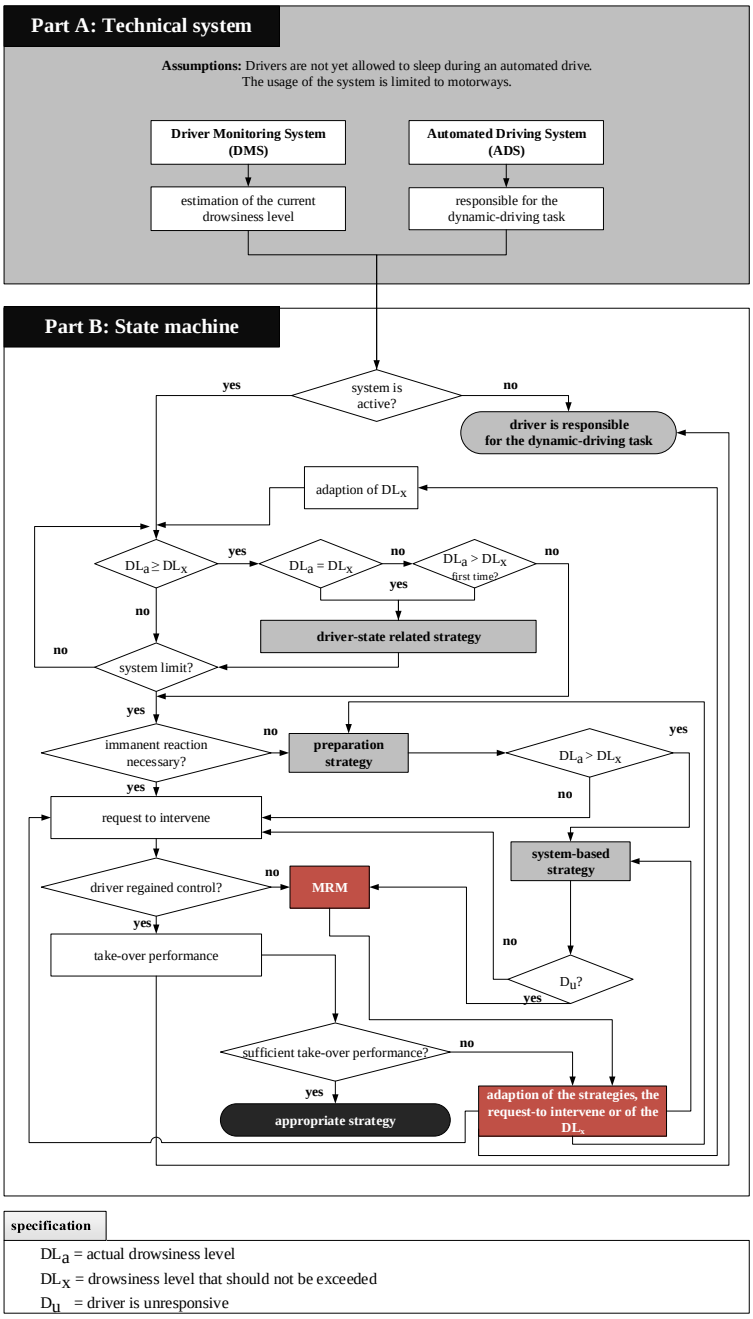


Figure 2. Framework for a drowsiness management concept

Method

Sample

The sample consisted of 31 employees of the AUDI AG (females: n = 12 and males: n = 19). On average, participants were 31 years (SD = 8) old. Data of one participant were excluded from the analysis due to constantly narrowed eyes, which made an assessment of the drowsiness level impossible. Data of another participant could not be used for the analysis of subjective assessments of the system-based strategy due to missing data.

Test vehicle, test track, drowsiness generation and assessment

A right-hand drive vehicle equipped with pedal and steering-wheel dummies (see figure 3) was used to simulate an AD in a real traffic environment. The study was conducted on the A9 autobahn in Germany from Lenting to the Nürnberg-Ost intersection and back again, representing a maximum test drive duration of 1h 30 min. Participants were informed that the automated system was simulated by an investigator. During the test drive, a curtain separated and hid the driver (investigator) from the participant. The adaptive-cruise control and lane-keeping systems were not used during this study. The maximum speed was 130 km/h. In addition, lane changes were performed very cautiously. Participants were not able to intervene in the real driving process at any time.

Figure 3. Test vehicle

In order to generate drowsiness, participants were asked not to drink caffeinated beverages for 5 hours prior to the examination. Furthermore, relaxing music was played during the simulated AD. Participants were informed that they should avoid closing their eyes and falling asleep during the entire test drive. To assess participant's drowsiness level, four cameras were integrated into the test vehicle and displayed on one screen at the back seat. Two investigators sitting in the rear assessed the participant's drowsiness level every two minutes during the test drive. The observer rated sleepiness scale used was originally developed by Wierwille and Ellsworth (1994). For further information see also Weinbeer et al. (2017).

The test drive was completed when the end of the test route was reached or when participants had reached the highest drowsiness level on the Wierwille and Ellsworth scale (1994) and had performed subsequent response-time tasks.

Purpose of the study and questionnaire

This study aims to gain initial insights into the acceptance of various options of a driver-state related and of a system-based strategy. These different options are presented in tables 1 and 2. These collections are derived from different existing measures and supplemented by some of the options that are possible due to the vehicle automation. These were assessed on a five-point Likert Scale: 1 (strong support), 2 (some support), 3 (neither support nor rejection), 4 (some rejection) and 5 (strong rejection).

Table 1. Options of the driver-state related strategy

Options of a driver-state related strategy (DSRS)	
"Imagine that your drowsiness level increases constantly during a highly-automated drive. In order to keep the system going as long as possible, your drowsiness level needs to be kept at a low level. Please rate how far you would support or reject the following adjustments."	
DSRS-O1:	The vehicle opens the window slightly in order to allow fresh air into the vehicle.
DSRS-O2:	The vehicle emits a scent to stimulate you.
DSRS-O3:	The vehicle increases the volume of the radio.
DSRS-O4:	The vehicle moves the seat into an upright position.
DSRS-O5:	The vehicle adjusts the interior lighting.
DSRS-O6:	The vehicle offers a specific selection of non-driving related tasks (for example a quiz) during the automated drive.

Table 2. Options of the system-based strategy

Options of a system-based strategy (SBS)	
"Imagine that your drowsiness level increases constantly during a highly-automated drive. In order to ensure your safety the system adapts at a certain drowsiness level. Please rate how far you would support or reject the following adjustments."	
SBS-O1:	The vehicle ceases to change lanes and drives on the right lane so that the vehicle can come to a safe stop on the hard shoulder should you fall asleep.
SBS-O2:	The vehicle hands the driving task back to you. After that the system will no longer be available. You take full responsibility for the subsequent drive without the system.
SBS-O3:	The vehicle drives to the next rest area. The system will be available again after a break, depending on your level of drowsiness.
SBS-O4:	The vehicle reduces the maximum speed to give you more time to take control in case of a request to intervene.
SBS-O5:	The vehicle drives without any adjustment. When it recognises that you have fallen asleep, it brakes, coming to a stop on the hard shoulder.
SBS-O6:	The vehicle drives without any adjustment. When it recognises that you have fallen asleep, it brakes, coming to a stop on the lane you are in.

In addition, evaluations were conducted into whether suffering drowsiness led to a change in the subjective assessment of the driver-state related and system-based strategies. These strategies and the 5-Point-Likert Scale were translated from German into English in order to present the results. The preparation strategy is not evaluated in this study as it represents a combination of the driver-state related and system-based strategy. Participants were also asked to declare the option that they would accept most from the driver-state related and system-based options. In addition, participants were asked to declare the most effective driver-state related option.

The collections of the driver-state related (see table 1) and system-based options (see table 2) were assessed before (S1) and after the test drive (S2) (see table 3).

Table 3. Experimental design

Test procedure	*Subjective assessment*	*RtI in dependence of the drowsiness level*			*Subjective assessment*
		DL 1	*DL 4*	*DL 6*	
Group A n = 16	**S_1** (before the test drive)	Group A (DL1)	Group A (DL4)	Group A (DL6)	**S_2** (after the test drive)
Group B n = 15		x	Group B (DL4)	Group B (DL6)	

Furthermore, the effectiveness of the drowsiness manipulation procedure and the influence of different drowsiness levels on take-over-time aspects were assessed in this experimental setting. As the presentation of these results is beyond the scope of the present article, the results are reported in a separate paper (Weinbeer et al., 2017).

Results

Driver-state related strategy (DSRS)

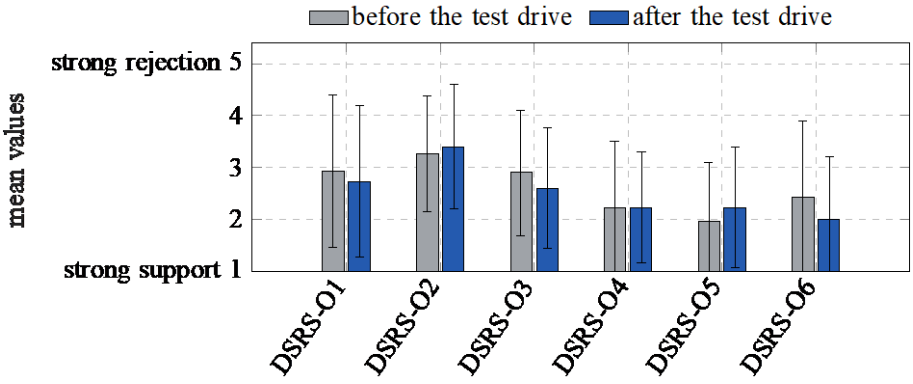


Figure 4. Subjective assessments of various options of a DSRS (Mean ± 1SD)

The mean values of the options assessed before and after the test drive are presented in table 4. After the test drive a targeted offer of non-driving related tasks (e.g., a quiz) received most support (see figure 4). Due to the multiple comparisons, the significane level was adjusted to $p < .008$. A Wilcoxon signed-rank test revealed no

significant differences between the ratings before and after the test drive for the options of a DSRS.

When asked which type of option would be most widely accepted, DSRS-O6 was seen to be most popular, with 26% mentions before the test drive and 30% after it. Participants also assessed DSRS-O6 as the most effective option with 30% mentions before the test drive and 40% after it. These results are presented in tables 5 and 6.

Table 4. Subjective assessment of various options of a driver-state related strategy

N = 30		*DSRS-O1*	*DSRS-O2*	*DSRS-O3*	*DSRS-O4*	*DSRS-O5*	*DSRS-O6*
Before the	*M*	2.93	3.27	2.90	2.23	1.97	2.43
test drive	*SD*	1.46	1.11	1.21	1.28	1.13	1.46
After the	*M*	2.73	3.40	2.60	2.23	2.23	2.00
test drive	*SD*	1.46	1.19	1.16	1.07	1.17	1.20
	z	-1.90	-1.27	-1.70	-0.04	-2.14	-2.41
	p-value	.058	.206	.089	.971	.033	.016

Table 5. Which driver-state related adjustment would you accept most? - Place 1

N = 30	*DSRS-O1*	*DSRS-O2*	*DSRS-O3*	*DSRS-O4*	*DSRS-O5*	*DSRS-O6*
Before the test drive	16.7%	6.7%	10.0%	20.0%	20.0%	26.7%
after the test drive	16.7%	6.7%	13.3%	20.0%	13.3%	30.0%

Table 6. Which kind of DSRS-O do you believe is most effective (most reactivating)? - Place 1

N = 30	*DSRS-O1*	*DSRS-O2*	*DSRS-O3*	*DSRS-O4*	*DSRS-O5*	*DSRS-O6*
Before the test drive	23.3%	0.0%	20.0%	20.0%	6.7%	30.0%
after the test drive	33.3%	3.3%	6.7%	10.0%	6.7%	40.0%

System-based strategy

The mean ratings of the different options of a system-based strategy before and after the test drive are presented in table 7.

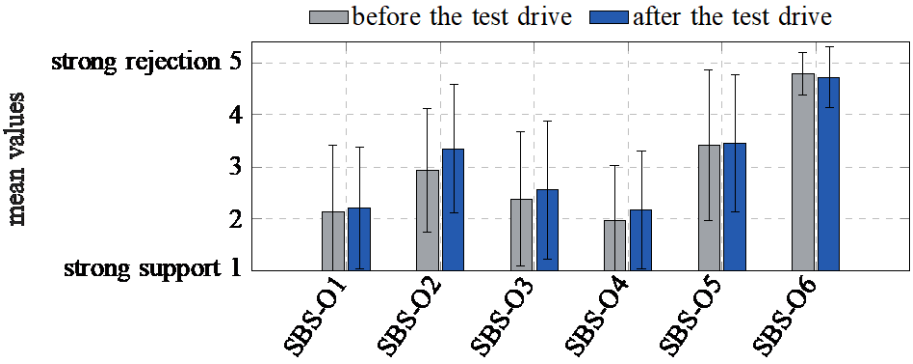


Figure 5. Subjective assessment of various options for a SBS (Mean±1SD)

After the test drive, SBS-O4 (reduction in maximum speed) was given most support, followed by SBS-O1 (no further lane changes and a move to the slow lane). The Wilcoxon signed-rank test revealed no significant difference between the ratings before and after the test drive for the options of a SBS (adjusted significance level

$p < .008$). SBS-O1 was most widely accepted with 37.9% mentions before the test drive and 31.0% afterwards. After the test drive, support for SBS-O4 and SBS-O3 (rest area) was the same. These results are presented in table 8.

Table 7. Subjective assessment of various options for a system-based strategy

N = 29		*SBS-O1*	*SBS-O2*	*SBS-O3*	*SBS-O4*	*SBS-O5*	*SBS-O6*
Before the	*M*	2.14	2.93	2.38	1.97	3.41	4.79
test drive	*SD*	1.27	1.19	1.29	1.05	1.45	0.41
After the	*M*	2.21	3.34	2.55	2.17	3.45	4.72
test drive	*SD*	1.18	1.23	1.33	1.14	1.33	0.59
	Z	-0.25	-1.96	-0.79	-1.26	-0.18	-0.63
	p-value	.799	.049	.431	.207	.858	.527

Table 8. Which kind of system-based adjustment would you accept most? - Place 1

N = 29	*SBS-O1*	*SBS-O2*	*SBS-O3*	*SBS-O4*	*SBS-O5*	*SBS-O6*
Before the test drive	37.9%	10.3%	17.2%	20.7%	13.8%	0.0%
after the test drive	31.0%	10.3%	24.1%	24.1%	10.3%	0.0%

Discussion and limitations

Of the driver-state related options, DSRS-O4 (upright seat position), DSRS-O5 (interior lighting) and DSRS-O6 (targeted offer of non-driving related tasks) received the most support (see table 4). The differences between these options were small when subjects were asked whether they support or reject these adaptions. However, when asked which of these options one would accept most, DSRS-O6 was mentioned most frequently (30%) and rated to be the most effective by 40% of the sample. Based on these results, it can be concluded that offering non-driving related tasks in order to provide the automated driving system as longs as possible would be widely accepted. However, further research is needed in order to investigate various non-driving related tasks and the effectiveness of these in reality.

On average, SBS-O4 (reduction of the maximum speed) obtained most support at the end of the test drive (see table 7). However, when asked which of the system-based options would be most widely accepted, SBS-O1 (no further lane changes and a move to the slow lane) was selected more frequently (31.0 %) than SBS-O4 (24.1%). SBS-O3 (rest area and break) was also mentioned by 24.1% of the sample. The options SBS-O5 (vehicle comes to a stop on the hard shoulder if the driver falls asleep) and SBS-O6 (vehicle comes to a stop on the current lane if the driver falls asleep) were rejected by the majority of participants. However, it needs to be considered that the different system-based options also represent different levels of escalation. The present results show that higher levels of escalation were rejected by the majority of the participants representing the user perspective. However, the evaluation may be dependend on the point of view. For instance, the perspective of other road users (e.g., driver of the following vehicle) may differ from the users' perspective regarding the appropriate system-based option. Therefore, further research should focus on the comparison of the different perspectives. In case of contradicting evaluations system developers face a dilemma: on the one hand, they must develop systems that are safe and accepted by users, on the other hand, they must develop automated driving systems that are safe and accepted by other road

users. Consequently, a holisitic view is needed for developing safe and accepted systems.

As the assessment of the driver-state related and system-based options were very similar before and after the test drive, it can be assumed that experiencing drowsiness did not essentially influence the subjective ratings of the different options.

The drowsiness management concept developed presents a framework for managing driver drowsiness during an AD. However, it must be borne in mind that this concept expects the DMS to be able to assess the drowsiness level consistently and reliably. Hence, it is necessary to take the performance of a DMS into account because an incorrect timing of the different strategies could lower their effectiveness. Further research is needed to derive the requirements for driver monitoring systems and to identify the critical drowsiness level. In addition, it is necessary to investigate whether (and to what extent) this critical drowsiness level differs between drivers.

Conclusion

In this article, a drowsiness management concept illustrates the relationship between a driver's drowsiness level and possible strategies to deal with it. Subjective assessments revealed that a specific offer of non-driving-related tasks has the potential to be an accepted driver-state related option. However, further research is needed to investigate various non-driving related tasks and their real effectiveness. In the case of a system-based strategy, a reduction in maximum speed, an adjustment of driving behaviour (no further lane changes and driving on the slow lane) or a rest at a service station were rated highest. In contrast, a minimum risk manoeuvre that would stop the vehicle on the emergency or ego lane was rejected by the majority of participants. These results demonstrate that from a users' perspective higher levels of escalation should be avoided. However, the perspective of other road users still remains unclear. Therefore, it needs to be investigated whether and to what extent this perspective differs compared to the users' perspective. Further, the idea of a preparation strategy, the drowsiness management concept developed and the safety outcomes regarding the take-over and the subsequent driver performance for the strategies derived need to be assessed.

Acknowledgement

This work results from the joint project Ko-HAF - Cooperative Highly Automated Driving and has been funded by the Federal Ministry for Economic Affairs and Energy based on a resolution of the German Bundestag.

References

Bahram, M., Aeberhard, M., & Wollherr, D. (2015). Please take over! An analysis and strategy for a driver take over request during autonomous driving. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, (pp. 913–919).

Davidsson, S. (2012). Countermeasure drowsiness by design - using common behaviour. *Work, 41*, 5062–5067.

Drucker, P.F. (2006). *The Practice of Management. Burlington*. HarperBusiness.USA: New York.

Feldhütter, A., Gold, C., Schneider, S., & Bengler, K. (2017). How the Duration of Automated Driving Influences Take-Over Performance and Gaze Behavior. In C.M. Schlick et al. (Eds.), *Advances in Ergonomic Design of Systems, Products and Processes* (pp. 309-318). Germany: Berlin, Heidelberg. Springer Berlin Heidelberg.

Gaspar, J.G., Brown, T.L., Schwarz, C.W., Lee, J.D., Kang, J., & Higgins, J.S. (2017). Evaluating driver drowsiness countermeasures. *Traffic Injury Prevention, 18,* 58-63.

Goncalves, J., Happee, R., & Bengler, K. (2016). Drowsiness in Conditional Automation: proneness, diagnosis and driving performance effects. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 873–878).

Hargutt, V. (2002). Das Lidschlussverhalten als Indikator für Aufmerksamkeits- und Müdigkeitsprozesse bei Arbeitshandlungen. PhD thesis, Julius-Maximilians University of Würzburg. Germany: Philosophy faculty.

Jacobson, S.R. (2010). Aircraft Loss of Control Causal Factors and Mitigation Challenges. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100039467.pdf (checked on: 04.09.2017).

Jarosch, O., Kuhnt, M., Paradies, S., & Bengler, K. (2017). It's Out of Our Hands Now! Effects of Non-Driving Related Tasks During Highly Automated Driving on Drivers' Fatigue. In *Driving Assessment Conference*.

Johns, M. (1998). Rethinking the assessment of sleepiness. *Sleep Medicine Reviews, 2*, 3–15.

Johns, M.W. (2000). A sleep physiologist's view of the drowsy driver. In *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*, 241–249.

Martin, W.L., Murray, P.S., & Bates, P.R. (2012). The Effects of Startle on Pilots During Critical Events: A Case Study Analysis. https://research-repository.griffith.edu.au/handle/10072/54072 (retrieved 04.09.2017).

Oron-Gilad, T., Ronen, A., & Shinar, D. (2008). Alertness maintaining tasks (AMTs) while driving. In *Accident Analysis and Prevention*, *40*, 851–860.

Rowley, J.A. (2006). Measuring the ability to stay awake: role of motivation. *Sleep Breath*, *10*, 171–172.

Rumelt, R.P. (2013). *Good strategy, bad strategy. The difference and why it matters*. Profile Books. England: London.

SAE (2016). *J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. USA: SAE Society of Automotive Engineers.

Sagberg, F. (1999). Road accidents caused by drivers falling asleep. *Accident Analysis & Prevention*, *31*, 639–649.

Saper, C.B., Cano, G., & Scammell, T.E. (2005). Homeostatic, circadian, and emotional regulation of sleep. *The Journal of comparative neurology*, *493*, 92–98.

Schömig, N., Hargutt, V., Neukum, A., Petermann-Stock, I., & Othersen, I. (2015). The Interaction Between Highly Automated Driving and the Development of Drowsiness. *Procedia Manufacturing*, *3*, 6652–6659.

Vogelpohl, T., Vollrath, M., & Kühn, M. (2017). 'Übergabe von hochautomatisiertem Fahren zu manueller Steuerung - Teil 2 -', *Unfallforschung der Versicherer*, Forschungsbericht Nr. 47.

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 *8. Tagung Fahrerassistenz*.

Wierwille, W.W. and Ellsworth, L.A. (1994). Evaluation of driver drowsiness by trained raters. *Accident Analysis & Prevention*, *26*, 571–581.

Eye movements and verbal communication as indicators for the detection of system failures in a control room task

Carmen Bruder, Carolina Barzantny, & Dirk Schulze Kissing
German Aerospace Center, Department of Aviation and Space Psychology
Germany

Abstract

In modern control rooms, operators need to monitor visual information representing large technical systems. Operators usually monitor together in teams in order to detect abnormal system behaviour in time. It remains an open question which performance indicators are valuable for assessing a team member's capabilities of detecting abnormal system behaviour. The present study investigates the value of monitoring behaviour and communication behaviour for predicting the performance results of subjects attempting to detect system failures while executing a control room task. A simulation of a generic control room was implemented in order to enable synchronized measurement of monitoring processes in teams. The monitoring behaviour was measured by tracking the eye movements of the team members while they were monitoring for system failures. Simultaneously, the communication behaviour between team members was recorded. Eye-tracking data and communication data were analysed including the interaction with team members' performance in detecting system failures in time. Data from 21 three-member teams indicate that there are significant differences in communication and to some extent in eye-movement, between operators who detect system failures in time and those who fail to do so. The findings are discussed in the context of personnel selection and training team members in control rooms.

Introduction

This paper presents an eye-tracking study that investigates the monitoring and communication behaviour of operators while collaboratively supervising the dynamic processes of a control room simulation. In this study, monitoring behaviour was measured using eye tracking. By tracking the operator's eye movements, the visual attention processes while gathering relevant information as well as detecting abnormal system behaviour could be visualized. Furthermore, recording verbal communication behaviour between team members makes it possible to indicate the coordinative processes while monitoring together. By specifically investigating how monitoring and communication behaviour can be used to predict the performance of operators attempting to detect system failures, the goal is to provide initial indications for selecting and training operators in control room teams.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Collaborative monitoring in control room teams

The control room is an example of a working environment where operators supervise complex and dynamic processes together. Control rooms can be found particularly in domains where safety is of critical importance, such as airport operational centers, air traffic control centers, nuclear power plant control and military control centers, where human error can have severe consequences (Hauland, 2008; Salas et al. 2008). As monitoring is one of the core tasks in control rooms, teams of operators are required to monitor the system appropriately (Sharma et al., 2016). In control rooms, not only is the individual situation awareness relevant, but also the situation awareness of the team. Through interactions, operators in a team can dynamically modify each other's perceptual and active capabilities (Gorman et al., 2006). However, when monitoring a system, it is essential that team members work together effectively and cooperatively (Cooke et al., 2000; Salas et al., 2008). In order to coordinate their activities in such "centers of coordination," not only do individuals have to be aware of their own situation, but they must also be aware of their team members' situation (e.g. Suchman, 1997).

The importance of communication in control operations has been stressed by Carvalho et al. (2007). Communication as a "meta-teamwork process that enables the other processes" (Papenfuss, 2013, p. 319) provides indications for the coordinative activities while monitoring. Cooke et al. (2013) stressed that, especially in critical situations, "it is not only critical that teams correctly assess the state of the environment and take action, but how this is accomplished (p. 279)". As a consequence, recording the quality and degree of a team's communication provides insight into how the group deals with critical situations.

Measuring collaborative monitoring

A variety of studies support the idea that eye movements offer an appropriate means for measuring the efficient and timely acquisition of visual information (e.g. Findlay & Gilchrist, 2003; Underwood et al., 2003; for an overview see Holmqvist et al., 2011). Based on this research, eye movement parameters that reflect the human monitoring performance have been identified (Grasshoff et al., 2015; Hasse & Bruder, 2015). Bruder et al. (2014) investigated the link between these eye movement parameters and the monitoring behaviour of experts, compared the monitoring behaviour of experts with novices (Bruder et al., 2013), and used eye movements to research differences in monitoring behaviour resulting in detected automation failures and behaviour resulting in missed failures (Bruder & Hasse, 2016).

While the results of previous studies give valuable insight into eye movements during the process of monitoring individually, the present study focuses on collaborative monitoring behaviour in a team task. In this context, monitoring behaviour leading accurate failure detection will be compared with monitoring behaviour that leads to missed failures. Additionally, the communication behaviour while monitoring will be taken into account. The following research questions will be addressed: What are valuable performance indicators in a team task with respect

to communication quality and monitoring behaviour that differentiate between accurate failure-detection and missed failures?

Method

An empirical study was undertaken requiring collaborative monitoring while performing a control room team task.

Simulation of a generic control room

In the present study, the simulation of a generic control room, called ConCenT (Generic Control Center Task Environment), was used to enable synchronized measurement of monitoring processes in teams (Schulze-Kissing & Bruder, 2016). ConCenT replicates different control room tasks by simulating the production processes of several technical facilities spread over three locations, which are supervised by a team of three human operators. It simulates four different tasks: monitoring the distributed production processes, reporting system deviations (failures), diagnosing the sources of deviations and remedying the deviating processes by deciding between two alternative choices. These four tasks have to be managed within a team of three operators. Since this paper presents findings concerning the monitoring task and the reporting task, these two tasks are described in more detail. Figure 1 shows a screenshot of the monitoring screen of ConCenT.

Figure 1. Monitoring screen of ConCenT containing the displays of nine production facilities and three power stations, which are distributed over three locations

In the monitoring and reporting task, each team member had to observe nine of 27 gauges in total and three joint power station gauges with the objective of reporting deviations from standard processes within a time span of four seconds. Each of the 27 gauges represented the production processes of a single production line. Deviations could be recognized when one of the black arrows, indicating the current value on each of the 27 gauges, exceeded or fell below the tolerance range (marked green). Before a deviation occurred, a specific constellation of production processes indicated this kind of critical situation. Critical situations could only be identified when the distributed information on the production processes was communicated

between team members. As a consequence, the team was able to anticipate deviations in the production. Sharing all relevant information on the production processes therefore helped identify critical situations and anticipate as well as helped report any system deviations.

Eye tracking system

Each participant was seated in front of a 24-inch LCD computer display at a distance of approximately 60 cm. Eye movements were recorded remotely by using the Eye Follower System manufactured by LC Technologies, Inc. The system operated at 120 Hz and was combined with the simulation tool ConCenT to ensure that both systems used the same timestamp. The fixation-detection algorithm was set with a minimum sample for fixation detection of six gazes on a particular screen point – within the deviation threshold of 25 pixels.

Sample

The study was conducted with a sample size of $N = 63$. Of this total, 41 individuals were applicants for air traffic control training (ATC) at DFS (German Air Traffic Control), while the remaining 22 individuals were students and graduates from different universities. All participants were between 18 and 34 years old ($M = 21.57$, $SD = 3.39$) and 47.6% were female (52.4% male). ATC participants were recruited with a personal call from DLR (German Aerospace Center), Hamburg, and compensated €25 for their participation in the 2.5hrs experiment. Students were recruited via social media and with flyers posted on the campus of the University of Hamburg.

Procedure

The three participants in each team performed the experiment at the same time, each with a separate computer and eye tracking system. A room divider was installed between the participants to prevent direct communication and eye contact. Written instructions introduced participants to their general tasks as operators working in a control center, and explained their specific responsibilities while monitoring the system, diagnosing errors and solving problems. Following this, each team was guided through a practice scenario that lasted about ten minutes. Throughout the practice scenario, participants familiarized themselves with how to anticipate, detect and report deviations from standard processes in time. After the practice scenario, participants confirmed their understanding of the monitoring procedure and the other required tasks. The test scenario began with the ramp-up of the gauges and ended after 72 minutes. A manipulation check was done, and participants were required to complete a questionnaire regarding their attitudes towards teamwork. Finally, participants were asked to give their impressions of the study.

Design and measurements

The present study investigates the relationship between team members monitoring as well as communication behaviour and their capabilities of detecting system deviations. The dependent variables included the monitoring behaviour (tracking eye movements) and the quality of communication. The quasi-independent variable was

the performance level (deviation reported successfully vs. deviation missed). These two groups (cases of successful detection of deviations and cases of missed deviations) were created post-hoc. A deviation was successfully reported if a participant clicked on the button "Diagnose" next to the gauge within the corresponding time frame (4s). Each of the six deviations could either be detected (= successful detection of deviation) or not detected (= missed).

Measuring monitoring behaviour and communication quality

Eye movements were recorded while monitoring the distributed production processes as well as reporting system deviations. Afterwards, they were synchronized with the logged simulation events before and during the occurrence of deviations. At first, twelve areas of interest were defined for each team partner (A, B, C): nine gauges for the production processes and three gauges for the power stations. For each of the six deviations in the test scenario, AOIs were predefined according to where an operator's attention should be allocated within the interval before and while a deviation occurred. It was defined in advance, which gauges must be monitored to anticipate system deviations and this decision was based on the information necessary for detecting critical situations.

Regarding the timely allocation of attention on relevant AOIs when detecting deviations, four successive monitoring phases were defined (1. identification phase, 2. verification phase, 3. anticipation phase, 4. detection phase). Within each of these four monitoring phases, the team member had the opportunity to share their information in order to allocate their attention in an ideal way. In the first two phases, identification and verification, the team member had to share their information to find out whether or not there was a critical situation. In the third phase (anticipation), they had to anticipate the gauge where the deviation could happen. In the last phase (detection), the deviation could occur and had to be reported. The eye tracking parameters on the relevant AOIs were analysed for each monitoring phase, team partner and deviation.

The relative fixation count (rfc) was calculated in terms of the predefined, relevant AOIs for each of the four monitoring phases. The rfc is defined as the ratio between the number of fixations on relevant AOIs and all fixations within a given time span. Relative parameters ranged from 0 to 1, with 0 indicating that no eye movements fell on predefined AOIs within a time period, and with 1 indicating that all eye movements fell on the predefined AOIs within that time period.

During the test scenario, the verbal communication of each team member was recorded. An audio file logged the identities of each speaker, the content of the information exchanged, and the duration of this communication. Each audio file was analysed with respect to the necessary communication in all six intervals before a system deviation. This analysis provided the basis for determining the quality of communication. For each of the four monitoring phases, participants could score on a scale from 0 (no communication or wrong communication of necessary information) to 1 (right communication/no communication needed) in each of the 6 intervals before a system deviation.

Results

Data from 52 subjects were reported, each of whom experienced six deviations within the test scenario. Data were excluded from the reported results when a scenario was not completed due to technological problems (18.1%), if they failed the manipulation check the manipulation check was not passed (4.8%), and when eye movement data were missing or showed major inconsistencies (3.2%). For communication analyses, additional data were excluded when no communication was recorded by the system (14.8%). In sum, eye-tracking data, communication data and deviation-detection data from 212 deviations were included in the statistical analyses. On a scale from 0 to 6, an average of 4.33 (SD = 1.37) deviations were reported with an average response time of 2.17 seconds (SD = 0.56; see Table 1 for a detailed overview).

Table 1. Descriptive performance data (N = 52)

	Deviation detected		*Response time*	
Deviation	*n*	*%*	*M*	*SD*
1	25	48.1	2.79	0.74
2	25	48.1	2.68	0.76
3	44	84.6	1.95	0.70
4	42	80.8	2.01	0.87
5	42	80.8	2.00	0.67
6	47	90.4	1.76	0.70
All			2.17	0.56

Looking at the eye tracking data, the attention allocation of the test subjects implies that in the case of successful detection of deviations, relevant AOIs were focused on more intensively if the deviation was detected successfully (see Figure 2, which shows the second deviation in the test scenario as an example).

Figure 2. Comparison of attention allocation in a case of successful detection of deviation (left) and a missed deviation (right), illustrated by the eye tracking data (N = 52) during the anticipation phase of the second deviation in the test scenario (marked yellow)

A variance analysis with repeated measurement was conducted to compare the main effects of monitoring phase and the interaction effect between monitoring phase and performance in detecting deviations on the relative fixation count. The factors PHASE (four levels: identification, verification, anticipation, detection) and DETECTION (two levels: detected, not detected) were defined and analysed. See Table 2 for descriptive data of the eye tracking parameter. Multivariate tests showed

a significant effect for PHASE [F (3, 240) = 5.038, p < .005; Wilk's λ =.94, partial ε² = .059]. It could be shown that subjects fixated relevant AOIs most frequently within the identification phase (1) and the verification phase (2). No significant effect of the interaction between PHASE and DETECTION [F (3, 240) = 2.297, p = .078; Wilk's λ =.97, partial ε² = .028] on eye tracking parameter was found. Post hoc tests indicated that accurate deviation detection is only related to a higher frequency of fixations on relevant AOIs during the anticipation phase [t(305)=-2.22, p<.05)]. Concerning the identification phase, verification phase and detection phase, differences between cases of accurate and missed deviation detection were not significant [p > .05]. The interaction of DETECTION and PHASE on relative fixation counts on relevant AOIs is shown in Figure 3 (left).

Table 2. Descriptive data for the eye tracking parameter (relative fixation count) and communication quality parameter in the four monitoring phases (rows), separately for deviations detected and deviations NOT detected (columns).

	Deviation detected		*Deviation NOT detected*	
	M	*SD*	*M*	*SD*
Relative fixation counts				
Identification (1)	0.47	0.25	0.51	0.20
Verification (2)	0.48	0.26	0.44	0.28
Anticipation (3)	0.45	0.23	0.37	0.19
Detection (4)	0.44	0.20	0.38	0.18
Communication quality				
Identification (1)	0.99	0.11	0.86	0.38
Verification (2)	0.47	0.50	0.31	0.47
Anticipation (3)	0.74	0.44	0.55	0.50
Detection (4)	0.15	0.36	0.25	0.44

Following, a variance analysis with repeated measurement was conducted to compare the main effects of monitoring phase and the interaction effect between monitoring phase and performance in detecting deviations on communication quality. The factors PHASE (four levels: identification, verification, anticipation, detection) and DETECTION (two levels: detected, not detected) were defined and analysed. See Table 2 for descriptive data of the communication quality. Multivariate tests showed a significant effect for PHASE [F (3, 320) = 336.142, p < .001; Wilk's λ =.24, partial ε² = .759]. It could be shown that subjects communicated accurate information most frequently during the identification phase (1) and anticipation phase (3). The interaction between PHASE and DETECTION [F (3, 320) = 5.457, p < .005; Wilk's λ =.95, partial ε² = .049] on communication quality was found. Post hoc tests showed that accurate deviation detection is related to higher communication quality during the identification phase [t(98.11)=-3.20, p<.05)], verification phase [t(182.43)=-2.42, p<.05)] and anticipation phase [t(152.92)=-3.61, p<.05)], but not during the detection phase [p > .05]. The interaction of DETECTION and PHASE on communication quality is shown in Figure 3 (right).

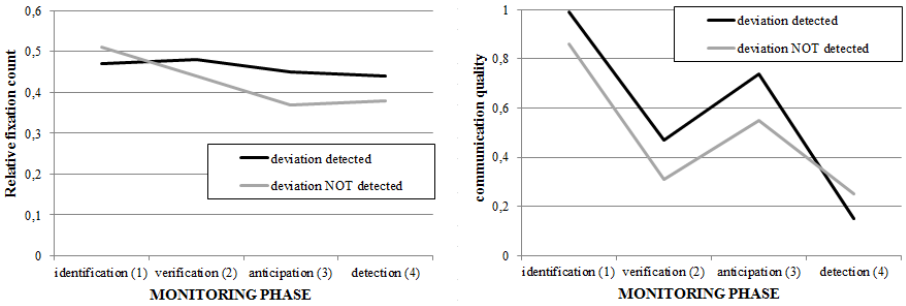


*Figure 3. Interaction effects of detection * time unit (estimated mean values) on the communication quality as the relative frequency of correctly communicated information (left) and on the relative fixation counts on relevant AOI (right)*

Discussion and further research

The present study investigated the role of monitoring behaviour and communication behaviour as performance predictors for the detection of failures (=deviations) in a control room team task. To the subjects were given the task of monitoring dynamic processes in a team of three operators with the objective of anticipating and detecting deviations from standard processes by communicating relevant information adequately. To summarize the results, data from 21 three-member teams indicate that there are significant differences in communication and to some extent in eye-movement, between operators who detect system deviations in time and those who miss the deviations. This is shown by the fact that successful failure detection is related to a higher frequency of communication and focusing attention on relevant information during the anticipation phase.

Comparing the predictive value of communication quality and monitoring behaviour, the relationship between the frequency of monitoring relevant information and the detection of system deviations is clearly weaker than the relationship between the frequency of communicating relevant information and the detection of system deviations. However, in the case of successful failure detection, relevant information is monitored more frequently shortly before the deviation occurs when the automation failure should be anticipated. This is quite understandable, because monitoring relevant information within the anticipation phase is only possible if the subject has identified the critical production system together with the team partners, thus leading to successful detection of system deviations in time.

Contrary to prior expectations, no substantial relationship between successful deviation detection and monitoring behaviour within the identification phase, information phase and detection phase was found. Besides this, the effect sizes on eye tracking parameters are small. These may be due to the fact that technical problems lead to losses of eye tracking data, but also to certain methodological shortcomings of predicting deviation detection by means of the eye movements of human operators. Further research will improve the reliability of eye-movement

indicators by adjusting the definition of information that is relevant for detecting deviations.

With respect to communication behaviour, the differences between detected and missed automation failures were highest when the system deviation could be verified, which happened in the second monitoring phase. This result implies that successful failure detection is highly related to adequate communication of relevant information at the beginning of an upcoming situation. A deviation can only be detected in time if the team members communicate the relevant information and identify the critical production system together with the team partners.

Predicting the detection of system failures in a team task within a dynamic setting using eye tracking and communication quality is an innovative strategy that enables the development of new approaches for personnel selection and training. Learning from the differences in monitoring and communication behaviour between successful and unsuccessful failure detection will be helpful in selecting successful trainees and providing them with appropriate training. Especially the monitoring and communication patterns related to successful detection may be useful in order to give trainees direct feedback on their own monitoring behaviour or to demonstrate "correct" monitoring behaviour.

Further research is replicating this study with a larger sample of 48 teams and prior technical problems are being reduced, which will lead to a significant gain in the volume of data. In contrast to the study reported here, in further research the effect of team coordination within a monitoring task is systematically investigated by comparing the monitoring behaviour of communicating teams to a control condition where all channels for oral communication are blocked.

References

Bruder, C., Eißfeldt, H., Maschke, P., & Hasse, C. (2013). Differences in monitoring between experts and novices. In *Proceedings of the HFES 57th Annual Meeting, 2013* (pp. 295-298). Sage, Thousand Oaks, CA.

Bruder, C., Eißfeldt, H, Maschke, P., & Hasse, C. (2014). A model for future aviation: Operators monitoring appropriately. *Aviation Psychology and Applied Human Factors, 4*, 13-22.

Bruder, C., Weber, P., & Hasse, C. (2016). To Look and (Not) See: Predicting the Detection of Automation Failures Based on the Eye Movements of Human Operators. In *Proceeding of the HCI-Aero '16 Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. New York, NY, USA: ACM Press.

Carvalho, P.V.R., Vidal, M.C.R., & de Carvalho, E.F. (2007). Nuclear power plant communications in normative and actual practice: A field study of control room operators' communications. *Human Factors in Ergonomics and Manufacturing, 17*, 43–78.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A. & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42*, 151-173.

Cooke, N.J., Gorman, J.C., Myers, C.W. & Duran, J.L. (2013). Interactive team cognition. *Cognitive Science, 37*, 255–285.

Findlay, J.M. & Gilchrist, I.D. (2003). *Active Vision*. Oxford (UK): Oxford University Press.

Gorman, J.C., Cooke, N.J., & Winner, J.L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics, 49*, 1312-1325.

Grasshoff, D., Hasse, C., Bruder, C., & Eißfeldt, H. (2015). On the development of a monitoring test for the selection of aviation operators. In D. Harris (Ed.), *Proceedings of Engineering Psychology and Cognitive Ergonomics, 12th International Conference EPCE 2015, held as Part of HCI International 2015 (*pp. 537-546). Berlin, Heidelberg: Springer.

Hasse, C. & Bruder, C. (2015). Eye Tracking Measurements and their Link to a Normative Model of Monitoring Behaviour. *Ergonomics, 58*, 355-367.

Hauland, G. (2008). Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. *The International Journal of Aviation Psychology*, *18*, 290-304

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van De Weijer, J. (2011). *Eye Tracking. A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.

Papenfuss, A. (2013). Phenotypes of teamwork - an exploratory study of tower controller teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp.319-323).

Salas, E., Cooke, N.J., & Rosen, M.A. (2008). On Teams, Teamwork, and Team Performance: Discoveries and Developments. *Human Factors*, *50*, 540-547.

Sharma, C., Bhavsar, P., Srinivasan, B. & Srinivasan, R. (2016). Eye gaze movement studies of control room operators: A novel approach to improve process safety. *Computers & Chemical Engineering*, *85*, 43-57.

Suchman, L. (1997). Centers of coordination: A case and some themes. In L.B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, tools and reasoning: Essays on situated cognition* (pp. 41–62). Berlin: Springer.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. *Ergonomics, 46*, 629-646.

A method to improve driver's situation awareness in automated driving

Yucheng Yang[1], Martin Götze[1], Annika Laqua[1], Giancarlo Caccia Dominioni[2], Kyosuke Kawabe[2], & Klaus Bengler[1]
[1]Chair of Ergonomics, Technical University of Munich
Germany
[2] Toyota Motor Europe NV/SA
Belgium

Abstract

In the future, raising automation levels in vehicles is an imaginable scenario. However, there will be situations, which cannot be handled by the automation and the driver should take-over the driving task within a specific time budget. With a level 3 system (according to SAE), the driver no longer has to monitor the driving environment and, therefore, could perform other non-driving related tasks; consequently, leading to lower situation awareness (SA) and possibly worse take-over performance. In this paper, two versions of new visual advanced driving assistance systems are presented, which display subliminal information about the system states and confidence levels of the automation system. The goal is to increase the SA during automation and improve the take-over quality while allowing the driver to perform secondary tasks without distraction and annoyance. In this mixed design experiment, 32 participants performed a visual-motor task on a smartphone under 20 min automated driving with either one or another version of the new advanced driver-assistance systems (ADAS). Relative to baseline, the results showed some trends to significant improvements in the take-over quality and eyes on road time, especially for young or inexperienced drivers. The reported systems are currently in the process of being patented.

Introduction

Highly automated driving is currently one of the most discussed innovative topics and likely to become a series product within the next few decades (Gold, 2016). The development of driver assistance systems was based on the premise that the driver is continuously in the control loop supported by technical systems to conduct the driving task, which corresponds to level 1 and level 2. From level 3 automation (SAE) on, the driver does not have to monitor the vehicle while driving constantly (SAE J3016, 2016), which means the driver can conduct non-driving related tasks and be out of the control loop. Non-driving related tasks (NDRT) are for example eating, texting, talking, relaxing and so on (Pfleging, Rang, & Broy, 2016), which may lead the driver to divert attention from the driving scenery. This out-of-loop scenario may cause loss of awareness of the state and processes of the system (Endsley, 1995).

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

However, level 3 automation systems require the driver to react appropriately if the systems request this when reaching their system limit (SAE J3016, 2016), so called "take-over request" (TOR). Since situation awareness (SA) is critical to effective decision making and human performance in dynamic systems (Endsley, 1993), it is reasonable to help the driver's mental model of the current system states and traffic situation to be updated, in other words, gain a higher SA, which supports an appropriate reaction in this time restricted situation.

Situation Awareness

Situation Awareness is a critical research theme in many domains, which involves human performance in dynamic or complex systems. It is widespread and exists in the military, air traffic, automobile driving and many more. There is no absolute definition and model of SA yet. Three different definitions and their associated theoretical perspectives dominate (Stanton & Young, 2000):

1) Three-level model (Endsley, 1995)
2) Perceptual cycle model (Smith & Hancock, 1994)
3) Activity theory model (Bedny & Meister, 1999)

The main difference lies in whether the SA refers to the process employed or to the product derived as a result of this process. The three-level model from Endsley comprised of three hierarchical levels describes SA as a product (Endsley, 1995). On the other hand, Smith and Hancock (1994) suggest the perceptual cycle model and define SA as adaptive, externally directed consciousness, which defines SA as a generative process of knowledge creation and informed action taking, not a snapshot of the agent's current mental model. Bedny and Meister proposed that SA is part of cognitive activity that is intensely dynamic (Bedny & Meister, 1999).

Measurement of SA

Salmon, Stanton, Walker, and Jenkins (2009) listed several SA measurement methods: SA requirements analysis, freeze probe technique, real-time probe technique, self-rating technique, observer-rating techniques, performance measures (direct / indirect), process indices (eye tracking), team SA measurefas.

Performance measures allow an indirect assessment of SA, which may be hits, crash avoidance during a simulated driving task or detection of hazardous events (Gugerty, 1997). Those measures are simple to obtain and are non-intrusive as they are generated through the natural flow of the task. It may be that efficient performance is achieved despite an inadequate level of SA, or that deficient performance is achieved regardless of a high level of SA. This has to be taken into account (P. Salmon, Stanton, Walker, & Green, 2006). Process indices involve recording the process in order to develop SA during the task under analysis, e.g. eye movement during task performance (Smolensky, 1993). The data can be used to assess which situational elements the participant fixated upon during task performance, and has been extensively used in SA assessment exercises (P. Salmon et al., 2006). The use of an eye-tracking device in the field is difficult but recommended for simulator studies. However, the disadvantage of "look but do not

see" phenomenon should be considered (Brown & Great Britain. Department for Transport., 2005)

Goal

In the experimental study, different countermeasures to the loss of SA were developed, implemented and evaluated. The stimulation used should raise the SA of the driver to a certain level, which ensures a better take-over performance (red arrows in Figure 1). The stimulation should carry certain information to the drive, but it should not be a warning and not be annoying. The scenario is a level 3 automation (SAE J3016, 2016), which is defined as conditional automation.

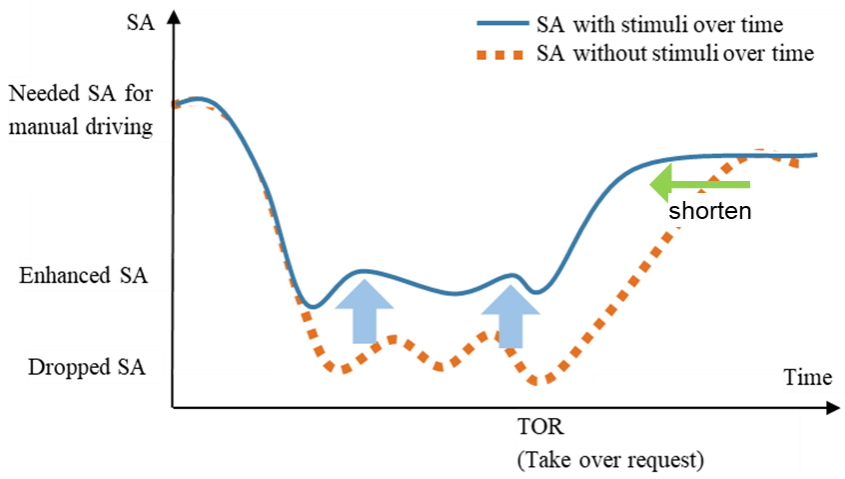


Figure 1. SA drops during automated driving and the stimuli should help the driver have higher SA (modified from Toyota Motor Europe NV/SA)

Method

Stimuli Design

Visual, auditory, tactile, and haptic stimuli are applied to interactions between human and machine (Schenk & Rigoll, 2010).

To rate the suitability of a specific modality of a stimulus in the vehicle, Hoffmann and Gayko (2012) used the following categories: "content of information", "coverage rate", and "forgiveness rate". In this work, a heuristic approach with various additional categories was conducted with two ergonomic experts. In addition to those mentioned above, the following categories are introduced concerning the usage and design purpose of the stimuli. These factors are the "perceptibility", "interpretability", "limitability", "interference potential", and "localisability". Table 1 shows the evaluation result.

Table 1. Heuristic evaluation of modalities of a stimulus

	Modality				
Category	**Visual**	**Auditory**	**Haptic**	**Thermal**	**Olfactory**
Content of Information	++	++	**o**	-	**o**
Coverage Rate	+	+	-	--	**o**
Forgiveness Rate	**o**	--	+	+	+
Perceptibility	+	++	**o**	**o**	+
Interpretability	++	+	**o**	--	-
Limitability	++	++	+	--	--
Interference Capability	+	-	**o**	+	**o**
Localisability	+	+	**o**	-	-

[++] very good [+] good [o] neutral [-] bad [--] very bad

As evaluated in the Table 1, the visual channel can display very detailed and various information at once. It can be modified in many ways like varied colours, sizes or brightness, therefore "content of information" is [++]. The coverage is good overall but the visual capacity or visual attention might be limited by one or another scenario, therefore "coverage rate" is [+]. False alarms are quite forgivable because they are not as intrusive as other modalities. On the other hand, most visual stimuli can even be seen on the periphery and will be perceived, therefore "forgiveness rate" is [o]. The relevance or significance of information displayed can be perceived in most ways; in some use cases, the periphery would be ignored though, therefore "perceptibility" [+]. Since visual stimuli are modifiable in a lot of ways (format, brightness, colour, animation, etc), it can be designed with a very high interpretability, therefore "interpretability" is [++]. The period (time) as well as the area (space) of the stimuli can be designed very precisely with clear boundary, therefore "limitability" is [++]. The driver can decide to look or not or even ignore the given stimuli. Still, he/she will be peripherally stimulated in most cases by one or more visual stimuli, therefore "interference potential" is [+]. The feedback pointing at a specific scenario can be directly linked to events outside well in most cases. Still, the position of stimuli influences and limits its localisability, therefore "localisability" is [+]. As a result, visual stimulus is the suitable balance between information carrier and subliminal stimuli.

Furthermore in the literature, visual stimuli have some advantages over other modalities: the foveal perception (driving scene for the driver) will not be restricted

by the stimuli in the periphery (Posner, 1980; Wickens, 2008). Visual perception in the periphery does not need direct transition of attention (Maier, Kathrin ; Sacher, Heike ; Hellbrück, Jürgen ; Meurle, Jürgen ; Widmann, 2011), which is on the primary task. Information can be transmitted without an explicit concentration on the stimuli (Utesch, 2015). Ambient light can catch the user's attention and raise awareness for an upcoming event unobtrusively (Müller, Kazakova, Pielot, Heuten, & Boll, 2013). Additionally, visual stimuli can be ignored on request so that the stimuli could not be annoying. As a result, visual stimuli have been chosen considering the requirements of raising the SA and not being annoying.

In this work, as a visual stimulus, an LED bar at the bottom of the windscreen is chosen and implemented from the bottom of the left A-pillar to the bottom of the right A-pillar of the static driving simulator (Figure 2). The 0.15 Hz pulse is defined as the basic pulse, which should generate a calm and natual felling and corresponds to the frequency of the calm, natual human respiratory (9 times/minute) (Lehrer, Vaschillo, & Vaschillo, 2000).

There are three different configurations of the stimuli:

1) Pulse Only (PO): the frequency of the pulse is 0.15 Hz, the colour is white. (Figure 2)

2) Pulse Event (PE): the frequency and colour of the pulse depends on the confidence level of the automation system. When the system is at its:

   a. …high confidence level: the pulse is 0.15 Hz in white. (Figure 2)
   b. …medium confidence level: the pulse is 0.50 Hz in white. (Figure 2)
   c. …low confidence level: the pulse 0.50 Hz in blue. (Figure 3)

3) Take-over Request (TOR): the frequency of the pulse is 1 Hz, the colour is red. (Figure 4)

Figure 2. White pulse, 0.15 Hz or 0.5 Hz

Figure 3. Blue pulse, 0.5 Hz

Figure 4. Red pulse, 1 Hz

Hypothesis

The visual stimuli

1) …will be accepted by the participants in terms of modality, position, colour and frequency.
2) …will improve the reaction time (RT) after a take-over request.
3) …will increase the minimal time to collision (TTC_{min}) to a dangerous obstacle.
4) …will improve the manual driving directly after the take-over.
5) …will enhance the SA by increasing the eyes on road time/frequency.

Experimental Set-ups

To examine the hypothesis, a mixture within-between experiment was conducted. Two configurations of the stimuli are the between factors and the within factors are with stimuli or the baseline without stimuli. The sequences of all variance were all counterbalanced (Table 2).

Table 2. Experimental Design and counterbalancing

		Within ↻	
Between ↻	(n=16)	Baseline	Pulse Only
	(n=16)	Baseline	Pulse Event

The experiment was conducted in the static driving simulator consisting of a complete vehicle mock-up. Seven projectors provided a front view of about 180 degrees as well as the views of all mirrors. The simulation software SILAB from WIVW (Würzburger Institut für Verkehrswissenschaften GmbH) was used to create the driving environment. The SILAB logs all relevant driving parameters and allows the LED-strip as well as the Dikablis 2 eye-tracking system to be controlled.

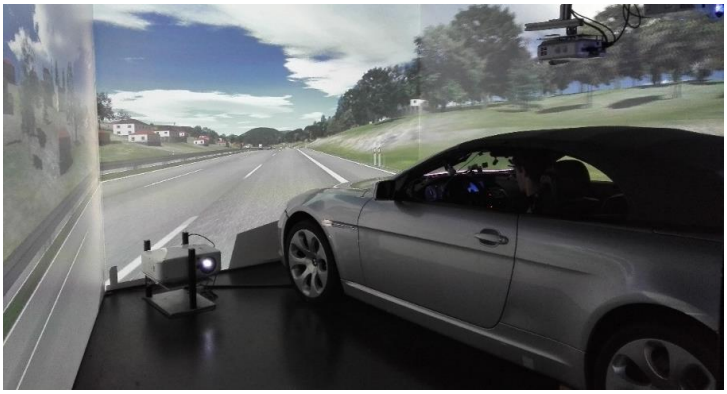


Figure 5. Static driving simulator

Tracks

To minimalise the learning effect, two different tracks and TOR scenarios were built. On the other hand, parameters like traffic density, time budget of the TOR and possible take-over manoeuvre are kept identical to ensure the comparability. Both tracks are 16 minutes long, the automation takes around 15 minutes. The route consisted of three parts (Figure 6). "Boring scenarios" simulates a monotony drive with occasional overtaking traffic as well as one overtaking scenario of the ego-vehicle. The second part has higher traffic density including manoeuvres around mobile construction vehicles. The third part contained a take-over scenario caused by system boundaries, in which the first hint of the danger appears when Time To Collision (TTC) is 7s while the TOR occurs with two-beep tone when TTC = 6s and the entire danger shows up.

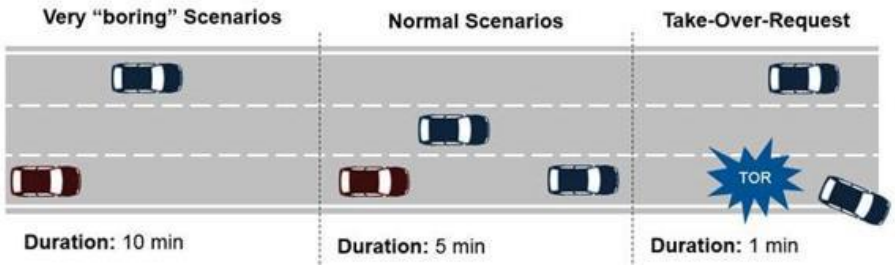


Figure 6. Test tracks with 3 parts and a TOR scenario

Non-Driving Related Tasks (NDRTs)

Pfleging et al. (2016) identified several NDRTs that people will conduct in transportation. Apart from those, standard NDRTs are considered in this work concerning their controllability, reproducibility and clear separation of necessary modal resource. Table 3 shows the summary of some standard NDRTs and their characteristics.

Table 3. Analysis of standardised NDRTs

	Modality						**Paced**	
	Visual	Motoric	Acoustic	Haptic	Verbal	Cognitive	User	System
SuRT	X	X					X	
CTT	X	X						X
n-Back-Task			X		X	X		X
20-Questions Task			X		X	X	X	
Shape-sorter ball		X				X	X	
DRT (Visual)	X	X				X		X
DRT (Haptic)		X		X		X		X
DRT (Acoustic)		X	X			X		X
Pointing Task	X	X	X					X
Counting/Calculating	X		X		X	X	X	X
Cognitive Task	X		X		X	X		X

The Surrogate Reference Task (SURT) (ISO/TS 14198, 2012) is a visual-motoric, user-paced task with various levels of difficulty, which was chosen in this experiment to simulate the daily smartphone usage. The participants should report an unusual item (target) in an array of similar items (distractor), usually an array of symbols, forms, colours or words. The similarity, which can be manipulated, influences the time for the participants to react. The more similar the distractors are to the target, the longer the reaction time is mostly. For the participant to be able to select the target, the display is divided into evenly distributed vertically arranged rectangular areas. The target is placed in one of those areas (ISO/TS 14198, 2012). This simulates a common use case of using a cell phone. To encourage participants to engage in the NDRT instead of monitoring automation, a real-time scoring bar was implemented on the screen, which shows the current performance of the user.

Measurements

Regarding the analysis of the SA measurement earlier, in this driving simulator study, performance and eye-tracking are evaluated as well as the acceptance of the subjects through questionnaires. The evaluation metrics include eye-tracking data before take-over scenarios, take-over time (Reaction Time (RT)), take-over quality, which consists of minimal Time To Collision (TTC_{min}) and Standard Deviation of Lateral Position (SDLP).

Results

The total number of participants in the study is 35. Because of technical problems during the experiment, there are 32 data sets available. As for eye tracking, due to technical failures on marker recognition, camera focusing and pupil detection, only 22 out of the 32 data sets could be analysed. There is no statistical significance ($\alpha = 0.05$) found in terms of TOR performance and eyes on road time (EoRT), since most participants had already performed very well. Nevertheless, in case of $0.5 < p < 0.1$, tendency to significance is reported.

Participant statistics

There were 7 female and 25 male participants. Average age was 25.63 years (SD = 4.43). All participants had a valid driving licence, mean = 8 years (SD = 4.10). 56% of them had already taken part in an experiment with a driving simulator, 22% even multiple times. 44% of the participants drove maximum 5,000 kilometres per year. 25% had between 5,000 and 10,000 kilometres; 22% between 10,000 and 20,000 kilometres. 60% had advanced or expert level knowledge of HAD/ADAS.

Driving Performance

Reaction Time (RT)
The reaction time is defined as the period, which starts from the TOR and ends with the first conscious engagement of the driver. Conscious engagement is present when the steering is turned (left or right) more than 2 degrees or the brake pedal is pushed over 10 % of its maximum.

Comparing the mean RT of the baseline groups with both stimuli groups combined. No significant difference was found ($p = 0.2$). However, there is a descriptive smaller mean and smaller SD, which suggests a lower variance (Figure 7 left). Generally, most participants performed already very well in terms of reaction time (around 2 seconds) in the baseline.

The mean RT of the PO group showed a tendency to be significantly faster than its baseline ($p=0.095$). Additionally, the standard deviation is smaller. For the PE group, no significant difference was found ($p=0.98$) (Figure 7 right). Furthermore, PO helped 3 out of 4 "worst performers" (25^{th} percentile) to get better, furthermore PE helped all the "worst performers" to improve their RT.

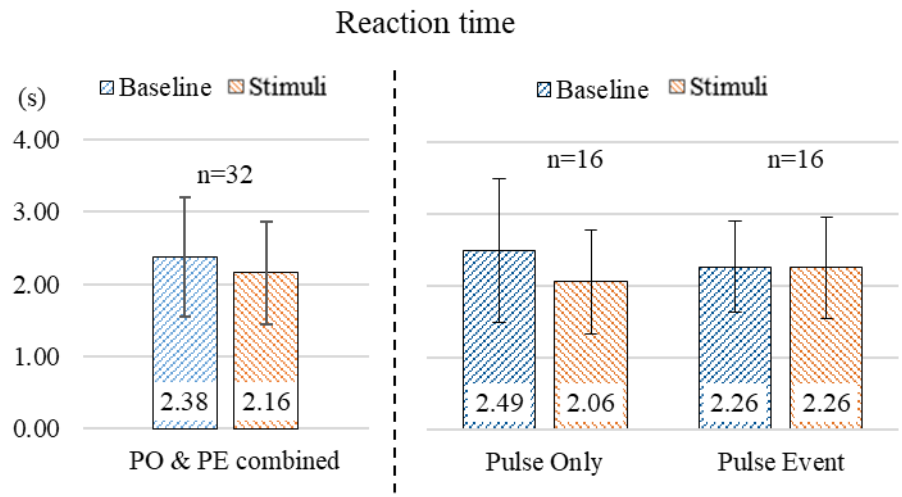


Figure 7. Comparison of RT. Left: PO and PE combined as LED; right: PO and PE groups [1]

In addition, six sub-groups are built according to the participants' self-reported characteristics of themselves to compare the RT:

1) LED noticed or not: whether they have noticed the changing pattern of the Stimuli;
2) Driving simulator experience;
3) Drive experience;
4) Age;
5) Practical experience with Active Cruise Control (ACC)/ Lane Change Assistance (LCA);
6) Knowledge about ADAS/ Highly Automated Driving (HAD).

Because of the small number of subgroups, only the combined stimuli (PO+PE) are compared with the combined baseline. The stimuli showed a clear positive effect (less RT) for those participants:

1) …who have not actively noticed the stimuli. They also had a lower eyes on road time (EoRT), which means the stimuli had positively affected them in a subliminal way.
2) …who are younger (16-25 years old) and have less driving experiences (<5,000km/year).

Additionally, some small positive effects were found for participants with less simulator experience and no practical experience or knowledge about ADAS and HAD.

Minimal Time To Collision

The minimal time to collision is defined as the minimum value of all the TTC values within the measured time interval, for each time frame of measurement:

[1] All error bars in the diagram in this work are standard deviations.

$$TTC = \frac{v_{ego} - v_{obstacle}}{distance\ to\ obstacle} \quad (when\ v_{ego} > v_{obstacle})$$

In cases of $v_{ego} \leq V_{obstacle}$, TTC = ∞. Having a lower minimal TTC complies with a bad take-over performance (higher danger).

In this work, the measurement interval starts from the TOR until the last moment when the centre of the car crosses the lane mark, if there is a lane-change manoeuvre, which is demonstrated as the red arch in Figure 8.

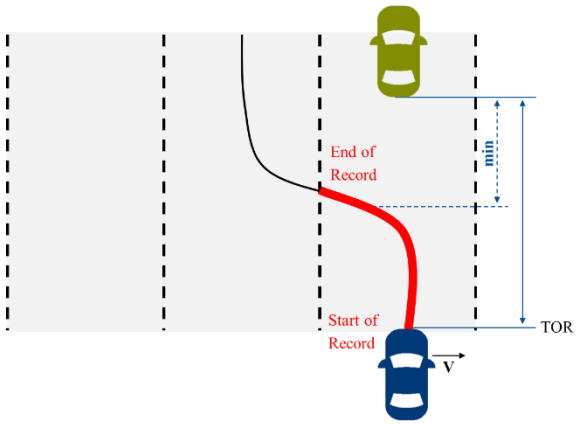


Figure 8. Demonstration of the calculation of TTC_{min}

Comparing the TTC_{min} performance of the combined baselines with both stimuli combined, no significant difference was found (p = 0.87) (Figure 9 left). For the same reason as the RT, participants could not improve much since with a TTC_{min} > 2 s the performance is already very good and they are far from danger.

For the specific analysis, TTC_{min} of the PO group showed no significant difference (p = 0.37) to the baseline, but a higher mean and a much smaller SD (Figure 9 right). Additionally, no significant difference for the PE group was found (p = 0.31), but a slightly lower mean with similar SD, which indicates even a negative effect of the PE stimuli. The explanation could be that participants reported that they misunderstood the stimuli as a warning system, which will warn them in any dangerous case, which may lead to over trust, and a delayed reaction to the danger, therefore smaller TTC_{min}.

In the "best/worst performers" analysis, PO increased all TTCs to at least 2 s: "worst performer" (TTC < 2 s) all got better and passed the 2 s TTC mark. Some good performers got worse but only because they were already at a very good level (still > 2 s TTC). PE helped the worst performing participants (TTC < 2 s) to get better or not worse. However, most good performing participants (TTC > 2 s in the baseline) got worse, one participant had below 1 s TTC and one crashed in the PE condition.

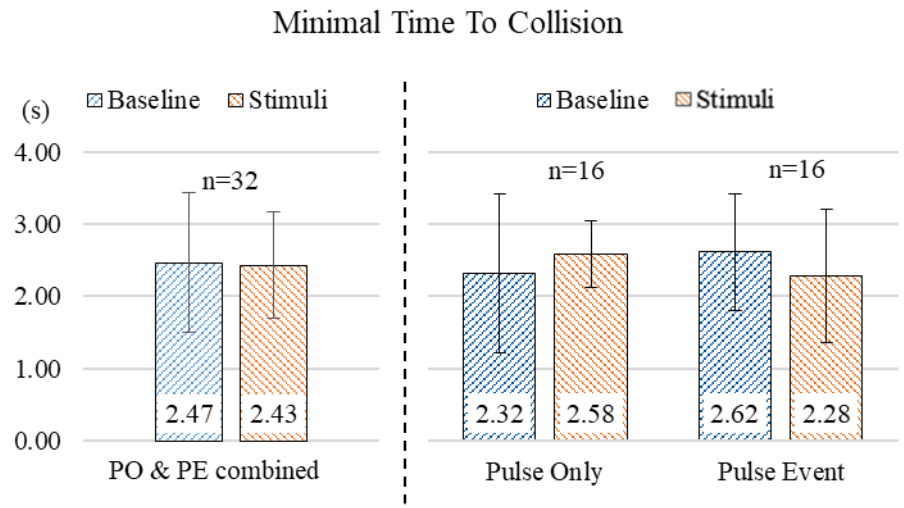


Figure 9. Comparison of TTC_{min}. Left: PO and PE combined as LED; right: PO and PE groups

Like the subgroup analysis of the RT, the stimuli showed again a positive effect to TTC_{min} for:

1) …those who have not actively noticed the stimuli (≈subliminal).
2) …those who are younger (16-25 years old) and have less driving experiences (<5,000km/year).

Additionally, some smaller positive effects were found for participants with less simulator experience and no practical experience or knowledge about ADAS and HAD. Slightly negative effects were found for participants with much driving experience and theoretical and practical knowledge of ADAS and HAD. This could be due to participants considering the PE (and PO) as a warning system, since they were familiar with many ADAS systems as they claimed.

Standard Deviation of Lateral Position (SDLP)

SDLP, an index of 'weaving', is a stable measure of manual driving performance with high test–retest reliability (Verster & Roth, 2012). The lateral position in this work is a value of the distance d (Figure 10) between the centre of the ego vehicle and the middle of the driven lane. Since it's not meaningful to calculate the SDLP when lane changing, the measurement intervals therefore start after the lane change process has finished, which is defined when the centre of the ego vehicle is at first closer to the middle of the 2nd lane than 0.1 m. The measurement will last for 5 seconds due to the length of an overtaking process.

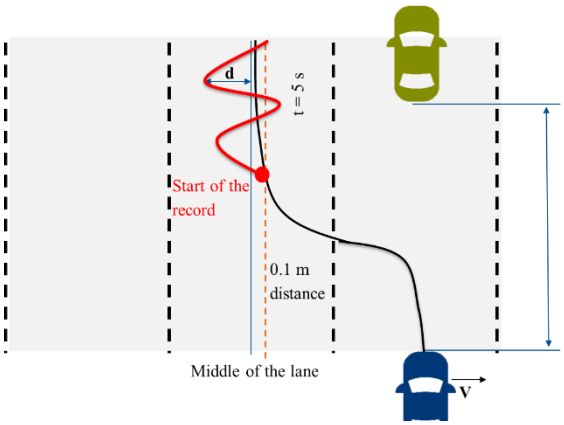


Figure 10. Demonstration of the calculation of SDLP

In the SDLP analysis, n=17 because it includes only those who changed lane to the middle lane in both trials (for baseline and PO, n=7; for baseline and PE, n=10).

Between the baseline and stimuli (PO/PE combined), there was no significant difference ($p = 0.27$) (Figure 11 left). Still, the descriptive mean of the SDLP with the stimuli is smaller and the SD is slightly smaller too (Figure 11 left).

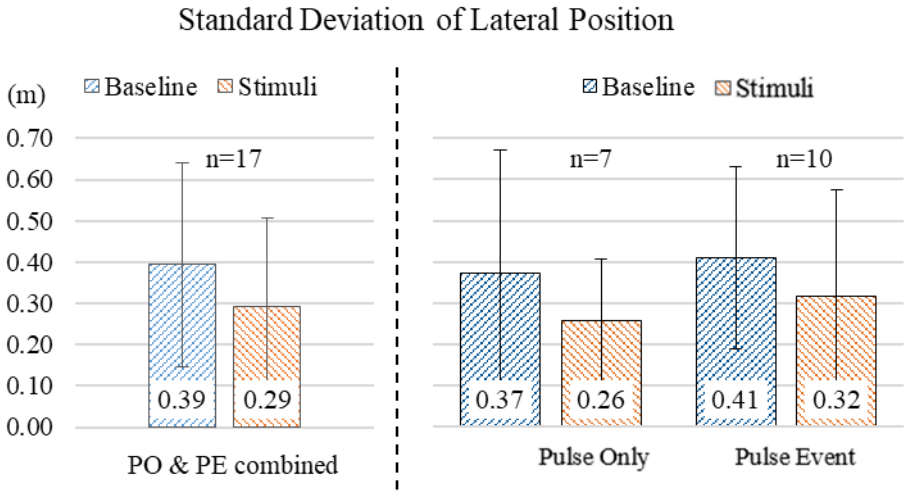


Figure 11 Comparison of SDLP. Left: PO and PE combined as LED; right: PO and PE groups.

The PO condition showed a descriptive lower mean of SDLP with lower variance but no significance, which is similar to PE (Figure 11 right).

Both stimuli conditions showed a tendency towards better manual driving directly after TOR but there is no significance due to the number of participants.

Eye-tracking

As showed in Figure 12, three different Areas Of Interest (AOIs) were investigated:

Figure 12. Demonstration of AOIs

1) Blue area: road, driving scene. This is to check the eyes on road time (EoRT).
2) Purple area: Tablet screen of NDRT: Sony Xperia Z Ultra, 6.4 inches tablet. This is to check how well they engaged in the NDRTs.
3) Yellow area: left side mirror. This is to check the quality of the take-over manoeuvre, how well they check the left lane before changing lane.

The first measurement interval starts from activation of the automation until the TOR (about 14 min). The second measurement interval starts from the TOR and lasts 10 seconds under the consideration of period including take-over manoeuvre and over-taking manoeuvre. The third measurement is the time until the first glance on the road after TOR. The recorded data are the following three:

1) Total glance time in [%] or [sec] is defined as the sum of all time on a specific AOI.
2) Mean glance duration in [sec] is the mean time of each glance on a specific AOI.
3) Number of glances is the total number of glances on a specific AOI.

Generally, 13 out of 22 participants' percentage of EoRT increased to a meaningful level with LED, 5 of them did not change much. There are four participants that looked much less at the road with the LED, which had all the same LED conditions as the first trial, which may be probably due to a strong sequence effect.

It is found that the participants looked more often and longer at the road in the second trial regardless of being with or without stimuli. Because of this "sequence effect", only the eye tracking data from first trials of each participant will be investigated.

In Figure 13 (left), there is a tendency to a significant longer "mean glance duration on the road" ($p = 0.07$). The total glance time on the road shows a higher mean but

no significant difference (Figure 13 (middle)), 3 participants as outlier (16.3%; 22.3%; 31.4%) are out of consideration. Finally, the mean of the number of glances on the road increased as well but without significance (Figure 13 (right)), 1 participant as an outlier (340) is out of consideration. These facts indicate with stimuli, participants tend to watch longer for each glance and longer in total and more often at the driving scene, which is a way to gain higher SA.

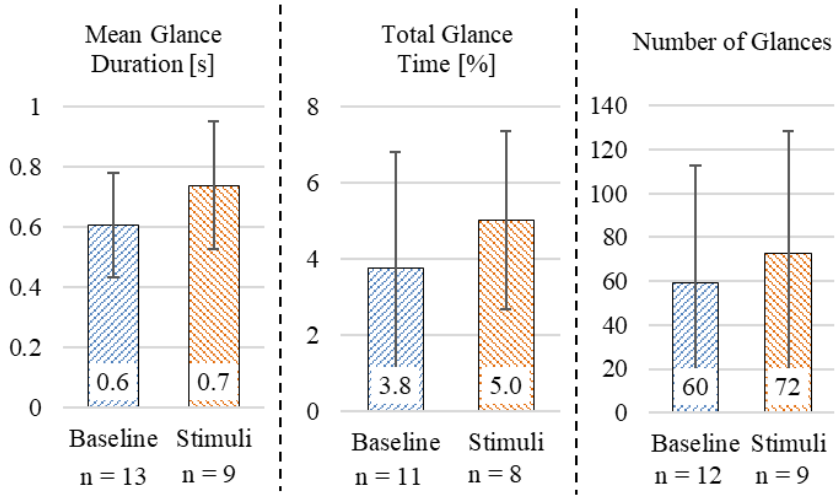


Figure 13. Comparison eyes on road data in three aspects, PO and PE combined

This is also supported in Figure 14, both stimuli (PO and PE) increased mean glance duration, as well as the total glance time and number of glance.

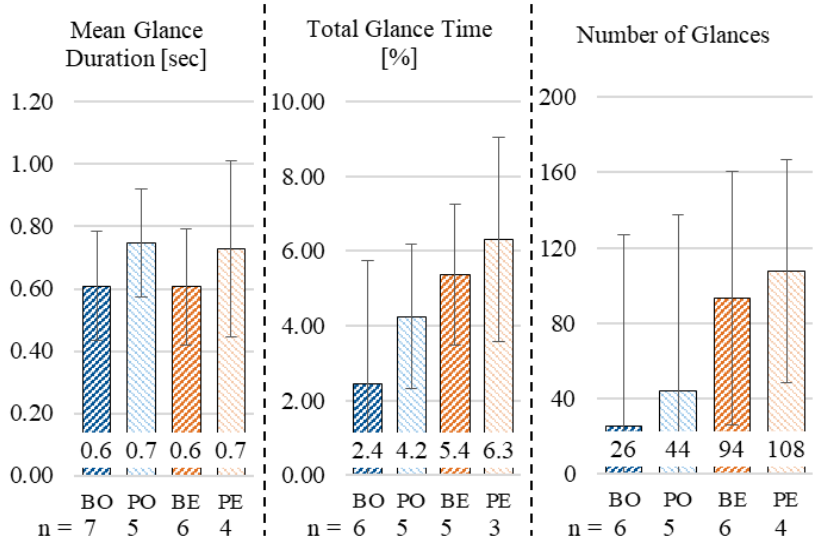


Figure 14. Comparison eyes on road data in three aspects, PO and PE separated

In addition, novice (< 5,000 km / year) and young drivers (16-25) showed no change in glance patterns, seem not to be affected by the stimuli while experienced (both 5,000-10,000 km and > 10,000 km groups) and older drivers (26-37) increased their mean glance duration on the road by using the stimuli and therefore they had a better chance to gain higher SA. That might because of the over-trust (in automation) of young and less experienced driver due to less expertise of dangerous situations.

Furthermore, as expected, eye tracking data does not correlate well with driving performance. Participants with bad RT did not have lower EoRT than those who performed well. The "Look but not see" problem occurred.

Subjective Evaluation

A questionnaire after each experiment about the colour, position, frequency, and visual modality of the stimuli indicates that:

1) Colour: 93.75% of participants rated white for the constant stimuli in the "automation mode" as (very) proper, 62.5% of participants are for the blue stimuli when events occur, finally 100% rated red very proper for the TOR. When alternatives were asked, it's reported green could replace white, and yellow could be used instead of blue, because they understood the stimuli as a warning system, and therefore they were strongly influenced by the traffic light colour concept.
2) Frequency: In the PO condition, 87.5% of participants rated the frequency of 0.15 Hz for the white pulse at least as properly designed; 93.75% rated the 1 Hz red pulse at least as proper. In the PE condition, 75% of participants rated the frequency of 0.15 Hz for the white pulse (high confidence) at least as properly designed; 68.75% are for the 0.5 Hz white pulse (medium confidence); 68.75% are for the 0.5 Hz blue pulse (low confidence); 100% rated 1 Hz of the red pulse (TOR) at least as proper.
3) Modality: 94% prefer visual modality in such application.
4) Position: 88% think that the applied position of stimuli (as Figure 3-5) is (very) proper.
5) Acceptance: 75% of participants would wish to have such a system (PO as well as PE) in an automated vehicle. Reasons such as the system supports them to understand the situation; helps to build trust; allows passengers also to be aware of the system states regardless of the weak and ignorable intensity of the stimuli were mentioned for PO. Reasons such as the feedback of events helps build the trust, which would relieve the driver from the monitoring task, was mentioned.

Limitations

Due to the small number of participants (n=32) and many test variances of concepts (baseline, PO and PE) on different conditions (track 1 and track 2), the results are not statistically significant. The stimuli itself could be regarded as warning system, but then too weak, too regular and too general as warning. Due to its position, the stimuli may not be perceived when the driver conducts NDRTs. The PO version might be perceived as monotonic and therefore be superfluous and useless.

Conclusion

There is no statistical significance in terms of TOR performance. However, the stimuli did help the worse drivers to shorten the gap. Specifically, the stimuli (PO+PE):

1) ...had a high acceptance in terms of the modality, position, colour and frequency.
2) ...helped bad performers to improve with the RT and increase the TTC.
3) ...showed a positive effect (better RT and TTC) in this sample for participants who reported having not noticed the LED (≈ subliminal), having less simulator experience and no practical knowledge about ADAS and HAD compared with those who had.
4) ...improved the manual driving directly after the take-over (SDLP).
5) ...increased the mean glance duration, eyes on road time and number of glances.

It seems that in this study the PO condition was slightly better than the PE one. This might be due to participants misinterpreting the stimuli as a warning system (and not a likelihood information), which might lead to over-trust.

Overall, the stimuli showed a high potential to raise driver's SA and to improve possible take-over performance without annoying the driver. Adding an additional modality such as auditory or haptic for a multi-modal approach might improve the performance even more.

Acknowledgement

This work was a collaboration between the Chair of Ergonomics at Technical University of Munich and Vehicle Performance Engineering Division, HMI Department, Toyota Motor Europe NV/SA. The experiment was conducted at the Chair of Ergonomics as part of a semester thesis. The concept of the stimuli has been protected as an international patent since May 5th, 2017.

References

Bedny, G., & Meister, D. (1999). Theory of Activity and Situation Awareness. *International Journal of Cognitive Ergonomics*, *3*, 63–72. https://doi.org/10.1207/s15327566ijce0301_5

Brown, I.D. / Department for Transport. (2005). *Review of the "Looked but failed to see" accident causation factor*. Dept. for Transport. Retrieved from https://trid.trb.org/view.aspx?id=1156399

Endsley, M.R. (1993). Situation awareness in dynamic human decision making : theory and measurement. In *the First International Conference on Situational Awareness in Complex Systems*. Orlando, FL.

Gold, C. (2016). Modeling of Take-Over Situations in Highly Automated Vehicle Guidance.

Gugerty, L.J. (1997). Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied*, *3*, 42–66. https://doi.org/10.1037/1076-898X.3.1.42

Hoffmann, J., & Gayko, J. E. (2012). Fahrerwarnelemente. In *Handbuch Fahrerassistenzsysteme* (pp. 343–354). Wiesbaden: Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-8619-4_25

ISO/TS 14198. (2012). Road Vehicles -- Ergonomic aspects of transport information and control systems -- Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems. International Organization for Standardization. Retrieved from httpswww.iso.orgstandard54496.html

Lehrer, P.M., Vaschillo, E., & Vaschillo, B. (2000). Resonant frequency biofeedback training to increase cardiac variability: Rationale and manual for training. *Applied Psychophysiology Biofeedback*, *25*, 177–191. https://doi.org/10.1023/A:1009554825745

Maier, K., Sacher, H., Hellbrück, J., Meurle, J., & Widmann, U. (2011). Multimodaler Warnbaukasten – eine neue Warnphilosophie für Fahrerassistenzsysteme. *Der Fahrer Im 21. Jahrhundert Fahrer, Fahrerunterstützung Und Bedienbarkeit.* Retrieved from http://edoc.ku-eichstaett.de/15119/

Müller, H., Kazakova, A., Pielot, M., Heuten, W., & Boll, S. (2013). Ambient timer - Unobtrusively reminding users of upcoming tasks with ambient light. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8117 LNCS*(PART 1), 211–228. https://doi.org/10.1007/978-3-642-40483-2_15

Pfleging, B., Rang, M., & Broy, N. (2016). Investigating User Needs for Non-Driving-Related Activities During Automated Driving. *15th International Conference on Mobile and Ubiquitous Multimedia (MUM '16)*, (1), (pp. 91–99). https://doi.org/10.1145/3012709.3012735

Posner, M.I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*, 3–25.

SAE J3016. (2016). SAE J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles RATIONALE. Retrieved from http://standards.sae.org/j3016_201609/

Salmon, P.M., Stanton, N.A., Walker, G.H., & Jenkins, D P. (2009). *Distributed situation awareness : theory, measurement and application to teamwork*. Ashgate.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, *37*, 225–238. https://doi.org/10.1016/j.apergo.2005.02.001

Schenk, J., & Rigoll, G. (2010). *Mensch-Maschine-Kommunikation*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-05457-0

Smith, K., & Hancock, P. A. (1994). Situation awareness is adaptive, externally-directed consciousness. *Human Factors*, *37*, 137–148. https://doi.org/10.1017/CBO9781107415324.004

Smolensky, M. W. (1993). Toward the physiological measurement of situation awareness: The Case For eye movement measurements. In *Human Factors and Ergonomics Society 37th Annual Meeting*. Seattle.

Stanton, N.A., & Young, M.S. (2000). A proposed psychological model of driving automation. *Theoretical Issues in Ergonomics Science*, *1*, 315–331. https://doi.org/10.1080/14639220052399131

Utesch, F. (2015). Unscharfe Warnungen im Kraftfahrzeug Eignet sich eine LED-Leiste als Anzeige für Fahrerassistenzsysteme. *Statewide Agricultural Land Use Baseline 2015*, *1*. https://doi.org/10.1017/CBO9781107415324.004

Verster, J.C., & Roth, T. (2012). Predicting psychopharmacological drug effects on actual driving performance (SDLP) from psychometric tests measuring driving-related skills. *Psychopharmacology*, *220*), 293–301. https://doi.org/10.1007/s00213-011-2484-0

Wickens, C.D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*, 449–455. https://doi.org/10.1518/001872008X288394

Persuasive assistance for safe behaviour in human-robot collaboration

Matthias Hartwig, Vanessa Budde, Alissa Platte, & Sascha Wischniewski
Federal Institute for Occupational Safety and Health
Germany

Abstract

In a working context, conflicts between working safe and working fast can lead to deliberate violations of safety rules. Modern computer-human interfaces can create new opportunities to reduce these violations by influencing the user. Technologies deliberately used to influence attitudes and/or behaviour of users are called persuasive technologies and often make use of nudging strategies. In a randomized experiment, 90 participants had the task to collaborate with an industrial robot in a conflict between meeting the safety instructions and monetary incentives for working fast. An intervention group received emotional computer generated feedback on their safety behaviour, while a control group did not. Violations committed by the participant during and after the intervention were measured as well as intention towards the safety behaviour. Results show that participants receiving feedback on their behaviour committed only half as many violations as participants in the control group, a tendency that was also visible after the intervention ceased. Interestingly, subjective behaviour intention was nearly identical between the groups, which hint to a less deliberate form of behaviour impact of the feedback. Results suggest considering nudges as complementary action to promote safe behaviour at work besides giving information and penalising.

Introduction

Concerning occupational safety and health, there is a gap between extensive knowledge about hazards, regulations to minimize these and their implementation in operational practice. One level, where safety regulations are to be applied, is the individual level, where the reasons for safety violations are divergent.

Reason (2008) classifies “unsafe acts” in unintentional errors, like slips and mistakes, and intentional violations, which are based on a conscious decision to act against the regulation. These conscious deviations often pose an especially high risk, because they commonly form a habit and will most likely be repeated in every similar situation.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Modern man-machine systems, especially in the area of man-robot interaction, create an increasing degree of direct collaboration with interlocking working steps between user and machine. Therefore, on the one hand these interfaces evoke new demands for safe individual behaviour, while on the other hand they may have the potential to provide assistance by facilitating safe behaviour.

The second aspect is especially true for those oriented to the technology vision of ambient intelligence, characterized by Aarts and De Buyter (2009) by the central features of context awareness, personalization, adaptive behaviour and anticipation. In a working environment, these systems are called adaptive work assisting systems (AWAS; Windel & Hartwig, 2012). These features possibly enable the system to reduce violations by (1) being aware of the behaviour of the user, (2) evaluating it autonomously regarding violations and (3) presenting evaluative feedback or reminder that change the user's behaviour. Such computer interfaces, purposely designed to change the behaviour of the user, can be subsumed under the term persuasive technology (Fogg, 2002).

Persuasive technology is a technology-based form of nudging. This concept by Thaler and Sunstein (2009) encompasses any form of choice architecture that changes the behaviour of a person in a predictable way without forcing choice or economic incentives. The concept relies on the assumption that human decision-making is influenced by cognitive biases based on cognitive boundaries, routines and habits. Nudges use these mechanisms to influence decisions in an intended direction (Hansen, Skov, & Skov, 2016).

In a predecessor study on work assistance systems (Hartwig & Windel, 2013), a manually triggered anthropomorphic agent was proved to be effective to influence user's behaviour, using different emotional facial expressions. In the present study, the same virtual agent is implemented in a work assistance system to autonomously improve individual safety behaviour in a man-robot-collaboration setting. To gain insights into the best form of assistance, different types of persuasive strategies were applied: persuasive feedback that reacts to the participants' behaviour and a persuasive reminder that occurs at the moment the target behaviour becomes relevant. Therefore, we first hypothisized these forms of persuasive interventions to reduce safety violations when working in a man-robot interaction simulation compared to a control group that receives no persuasive assistance. Furthermore, the study aims at identifying the psychological mechanisms of the intended behaviour changes by investigating attitude towards behaviour and subjective social norm concerning the safety behaviour as two key sources for behaviour decisions in the theory of planned behaviour (TPB; Ajzen, 1991).

Method

The study sample included 90 participants, 45 men and 45 women. The participants were on average 24.5 (SD = 3.33, range 20-34 years) years old at the time of the investigation. All participants were students, recruited at nearby universities. In a randomized experiment the work task to assemble circuits in collaboration with an industrial robot was given to all participants. This task required the positioning of empty plug boards and the corresponding components in specific holders (Figure 1).

Afterwards, the participants started the robot, which autonomously connected the individual parts and then tested the assembled electric circuit for operability. This procedure was repeated 14 times, assembling one operational plug board each.

The behaviour of the participants during the working phase of the robot was the primary dependent variable. All subjects were instructed to wait after starting the robot until it finished the assembling and the testing. Then the participants received a safety clearance message and are allowed to proceed with the next plug board.

Figure 1. Participant placing the components.

Working within the robots reach prior to the security clearance is recorded by a light barrier installed in the workplace and counted as a safety violation. The recording is done unnoticed by the participants. The instruction explains this working sequence as a necessary safety procedure to avoid collisions with the moving robot. In fact, the implemented work system is designed for direct collaboration with the user, so there is no actual threat, regardless of the participant's behaviour. However, constituting a credible threat in the experimental setup is crucial to simulate a realistic decision for or against safety behaviour. To simulate the surrounding that often leads to safety violations in operational practice, a financial bonus of €10 for fast task completion is promised the participants.

Since violating the waiting process and prematurely working on the subsequent board was speeding up the working task substantially, a conflict between profitable and safe behaviour was created, as it exists in operational practice as well. Participant's safety behaviour as the primary dependent variable was measured by recording the number of plug boards on which safety violations were committed. Safety violations were operationalised as reaching into the work area of the robot before the safety clearance. This action was recorded automatically by a laser barrier installed in the experimental set-up.

Participants were able to commit safety violations on all 14 plug boards at most. Furthermore, the variables "Attitude towards Behaviour" and "Perceived Social Norm" were measured as two of the deterministic antecedents for behaviour according to the TPB to gain first insights into the psychological mechanisms of

behaviour change. A questionnaire was created according to Ajzen's (2002) guidelines, the wording of the items matching the specific behaviour.

As independent variables, the participants received different interactive assistant systems that should animate them to wait for the safety clearance. The control group worked on the task without any assisting system. The group "Reminder" was reminded by an anthropomorphic virtual agent (see Figure 2) to wait for safety clearance each time they start the robot. In the group "Feedback" the same agent giving negative feedback was presented every time the participant worked prior to the security clearance of the active robot. In addition, positive feedback was presented to participants after the fourth and tenth plug board, if they had not committed any violation until that point. Positive Feedback was given by the same virtual agent, showing a friendly emotion, underlined by an affirmative text message "Very good! You complied with the safety clearance."

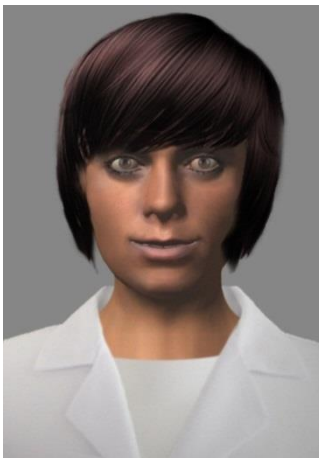


Figure 2. Anthropomorphic agent.

Results

The first hypothesis assumed a difference in the number of safety violations/premature intrusions between the different experimental conditions. Figure 3 shows the mean number of boards in which safety violations were committed. The mean value in the "Feedback" condition was M = 2.69 (SD = 3.17), in the "Reminder" condition M = 4.71 (SD = 5.65) and in the "Control Group" M = 4.91 (SD = 5.12) violations. The results of the one-factorial analysis of variance showed no significant differences between the number of boards with violations and the experimental groups $F(2, 65) = 1.59$ $p > .05$. Therefore, Hypothesis 1 was declined.

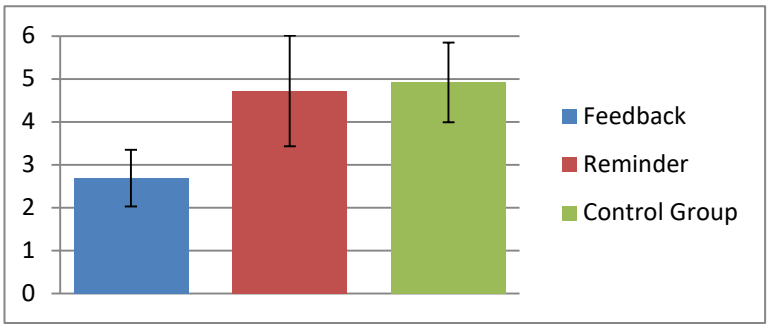


Figure 3. Number of violations in the different groups. Error bars reflect Standard Error of the mean.

Hypothesis 2 investigated differences between the experimental groups concerning the attitude and perceived social norm towards the safety behaviour. The average attitude towards safety behaviour of the feedback group was M = 17.13 (SD = 4.19), of the reminder group M = 19.86 (SD = 3.84) and of the control group M = 16.96 (SD = 4.42), see Figure 4.

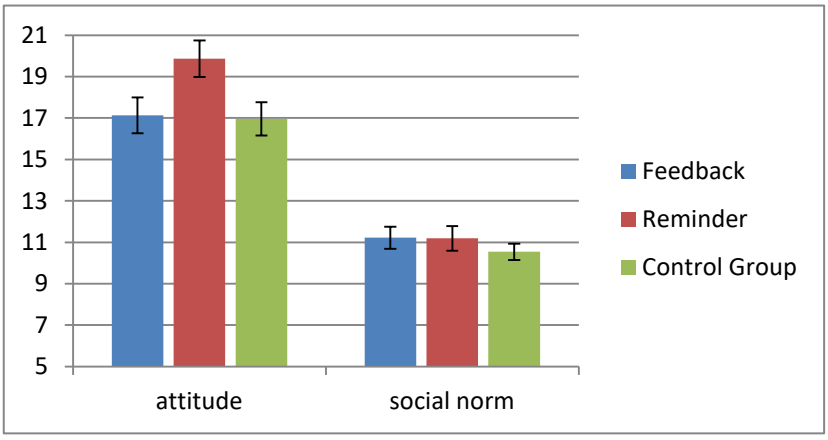


Figure 4. Average attitude and social norm towards safety behaviour. Error bars reflect Standard Error of the mean.

With regard to social norm towards safety behaviour, average score of the feedback group was M = 11.22 (SD = 2.52), of the reminder group M = 11.19 (SD = 2.62) and of the control group M = 10.54 (SD = 2.12).

To examine this hypothesis, two one-way analyses of variance were performed. The results show there was no significant difference between the three tested experimental groups with regard to the setting concerning attitude towards behaviour $F(2, 65) = .04$, $p > .05$ and concerning social norm $F(2, 65) = .58$, $p > .05$. Thus, hypothesis 2 showed no differences between the experimental groups regarding attitude towards safety behaviour or corresponding social norms.

Discussion

The presented experiment investigated the effects of different persuasive techniques on safety behaviour and subjective attitudes. The results show a substantially higher number of violations in the control group than in the two intervention groups where behaviour was assisted by different persuasive strategies. However, the statistical inference shows no significant result of the Anova concerning the safety violations, which indicates that the treatment had no effect on participant's behaviour. This result is in contrast to a predecessor study (Hartwig & Windel, 2013) and therefore hints more at a failure of the experimental setup rather than an overall ineffectiveness of the persuasive strategies which were used in both studies. Finally, there is also a probability of an existing systematic difference between the groups that the anova failed to detect (2nd type error), which is here neglected for the benefit of conservative hypothesis testing.

The different results of the two studies may partly be caused by the different test setting including a more realistic working task and the industrial robot. Looking at the absolute numbers, even the control group committed only a third of all possible safety violations, creating a ceiling effect. The low number of violations might be caused by the participants' unfamiliarity with industrial robots, resulting in a quite cautious behaviour. This could have been counteracted by a more intense training or recruiting participants experienced in working with robots. This, however, was not possible without exposing the cover story, as even moderate expertise in working with collaborative robots would reveal that there was no real danger because of the robot's integrated safety measures. Regarding future studies, the conflict between safe behaviour and the incentive for quick work should therefore be intensified by realistic conditions regarding the time constraints in everyday work, causing less cautious behaviour and more safety violations without the persuasive intervention.

A surprising finding is the discrepancy between the subjective personal perceptions towards safety behaviour and the actual behaviour. The numerical lowest violations occur in the feedback group, while the most positive attitude towards the behaviour is measured in the reminder group. Our initial assumption was that the persuasive techniques would change the subjective attitude and social norm, which in turn leads to less safety violations, but the data show no indication for this causal chain. The psychological mechanisms remain unknown for the time; subsequent studies should therefore put great attention to the psychological mechanisms that cause the intended behaviour change. Only by understanding why persuasive technology works, it will be possible to identify potential applications and limits of persuasive assistance systems that may contribute in safer behaviour at work.

References

Aarts, E., & De Buyter, B. (2009). New research perspectives on Ambient Intelligence. *Environments, 1*, 5-14.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

Ajzen, I. (2002). *Constructing a TPB questionnaire: Conceptual and methodological considerations.* (Working Paper). Amherst: University of Massachusetts. (available online at
http://www-unix.oit.umass.edu/~aizen/pdf/tpb.measurement.pdf).

Fogg, B. J. (1998). Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 225-232). New York, USA: ACM Press/Addison-Wesley Publishing Co.

Fogg, B.J. (2002). Persuasive technology: using computers to change what we think and do. *Ubiquity, 2002 (December)*, 5.

Hansen, P. G., Skov, L. R., & Skov, K. L. (2016). Making healthy choices easier: regulation versus nudging. *Annual review of public health*, *37*, 237-251.

Hartwig, M., & Windel, A. (2013). Safety and health at work through persuasive assistance systems. In V.G. Duffy (Ed.) *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Human Body Modeling and Ergonomics (*pp. 40-49). Heidelberg, Germany: Springer.

Reason, J.T. (2008). *The Human Contribution. Unsafe Acts, Accidents and Heroic Recoveries*. Farnham, UK: Ashgate Publishing.

Thaler, R. H., & C. R. Sunstein (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. London, UK: Penguin.

Windel, A., & Hartwig, M. (2012). New Forms of Work Assistance by Ambient Intelligence. In F. Paternò, B. de Ruyter, P. Markopoulos, C. Santoro, E. van Loenen, and K. Luyten (Eds.) *Ambient Intelligence* (pp. 348-355). Heidelberg Germany: Springer.

Comparing the effects of space flight and water immersion on sensorimotor performance

Bernhard Weber, Simon Schätzle, & Cornelia Riecke
German Aerospace Center, Institute of Robotics and Mechatronics, Oberpfaffenhofen-Weßling, Germany

Abstract

Several studies documented the detrimental effects of microgravity during spaceflight on human motor control (e.g., during aiming tasks). In addition to parabolic flight, water immersion has been used for simulating microgravity effects on earth. Until now, however, the validity of partial or full water immersion setups as test environments to explore effects on sensorimotor performance has not been tested. In the present paper, the results of three empirical studies were compared using the identical aiming task paradigm during forearm water immersion (N = 19), full body water immersion (N = 22), and during spaceflight (N = 3 astronauts). In line with prior research, slower aiming motion profiles were found during spaceflight (2 weeks in space) compared to the terrestrial experiments. Astronauts required substantially more time to approach target areas and for matching the targets precisely in space. Average motion speed and speed variance decreased significantly. Intriguingly, the same overall effect pattern was evident in both partial and full water immersion, although the effect sizes tended to be smaller. Altogether, results indicate that water immersion is a valid form of weightlessness simulation. However, effects solely present during spaceflight (such as vestibular dysfunction) additionally contribute to performance losses.

Introduction

Until today, the human capabilities and skills are crucial and indispensable for the success of many space missions. Onboard the ISS, astronauts perform challenging tasks such as manual docking of spacecraft, or control of complex robotic systems including the Canadarm 2. Candidates undergo an extremely strict selection process, and are intensively prepared for reliable performance, even under the adverse conditions of space flight. In Earth orbit, the effect of the gravitational force no longer acts on the human body, as a centrifugal force is generated by the orbiting spacecraft leading to a state of microgravity. As one of the most demanding aspects of space flight, human physiology (including the vestibular, cardiovascular, musculoskeletal, and sensorimotor systems) has to adapt to the novel condition of weightlessness which usually takes up to six weeks (Kanas & Manzey, 2008). Astronauts therefore receive extensive trainings during parabolic flights or underwater exercises to prepare for sensorimotor tasks during space flight. These environments, simulating the conditions of space flight, are not only important for

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

astronaut training but also for scientific research on sensorimotor performance in weightlessness. During parabolic flight, the aircraft is in free fall condition for 20-25s per parabola, causing short-term weightlessness. While under such conditions, space flight conditions can be achieved. Experiments are interrupted during each of the 30-60 parabolas by hypergravity (1.8g) and 1g episodes. Some subjects experience space motion sickness. In research on sensorimotor performance, water immersion studies have been conducted to simulate weightlessness by neutral buoyancy of the human body. Some key advantages include longer experimental periods, larger sample sizes, a higher control of experimental conditions and lower costs. However, the conditions substantially deviate from space flight conditions: 1) the gravitational force is unchanged (i.e., the vestibular system is not affected); 2) increased ambient pressure (e.g., 1.6 bar in 6m depth) leads to cognitive impairment in depths greater than 6m (e.g., Hancock & Milner, 1982); and 3) body motions are damped due to the dynamic viscosity of water.

In the present paper, the validity of water immersion studies for simulating the effects of microgravity on the human sensorimotor system is explored. There is only anecdotal evidence in prior research on the effects of water immersion vs. microgravity. Wang and colleagues (2015) compared general wrist and trunk activities in full water immersion and parabolic flight and reported divergent activity patterns. Whiteside (1960) investigated the sensorimotor performance during an arm pointing task during in water immersion (up to the neck) vs. parabolic flight and reported different results for the two setups - based on data from one subject. In a prior study of the authors (Weber et al., 2016), the effects of full water immersion and space flight were investigated with a zero-order manual pursuit tracking task and documented similar degradations of tracking accuracy. Yet, the experimental setups were not identical (simulation vs. real telerobotic task) and only one astronaut participated in the space experiment.

In the current series of studies, the impact of weightlessness during space flight, partial water immersion, and full body water immersion are compared using the same sensorimotor task. For the current experiments, an aiming task paradigm was chosen, as sensorimotor degradation for rapid, aimed motions has been reliably found in several empirical studies. Using a paper-and-pencil aiming task, Ross (1991) found at least a trend for longer movement times and significantly higher positional error during parabolic flight compared to 1g. Bock et al. (1992) also documented that subjects consistently overshot targets when performing aimed arm movements during parabolic flight. Crevecoeur et al. (2010) found that compared to normal gravity, motions slowed down during parabolic flight with lower peak velocities and higher movement durations during arm movements along the sagittal plane while holding a manipulandum. In the experiment of Newman and Lathan (1999), aiming performance was explored during an 8-day space mission with a joystick and a trackball as input devices. Compared to terrestrial performance, aiming times increased for both devices in microgravity. Results from three cosmonauts, performing pointing arm movements after 10, 140 and 172 days in space also revealed higher motion times, lower peak velocities and accelerations in all phases of the mission compared to the 1g baseline as reported by Berger et al. (1997). Similar kinematic changes during spaceflight (4-18 days in space) have been

documented by Sangals and his colleagues (1999) during a joystick controlled aiming task.

There have been several explanatory approaches for the general slowing of aimed movements in microgravity: 1) a distortion of human proprioception due to the reduced muscle resting tone (no anti-gravity stabilization is required) and hence changed muscle spindle activity (Lackner & DiZio, 2000), 2) an underestimation of limb mass due to the absence of weight (but not mass), see Bock et al. 1996, and 3) inadequate internal movement models (e.g., Crevecoeur et al., 2010).

Theoretically, all of the described mechanisms should be active during water immersion. The buoyant force counterbalances the gravitional force reducing the limb weight – in case of perfect balance – to zero. If the main mechanism behind the degradation of aiming performance is solely due to the weight change of the human limbs, it should also occur during partial immersion, with the respective limbs immersed in water. When controlling a hand-held joystick, for instance, multi-joint interactions involving the trunk, shoulder, upper and lower arm as well as the wrist occur. Berger et al. (1997) hypothesized that the slowing effect may be due to the attempt to reduce reaction forces on the trunk, which is difficult to stabilize in weightlessness. Provided that multi-joint and multi-limb destabilization additionally contributes to the reported overall effect, larger performance decreases should be observed during full body immersion compared to partial immersion. Researchers also suggested that the impairment of the vestibular system could play a crucial role for sensorimotor degradation in space (e.g., Mierau et al., 2008). Then, additional performance losses should be evident during space flight compared to water immersion. These assumptions are tested in a forearm vs. full body water immersion vs. space experiment with the same aiming task paradigm.

Methods

Study 1: Forearm water immersion

Sample. In the first study, N = 19 naïve subjects (4 females, 15 males; 1 left-, 18 right-handers) with an average age of M = 23.1 (*S.D.* = 1.16) years participated.

Apparatus. An underwater qualified joystick (2 axes, max. workspace of ± 20° each axis) with a padded armrest and an elbow strap was positioned in a 50x70 cm basin (see Figure 1). For all of the following studies, all GUI positions could be reached with max. deflections of ± 8 degrees on both axes (resulting in 2 cm motions at the upper joystick end). Thus, anterior/posterior motions could be performed without any (bio-)mechanical restrictions (wrist deflection, elbow motion on armrest). Please note that the hydrodynamic drag is about 15.8 g during aiming motions in the transversal plane (estimated with a C_DA value of 0.36, see Goldstein, 1969, and an average arm speed of 3 cm/ sec, 6 cm/sec at the hand and 0 at the elbow). The elbow strap was designed and attached in a way that allowed unrestricted movement, but also guaranteed a similar arm position for all subjects. The software ran on a real-time PC with a sampling rate of 50 Hz for data recording. The experimental GUI had the size of 20.6 x 17.7 cm for all experiments, displayed here on a 17” LCD monitor.

Figure 1. Underwater joystick, experimental setup, and experimental GUI.

Experimental task and design. On the experimental screen, crosshairs with black lines were shown on a grey background. Subjects had to move the circular cursor to the starting point at the centre of the crosshairs. Upon reaching the centre, the starting position had to be held for 2 sec, until the aiming task was started. There were four different target ring positions at the intersection points between the black circle and the vertical and horizontal axes (see Figure 1). The centre of the ring had to be matched as quickly as possible and held for 0.5 sec. Then, subjects had to move back to the start position, whereby the next aiming task was started. The order of the four target positions was randomly chosen. Each subject performed the experiment in filled (22° C water temperature) vs. empty basin, while the order of both conditions was counterbalanced across subjects.

Procedure. Subjects were seated at the water basin with a 70 cm distance to the monitor, positioned their right arm on the joystick armrest, attached the elbow strap and grasped the joystick. Subjects were instructed about the experimental task and procedure online. In the "Water" condition the complete joystick and the subjects' right forearms were fully immersed in the water. In the "Dry" condition the same setup was used in the empty basin. The two conditions were performed on different days, with a maximum interval of 8 days between both sessions. In each session, subjects performed a training trial with four aiming tasks prior to the main experiment. Subjects wore ear protectors to avoid any acoustical disturbances during the experiment. After completing an experimental condition, participants rated the physical effort during the experimental task ("How physically demanding was the last task?"; 20-point Likert scale ranging from "very low" (1) to "very high" (20), adapted from the NASA-TLX questionnaire; Hart & Staveland, 1988).

Study 2: Full body water immersion

Sample. N = 22 subjects, naïve to the experiment (3 females, 19 males; 2 left-, 19 right-handers; M = 27.8 (*S.D.* = 8.0) years of age) participated in the following study. All of them had at least basic diving experience.

Apparatus. The same underwater joystick as in Study 1 was used for this experiment. The joystick and a water-proof 15" LCD monitor (70 cm distance to subjects' head) were installed in an aluminium frame (see Figure 2). The experiment was conducted in an upright position and body posture was stabilized by a foot strap

and an additional holding grip for the left hand. In the underwater condition, the frame was set on the bottom of the 5 m deep pool. The average water temperature was 27° C. Oxygen was provided via a hose connected to a compressed air bottle on deck, i.e., divers did not have to wear a SCUBA jacket during the experiments.

Figure 2. Full water immersion setup.

Experimental task and design. The same experimental task and GUI as in Study 1 was used for Study 2. For the underwater condition, however, the window size was scaled down by 1/3 due to the refractive index (1.33) of the diving mask, leading to a magnification of object sizes. Following the same rationale as Study 1, subjects had to complete a "Dry" and a "Water" condition on different days (max. interval of 8 days), with both conditions being counterbalanced across individuals.

Procedure. In general, the same procedure was realized as in Study 1. In the "Dry" condition, the frame was located on deck and subjects wore ear protectors. Before starting with the underwater sessions, each subject put on a 7 mm short sleeves neoprene suit to avoid hypothermia, a conventional diving mask (prepared with anti-fog spray) and a belt with individually adjusted diving weights to achieve neutral buoyancy.

Study 3: Space flight

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

Apparatus. A space qualified joystick (2 axes, max. workspace ± 20°, 100 Hz sampling rate) was installed onboard the Russian Zvezda service module of the ISS (see Figure 3). The positional resolution of the ISS joystick was higher compared to the unterwater joystick, i.e., more fine-grained motions were recorded.

The body stabilization was similar as in Study 2: foot straps 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 laptop.

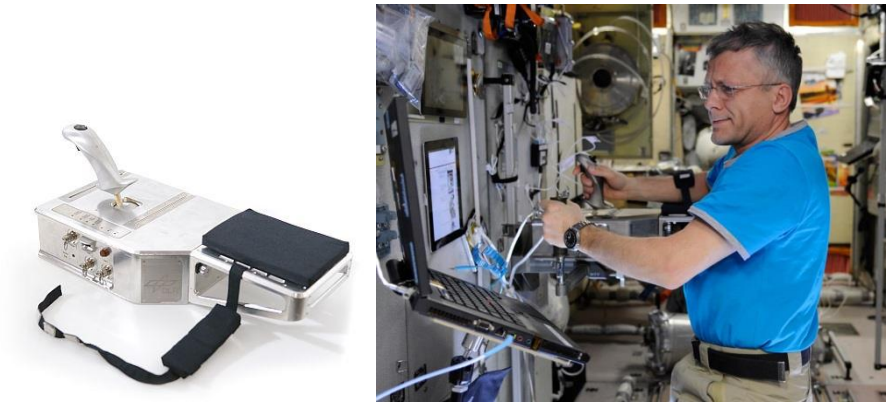


Figure 3. DLR space qualified joystick and experimental setup onboard the ISS.

Experimental design and procedure

All of the three cosmonauts performed the same experiments as in Study 1 and 2 during a pre-mission training session three months before their mission launch, onboard the ISS (exactly after 2 weeks in space), and a post-mission session, two weeks after completing their half-year space missions. The same procedure (instruction, experimental workflow, and questionnaire) as in Study 1 and 2 was carried out.

Results

The complete aiming motion was split up into two functionally meaningful task segments for a more detailed analysis: a gross motion part and a fine motion part. We recorded the gross motion part from initial motion onset (> 20 pixels (px) distance from start), from experiement start until reaching the target zone. The gross motion part was deemed completed after an interpenetration of 20 px into the target ring. Subsequently, slower and more finely graded motions were performed until the target position was precisely matched with a threshold of 3 px.

For all subtasks the required times were recorded. Additionally, kinematic parameters, i.e., the mean motion speed and the standard deviation of motion speed, were computed for the gross motion part, since effects of water immersion or microgravity should be most evident in the gross aiming motion. For Study 2 the data from one subject was omitted in the subsequent analyses due to the occurrence of several interruptions during the underwater session (problems with the diving mask).

The following statistics were calculated for each measure: arithmetic mean, standard deviation (in parentheses), p-value of the paired samples t-test, percentage change, and effect size (Hedges' g).

Table 1. Performance measurements for Studies 1-3

Study/ Measure	*Dry/ 1G*	*Water/ µG*	*Sign. (t-test)*	*Rel. Change/ Effect Size g*
Study 1 (n =19)	**Forearm Water Immersion**			
Gross Motion Time [s]	0.351 (0.091)	0.482 (0.210)	$p < .01$	+37.3%/ 0.79
Fine Motion Time [s]	1.772 (0.317)	1.847 (0.361)	*ns.*	+ 4.2%/ 0.22
Gross Motion Speed [px/s]	844.2 (171.5)	706.9 (209.5)	$p < .01$	-16.3%/ 0.70
Max Gross M. Speed [px/s]	1782.2 (685.0)	1408.2 (517.5)	$p < .01$	-21.0%/ 0.60
SD Gross M. Speed [px/s]	568.7 (190.6)	432.9 (229.5)	$p < .01$	-23.9%/ 0.63
Physical Demand [1-20]	3.684 (2.110)	3.474 (1.926)	*ns.*	-0.06%/ 0.10
Study 2 (n = 21)	**Full Body Water Immersion**			
Gross Motion Time [s]	0.361 (0.165)	0.442 (0.156)	$p < .05$	+22.4%/ 0.49
Fine Motion Time [s]	1.675 (0.499)	2.223 (0.688)	$p < .01$	+32.7%/ 0.89
Gross Motion Speed [px/s]	747.1 (225.6)	636.5 (164.3)	$p < .05$	-14.8%/ 0.55
Max Gross M. Speed [px/s]	1762.6 (562.6)	1510.3 (502.8)	$p = .06$	-16.7%/ 0.46
SD Gross M. Speed [px/s]	591.4 (221.9)	507.3 (193.6)	$p < .10$	-16.6%/ 0.40
Physical Demand [1-20]	3.833 (3.148)	4.167 (2.915)	*ns.*	+ 8.7%/ 0.11
Study 3 (n = 3)	**Space Flight/ Microgravity**			
Gross Motion Time [s]	0.308 (0.057)	0.432 (0.152)	--	+40.3%/ 0.86
Fine Motion Time [s]	2.359 (0.213)	3.017 (0.662)	--	+27.9%/ 1.07
Gross Motion Speed [px/s]	746.7 (93.0)	632.6 (106.7)	--	-15.3%/ 0.91
Max Gross M. Speed [px/s]	2035.0 (520.2)	1453.5 (232.7)	--	-28.6%/ 1.15
SD Gross M. Speed [px/s]	667.1 (174.6)	470.8 (116.9)	--	-29.4%/ 1.06
Physical Demand [1-20]	4.0 (1.323)	9.0 (3.464)	--	+125%/ 1.53

Comparing the performance measures in the "Dry"/ "1g" conditions across the three studies revealed no substantial differences, except for the fine motion times in Studies 1 and 2 vs. 3. The higher baseline level for the space flight experiment can be explained by the higher positional accuracy of the space joystick, making it more difficult not to exceed the 3 px threshold.

The *gross motion times* in all of the three studies were significantly increased in the "Water" or "Microgravity (µG)" conditions compared to the "Dry" or "1g" conditions. Large effect sizes were obtained for partial immersion ($g = 0.79$) and the space study ($g = .86$), while the corresponding effect size in the full immersion study only reached a moderate level ($g = .49$). Regarding the *fine motion times*, similar significant increases due to water immersion or microgravity were found in the full immersion study ($g = 0.89$) and the space study ($g = 1.07$), whereas no significant effect (and only a small effect size of $g = .22$) was evident in the partial immersion study.

A highly consistent result pattern was found when analysing the *gross motion speed*, which decreased significantly in all studies, with moderate effect sizes in the partial and full immersion study ($g = .70$ and $g = .55$) and a large effect size in the space study ($g = .91$). Consistently, the *maximum speeds* decreased in all studies with

moderate effects during water immersion (g = .60 and g = .46), as well as a large effect in the space study (g = 1.15). Please note, however, that the conventional level of significance was not reached in the full immersion study (p = .06). Consistently, the *standard deviation of speed* decreased in all studies. Results yielded significantly lower values and a moderate effect size for partial immersion (g =.63), and a large effect for the space study (g = 1.06). Again, no significant effect was observed for the full immersion study, where only a small effect size was found (p <.10; g = .40).

Finally, the subjective rating on *physical demand* during the aiming tasks was explored. Here, we found no effects of water immersion at all, but a large effect size when comparing space and terrestrial conditions (g = 1.53).

Discussion

In three empirical studies, the effects of weightlessness on aimed arm movements was investigated during forearm water immersion, full body water immersion in 5m depth and during spaceflight after 2 weeks in space.
In all of the three setups, the same result pattern for rapid, gross arm motions is evident: weightlessness caused significantly longer motion times (+22–40%), lower maximum (-17–29%), mean speeds (-15–16%) and speed variance (-17–29%). This overall pattern is consistent with prior research, which demonstrates that sensorimotor control is substantially degraded in weightlessness, resulting in decelerated motion profiles. The magnitude of this effect varies individually, as reported by Bock (1998). Maybe, different individual vulnerability to weightlessness also factor into the smaller effect sizes found in the full body immersion study.

Regarding fine motion, longer times are required in all setups, although large effects are only evident during full immersion (+33%) and space flight (+28%), while only a minimal effect emerges for partial immersion (+4%). Mean motion speed and variance of speed are also reduced for this aiming phase, as indicated by additional analyses. In the target zone, gross arm motion has to be decelerated abruptly and several motion reversals have to be performed until matching the target precisely. A plausible explanation for the above results can be that multi-limb coordination or stabilization play a significant role during these dynamic positional corrections. In the case of full immersion or space flight, the complex coordination of the inertial load and reactive forces of all limbs and joints involved is more difficult and thus the dynamic impulses are reduced.
Interestingly, the weightlessness of the human forearm seems to be sufficient to induce a slowing of gross aiming motion. It could be argued that this effect is a direct result of water viscosity. There are several facts contradicting this assumption: 1) the C_DA values for forearm motions in the sagittal plane (top and bottom aim) should be substantially lower than the corresponding values for arm motions in the transverse plane (left and right aim). However, no significant differences regarding maximum speeds for the both movement planes of the immersed forearm are found, 2) there is no significant correlation (r = 0.03) between the subjects' maximum speed in the "Dry" condition and the decrement of maximum speed in the "Water" condition, 3) the overall result pattern is very similar to the space flight results.

As discussed in the introduction, cognitive impairment may affect performance during full body immersion due to higher ambient pressure (1.5 bar in the present study). Moreover, it has been discussed that a higher cognitive load e.g. due to increased general stress level during a space mission has a detrimental effect on sensorimotor performance (Manzey et al., 2000). In additional analyses, we did not find any significant changes of response times in weightlessness, which would be an indication of reduced cognitive resources.

Comparing the water immersion setups with space flight revealed two main differences: 1) sensorimotor degradation is even more pronounced in space, with large effect sizes for all the performance measures and 2) the subjectively rated physical effort was significantly higher in space compared to the terrestrial sessions. Seemingly, the changed gravitational state further contributes to the degradation motor control. Lackner and DiZio (1992) emphasized that muscle spindle activity is also modulated by vestibular activity. The additional vestibular dysfunction explains the stronger effects during spaceflight and might also be the reason for higher physical efforts astronauts have to expend to stabilize their motions. Other studies investigating force production with an isometric joystick (i.e., the joystick is not deflected) successfully demonstrated that the changes of proprioception due to weightlessness can also be shown during water immersion (exaggerated peak and end forces; Dalecki, 2013). However, the specific effects attributed to vestibular dysfunction (higher initial forces) could not be documented in the underwater condition.

Altogether, promising results could be gathered showing that the general effect direction of weightlessness on sensorimotor performance can be effectively simulated by water immersion. Even for rapid aiming tasks - requiring joystick deflections - the water immersion analogue is able to simulate key aspects of space flight, making it a valuable tool for future research.

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., Arnold, K.E., & Cheung, B.S. (1996). Performance of a simple aiming task in hypergravity: I. overall accuracy. *Aviation, Space, and Environmental Medicine, 67,* 127-132.

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.

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

Dalecki, M. (2013). *Human fine motor control and cognitive performance in simulated weightlessness by water immersion.* PhD thesis, Köln, Germany: German Sport University Cologne.

Goldstein, S. (1969, October). Model design techniques and test methods for precision underwater simulation. In *Proceedings of th 4th Space Simulation Conference* (p. 1005).

Hancock P.A. & Milner, E. (1982). Mental and psychomotor task performance in an open ocean underwater environment. *Research Quartely for Exercise and Sport 53*, 247–251.

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.

Kanas, N. & Manzey, D. (2008). *Space psychology and psychiatry*. Dordrecht, The Netherlands: Springer.

Lackner, J.R. & DiZio, P. (1992). Gravitoinertial force level affects the appreciation of limb position during muscle vibration. *Brain Research, 592*, 175-180.

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., 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.

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.

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*, 387-394.

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

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

Wang, P., Wang, Z., Wang, D., Tian, Y., Li, F., Zhang, S., et al. (2015) Altered gravity simulated by parabolic flight and water immersion leads to decreased trunk motion. *PLoS ONE, 10(7): e0133398.*

Weber, B., Schätzle, S., Riecke, et al. (2016). Weight and weightlessness effects on sensorimotor performance during manual tracking. In F. Bello, H. Kajimoto, and Y. Visell (Eds.), *Haptics: Perception, Devices, Control, and Applications*, LNCS 9774 (pp. 111-121). Springer International Publishing.

Whiteside, T.C.D. (1960). *Hand-eye co-ordination in weightlessness*. Flying Personnel Research Committee.

Analysis of potentials of an HMI-concept concerning conditional automated driving for system-inexperienced vs. system-experienced users

Kassandra Bauerfeind[1], Amelie Stephan[1], Franziska Hartwich[2], Ina Othersen[1], Sebastian Hinzmann[1], & Lennart Bendewald[1]
[1]Volkswagen Aktiengesellschaft, Wolfsburg
[2]Chemnitz University of Technology, Chemnitz
Germany

Abstract

Conditional automated driving (CAD) functions will be one of the key technologies promising comfort and efficiency in personal transportation. This work addresses the importance of a user-centered and variable Human-Machine-Interface (HMI) for CAD in consideration of different levels of trust. The question arises as to how the level of trust, presumably caused by system-experience with an automated system, modulates information needs. The variable HMI-concept was tested with a panel of 47 subjects in a driving simulator. Effects on system evaluation in terms of experience with a conditional automated system (between; system-inexperienced vs. system-experienced users) and the HMI (within; maximal-HMI with higher informational content vs. minimal-HMI with lower informational content) were examined. The gaze behaviour showed that the system-experienced users trusted the system more and monitored the system less frequently than the system-inexperienced users. System-experienced users focused on a non-driving-related task more often than system-inexperienced users. Even though, both user groups trusted the system more using the maximal-HMI than using the minimal-HMI, it is assumed, that long-term use will modulate the level of trust and the resulting information needs. This study supports the idea of adaptability of the HMI depending on the level of trust and the information needs.

Introduction

Besides the need for security (Benmimoun, Zlocki, Aust & Faber, 2011) and comfort, the wish for a flexible mobility rises. With technical progress in sensor technology as well as digitalization, a flexible mobility as promised by conditional automated driving will be possible in the near future (Federal ministry for traffic and digital infrastructure, 2015). When using such a new technology, it is unclear how the human and the machine will interact to prevent misunderstandings. Especially during initial contact, the HMI might adopt the role of a teacher, introducing the user into the system. The user has to become familiar with his task as some kind of a co-driver, trusting the system's ability to safely perform the main driving task. Trust in automation (Lee & See, 2004) represents an important factor while interacting with the vehicle (Beggiato, Hartwich, Schleinitz, Krems, Othersen & Petermann-Stock,

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

2015; Hergeth, Lorenz, Vilimek & Krems, 2016). This fact stresses the importance of a trustworthy human-machine-interface (HMI) (Bendewald, Stephan, Petermann-Stock & Glaser, 2015). The HMI should be user-orientated and therefore guarantee system transparency, predictability and comprehensibility for all upcoming manoeuvres (Beggiato et al., 2015). To match the users' requirements, Nielsen (1993) recommends the distinction between two user groups: people, who have never interacted with a conditional automated system before (system-inexperienced users, SIUs) versus people, who have already got to know such a system (system-experienced users, SEUs). Using such an automated system requires trust in technology (Lee & See, 2004), while for trust in technical systems, system experience plays a major role (Muir, 1994). It is assumed that trust in the system is a result of system experience. Hergeth et al. (2016) observed that higher trust in the automated vehicle results in reduced control gazes while focusing on a non-driving-related task (NDRT). In general, system users tend to focus on a NDRT more often, if they trust the system (Beggiato et al., 2015). In contrast, SIUs are expected to have a higher need to monitor and control the system than SEUs. Results of a driving simulator study (Beggiato et al., 2015) demonstrated that this user type wishes to have detailed system relevant information especially while initially getting in contact with the system. If users are informed about any system decision, they are able to understand those decisions, match them with the environment and develop a system comprehension. To guarantee this transparency, detailed as well as redundant information concerning system decisions and manoeuvres should be available. It is assumed that the users will expect this information in the instrument cluster (FPK) and in the Head-up Display (HUD), which are seen as usual information sources to fulfil the main driving task.

In contrast to SIUs, SEUs are expected to already possess a developed system understanding. They might have less uncertainty trusting the system's ability to safely perform the main driving task compared to SIUs. Beggiato et al. (2015) and Hergeth et al. (2016) claim that by the gain of trust in the automated system, the need to control is shrinking. Hence, SEUs do not want to monitor the system as strongly as SIUs. Concluding from the participants' statements of the driving simulator study of Beggiato et al. (2015), users want to have the possibility to obtain system relevant information concerning system decisions, manoeuvres and status, but do not want to be confronted with these at any time. SEUs might use the chance to give away the main driving task to turn towards a NDRT, like the infotainment.

The question arises as to how a standardized HMI serves different types of users: a user, who has never interacted with an automated system before versus a user, who has. This paper focusses on a user-centered HMI-concept for conditional automated driving. The special feature of this HMI-concept is the scalability, which allows adjusting to the user's level of trust and his need for information. Hence, the objective of the present research was to examine how the HMI meets users' requirements best. Specifically, this driving simulator study addresses the control behaviour and the information needs of two different user groups driving conditionally automated. It is assumed that SIUs will monitor the system stronger as they have little trust in the system and therefore wish for detailed and redundant system information. SEUs, however, are expected to prefer reduced information and

a comfortable sitting position to enjoy NDRT since they trust the system more. This was realised by examining the effects on system evaluation in terms of experience with a conditional automated system and the HMI.

Experimental user study

Experimental setup and design variations

The experimental setup consisted of a static driving simulator with an AUDI mock-up of the Group Research of Volkswagen Aktiengesellschaft, equipped with an automatic gear and three projection screens 3.5 m in front of the mock-up. The projections screens' width were 3.05m each and the resolution of the projector was 1920 x 1200 pixels. A field of view of 140° was covered. The simulation was implemented with the software Virtual Test Drive (VTD) and was projected onto three screens.

The study was conducted using a 2x2-mixed factors design with the factors system experience and HMI with two characteristics each. Participants differed in system experience (between-subjects factor): users, who had never interacted with a conditional automated system before (SIUs) versus users, who have already got to know such a system (SEUs). The SEUs took part in a previous study for conditional automated driving, where they learned how to activate and deactivate this automated system. Furthermore, they attended a separate training, where they received a detailed system description and practiced to operate the two HMIs. All participants tested two HMIs for conditional automated driving (within-subjects design) in randomized order, which were designed to fit the respective user group's needs. There was the maximal-HMI with detailed as well as redundant system relevant information, which supports monitoring the system, potentially appropriate for the SIUs. In contrast to this, the minimal-HMI with rather reduced system relevant information, allowing the user to lean back due to moveable hardware elements. Effects of experience with a conditional automated system and the HMI on system evaluation were examined. For this evaluation, data of trust, control behaviour and information needs were gathered.

User-centered HMI-concept for conditional automated driving

To take the individual level of trust and the resulting information needs into account, the Volkswagen Group Research has developed a user-centered HMI-concept for conditional automated driving. The special feature of this human-machine-interface is the scalability, which allows adjusting to the user's need by varying the amount of information in the different displays (see figure 1) and the sitting position while driving in conditional automation. The comfortable sitting position is realised with the movement of hardware elements: the steering wheel will move towards the instrument panel and the driver seat (inclusively the operating element for the infotainment) will move backwards for a bigger legroom. This concept is a further development of the HMI shown in the test vehicle *Jack* (Bendewald, Stephan, Petermann-Stock & Glaser, 2015). Concerning the HMI, this paper's scope is on display content and the movement of hardware elements. The purpose of other HMI

elements is explained in a detailed potential analysis of this HMI-concept (Bauerfeind, 2016).

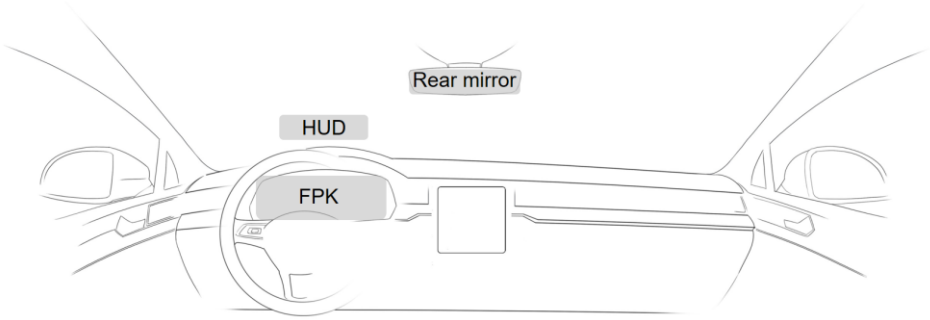
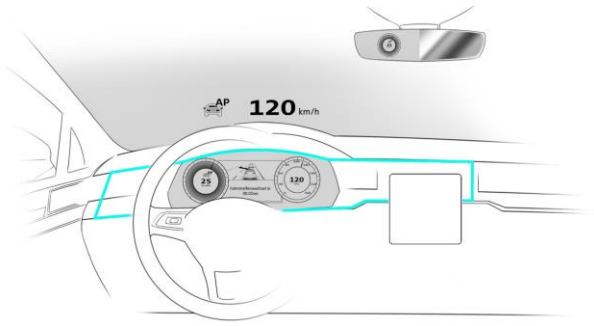


Figure 1. The amount of system relevant information on different displays is scalable.

Since it is assumed that the needs of SIUs and SEUs differ, it was necessary to develop two different HMIs: The HMI for the SIUs (maximal-HMI, see figure 2, picture 1) and the one for the SEUs (minimal-HMI, see figure 2, picture 2). Each HMI contains a certain amount of system relevant information on different displays and the adjustment of scalable hardware.

The maximal-HMI, which potentially serves the need of the SIUs, contains detailed as well as redundant system information, especially located in the HUD and the FPK. Except for the steering wheel, the hardware elements will not move, enabling the user to stay in his driving position able to monitor the system. The movement of the steering wheel conveys the system status *automated driving*: By moving away from the driver, the system seems to announce the taking over of the main driving task. To remind the driver of taking back control, the steering wheel moves towards him. In contrast, the minimal-HMI, which should be appropriate for the SEUs, gives the chance to turn away from the main driving task to enjoy NDRTs. The user will be presented with reduced system relevant information. Since the user's attention is expected to be on the infotainment in the centre console display, this user type might enjoy receiving system relevant information close to this area. This is why the rear mirror, as the nearest display to the infotainment, serves as an informing display. With moving hardware elements (driver seat and steering wheel) the SEUs are provided with a comfortable sitting position to enjoy NDRTs.

Picture 1: maximal-HMI



Picture 2: minimal-HMI

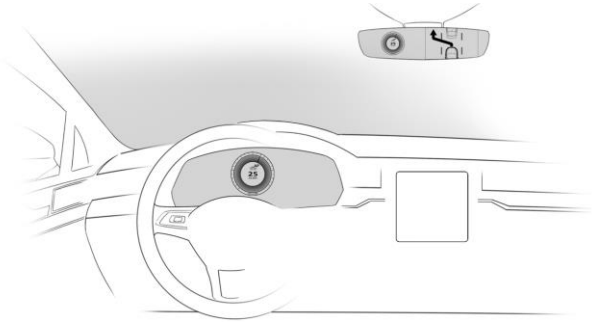


Figure 2. System relevant information on three different displays (FPK, HUD and rear mirror) in the two HMIs: the maximal-HMI (with detailed and redundant information, picture 1) and the minimal-HMI (reduced information, picture 2). Other HMI elements are explained in the detailed HMI specification (Bendewald et al., 2015) and in the potential analysis (Bauerfeind, 2016).

Procedure

All participants were presented with the two HMIs in a randomized order. The driven interstate route of 25 km with other traffic was the same for both HMIs. The gaze behaviour was measured with an eye-tracking system from Ergoneers GmbH using the software Dikablis 2.5 (Ergoneers GmbH, 2016). After receiving the instruction how to activate and deactivate the system without getting further information about the content of the different displays, all participants fulfilled two short trainings. Before the main drive, participants were informed to have the possibility to watch videos on the infotainment display.

The participants were confronted with either a scenario that included an accident, which was the cause for a take-over request (TOR) or the TOR was triggered due to an obstacle on the street. It was randomly choosen which HMI was tested with

which scenario. The participants started driving manually to the interstate, where the system became available and could be activated to fulfil the complete driving task. Due to other road users, participants experienced lane changes done by the automated system. In the middle of a 15 minutes drive, the system asked for a take-over by the participant because it could not handle the upcoming situation on its own (obstacle on the street or an accident). This TOR consisted of visual information in the FPK and a voice output one minute and also 15 seconds prior to the take-over. Participants were told that they had to take over the driving task and exit the interstate manually in the end of each session. After testing each HMI, participants completed the questionnaire *Trust in technical systems* on a 4-point likert scale (Wiczorek, 2011). This questionnaire listed a total of 16 items and Cronbach's Alpha was α=.91 for both HMIs. In the end of the study, the participants created their desired HMI for conditional automated driving. To facilitate this, a cockpit template was used that allowed participants to place cut-out pictures of system relevant information (system status, velocity, manoeuvre announcement, other vehicles, traffic lanes). By doing so, participants could personalise the cockpit, so they could demonstrate where they prefer to receive the different information. Furthermore, they indicated the desired movement of hardware elements. Choosing between the maximal- and the minimal-HMI, participants could also select their preferred HMI. They could also abstain from this decision.

Participants

The sample included 47 drivers, who were recruited from the test driver pool of Volkswagen Group Research. There were 24 SIUs (42% female) and 23 SEUs (44% female). The SIUs' mean age was 37.7 years (SD = 11.9 years; min = 22, max = 59) and the SEUs' mean age was 39.5 years (SD = 9.7 years; min = 22, max = 58). All participants were employees of the Volkswagen Aktiengesellschaft. The participants drove an average of 19298 km per year. Most of the participants had gained experience with an Adaptive Cruise Control (ACC), and a Cruise Control (CC) (see table 1). Most of the SIUs had not driven with Heading Control (HC).

Table 1. Percentages of user groups' experience with the driver assistance systems "Adaptive Cruise Control"(ACC), "Cruise Control" (CC) and "Heading Control" (HC)

	ACC	CC	HC
SIUs	67	71	42
SEUs	70	65	65

Results

Trust in the conditional automated system: subjective data

Subjective data of 47 participants were analysed using repeated measures Analysis of Variance (rmANOVA) to examine the effects of experience with a conditional automated system and the HMI on trust in the system. Data revealed that SEUs trusted the conditional automated system more than SIUs, $F(1, 45) = 8.47$, $p = .006$,

η_p^2 =.16. The system was rated as more trustworthy when participants drove with the maximal-HMI than with the minimal-HMI, $F(1, 45) = 5.11$, $p = .029$, $\eta_p^2 = .10$.

Control behaviour: gaze data

The aim was to investigate whether SIUs showed a stronger control behaviour than SEUs. Areas of interest had to be determined (see figure 3) to compute the participants' percentage distribution of gazes.

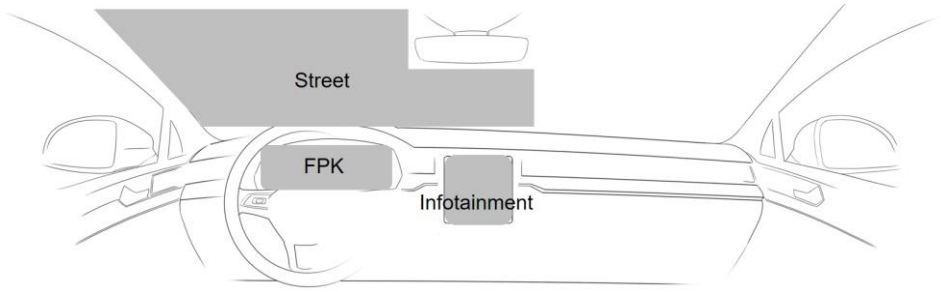


Figure 3. Areas of interest for the investigation of gazes. For the analysis of control behaviour, just the gazes on the street and in the FPK are reportet. Control gazes into mirrors are descriped in the detailed potential analysis (Bauerfeind, 2016).

One participant had to be excluded from the gaze analysis due to technical issues. Thus, gaze data of 46 participants were analysed using *t*-tests for independent samples (Mann-Whitney-test in case of missing normal distribution) to examine whether SIUs showed a stronger control behaviour compared to SEUs. There is a tendency that SIUs monitored the FPK more than SEUs (see figure 4 & table 2). Furthermore, SIUs monitored the street significantly stronger than the SEUs, in case of driving with the minimal-HMI. There is no difference in gaze frequency between the two user groups when using the maximal-HMI.

Gaze data showed that SEUs watched the infotainment in the infotainment display more often than SIUs; especially while driving with the minimal-HMI (see table 2). This tendency could also be observed while driving with the maximal-HMI.

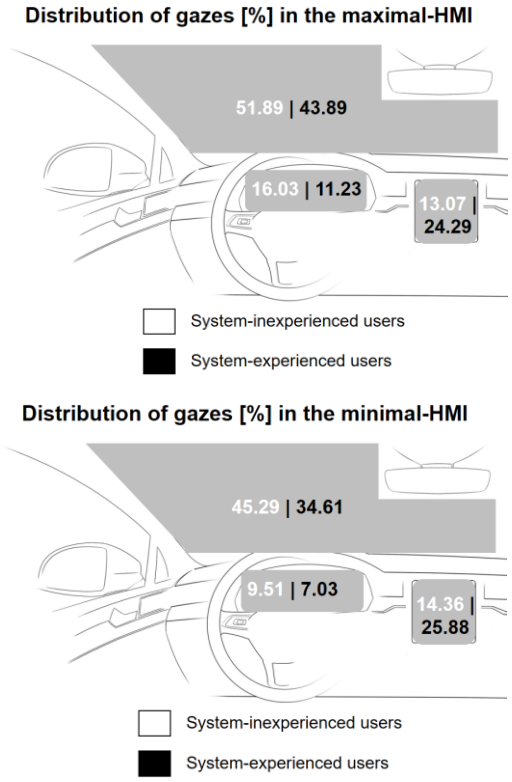


Figure 4. Average percentage distribution of gaze frequencies into different AOIs while driving in conditional automation for the two user groups and for the two HMIs. Control gazes into mirrors are descriped in the detailed potential analysis (Bauerfeind, 2016).

Table 2. Results regarding gaze frequencies for both user groups into different AOIs for the two HMIs.

	SIUs	SEUs					
	M (*SE*)	*M* (*SE*)	*t*	*U*	*z*	*p*	*r*
	Maximal-HMI						
FPK	16.03 (1.85)	11.23 (1.15)		187.00	-1.69	.090	-.25
Street	51.89 (2.78)	43.89 (4.36)	1.55			.131	.25
IT	13.07 (2.54)	24.29 (4.41)		187.00	-1.69	.090	-.25
	Minimal-HMI						
FPK	9.51 (1.05)	7.03 (0.92)		182.00	-1.80	.071	-.27
Street	45.29 (2.94)	34.61 (3.96)		157.00	-2.35	**.019**	-.35
IT	14.36 (2.20)	25.88 (3.51)	-2.83			**.007**	-.42

Note. SIU = system-inexperienced users, SEU = system-experienced users, IT = infotainment, *M* (*SE*) = mean with standard error, *t* = t-tests for independent samples (in case of normal distribution), *U* = Mann-Whitney-test (in case of missing normal distribution), for a simplified comparison just the means are reported.

Desired HMI

The aim was to examine what the user groups' ideal HMI would look like concerning system relevant information and the movement of hardware elements. Descriptive data analysis of the desired HMI revealed that both user groups had a similar need for information. Half of all participants asked for rather detailed system relevant information while driving in conditional automation: 50% of the SIUs and 52% of the SEUs wanted to be informed about the current driving environment (other vehicles, traffic lanes). In regard of redundancy, information about the system status should be available on several displays according to most of the participants (67% of the SIUs, 70% of the SEUs). 58% of the SIUs (30% of the SEUs) also asked for redundant information about the current velocity. 79% of the SIUs and 74% of the SEUs liked the rear mirror as an informing display, even though 63% of the SIUs asked for a free mirror half without any information. 52% of the SEUs could imagine the rear mirror to serve as a full display.
Results made clear, that in comparison with SIUs, SEUs were more open for the movement of hardware elements while driving in conditional automation. The majority of all participants asked for the movement of the steering wheel (58% of the SIUs, 78% of the SEUs). Concerning the driver seat's flexibility (including operating element) the two user groups had different demands. In contrast to the 33% of SIUs who were against it, 61% of the SEUs asked for this movement to have a comfortable sitting position while driving in conditional automation.
In the end of the study, participants could choose their preferred HMI. They could also abstain from this decision. Results revealed that the majority of the SIUs (75%) liked the maximal-HMI most. In contrast, there was no tendency for the SEUs: 39% of this user group preferred the minimal-HMI and 43% chose the maximal-HMI.

Discussion and conclusions

The aim of this research was to investigate the effects on system evaluation in terms of experience with a conditional automated system and the HMI. This driving simulator study addresses the control behaviour, presumably caused by a lower level of trust and the information needs of two different user groups driving conditionally automated.

With regard to trust, SEUs trusted the conditional automated system more than the SIUs. Hence, these results support the findings of Muir (1994), who states that system experience plays a major role concerning trust in a technical system. The maximal-HMI with detailed system relevant information was rated as more trustworthy and more transparent as the minimal-HMI. It needs to be taken into account that in this study system experience was obtained by participating in a previous study based on this HMI, a seperate training, and a detailed system description. Therefore, SEUs did not have long-time experience with this system. The question is whether SEUs' evaluation will change after having driven the minimal-HMI for a longer time. They might prefer the minimal-HMI, because detailed and redundant information might not be seen as transparent and clear anymore, but rather as annoying (Beggiato et al., 2015). This can only be explored in long-term studies. In general, it has to be considered that participants might rate

the system more trustworthy while driving in a simulator rather than driving in a real car. They might feel more secure driving in a simulation than being exposed to a real traffic situation.

Concerning the control behaviour, SIUs were looking at the FPK more often than SEUs to monitor the system. This behaviour was shown for both HMIs. Concerning gazes to the street, just the minimal-HMI made a difference: SEUs watched the street driving with this HMI much rarer than SIUs. One reason could be that SIUs observed the street since there were minimal information on the displays. Another explanation is that the minimal-HMI's aim to convey the possibility of turning away from the driving task to a NDRT succeeded. The SEUs accepted this HMI and made use of this option. These results are consistent with the findings of Hergeth et al. (2016), who declares that higher trust in the automated vehicle leads to reduced control gazes while focusing on a NDRT.

In general, SEUs trusted the system more and showed a weaker control behaviour than the other user group. Nevertheless, all participants had similar information needs in this study. Both user groups asked for system transparency, predictability and comprehensibility. Information about the system status should be available on several displays according to most of the participants. Half of all participants preferred rather detailed system relevant information, which is requested for the initial contact with the system (Beggiato et al., 2015). It has to be discussed whether these results might be attributable to the user training applied to the members of the SEU group. This training might have been too short to gain a sufficient level of system experience, which has to be taken into account when interpreting these results. Nevertheless, the user groups' statements concerning the movement of hardware elements differed: SEUs were more open for the movement of hardware elements while driving in conditional automation compared to SIUs. An explanation might be that SIUs associate this movement with a loss of control. Instead of enjoying a comfortable sitting position while driving in conditional automation, SIUs might feel less secure by taking back the driving task while sitting leaned back. Thus, this training applied to the SEUs already had an impact on users' demands.
Descriptive data analysis showed that in contrast to the SIUs, who preferred the maximal-HMI, the SEUs did not show such a tendency. They rather wished for an HMI individually customized. The SIUs enjoyed to be led by the maximal-HMI, which had proved to be a suitable introduction for using a conditional automated system for the first time.

With regard to the methodology, one of the advantages of this experimental setup with a driving simulator is the high level of situation standardization. It could be guaranteed, that every participant experienced similar conditions. Furthermore, the implementation of a prototypical HMI-concept is easier to realise in a mock-up than in a test vehicle.

These results help to understand differences concerning the interaction between different user types and an automated system. Especially during initial contact, the HMI should be transparent in system decisions, which allows the user to gain trust in the technology. Overall, this study recommends the distinction between user

groups with different levels of system experience while developing an HMI for conditional automated driving. Furthermore, the idea of adaptability of the HMI depending on the level of trust and the need for information is suggested.

References

Bauerfeind, K. (2016). *Potenzialanalyse eines HMI-Konzepts im Kontext des hochautomatisierten Fahrens für erfahrene und unerfahrene Nutzer* [Analysis of potentials of an HMI-concept concerning conditional automated driving for system-inexperienced vs. system-experienced users] (Unpublished master's thesis). Chemnitz University of Technology, Chemnitz.

Beggiato, M., Hartwich, F., Schleinitz, K., Krems, J.F., Othersen, I., & Petermann-Stock, I. (2015). *What would drivers like to know during automated driving? Information needs at different levels of automation*. Paper presented on the 7th conference for driver assistance, November 2015, Munich.

Bendewald, L., Stephan, A., Petermann-Stock, I. & Glaser, E. (2015). *"Jack" - A holistic approach of designing a human machine interface for highly-automated driving*. In VDI Wissensforum GmbH (Ed.), 17th International congress ELIV (p. 453–467). Düsseldorf: VDI publisher GmbH.

Benmimoun, M., Zlocki, A., Aust, M. L. & Faber, F. (2011). *Impact assessment of active safety systems on safety, traffic efficiency and environment within the field operational test "eu-roFOT".* Retrieved from http://eurofot-ip.eu/download/papersandpresentations/its_world_congress/its_wc_a_m_paper.pdf

Federal ministry for traffic and digital infrastructure, (2015). Strategie automatisiertes und vernetztes Fahren [Strategy for automated and connected driving]. Retrieved from https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/broschuere-strategie-automatisiertes-vernetztes-fahren.pdf?__blob=publicationFile

Ergoneers GmbH. (2016). Dikablis Eye-Tracking System: Monokulare Head-Mounted System. Retrieved from http://www.ergoneers.com/wp-content/uploads/2014/09/Dikablis-Essential-Eye-Tracking-Glasses.pdf

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J.F. (2016). Keep Your Scanners Peeled Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving. *Human Factors, 58*, 509-519.

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

Muir, B.M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*, 1905–1922.

Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press.

Wiczorek, R. (2011). Entwicklung und Evaluation eines mehrdimensionalen Fragebogens zur Messung von Vertrauen in technische Systeme [Development and evaluation of a multidimensional questionnaire for the measurement of trust in technical systems]. *Reflexionen und Visionen der Mensch-Maschine-Interaktion–Aus der Vergangenheit lernen, Zukunft gestalten, 9*, 621-626.

Canary in an operating room: integrated operating room music

Alistair MacDonald[1] & Joseph Schlesinger[2]
[1]Saint Patrick Hospital, [2]Vanderbilt University Medical Centre
USA

Abstract

Loud music in the operating room may lead to missed alarms and deleterious patient outcomes. A music volume controller that integrates operating room music with vital sign data from the anaesthesia monitor was tested in a clinical environment with twenty-one anaesthesiologists and nine operating room personnel. Background music volumes were reduced or silenced based on flexible algorithms for heart rate, oxygen saturation or blood pressure. After implementation, study participants completed a survey to assess the performance and usefulness of the device. The results indicated the music volume controller was functional and clinically useful and may promote patient safety.

Introduction

Music is an integral part of surgery today. While studies have demonstrated that music reduces surgeon stress and improves the speed and quality of surgical closures, there is evidence that music poses a distraction hazard and contributes to intraoperative noise pollution which may mask an impending emergency (Lies & Zhang, 2015). The American College of Surgeons (ACS), the American Society of Anaesthesiologists (ASA) and the Association of periOperative Registered Nurses (AORN) and The Joint Commission (TJC) have issued independent statements regarding distraction and noise in the operating room (ACS, Statement on Distractions in the Operating Room, 2016, AORN, Position on Managing Distractions and Noise, 2015, ASA, Statement on Distractions, 2015, TJC, Minimizing Noise and Distractions in the OR and Procedural Units, 2017).

The acoustics in the operating room are generally poor and noise levels frequently exceed Occupational Safety and Health Administration (OSHA) safe exposure standards. Powered orthopaedic saws and drills, forced air warmers, fluid collection suction systems, clanging metal instruments, conversation, electronic equipment and music all contribute to high levels of noise pollution. In one study, the noise in the operating room measured over 100dB for 40% of the time during orthopaedic and neurosurgery procedures, levels comparable to those of a busy freeway (Katz, 2014). The signal-to-noise ratio required for speech discrimination in the operating room is greater both because hard flat walls cause sound reverberation and surgical masks preclude lip reading. As a result, communication is challenging and conversations

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

routinely exceed ambient noise levels. The addition of music helps surgeons ignore distracting sounds, but raises the overall level of ambient noise in the room and further impairs communication, alarm detection, and cognitive processing (Stevenson, 2013). As operating rooms have evolved from cassette players at the head of the bed to central streaming music systems, anaesthesiologists have a lesser degree of control of the acoustic environment in the operating room (Schlesinger, 2015). The current state of practice entails the anaesthesiologist perceiving the pitches and tones of the anaesthesia monitor (e.g. pulse oximetry) over the noise in the room, recognizing there is a problem, and requesting the circulating nurse to interrupt his or her duties to turn the music down or off. In an emergency that requires clear communication, delays in minimizing noise can be critical (Weldon, 2015).

We hypothesized that building intelligence into the operating room music system was feasible and would be useful to the anaesthesiologist and operating room personnel. As an example of an intelligent audio system, modern car stereos now restrict the volume of music until seatbelts are fastened. The precondition of an acceptable pulse oximetry measurement for operating room music could be compared to that of the fastened seatbelt for a car. Similarly, a car's radar, lasers, and cameras can detect an impending collision and integrate with the vehicle's braking, steering and audio systems as a mitigating 'pre-crash system'. A slowing heart rate, diminishing blood pressure, or declining oxygen saturation could be deemed an 'impending collision' that requires a quieter environment for the surgical team to communicate and concentrate on the patient. Our objective was to test a system with both preconditions and automatic music volume reductions based on user-controlled thresholds for heart rate, systolic blood pressure, and oxygen saturation in a clinical environment.

Methods

The study was approved by the Providence Saint Patrick Hospital Joint Investigational Review Board and written informed consent was obtained from patients. The study involved the use of one music volume controller in one operating room at Saint Patrick Hospital. The music controller (CanaryBox™, Canary Sound Design LLC, USA) was interfaced with a Philips Intellivue™ monitor using the RS232 data port and connected to a music source and the operating room audio system (Figure 1).

The study was designed to assess the preferences and experiences of the anaesthesiologist and operating room staff after using the music controller for one day. As usual, music selection and listening volumes were at the discretion of the surgeon and operating room staff. The target sample size was 30 users for a minimum of one day. All patient and cases were eligible for inclusion.

Table 1. The vital sign ranges for music volume adjustments and "time-in-zone" delays to prevent nuisance triggers

	Full volume	*Half volume*	*Music off*
Oxygen saturation (SpO_2) %	$90 \leq SpO_2$	$85 \leq SpO_2 < 90$	$SpO_2 < 85$
SpO_2 delay		20 seconds	10 seconds
Heart Rate (HR) bpm	$50 \leq HR \leq 130$	$40 \leq HR < 50$, or $130 < HR \leq 150$	$HR < 40$, or $HR > 150$
HR delay		20 seconds	10 seconds
Systolic Blood Pressure (SBP) mm Hg	$80 \leq SBP \leq 170$	$70 \leq SBP < 80$, or $170 < SBP \leq 190$	$SPB < 70$, or $SPB > 190$
SBP delay		60 seconds	30 seconds

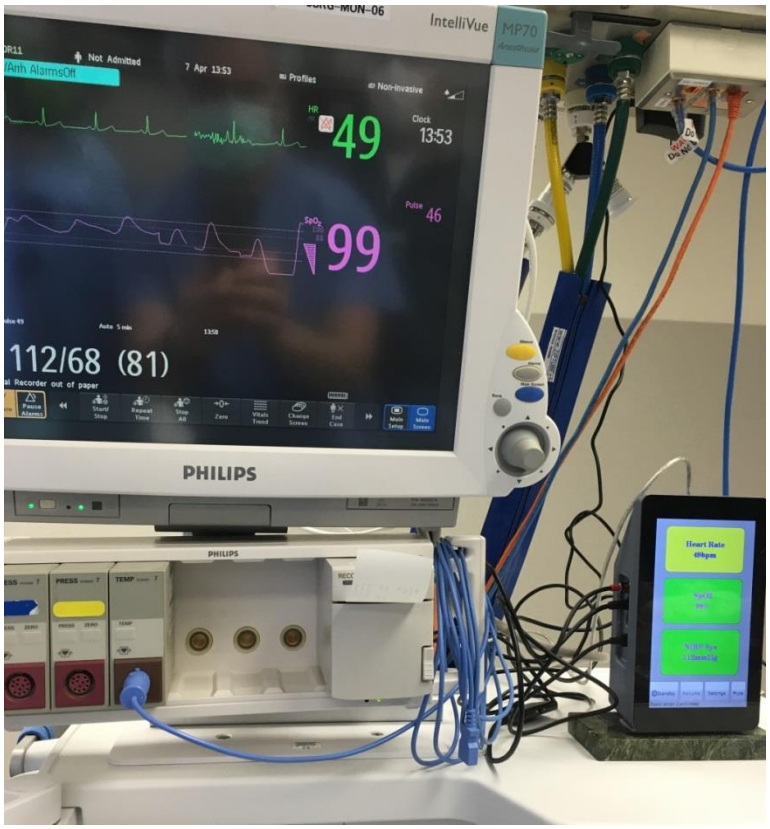


Figure 1. Music Volume Controller Interfaced with Anaesthesia Monitor. The photo shows the position of the music volume controller next to the anaesthesia monitor. The heart rate of 49 beats-per-minute triggers a music volume reduction and this is indicated by a colour change on the controller screen from green to yellow.

The device parameters were adjustable so they could be customized to the patient and the procedure. Default settings of the device for adult patients were as follows:

1) music at full volume if oxygen saturation (SpO_2) >90% and heart rate (HR) between 50 and 130 beats per minute and systolic blood pressure (SBP) between 80mm Hg and 170mm Hg; 2) music at half volume if SpO_2 between 85 and 90%, or HR between 40 and 50bpm or 130 and 150bpm, or SBP between 70 and 80mmHg or 170 and 190mm Hg; and 3) music off if SpO_2 85%, or HR <40 or >150bpm, or SBP <70 or >190mm Hg. To minimize nuisance triggers delay periods ('time in zone') were set for SpO_2 and HR at 10 seconds for full mute events and 20 seconds for half volume events. For SBP, delay periods were set at 30 seconds for full mute events and 60 seconds for half volume events. All volume changes were gradual (fade-in/fade-out) to not startle the surgeon (Strickland, 2015).

Following clinical use, anaesthesiologists and staff were asked to complete a survey assessing performance and usefulness of the music controller.

Results

The target sample size of 30 users was reached and included 21 anaesthesiologists and 9 operating room personnel. *Table 1* - shows the results of the survey including a condensed selection of written responses to open ended questions. Twenty-nine participants responded that they would use the controller again.

Table 1. Survey Data: Responses from 21 anaesthesiologists, 6 nurses, 1 surgeon, 1 surgical technician, and 1 physician assistant.

How much did your room use the music volume controller?

Number of hours	**Number of responses**
>4	23
2-4	6
<1	

How did the volume controller function?

Rating	**Number of responses**
well	27
okay	3
poorly	0

Sample of written responses

What worked well?	**What did not work well?**
Music stopped when O_2 sat went to 80%	
Liked the ability to mute and suspend	Hard to hear a 50% reduction in music
Silenced music when SpO_2 was disconnected	Would like default profiles - adult, paediatric
Responded appropriately to bradycardia	Needs better fixed presets
Intuitive, easy to use	Pandora™ 'timed out' - thought it was controller
Turned off music during hypotension	

Discussion

A majority of anaesthesiologists feel music is a distraction if a patient is having anaesthetic-related problems, so it is important for the anaesthesiologist to have the ability to quickly and easily minimize this source of intraoperative noise (Strickland, 2015). This study tested the feasibility of implementing a volume controller that reduces or silences music automatically based on adjustable vital sign algorithms. Based on experiences of thirty users, the system was found to be functional and clinically useful. Limitations of the study include majority of anaesthesia over surgical clinician response and involvement of only one hospital. Future research on integrated operating room music in a multicentre trial may be useful to assess the effects of automated noise reduction on clinical performance and patient safety.

Disclosure footnote

Dr. MacDonald is the developer of CanaryBox™ and co-founder of Canary Sound Design, LLC.

References

American College of Surgeons. Statement on Distractions in the Operating Room. ACS website. www.facs.org/about~acs/statements/89-distractions. Accessed October 1, 2016.

American Society of Anesthesiologists Committee on Quality Management and Departmental Administration (QMDA). Statement on distractions. ASA website: www.asahq.org/~/media/Sites/ASAHQ/Files/Public/Resources/standards-guidelines/statement-on-distractions.pdf. Accessed January 27, 2016.

AORN Position Statement on Managing Distractions and Noise During Perioperative Patient Care (2015). *Association of Perioperative Registered Nurses Journal*, 99, 22-26.

Katz, J. (2014). Noise in the Operating Room. *Anesthesiology*, 121, 894-898.

Lies, S.R., & Zhang, A.Y (2015). Prospective randomized study of the effect of music on the efficiency of surgical closures. *Aesthetic Surgery Journal*, *35*, 858-863.

Schlesinger J (2015). In response: smart operating room music. *Anesthesia & Analgesia*, *121*, 836.

Stevenson, R.A., Schlesinger, J.J., & Wallace, M.T (2013). Effects of divided attention and operating room noise on perception of pulse oximeter pitch changes: a laboratory study. *Anesthesiology*, *118*, 376-381.

The Joint Commission. Minimizing Noise and Distractions in the OR and Procedural Units. The Joint Commission website. https://www.jointcommission.org/assets/1/23/Quick_Safety_Issue_35_2017_Noise_in_OR_FINAL.pdf. Accessed November 1, 2017.

Relevant eye-tracking parameters within short cooperative traffic scenarios

Jonas Imbsweiler, Elena Wolf, Katrin Linstedt, Johanna Hess, & Barbara Deml
Karlsruhe Institute of Technology,
Germany

Abstract

In everyday road traffic, communication between road users plays an important role – especially in traffic situations where cooperation is necessary. In order to ensure successful future communication between human road users and autonomous vehicles, the communication between human road users must be better understood and modeled for automatic traffic. A relevant parameter in the analysis of cooperative scenarios is gaze behaviour. In contrast to e.g. mental workload, no specific parameters have been identified for analyzing cooperative scenarios so far. As a method, on a traffic-training-center, two experiments were conducted for cooperative situations implementing a narrow-passage (N=21) and a specific t-junction-scenario with three road users (N=20) to investigate cooperative behavior. In both experiments, the subjects were confronted with offensive or defensive approaching behaviours and the decision-making behaviour was investigated. Aim of the analysis was to identify relevant gaze parameters for cooperative scenarios. The results show that for different scenarios different parameters become relevant. For a complex scenario saccadic parameters are more important than fixation parameters. In contrast fixation-metrics show higher importance in simple scenarios.

Introduction

Cooperation between road users is necessary for maintaining the traffic flow (Benmimoun et al., 2004). Generally, cooperation occurs in situations in which it is not clear who is allowed to drive first. Specifically in city situations cooperation is needed, e.g. at equal narrow-passages or with specific trajectory combinations of road users in t- or x-junctions (figure 1).

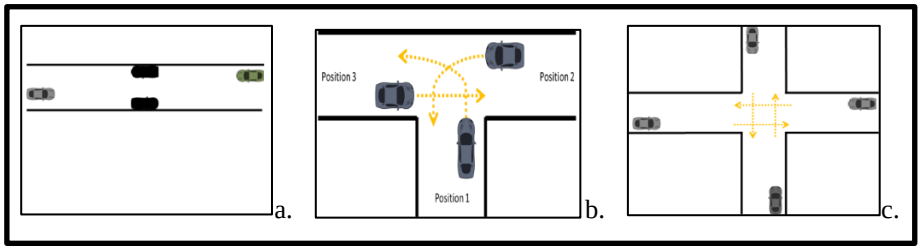


Figure 1. Examples of deadlock situations.a. narrow passage, b. t-junctions, c. x-junction

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In every of the three situations (figure 1) the road users approach simultaneously which creates a deadlock. In Germany in particular, the traffic regulations demand to negotiate who drives first if the situation is not regulated by the traffic law (§ 11(3)). Hoc (2001) defines a situation as cooperative if at least two agents are involved in a way in which both have similar goals and can interfere with the resources and procedures of the other agent. Furthermore, the agents try to solve the problem with the goal to facilitate their own actions. In a cooperative situation the cooperation-partners are forced to solve the problem within the situation.

Cooperation between road users is only partially well investigated and there exists much research for cooperation between car drivers and vulnerable road users (e.g. Witzlack, 2016). But there is still a lack of research addressing cooperation between car drivers, in particular regarding the eye-tracking parameters.

Imbsweiler et al. (2016) conducted an observation study to identify interactive and cooperative behavior at three different intersections in which the described situations could occur. They were able to define six different kinds of approaching behaviour for the t-junction-scenario and for the narrow-passage (table 1 and 2). The approaching behaviour was classified into offensive and defensive depending on the resulting order of driving: When a person displays an offensive behaviour, he or she is more likely to drive sooner than the cooperation partner. Defensive behaviour on the other hand promotes sooner driving of the cooperation partner

Table 1. Approaching behaviour for the narrow-passage.

Approaching behaviour	Driver behaviour	Behaviour classification
1	Driver stops distinctively	Defensive behaviour
2	Driver stops and uses the flasher	Defensive behaviour
3	Driver decelerates and uses the flasher	Defensive behaviour
4	Driver maintains speed	Offensive behaviour
5	Driver accelerates	Offensive behaviour
6	Driver decelerates	Offensive behaviour

Table 2. Approaching behaviour for the t-junction.

Approaching behaviour	Position of drivers	Driver behaviour	Behaviour classification
1	3	Driver 1 decelerates, stops and use the flasher	Defensive behaviour
	1 or 2	Driver 2 direcetion indicator and stops	
2	3	Driver 1 decelerates and stops	Defensive behaviour
	1 or 2	Driver 2 2 direcetion indicator and stops	
3	1 or 2	Driver 1 decelerates and indicate	Defensive behaviour
	1 or 2	Driver 2 decelerates, 2 direcetion indicator and use the flasher	
4	3	Driver 1 maintains the speed	Offensive behaviour
	1 or 2	Driver 2 direcetion indicator and decelerates	
5	3	Driver 1 decelerate	Offensive behaviour
	1 or 2	Driver 2 2 direcetion indicator and decelerate	
6	1 or 2	Driver 1 decelerates and 2direcetion indicator	Offensive behaviour
	1 or 2	Driver 2 direcetion indicator, than decelerates and	

In order to initiate the cooperation-process the road users have to communicate. Risser (1985) distinguishes between implicit and explicit communication. Implicit communication refers to the driving behaviour and includes the acceleration, deceleration, trajectory or the position on the street of a road user. In contrast, explicit behaviour refers to communication signs like hand gesture, direction indicator or light flash. Ba et al. (2015) investigated the explicit communication signs for interactive scenarios by using eye-tracking data and focusing on fixation-metrics. Their results indicated that in interactive scenarios the fixation time, mean fixation duration, and frequency is higher when the interaction partner uses a communication sign. Furthermore, a signal helps to underline the intention in an interactive or cooperative scenario. But the sign depends on the context. The investigated scenarios included different kinds of overtaking scenarios in the city and on the motorway. These scenarios are not comparable to the presented scenarios from our study. Furthermore, only the fixation metrics in general were investigated. In the context of human-robot-interaction more research on communication is conducted. Sakita et al. (2004) could show that in human-robot-interaction the last fixation is an important cue for an operator's intention. Most studies in the human-

machine-interaction context address mental load (e.g. Schneider, 2017) or fatigue (Manzey & Lorenz, 1997) though.

Apart from mental load or fatigue there exists no scientific base for eye-parameters which are important for analyzing cooperative situations. A further challenge is that cooperative scenarios can be as short as five seconds. An analysis of the behaviour over the time is indeed possible but the results of the analysis of e.g. mental workload cannot be transferred in any simple way. As for deadlock situations we could only find empirical results regarding the perception of road users (Imbsweiler et al., 2017).

Imbsweiler et al. (2017) investigated the equal narrow-passage in an experiment and analyzed the subjective perception. They found out that road users feel more confident to drive first and assess the cooperation-readiness higher if the opponent drives in a defensive way. In this case, no situation was longer than 15 seconds. The results show that the subjects feel more confident if they could drive first. The narrow-passage is one of the simpler deadlock situations in contrast to the t-junction.

The present paper refers to the following research questions: Is it possible to identify eye-tracking parameters which indicate a specific behaviour in short cooperative situation? Is there a difference between simple and complex cooperative situations?

Method

Two experiments have been conducted. The design of the experiments is mainly similar, while once addressing a narrow-passage, and once a t-junction-scenario.

Subjects

42 subjects aged from 20 to 28 years were recruited for the experiments. Of these N=22 subjects drove the equal narrow-passage and N=20 subjects drove the t-junction-scenario. The subjects, mainly students, drove between 20 km and 500 km per week (M=141.82, SD=156.11). For their participation, a financial compensation was granted. There were N=5 examiners (male=3, female=2) involved as confederates at an average age of 23.00 years (SD=2.24 years).

Material

The test-vehicle was a VW Passat 2.0 TDI Variant equipped with lidar sensor (Velodyne VLP-16), eye-tracking cameras (SmartEye Pro 5.9) and CAN-recording. The confederates drove a Ford Fiesta and a VW Passat for the t-junction. For the narrow-passage the usage of the car was randomized. The confederates used scripts to interact with the subjects. These scripts were based on the observational study and are described in table 1 and, in more detail, in Imbsweiler et al. (2016). For the experiments a survey was conducted. It consisted of questions on a seven-point Likert-Scale and addressed the cooperativeness of other drivers as well as the overall cooperation-intensity of the situation. Furthermore, it was asked how confident the participants were to pass first, second or third – in the case of the t-junction.

Additionally they were instructed to rate the risk of a potential accident and they completed the abridged version of the NEO-FFI (Borkenau & Osterdorf, 1993).

General procedure

The experiments were conducted at a traffic-training-center. There were at least three examiners involved: one examiner (EX-1.1) drove the participating car in the narrow-passage-scenario, whereas two examiners (EX-1.1/1.2) were involved in the t-junction-scenario. Another examiner (EX-2) was placed in the rear of the test-vehicle to monitor the measurement and to take care of instructing the participant. Another examiner (EX-3) monitored the whole experimental situation from an office located at the training centre. If the timing of an approaching behaviour did not work well, the examiners, especially EX-3, requested to repeat the cooperation. Both for the narrow-passage and the t-junction six behaviour scripts were available (table 1 and 2), which were divided into offensive and defensive scripts.

At the beginning of the experiments the subjects had drive a standardized course to become acquainted with the car. Then the eye-tracking system was calibrated. For the t-junction every position was driven two times and every script was repeated depending on the position of the subject (figure 1). The same procedure was applied for the narrow-passage, only differing in terms of the number of positions that had to be tested (figure 1).

Specific procedure for EX 1.1 and 1.2

The examiners had to work through six different driving scripts, which regulated how to behave at the narrow-passage or the t-intersection, respectively. After every script the examiners assessed the situation. With a dictaphone the examiners rated the following aspects on a seven-pointed Likert-scale: the implementation of the script, the cooperation-readiness of the subject, and the cooperation-intensity. Additionally they had to record the communication signs given by the subjects and the order of driving. After all situations EX-1.1, 1.2, and EX-2 had to answer an overall survey for all situations, including the overall cooperation-readiness, the overall cooperation-intensity, and the perceived driving style of the subject.

Data analysis

In a first step the eye-tracking data were analysed in a 2x6 repeated measurement ANOVA, with factor 1 regarding the passage and factor 2 regarding the scripts for the narrow-passage. For the t-junction the data was analysed in a 2x4 repeated measurement ANOVA.

In order to detect a systematic pattern, the significant parameters were then clustered by a fuzzy-method (Hatzinger, Hornik & Nagel, 2011). It was expected that some parameters behave in similar ways to each other during different behaviour scripts making it possible to distinguish between those scripts. The fuzzy-approach has the advantage that every variable is allocated with a specific weight to every cluster. Thus it is possible to estimate how strong the cluster and variables are.

The software R 3.4.1 and the package "cluster" (Maechler, 2017) was used. Additionally the question was addressed with which survey variables the parameters correlate. The research question follows an explorative approach.

For the eye-tracking parameters various variables, describing in detail, fixations, saccadic eye-movements, or blinks were analyzed (table 3):

Table 3. Table of eye-tracking parameters.

Parameter category	Parameter
Blink-Parameters (B)	Nearest Neighbor Index (Di Nocera et al. (2006) (NNI,1), Percentage closure of eyes (Lal & Craig, 2001) (PERCLOS,2), blink duration in ms (3) , blink rate (4)
Pupil Diameter Parameters (PD)	pupil-diameter (5), the mean pupil diameter (6), the median of the pupil diameter (7), the frequency of the pupils diameter (8), the amplitude of the pupil diameter (9), the highest peak of the pupils diameter (10)
Fixation-Parameters (FX)	Fixation-durations metrics (maximum, mean, median) (11,12,13), the sum of the fixations (14), the ratio between fixation and saccades (15)
Saccadic-Parameters (SC)	saccades-velocity metrics (mean, max, median) (16, 17, 18), saccades-amplitude-metrics (mean, maximum, median) (19, 20, 21), saccade-duration-metrics (mean, maximum, median) (22, 23, 24, 25), Saccade sum (26)

Results

The analysis of the data focuses on the results of the narrow-passage first, while the outcome for the t-junction is presented next.

To reduce the amount of data only the parameters significant in the ANOVA will be reported and summarised. The alpha error accumulation was considered. As the physiological data did not follow a normal distribution outliers ± 2 SDs were replaced by the mean value of each condition to prepare a box-cox transformation. Values ± 2 SDs were measurements of low quality (below 75 % of the quality-index). Then the parameters were box-cox transformed (Box & Cox, 1964), using the box-cox-function of the "MASS"-package (Ripley et al., 2017) of R-Studio.

Narrow-passage

The parameters of the narrow-passage with significant results in the ANOVA are reported in table 4. Only significant variables are reported.

Table 4. Significant parameters of the narrow-passage.

Category	Parameter	DF	SSn	F	p	Partial-eta-square
B	NNI	5	.476	2.674	<.05	.067
B	PERCLOS	5	2.401	10.002	<.05	.144
B	Blink-Duration	5	5.602	7.860	<.05	.129
B	Blink-Rate	5	49.264	3.857	<.05	.069
PD	Diameter-Frequency	5	141.528	2.604	<.05	.047
PD	Diameter-Peaks	5	6.414	2.347	<.05	.048
FX	Fixation-Duration-Mean	5	4.884	2.459	<.05	.039
FX	Fixation-Sum	5	.035	10.965	<.05	.234
FX	Fixation-Saccade-Ratio	5	1.378	2.377	<.05	.038
SC	Fixation-Saccade-ratio Sum	5	1.836	2.544	<.05	.051
SC	Saccade-Amplitude-Maximum	5	1.176	5.243	<.05	.089
SC	Saccade-Amplitude-Mean	5	.269	3.028	<.05	.014
SC	Saccade-Duration-Median	5	547.416	3.069	<.05	.051
SC	Saccade-Sum	5	.001	13.259	<.05	.235

Bortz (2005) categorized the effect of the partial-eta-square as low for <.1, as middle for >.25, and as high for > .4. Thus, only parameters with a high partial-eta-square> .040 were considered (table 4).

For determining the appropriate number of clusters, the “pamk” function of the package “fpc” was used (Henning, 2015), which refers to the theory of Duda and Hart (1973). The pam function searches for representative objects or medoids in the data set which minimize the sum of dissimilarities (Henning, 2015). As there are three different kinds of eye-metrics, fixations, saccadic eye-movements, and blinks, it is reasonable to assume that there are also three clusters. The results of the cluster analysis also show three clusters, two of which can be allocated to fixations (Cluster-3) and saccades (Cluster-2). Cluster-1 cannot be allocated to a parameter category. The Dunn coeffient is high with a value of .84.

Table 5. Fuzzy-analysis-membership-coefficients for the narrow-passage.

Category	Parameter	Mem	Membership coefficient 2	Membership coefficient 3
FX	NNI (1)	.99	.01	.00
B	Blinkduration (3)	.97	.02	.01
SC	Saccade-Amplitude-Maximum (20)	.96	.03	.01
B	Perclos (2)	.08	.86	.07
B	Blinkrate (3)	.15	.73	.12
SC	Saccade-Duration-Median (24)	.01	.89	.10
SC	Saccde Sum (26)	.01	.78	.21
FX	Fixation Sum (14)	.00	.00	1
FX	Fixation-Saccade Ratio (16)	0	0	1

For every cluster one typical plot is presented (figure 2-4): Cluster-1 summarises very different kinds of parameters and thus, it is difficult to be interpreted reasonably. As it reveals only weak effects, this cluster shall not be regarded further and it is referred to as “noise cluster” (figure 2).

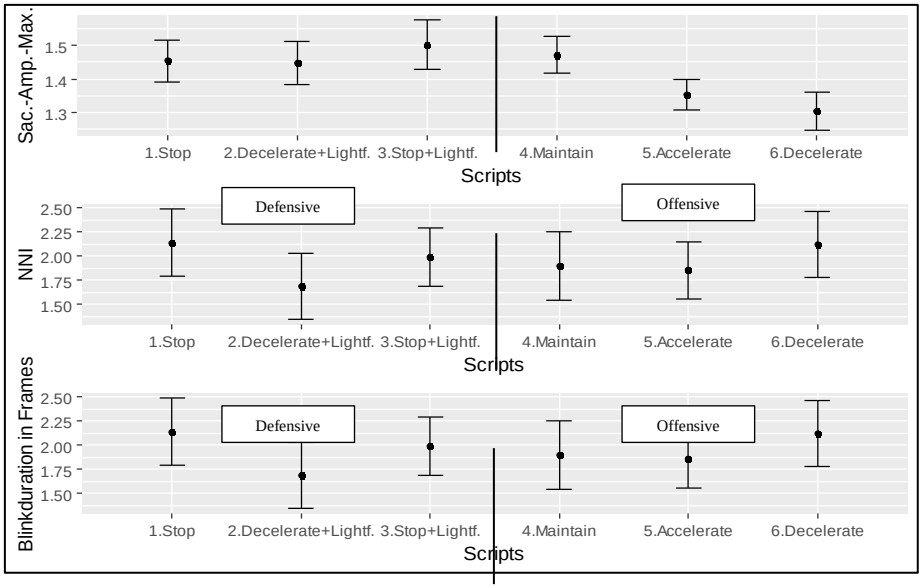


Figure 2. Mean-graph of the cluster-1 parameters for the narrow-passage.

Cluster-2 consists of saccadic parameters with middle effect sizes (figure 3), whereas cluster-3 summarises various fixation metrics with both strong and weak effects (figure 4).

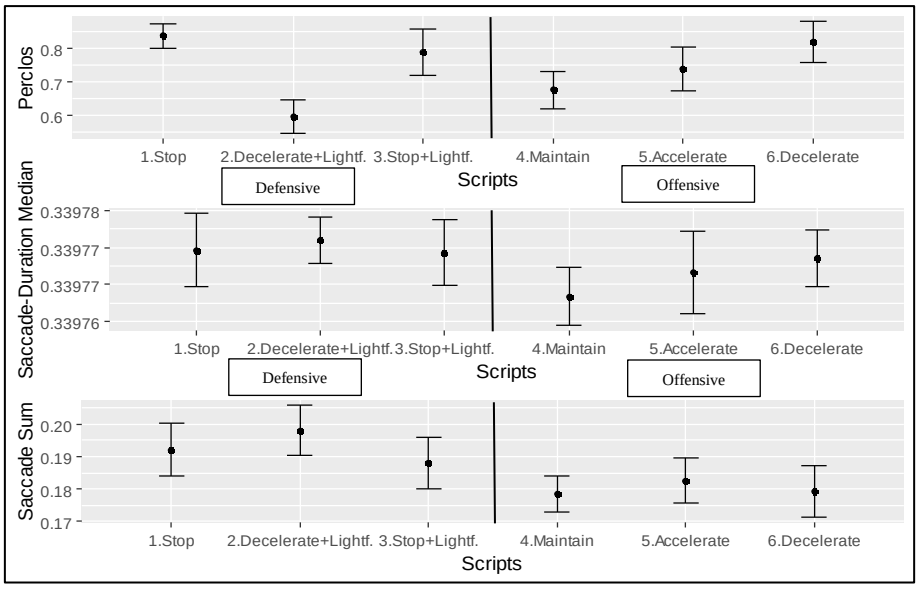


Figure 3. Mean-Graph of the Cluster-2 Parameters of the narrow-passage.

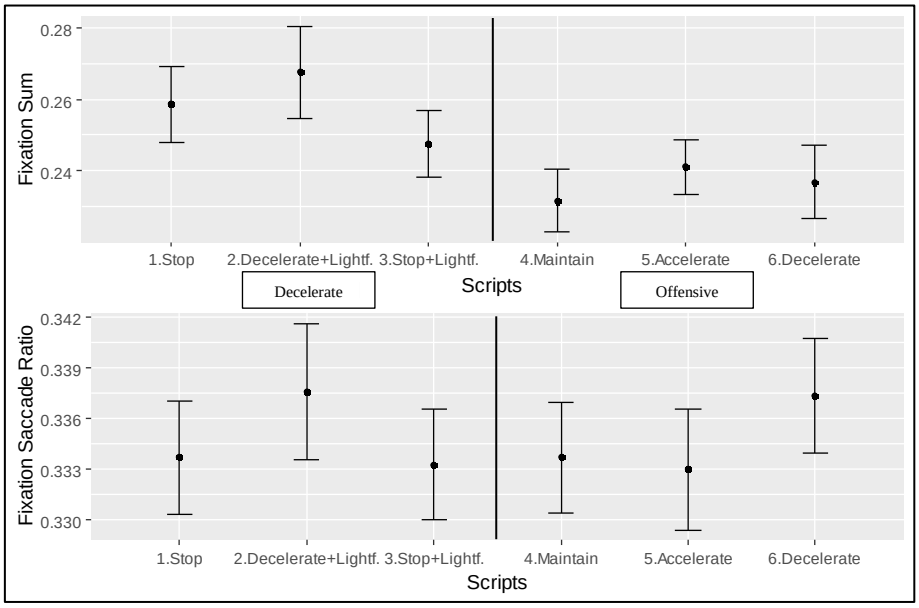


Figure 4. Mean-Graph of the Cluster-3 Parameters of the narrow-passage.

Finally, in order to get a complete picture for the narrow-passage-scenario, the significant eye-tracking parameters are correlated with the survey data (table 6).

Table 6. Correlation-Matrix: eye-tracking parameters and survey data for the narrow-passage.

Category	Parameter	Confidence of driving	Accidental risk	Cooperation-Reac	Cooperation-Inter
FX	NNI	-.140	-.128	.049	-.140
B	Perclos	.030	-.005	.138	-.130
B	Blinkduration	.072	.005	-.115	-.155
B	Blinkrate	.037	-.053	.067	.095
PD	Diameter-Freq	-.014	-.062	-.133*	-.005
PD	Diameter-Peaks	-.077	.146*	-.222**	-.109
FX	Fixation-Duration-Mean	.108	-.076	.002	-.125*
FX	Fixation-Sum	.059	-.089	.145*	.089
FX	Fixation-Saccade Ratio	-.640	.109	-.091	-.179**
FX	Fixation-Saccade ratio Sum	-.039	.074	-.016	-.006
SC	Saccade-Amplitude-Maximum	.103	-.157*	.005	.157*
SC	Saccade-Amplitude-Mean	-.120	.036	-.062	-.029
SC	Saccade-Duration-Median	.037	-.032	.030	-.229**
SC	Saccade-Sum	-.530	-.140	-.700	-.106

p <.000 ***, p < .001 **, p < .005*, p < .01.

It can be seen that there is no correlation between the variable “confidence of driving” and any other parameter, the variable “accidental risk” correlates with the Diameter-Peaks (r=.146, p< .05) and with the saccade amplitude-maximum (r= –.157, p< .05). The variable “cooperation-readiness” correlates with the Diameter-Frequency (r= –.133, p< .05), Diameter-Peaks (r= –222, p< .05), and Fixation-Sum (r= –.145, p< .05).

t-junction

The statistical procedure for the t-junction data was the same with the exception that only the left turning positions were analysed and script-5 and -6 will be ignored because they belong to position-3. Position-3 is the going straight position and requires different eye movements.

Again the analysis starts with the ANOVA. Only significant results are summarised in table 6.

Table 7. Significant eye-tracking parameters of the t-junction as given by an ANOVA.

Category	Parameter	DF	SSn	F	p	Partial-eta-square
FX	NNI (1)	3	2.836	2.929	.05	.005
FX	Fixation-Duration-Mean (12)	3	6.175	2.920	.05	.029
FX	Fixation-Sum (14)	3	1.608	3.532	.05	.038
FX	Fixation-Saccade Ratio (15)	3	.001	2.907	.05	.029
FX	Fixation-Saccade Ratio Sum (16)	3	.020	2.956	.05	.045
SC	Saccade-Amplitude-Maximum (20)	3	1.462	5.535	.05	.074
SC	Saccade-Amplitude-Mean (21)	3	.291	3.043	.05	.031
SC	Saccade-Duration Mean (22)	3	.002	1.716	.05	.056
SC	Saccade-Duration Median (23)	3	.006	3.275	.05	.043
SC	Saccade Velocity Maximun (17)	3	.002	5.489	.05	.066
SC	Saccade Velocity Median (18)	3	.009	2.893	.05	.020
SC	Saccade-Sum (26)	3	6.157	3.768	.05	.040

In the next step, the results of the cluster analysis will be presented. Thereby a model with two clusters was tested, which was based on the before mentioned “pam”-function (table 7).

Table 8. Fuzzy-Analysis-Membership-Coefficients for the t-junction.

Category	Number	Parameter	Membership Coefficient 1	Membership Coefficient 2
Saccadic	20	Saccade-Amplitude-Maximum	1.00	.00
Saccadic	16	Fixation-Saccade Ratio Sum	.29	.71
Saccadic	22	Saccade-Duration Mean	.00	1
Sacadic	23	Saccade-Duration Median	.01	.99
Saccadic	17	Saccade Velocity Maximum	.02	.098
Saccadic	26	Saccade-Sum	.00	1.00

The mean-graphs are presented for the two different clusters (fig. 4 and 5).

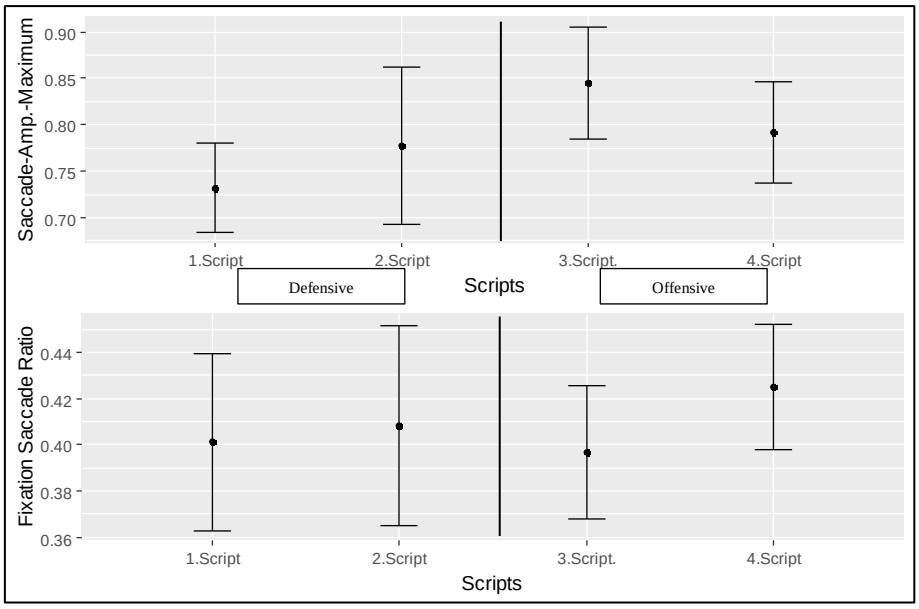


Figure 4. Mean-Graph of the Cluster-1 parameters of the t-junction.

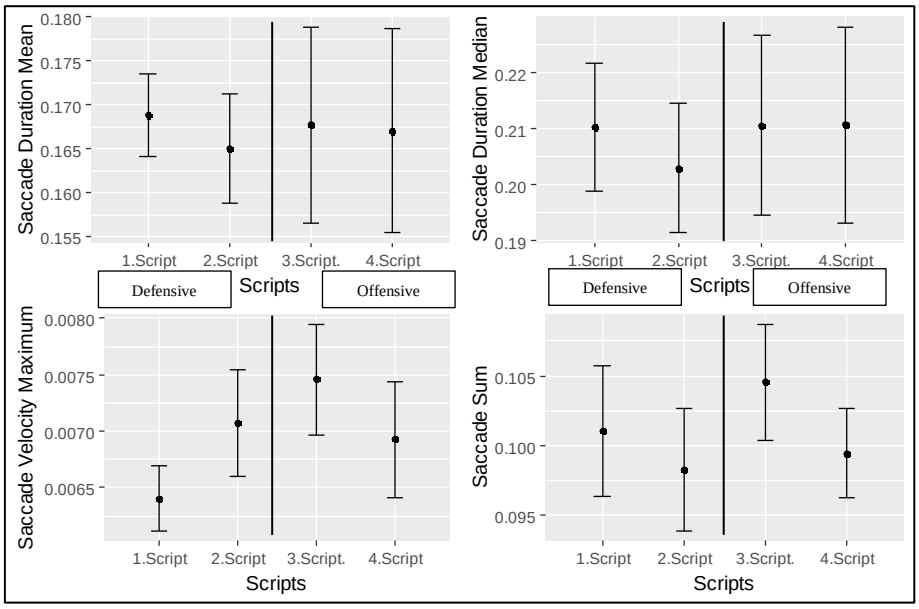


Figure 5. Mean-Graph of the Cluster-2 parameters of the t-junction.

Analogous to the analysis of the narrow-passage, finally, the result of the correlation analysis is presented (table 8).

Table 9. Correlation-Matrix: eye-tracking parameters and survey-Data.

Category	Parameter	Confidence	Accident risk	Cooperation-Readiness 1	Cooperation-Readiness 2	Cooperation-Readiness 3	Cooperation-Intensity
FX	NNI	.013	.031	-.152	.061	-.110	-.833**
FX	Fixation-Duration-Mean	.025	.000	-.88	.030	-.070	-.201**
FX	Fixation-Sum	-.125	-0.320	.027	.096	-.040	.091
FX	Fixation-Saccade Ratio	-.077	.051	.012	.248**	-.120	-.085
FX	Fixation-Saccade Ratio Sum	-.027	-.085	.087	.077	-.020	.154*
SC	Saccade-Amplitude-Maximum	-.155*	.083	.088	.184*	-.009	.085
SC	Saccade-Amplitude-Mean	.066	.100	.026	.173	-.207	.060
SC	Saccade-Duration Mean	.168*	.035	.053	.050	-.108	.032
SC	Saccade-Duration Median	.144*	.057	.105	.020	-.060	.085
SC	Saccade Velocity Maximun	.054	-.053	-.044	-.149	.057	.210
SC	Saccade Velocity Median	.044	.013	.100	.074	-.060	.052
SC	Saccade-Sum	.086	-.078	.125	.055	.037	.141*

p <.000 ***, p < .001 **, p < .005*, p < .01.

Discussion

First, the outcome of the narrow-passage and t-junction shall be discussed, then implications will be derived. Due to the large amount of significant results it may be assumed that the eye-tracking parameters seem to be quite meaningful in general. Based on the partial-eta-square value, nine of the 14 parameters could be considered for a further cluster analysis. The resulting clusters show that in cooperative traffic situations different aspects are relevant. In cluster-1 it becomes apparent that the NNI, maximum amplitude of a saccade, and blink duration reveal a similarity. In general, the values are lower for the offensive scripts than for the defensive scripts. This might be due to the fact that within the offensive scnearios the subjects had to initiate driving maneuvers themselves and not only react on the behaviour of the cooperation partner as it is the case within the more defensive maneuvers. Moreover the lowest manifestation occurs in script-5, which allows the clearest interpretation of the behaviour for all subjects. Cluster-1 is interesting because the parameters cannot be allocated to one parameter-category.

Cluster-2 consists of the variables PERCLOS, median saccade duration, and sum of the saccades. For this reason, it may be concluded that the saccadic parameters behave in a similiar way. The lowest manifestation is in script-4 (maintain speed), the highest manifestation in script-2 (Decelerate and using the flash of headlight). Furthermore the defensive scripts show higher values in general. It can be assumed that for the narrow-passage there are more and longer gaze movements within defensive scripts than within offensive scripts. For the parameter PERCLOS the opposite holds true – if the eye is closed, it will not be possible to use gaze. To sum up it can be said that cluster-2 is the saccade cluster.

Cluster-3 contains the parameters sum of fixations and the fixation-saccade-ratio. Both parameters are very similar to cluster-2 and show the same pattern: There is a lower manifestation in the offensive scripts and a higher manifestation in the defensive scripts. When a driver takes a more active role, both the sum and the duration of saccades and fixations are reduced; for a more passive driver the opposite holds true. Furthermore cluster-3 shows for the effect sizes one important point: The fixation-saccadio ratio (η^2p=.051) has in contrast to the sum of the fixation (η^2p=.234) a very weak effect size.

The correlations between the eye-tracking parameters and the assessment of the situation by the subjects show one interesting abnormality. Most of the eye-tracking parameters do not correlate with the subject's assessment except the variable cooperation-intensity. It correlates in a negative way with the mean of the fixation-duration, fixation-saccade ratio and the saccade-duration median and in a positive way with the saccade amplitude maximum. This means the more cooperative a situation is perceived, the fewer fixations and saccades are to be observed. Thus, a clearer communication leads to a faster understanding of the situation and in consequence, the drivers do not have to look longer and more intensively to several objects. This pattern is found in the saccadic parameters as well.

Within the t-junction-scenario a lot of significant results were observed, too; although only few of them reveal at least a middle sized effect. As most of the parameters show only a very small effect, the results should not be considered further and thus, only a two cluster analysis was reported. Cluster-1 contains the NNI and partially the fixation-saccade-ratio. For the offensive scripts corresponding patterns could be observed. The values are higher for script-3 than for script-4, but for the defensive scripts opposite holds true. Unfortunately, for cluster-1 it is not possible to draw conclusions.

Cluster-2 contains the mean and median of the saccade duration, maximum of the saccade velocity and sum of the saccades. The saccade duration parameters show similar patterns. Furthermore in script-3 and 4 they do not show any differences. In the offensive scripts the subjects have to react, to stop and not to drive first. So they can concentrate on what the other driver is doing. The maximum of the saccade velocity and the allocated sum show a similar pattern regarding differences between offensive and defensive scripts. Nevertheless, there is a small difference because in script-4 they show the same manifestation as the defensive scripts. This pattern should be tested for other parameters within future research.

Moreover the results of the correlation analysis show on one side the same effect as the narrow-passage regarding the cooperation-intensity. But there is a further correlation pattern. If the drivers feel more confident in making their decision when to drive, the saccade duration is higher. It seems like it is easier for road users to make a decision if they have more time to go from one fixation to the next. Correspondingly, there is a negative correlation between confidence of driving and maximum saccade amplitude which underlines the explanation. If the data of the narrow-passage is compared to the t-junction-scenario, it can be seen that in both it is possible - on the base of the eye-tracking data - to distinguish between situations in which the participant drives first or the cooperation partner. Furthermore, there is

a negative correlation between the cooperation-intensity and fixations duration and saccade duration. If a road user does not have to look long at an object the situation is clearer and so the perceived cooperation-readiness seem to be higher. Moreover the t-junction shows that there is a correlation between confidence of driving and saccade duration. It should be mentioned that an unsystematic behaviour of various variables is from the methodological point of view critical. The results should be seen as tendency and not as hypothesis-tested results.

The main differences between the narrow-passage and the t-junction are that in the t-junction-scenario the subjects had to turn left and concentrate on two opponents in contrast to the narrow-passage with two obstacles and a straight driving direction. Because of these characteristics the t-junction is more complex than the narrow-passage and takes longer overall to master. This leads to the results of the fixation parameters being more important for the narrow-passage than the saccadic parameters. For the t-junction the saccadic parameters are more important. Furthermore it is possible to see if a road user had to react or act based on the eyetracking data.

Conclusion

It can be said that it is possible to identify specific eye-tracking parameters for different driving behaviours. Furthermore it was shown that not every significant parameter is a "good" parameter because the effect sizes are spreading very widely.

Moreover it is possible to distinguish between simple and complex scenarios. For complex short cooperative traffic scenarios the saccadic parameters seem to be more important and for simple scenarios the fixation parameters seem to be more important. More studies should test this pattern.

Acknowledgments

This project has been funded within the priority program 1835 „Cooperatively Interacting Automo-biles "by the German Research Foundation (DFG). The authors thank the project partners for the fruitful cooperation. Furthermore thank the authors Simon Stache for helping with the clusteranalysis.

References

Ba, Y., Zhang, W., Reimer, B., Yang, Y., & Salvendy, G. (2015). The effect of communicational signals on drivers' subjective appraisal and visual attention during interactive driving sce-narios. *Behaviour & Information Technology*,*34*(11),1107–1118. https://doi.org/10.1080/0144929X.2015.1056547

Benmimoun A., Neunzig, D. & MAAG, C. (2004). *Effizienzsteigerung durch professionelles/partnerschaftliches Verhalten im Straßenverkehr (No. 181)*. Frankfurt/Main Forschung-dvereinigung Automobiltechnik e. V.

Borkenau, P, Ostendorf, P. (1993). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae.: Handanweisung*. Göttingen: Hogrefe.

Box, G.E.P & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B 26*(2). P. 211–252.

Di Nocera, F., Terenzi, M., & Camilli, M. (2006). Another look at scanpath: Distance to nearest neighbour as a measure of mental workload. In D. de Waard, K.A.Brookhuis, & A. Toffetti (Eds.), *Developments in human factors in transportation, design, and evaluation*, (pp. 295–303). Maastricht, the Netherlands: Shaker Publishing.

Duda, R.O. & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Henning, C (2015). *Flexible Procedures for Clustering. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/

Hoc, J.M. (2001). Towards a cognitive approach to human–machine cooperation in dynamic sit-uations. *International Journal of Human-Computer Studies, 54*(4), 509–540. https://doi.org/10.1006/ijhc.2000.0454

Imbsweiler, J., Ruesch, M., Palyafári, R., Deml, B. & Puente León, F. (2016). Entwicklung einer Beobachtungsmethode von Verhaltensströmen in kooperativen Situationen im innerstädti-schen Verkehr. In *32. VDI/VW-Gemeinschaftstagung Fahrerassistenzsysteme und auto-matisiertes Fahren*, Wolfsburg, 8-9 November 2016.

Imbsweiler, J., Palyafári, R., Puente León, F. & Deml, B. (2017). Untersuchung des Entscheidungsverhaltens in kooperativen Verkehrssituationen am Beispiel einer Engstelle. *at-Automatisierungstechnik, 65*, 477-488. https://doi.org/10.1515/auto-2016-0127.

Lal, S. K. & Craig, A. A. (2001). A critical review of the psychophysiology of driver‘s Fatigue, *Biological Physiology, 55*, 173-194.

Maechler, M. (2017). *Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/

Manzey, D., & Lorenz, B. (1998). Mental performance during short-term and long-term spaceflight. *Brain Research Reviews, 28*(1-2), 215–221. https://doi.org/10.1016/S0165-0173(98)00041-1

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A. & Firth, D. (2017). *Support Functions and Datasets for Venables and Ripley's MASS. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/

Risser, R. (1985). Behaviour in traffic conflict situations. *Accident, Analysis & Prevention, 2*(17), 179–197.

Schneider, M. (2017). *Blickbasierte Beanspruchungsmessung : Entwicklung und Evaluation eines Kalibrierungssystems zur individuellen Bewertung der mentalen Beanspruchung in der Mensch-Technik-Interaktion*. PhD dissertation, Karlruher Institut für Technologie, urn:nbn:de:swb:90-700251

Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., & Ikeuchi, K. (Eds.) (2004). Flexible cooperation between human and robot by interpreting human intention from gaze information -, Proceedings. 2004 IEEE/RSJ International Confer. : Vol. 4: IEEE.

Witzlack, C., Beggiato, M., & Krems, J. (2016). Interaktionssequenzen zwischen Fahrzeugen und Fußgängernim Parkplatzszenario als Grundlage für kooperativ interagierende Automati-sierung. *In VDI (Eds.). Fahrerassistenz und automatisiertes Fahren, VDI-Berichte 2288* (p. 323-336). Düsseldorf: VDI-Verlag.

Modelling driver styles based on driving data

Peter Mörtl, Andreas Festl, Peter Wimmer, Christian Kaiser, & Alexander Stocker
VIRTUAL VEHICLE Research Center
Austria

Abstract

Driving styles are habitual ways of driving that are characteristic for groups of drivers and represent an important topic of research for advanced automation in the vehicle of the future. Relatively little knowledge exists concerning the connection between driving styles and the underlying cognitive-psychological aspects of the driver. To better understand this connection, we investigate driving style indicators in a driving simulator and create a cognitive model of the underlying cognitive-psychological processes that we compare with the empirical data. The cognitive model produces steering behaviour that approximates the lateral deviations of human drivers while also producing similar rates of steering wheel direction reversals. These results confirm the utility of this approach for representing individual driving styles and states for advanced vehicle automation.

Introduction

Research in driving styles has a long tradition and recently experienced a new focus of interest because modern automotive technologies have become smart enough for personalised in-vehicle interventions. The term driving style has been defined by (Sagberg, Selpi, Piccinini, & Engström, 2015) as a *"habitual way of driving, which is characteristic for a driver or a group of drivers".* Driving styles also stay constant for a given driver across different driving contexts. Global driving styles combine multiple driving indicators (such as aggressive, calm, or careful driving) and specific driving styles are measured by one or two indicators.

While there are many driving metrics, there are currently few models that tie them to their underlying psychological processes. Such models would be important because of several reasons. First, they may allow tying multi-sensory observational data streams together to inform driver state inferences. Second, such integrative models could be used in virtual safety assessments to represent the human driver. Such virtual safety assessments are needed to determine whether higher levels of automated driving are safe because repeated extensive real-world studies would take too long and be too cost-intensive. And third, such models could be used to better personalise automated driving so that the self-driving car has knowledge of the human driver and adjust to his or her style. Also, personalised in-vehicle support could detect a need for intervention based on observing the driving behaviour for fatigue or distraction. The detection of such driver states depends on representing the driver's individual driving style which is the focus of this paper.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In this paper we first briefly describe some driving style indicators. Then we describe an empirical study to collect driving data in a driving simulator and investigate four driving style metrics: steering, speed, acceleration, and distance keeping. We then describe the development of a psychological driving process model to predict lateral lane deviations and compare them to human participants. We validate the model for one of the four driving metrics, the steering metric, and describe the results.

Specific driving styles

Steering behaviour represents a thoroughly investigated driving style metric. Li et al. (2017) use measures of entropy and variation in timely sequences of steering wheel angles to estimate driver drowsiness. Fairclough & Graham (1999) measured steering wheel reversal (SWR) differences between a control group and partially and fully sleep-deprived drivers in a simulator study. They find a reduction of SWR from about 15 to 11 per minute for the sleep-deprived group compared to the control group (see also McLean & Hoffmann, 1975). Thiffault & Bergeron (2003) examined the influence of fatigue on various steering measures including the mean steering angle amplitude, frequency of larger steering wheel movements, and their standard deviation. Otmani et al. (2005) investigated the influence of fatigue on the mean steer wheel angle changes and found that they amount to about between 0.5 and 5 degrees, averaged over 1 to 10 min driving periods. Similarly, Yan, Radwan, & Guo (2007)) and Ungoren & Peng (2005) report that individual drivers differ in their steering behaviour.

Another well investigated specific driving style metric are driving speed and accelerations. Ericsson (2000) reports on traffic dependent as well as individual differences in driving speed, for example between males and females (see also Brundell-Freij & Ericsson, 2005; Ericsson, 2001 who also investigated driving acceleration). Af Wåhlberg (2007) investigated the variability and amount of driving accelerations and found that they could serve as precursors of accidents (see also Af Wåhlberg, 2008; af Wåhlberg & Dorn, 2007). Bagdadi & Várhelyi (2011) investigated jerky driving as predictor of accidents, see also (Murphey, Milton, & Kiliaris, 2009). Desai & Haque (2006) investigated pressure on the acceleration pedal as individual parameters for driver alertness.

Another field of specific driving style metrics is the headway distance to the vehicle ahead (see, e.g. Shinar & Schechtman, 2002). Taieb-Maimon (2007) and Taieb-Maimon & Shinar (2001) investigated the impact of training on improving inter-vehicular distance.

Data Collection

We collected driving data in a non-motion based driving simulator. After an initial warm-up, 16 participants drove a car with automatic transmission on a curvy road of 10 km length. The road segment was extracted from satellite imagery recreated in the simulator and eight metres wide. Participants were between 20 to 60 years old, 12 were male and four female. All had driver licences and drove between 5,000 and 20,000 km per year. Participants completed two scenarios on two different roads.

In the first scenario ("driving and passing scenario"), they could select their driving speed and encountered oncoming traffic approximately every 25 seconds. They encountered slower vehicles that they could pass if desired. Participants were encouraged to drive as close as possible to how they drove in the real world. When passing a vehicle they were asked to indicate their intent to pass by actuating a lever on the left side of the steering wheel. In a second scenario ("steering only scenario"), participants only steered their vehicle that drove at a constant speed of 90 km/h. There were no opportunities to pass other vehicles. The order of the scenarios was randomized and counterbalanced so that 50 % of the participants experienced scenario 1 before scenario 2 and vice versa to avoid order effects.

Before the simulation started, participants completed a questionnaire (see Table 1) that assessed some aspects of their general driving style. After completion of both driving scenarios they completed a short questionnaire about their driving. The response scale to all questions was a 7-item Likert scale.

Table 1. Questionnaire items

General Driving Style	**Scenario Specific Driving**
I frequently pass cars.	I frequently passed cars.
I usually do not have to brake prior to curves.	I usually did not have to brake prior to curves.
I sometimes cross red lights.	I think that I drove safely.
People say that I am a safe driver.	I think that I drove very carefully.
I usually drive very carefully.	I drove as fast as was possible.
I sometimes drive as fast as possible.	I was braking hard at least once.
I never "chase" yellow lights.	I drove "sportily".
I often have to brake hard.	
When driving, cars often pass me.	
People say that I am a "sporty" driver.	

Driving style characterisation

We first present the results of the observed driving style metrics for scenario 1 where participants could freely choose their speed. Driving data during periods of free driving (i.e. without a vehicle ahead) were separately analysed from periods when they had to adjust their speed because of a vehicle ahead. The road of 10,000 m was divided into 100 segments that were each 100 m long. The first segment was removed because all drivers accelerated the vehicle. Each of the remaining segments was classified for each participant either as following another car (if it came within 50 m of the vehicle ahead), as driving freely, or overtaking a car. Only free and following driving segments were considered further in the analysis. Following driving metrics were investigated:

1. Count of the steering wheel direction reversals per second over “free” and “following” segments.
2. Mean vehicle speed in the “free” driving segments.
3. Mean following distance in the “following” driving segments.
4. Count of accelerations in all “free” and “following” segments.

Figure 1 shows the identified variability of the four driving metrics that differed considerably between them.

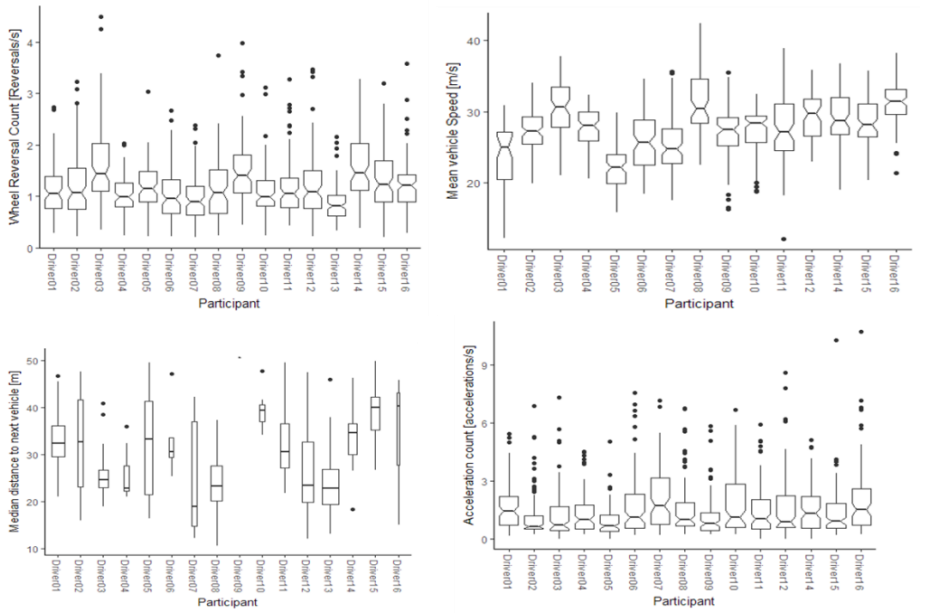


Figure 1. Driving metrics for Study 1.

Several of the metrics correlate with each other: The faster drivers accelerated more frequently, however only when driving freely (i.e. $r=0.23$, $p < 0.001$ when driving freely but only $r=0.09$, $p > 0.1$ when following). Drivers who drove faster also more frequently reversed the steering wheel direction ($r=0.21$, $p < 0.001$) regardless whether they followed or drove freely. When following a vehicle, people who kept more distance to the vehicle tended to reverse the steering wheel direction more frequently ($r = 0.18$, $p < 0.001$) and tended to drive faster ($r=0.29$, $p < 0.001$). No other correlation among the driving metrics reached significance. Also, we found no correlation between driving metrics and the participants’ responses on the questionnaires.

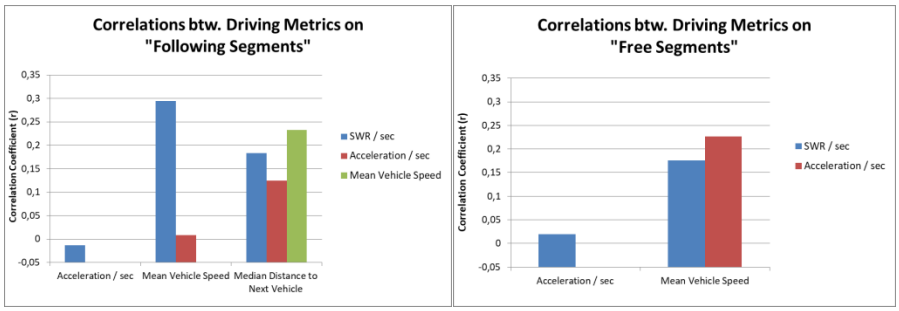


Figure 2. Correlations between driving metrics.

To determine to what extent these variations reflect just random noise or differed statistically between drivers and road segments, we utilised a two-way random effects model (Searle, Casella, & McCulloch, 1992), which is a special case of the linear mixed effects model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

Formula 1. Two-Way Random Effects Model.

where the indices $_i$ and $_j$ indicate the drivers and road segments and the variables are defined in the following way:

- Y_{ij} is the response variable, i.e. the specific driving metric.
- μ is the overall mean of the driving metric.
- α_i is the random effect of the driver which is normally distributed around 0 with a variance of σ_α^2 i.e. $\alpha_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$.
- β_j is the random effect of the road segment which is normally distributed around 0 with a variance of σ_β^2 i.e. $\beta_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_\beta^2)$.
- ϵ_{ij} is the random noise that cannot be attributed to the driver or road section. The variation of possible other factors of influence are also captured in this term.

No interaction $(\alpha\beta)_{ij}$ is included because each driver drove each road segment only once. That means we had to assume that the drivers responded equally to each road segment.

The results of this analysis are shown in Table 2. The variances of all random effects are highly significant, indicating statistically significant differences in the driving metrics among the drivers and the road segments. Knowing the driver and the road segment reduces the overall random noise as indicated Table 2. For example, the

variance for the driving metric “median distance to the next vehicle” reduces from 15.06 m to 6.6 m when subtracting driver and road segment effects.

Table 2. Driving style metric variability

	Median Distance to Next Vehicle (m)	Mean Vehicle Speed / sec	Steering Wheel Reversals / sec	Acceleration Count
Driver	5.5510*	2.4920*	0.2033*	0.3270*
Road Section	2.8630**	2.4700*	0.2126*	0.4254*
Residual	6.6450	2.7960	0.5173	1.3250
Total	15.0590	7.7580	0.9332	2.0774

* $p < 0.001$; ** $p < 0.01$

In Figure 3 the proportions of the explained variances are compared with each other. We find that the greatest amount of overall variation is explained by the driver’s median distance to the next vehicle (36.86%). This is somewhat expected because distance keeping should more or less reflect a conscious decision by the driver. Road characteristics also contribute significant variability though the driver influence is almost double of the road segments. Road and drivers explain about equal variances of the mean vehicle speed, similar is the case for SWRs. In terms of number of accelerations, knowing the road segment has a bigger impact on reducing the overall variance than knowing the driver.

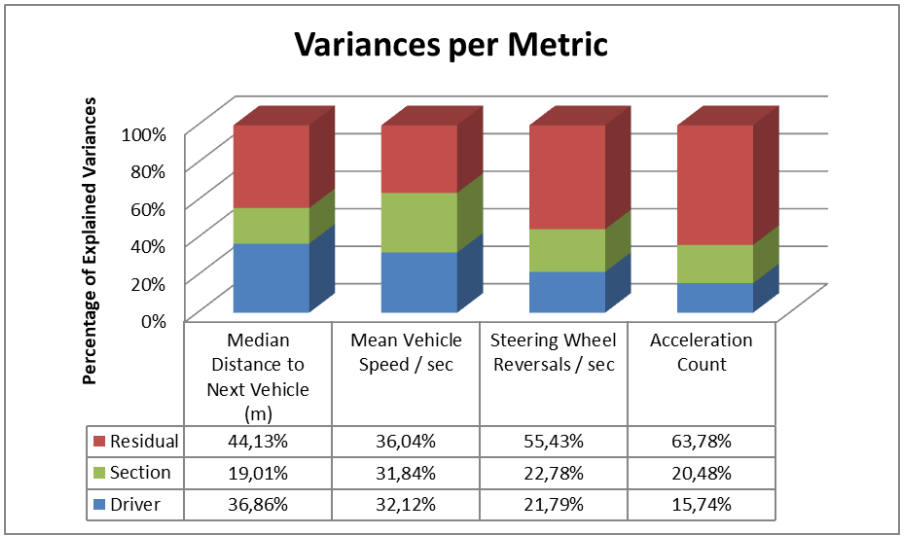


Figure 3. Relative contribution of the variances within each driving metric

Cognitive Modelling Architecture

There are many cognitive modelling approaches that have been applied to driving (e.g. Anderson et al., 2004; Bubb, Bengler, Grünen, & Vollrath, 2015; Deml, Neumann, Müller, & Wünsche, 2008; Kieras & Meyer, 1997; Laird, Newell, & Rosenbloom, 1987; Lewandowsky & Farrell, 2011; Liu, Feyen, & Tsimhoni, 2006; Salvucci, 2006) and it is beyond the scope of this paper to provide an overview. Primarily we were searching for a general purpose, modular architecture that would help to represent human psychological processes for engineering tasks. For this we came to utilise a cognitive modelling architecture that we describe in more detail in Moertl, Wimmer, & Rudigier (2017). In this approach we adopted the basic elements of the human cognitive architecture by Card, Moran, & Newell (1986), the Model Human Processor (MHP) and adapted it to specific driving tasks, see figure 4.

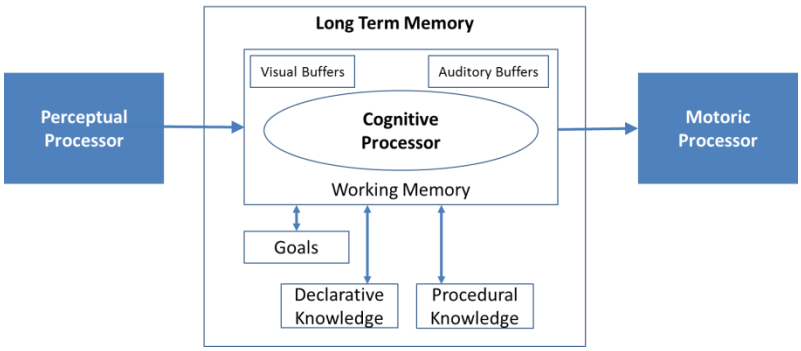


Figure 4. Cognitive modelling architecture.

In the context of the driving task, the MHP cognitive processor receives information from the perceptual processor about the external world and executes motoric tasks to control the vehicle. Each process takes a certain execution time and, dependent on serial or parallel processing, determines the overall task duration and timing of interactions. The central cognitive task is to determine the next driving action. In the model that we utilised in this study all tasks were executed serially but could also be executed in parallel. The transfer of information into memory and from memory involves buffers. However, in the simple model that we utilise in this study all the information was visually available in the environment and therefore did not require the use of explicit buffers.

Steering Modelling Method

With this model we implemented Salvucchi's steering model (see Salvucci, 2006; Salvucci & Gray, 2004) where steering is a direct result of perceptual fixations of certain points on the road ahead and the amount of time between subsequent environmental scans. The amount of time between scans is directly proportional to the number of steering corrections and inversely proportional to their size. Therefore, the more time is dedicated to steering and perceiving, the higher the number of control actions and the smaller their size. On the other hand, if only

limited time is available for perceiving and steering such as when multitasking and searching for signs or talking on the cell phone, the less frequent should be the steering control actions and the larger their size.

The model that we implement based on Salvucchi's work uses perceptual information that is available to human drivers when driving. The model has received empirical validation from visual occlusion experiments where human drivers drove in a driving simulator while some areas of the visual scenery were obscured (Land & Horwood, 1995; Land & Lee, 1994). This steering model utilises both a far point and a near point that are both ahead of the driver's own vehicle: the near point is a constant distance ahead and is located in the middle of the driving lane. The far point is further ahead and consists alternatively of the tangent point of an upcoming curve, the vanishing line of a straight road segment, or a vehicle that is driving ahead. The far point is intended to allow steering the vehicle into and out of curves whereas the near point helps to centre the vehicle on the driven lane.

We implemented this steering model in our cognitive architecture by only considering three cognitive processes that are executed in turn: a perception, a cognitive, and a motoric process. Each process was assumed to take 50 ms see e.g. (Card et al., 1986), so that one full cycle of steering updates takes 150 ms. We also updated the main parameters of the model: Whereas Salvucci (2006) suggested the three model parameters as k_{far} = 16, k_{near} = 4.0, and k_l was 3.0 we adjusted them to k_{far} = 1.6, k_{near} = 0.4, and k_l =0.09 to achieve better performance.

Results

Comparison lateral performance

We compared the steering quality of human participants with the psychological driver model in scenario 2. The red solid line in Figure 5 shows the lateral deviations of the model versus the 90 percentile of human steering over the whole 10 km. The model was 79.4 % of the time within the human driving boundaries and correlated on average with r = 0.36 with the human drivers.

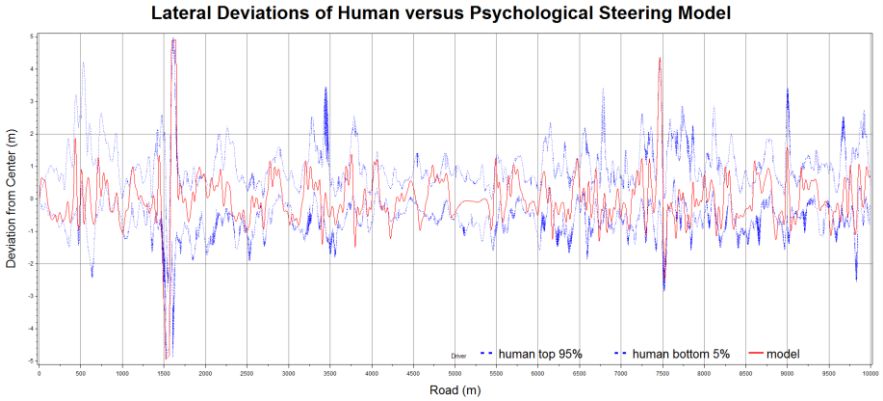


Figure 5. Comparison of lateral deviation between humans and psychological steering model.

Comparison steering wheel reversals

Given that the psychological model steered similar to human drivers, we compared the number of SWRs between humans and the model. It is important to note that we did not specifically attempt to fit SWR: we only fitted the model to the lateral lane deviations by adjusting the above mentioned three model parameters. Figure 6 shows the results for both scenarios 1 and 2. The bars indicate the mean number of steering wheel reversals per minute for our 16 participants. The blue line indicates the steering wheel reversals of our psychological driving models. On the left, drivers controlled speed and were passing other cars (scenario 1) whereas in the right figure, they only steered the vehicle (scenario 2). The roads were different between scenarios 1 and 2.

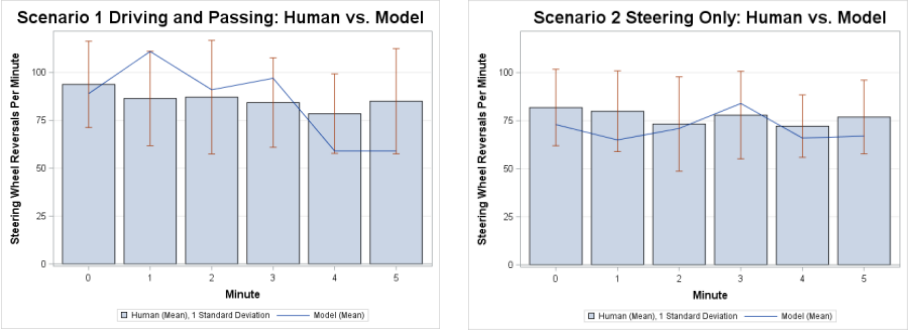


Figure 6. Comparison of steering wheel reversals between humans and model in two scenarios.

The results indicate that the psychological model produced very similar amounts of SWR to our human participants in both scenarios. All their counts fell within one standard deviation of the human participants. The model was similarly accurate in its steering wheel reversals in both scenarios as it produced, similar to human drivers, relatively more SWRs on scenario 1 (84.3 SWR/min for the model versus 85.7 SWR /min for the human) than on scenario 2 (71.0 SWR/min for the model versus 77.0 SWR/min for the human drivers). This indicates that by modelling the underlying cognitive processes the model was able to match human performance on two different dimensions despite having been fit only on lateral lane deviations.

Conclusions

The results of our study indicate that knowledge of the individual driver and road segment could explain up to 64% of the overall variability of the investigated driving metrics. The remainder of the variance apparently represents random fluctuations that would exceed driver modelling. This knowledge helps bound the expectations of how well driving models can approach real human driving.

After review of relevant literature we derived a relatively simple and modular cognitive modelling architecture in which we implemented as first step a psychologically plausible steering algorithm by Salvucci (2006). With some

moderate amount of fitting that basically consisted of adjusting the model's three critical steering parameters, the model not only resembled the lateral driving deviations of our 16 human participants but also approximated their rate of steering wheel direction reversals. This demonstrates the principal benefit of models that not only represent outcomes but underlying structure for the applications in the vehicle of the future: new and valid behavioural predictions can be derived from the model structure rather than having to base each prediction on an extensive learning process of stimulus-response. Such power of generalization is essentially missing in pure machine learning algorithms but seems crucial to better fit the contextual needs of the driver and help allow for inferential processes to ascertain whether a system is safe.

Much remains to be done to establish psychological driver modelling as a standard tool for human-centred automotive developments. First we will need to confirm that our psychological models are not only valid for simulation studies but also for real world driving. Then we need to test how the psychological model can be adapted to capture individual driver styles and states, such as, for example, driving distraction. Finally, we will extend our modelling to other driving aspects, specifically braking, distance keeping, and speed selections.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732189. The paper expresses the views of the authors.

References

Af Wåhlberg, A. E. (2007). Aggregation of driver celeration behavior data: Effects on stability and accident prediction. *Safety Science*, *45*, 487–500. https://doi.org/10.1016/j.ssci.2006.07.008

Af Wåhlberg, A. E. (2008). Driver celeration behaviour and accidents – an analysis. *Theoretical Issues in Ergonomics Science*, *9*, 383–403. https://doi.org/10.1080/14639220701596722

Af Wåhlberg, A. E., & Dorn, L. (2007). Culpable versus non-culpable traffic accidents; What is wrong with this picture? *Journal of Safety Research*, *38*, 453–459. https://doi.org/10.1016/j.jsr.2007.01.013

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*, 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Bagdadi, O., & Várhelyi, A. (2011). Jerky driving—An indicator of accident proneness? *Accident Analysis & Prevention*, *43*, 1359–1363. https://doi.org/10.1016/j.aap.2011.02.009

Brundell-Freij, K., & Ericsson, E. (2005). Influence of street characteristics, driver category and car performance on urban driving patterns. *Transportation Research Part D: Transport and Environment*, *10*, 213–229. https://doi.org/10.1016/j.trd.2005.01.001

Bubb, H., Bengler, K., Grünen, R. E., & Vollrath, M. (2015). *Automobilergonomie*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-8348-2297-0

Card, S. K., Moran, T. P., & Newell, A. (1986). *The Model Human Processor: An Engineering Model of Human Performance* (No. UIR-R-1986-05).

Deml, B., Neumann, H., Müller, A., & Wünsche, H. J. (2008). Fahrermodellierung im Kontext kognitiver Automobile Driver Modelling within the Context of Cognitive Automobiles. *At - Automatisierungstechnik*, *56*(11). https://doi.org/10.1524/auto.2008.0735

Desai, A. V., & Haque, M. A. (2006). Vigilance monitoring for operator safety: A simulation study on highway driving. *Journal of Safety Research*, *37*, 139–147. https://doi.org/10.1016/j.jsr.2005.11.003

Ericsson, E. (2000). Variability in urban driving patterns. *Transportation Research Part D*, *5*, 337–354.

Ericsson, E. (2001). Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. *Transportation Research Part D: Transport and Environment*, *6*, 325–345.

Fairclough, S. H., & Graham, R. (1999). Impairment of driving performance caused by sleep deprivation or alcohol: a comparative study. *Human Factors*, *41*, 118–128.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, *12*, 391–438.

Laird, J., Newell, A., & Rosenbloom, P. (1987). *SOAR: An Architecture for General Intelligence* (No. AJP-9). Carnegie Mellon University.

Land, M., & Horwood, J. (1995). Which parts of the road guide steering? *Nature*, *377*, 339–340.

Land, M., & Lee, D. (1994). Where we look when we steer. *Nature*, *369*(6483), 742–744. https://doi.org/10.1038/369742a0

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: principles and practice*. Los Angeles: Sage.

Li, Z., Li, S., Li, R., Cheng, B., & Shi, J. (2017). Online Detection of Driver Fatigue Using Steering Wheel Angles for Real Driving Conditions. *Sensors*, *17*, 495. https://doi.org/10.3390/s17030495

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP): A computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *13*, 37–70.

McLean, J. R., & Hoffmann, E. R. (1975). Steering reversals as a measure of driver performance and steering task difficulty. *Human Factors*, *17*, 248–256.

Moertl, P., Wimmer, P., & Rudigier, M. (2017). Praktikable Fahrermodelle mit psychologisch fundierten Prozessannahmen. In *Submitted to Conference Proceedings of the 9e VDI Tagung: Der Fahrer im 21en Jahrhundert*. Braunschweig.

Murphey, Y. L., Milton, R., & Kiliaris, L. (2009). Driver's style classification using jerk analysis. In *Computational Intelligence in Vehicles and Vehicular Systems, 2009. CIVVS'09. IEEE Workshop on* (pp. 23–28). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/4938719/

Otmani, S., Pebayle, T., Roge, J., & Muzet, A. (2005). Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers. *Physiology & Behavior*, *84*, 715–724. https://doi.org/10.1016/j.physbeh.2005.02.021

Sagberg, F., Selpi, Piccinini, G. F. B., & Engström, J. (2015). A review of research on driving styles and road safety. *Human Factors*, *57*, 1248–1275.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*, 362–380.

Salvucci, D. D., & Gray, R. (2004). A Two-Point Visual Control Model of Steering. *Perception*, *33*, 1233–1248. https://doi.org/10.1068/p5343

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). Fixed or Random. In *Variance Components*. New York: Wiley.

Taieb-Maimon, M. (2007). Learning Headway Estimation in Driving. *Human Factors*, *49*, 734–744. https://doi.org/10.1518/001872007X215809

Taieb-Maimon, M., & Shinar, D. (2001). Minimum and comfortable driving headways: Reality versus perception. *Human Factors*, *43*, 159–172.

Thiffault, P., & Bergeron, J. (2003). Monotony of road environment and driver fatigue: a simulator study. *Accident Analysis & Prevention*, *35*, 381–391.

Ungoren, A., & Peng, H. (2005). An adaptive lateral preview driver model. *Vehicle System Dynamics*, *43*, 245–259. https://doi.org/10.1080/00423110412331290419

Yan, X., Radwan, E., & Guo, D. (2007). Effects of major-road vehicle speed and driver age and gender on left-turn gap acceptance. *Accident Analysis & Prevention*, *39*, 843–852. https://doi.org/10.1016/j.aap.2006.12.006.

Graded auditory feedback based on headway: An on-road pilot study

Pavlo Bazilinskyy[1], Jork Stapel[1], Coert de Koning[1], Hidde Lingmont[1], Tjebbe de Lint[1], Twan van der Sijs[1], Florian van den Ouden[1], Frank Anema[2], & Joost de Winter[1]
[1]Delft University of Technology, The Netherlands,
[2]SD-Insights, The Netherlands

Abstract

Auditory feedback produced by driver assistance systems can benefit safety. However, auditory feedback is often regarded as annoying, which may result in disuse of the system. An auditory headway feedback system was designed with the aim to improve user acceptance and driving safety. The algorithm used a graded approach, which means that it delivered a more urgent warning if the time headway was smaller. In an on-road test, we compared this design with a conventional binary headway warning system. Participants drove a test vehicle on the highway, once with our graded feedback and once with conventional feedback. User acceptance was assessed through a questionnaire and interview. An inspection of the time headway distributions suggested that participants responded to the auditory feedback for both systems. There were substantial individual differences in time headway, and extremely short headways were rare. These findings suggest that long-term naturalistic trials are needed to assess the safety-effectiveness of graded auditory feedback.

Introduction

Car driving is safer than ever before (Stipdonk, 2017). The growing number of advanced driver assistance systems (ADAS), such as forward collision warning systems (FCW), may contribute to a further reduction of accidents. Auditory feedback is an attractive modality for in-vehicle warning systems because auditory feedback interferes little with the visually demanding driving task and can convey informative messages with different levels of urgency (Bazilinskyy & De Winter, 2015; Stanton & Edworthy, 1999).

ADAS often employ auditory feedback. Typically, the momentary safety margin (e.g., time to collision [TTC] or time headway [THW]) is used as an index to determine when feedback should be provided to the driver. A disadvantage of such discrete auditory warnings is that they may annoy the driver due to their saliency, repetitiveness, or binary nature without a clear indication of the reason for issuing feedback (Gonzalez et al., 2012; Parasuraman et al., 1997).

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

There is a balance between delivering feedback and maintaining user acceptance: if the decision threshold (criterion) is set so that auditory warnings are provided late, the warnings may be ineffective because the driver is caught by surprise or has little time to respond. Conversely, if the decision threshold is set so that warnings are provided early, the driver may become annoyed by the frequent warnings, and he/she may ignore or disable the warning system (Parasuraman & Riley, 1997). According to Sarter (2005), graded notifications, defined as "notifications that consist of signals that are proportional to the degree of urgency" are a promising yet underutilized means of supporting operators. Indeed, auditory warnings are sometimes not well-accepted (Parasuraman & Riley, 1997; Parasuraman et al., 1997; Wiese & Lee, 2004).

Several approaches exist to improve the acceptance of warning systems. One strategy is to provide individualization through adaptable or adaptive settings based on the driver's behaviour and driving style (e.g., Wang et al., 2013). Although this may improve acceptance, varying thresholds may also be a source of confusion for the driver. Providing the driver with information about why the warning is given, or providing clues that allow the driver to resolve the situation before the warning is triggered may also benefit acceptance.

The Dutch Institute for Road Safety Research (SWOV) sees any headway under 2.0 s as unsafe (SWOV, 2012), whereas the National Highway Traffic Safety Administration (NHTSA) reports that headways under 1.2 s are unsafe (NHTSA, 2004). In practice, however, drivers may adopt considerably shorter headways: highway observations showed that many drivers adopt a THW below 1 s (Hoogendoorn & Botma, 1997; Brackstone & McDonald, 2007; Treiber et al., 2006). An increase of minimal headway may improve safety (Ohta, 1993; Saffarian et al., 2017), whereas a reduction of variance of headway stabilizes traffic flow on the highway (Xie et al., 2008; Ye & Zhang, 2009).

In this study, we designed a new type of auditory feedback system and compared it to a conventional system. We propose auditory feedback that becomes more urgent (and therefore having a higher potential for annoyance) when the level of risk (operationalized in terms of three THW stages) is higher. An on-road measurement was conducted to pilot-test whether the system worked as it should, and how drivers responded to it.

Methods

Auditory feedback design: Survey study

As a first step to develop a feedback system based on headway, we performed an online survey among the student community and family members (N = 69). This survey compared user preferences for different types of earcons informing about the time headway. The sample consisted of 50 males and 18 females (one person preferred not to specify their gender). They had a mean age of 26.2 years (SD = 11.8).

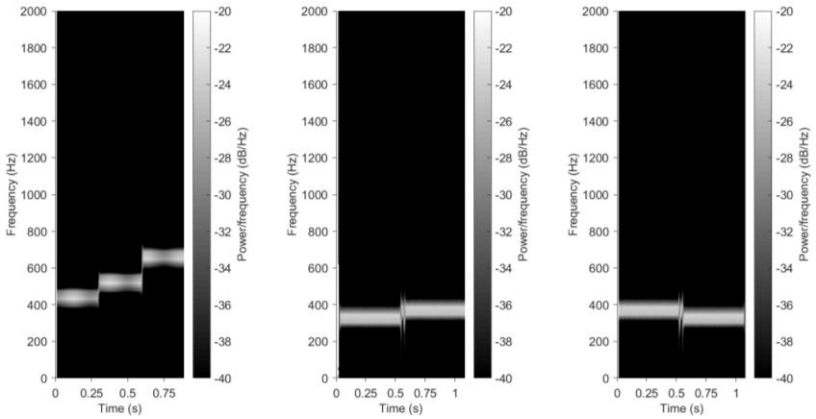


Figure 1. Spectrograms of Earcon 1 (left), Earcon 2 (centre), and Earcon 3 (right) from the online survey.

The respondents were asked to select the most suitable earcons for warning that the distance to lead vehicle is too short. They did this by ranking three selected earcons according to their preference. The earcons consisted of melodies that were assumed to be non-annoying. Figure 1 shows the spectrograms of the three sounds. The sounds were provided in three short clips (Figure 2). Each clip showed the same dash camera clip in which the driver was approaching another car. All earcons stood out from the highway traffic noise. The first earcon (Earcon 1) was a 900 ms three-note climbing tune (440 Hz, 523 Hz, 659 Hz), each note lasting 300 ms. The second and third earcons consisted of a two-note melody (330 Hz, 370 Hz), both tones lasting 500 ms. In Earcon 2 the lower frequency tone was presented first, followed by the higher frequency tone. In Earcon 3, the higher frequency tone was presented first, followed by the lower frequency tone. The earcon was provided when the THW in the video was approximately 0.5 s.

Figure 2. Video used in the online survey to study preferences for types of earcons informing about the headway (headway about 0.5 s).

The results are shown in Table 1. 68.1% of the respondents selected Earcon 1 as their first preference. Choices of the participants were assigned with ratings, where the first choice received 3 points, the second choice 2 points, and the third choice 1 point. The earcon with the highest rating (i.e., Earcon 1) was selected for use in the on-road study.

Table 1. Left: Reported orders for offered earcons. Right: rating of the earcons, where the first choice gets 3 points, second choice 2 points, and third choice 1 point.

Order or preference	*Percentage*
Earcon 1 – Earcon 2 – Earcon 3	44.9%
Earcon 1 – Earcon 3 – Earcon 2	23.2%
Earcon 2 – Earcon 3 – Earcon 1	14.5%
Earcon 2 – Earcon 1 – Earcon 3	5.8%
Earcon 3 – Earcon 1 – Earcon 2	5.8%
Earcon 3 – Earcon 2 – Earcon 1	5.8%

Earcon	*Rating*
Earcon 1	171
Earcon 2	132
Earcon 3	111

A question on the preferred headway at which to receive warnings was also asked. Participants were asked to rank the headways at which the cautionary warning should be given. Again, three videos were provided (same video as with Earcons 1–3), in which a neutral beep was played at three different THWs in this order: 0.5 s, 0.8 s, and 1.2 s. As above, choices of the participants were assigned with ratings, where the first choice received 3 points, the second choice 2 points, and the third choice 1 point. The results are shown in Table 2. The most preferred option was Timing 1 (0.5 s). Timing 2 (0.8 s) was almost as popular as Timing 1 (157 and 175 points, respectively). Timing 3 (1.2 s) was the least popular (82 points). In summary, the results suggest that feedback that is provided early is not preferred by participants.

Table 2. Left: preferred orders for headway timings. Right: rating of the earcons, where the first choice gets 3 points, second choice 2 points, and third choice 1 point.

Order or preference	*Percentage*
Timing 1 – Timing 2 – Timing 3	59.4%
Timing 1 – Timing 3 – Timing 2	2.9%
Timing 2 – Timing 3 – Timing 1	4.3%
Timing 2 – Timing 1 – Timing 3	27.5%
Timing 3 – Timing 1 – Timing 2	1.4%
Timing 3 – Timing 2 – Timing 1	4.3%

Timing	*Rating*
Timing 1	175
Timing 2	157
Timing 3	82

Conventional auditory feedback on headway

A Conventional feedback system was implemented. It produced an urgent sound (hereafter referred to as 'Sound 2') if the THW was smaller than 0.6 s. This sound was the same as Earcon 1 from the online survey, but the timbre was a square wave instead of a sine wave to convey a stronger sense of urgency.

Graded auditory feedback design

The above findings were used in the design of a 3-stage headway alerting system. A cautionary warning (Sound 1, identical to Earcon 1 from the online survey) was given the first time the THW dropped below 0.8 s. After this, between 0.8 s and 0.5 s (Stage 1), the informative message "Following distance too short" (Voice 1) in Dutch was played every 8 s. Based on recommendations from a previous survey on auditory in-vehicle interfaces (Bazilinskyy & De Winter, 2015) and an online experiment on the qualities of voice-based displays for cars (Bazilinskyy & De Winter, 2017), a computer-generated female voice was used for the voice-based warning. The 8 s timer was reset when the THW became larger than 1.0 s.

If the THW dropped below 0.5 s, another cautionary warning was provided once (Sound 2). As pointed out above, Sound 2 was identical to Sound 1, but had a more urgent sounding timbre. Between 0.5 and 0.3 s (Stage 2), an urgent voice (Voice 2) told the driver every 5 s in Dutch with Belgian accent to "Increase headway".

If the THW was shorter than 0.3 s (Stage 3), an imminent 659 Hz 300 ms alarm (Sound 3) was issued every 0.7 s until the THW increased. Figure 3 shows the spectrograms of the three sounds.

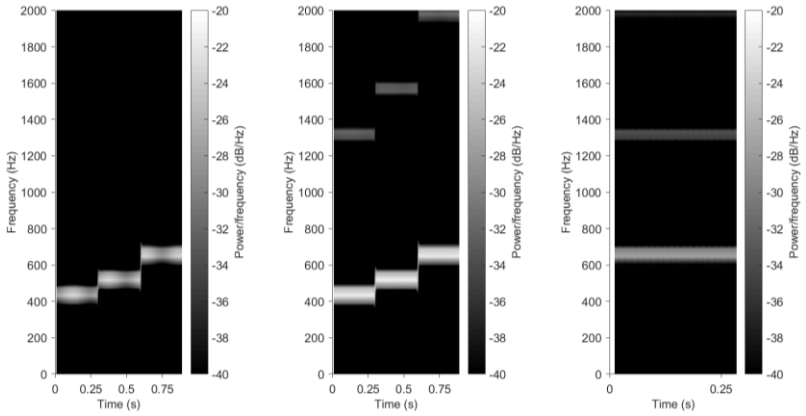


Figure 3. Spectrograms of Sound 1 (left), Sound 2 (centre), and Sound 3 (right) from the Graded auditory feedback.

A suppressing algorithm was implemented to reduce the occurrence of alarms in Stages 1 and 2. This algorithm suppressed all warnings (except in Stage 3) if the

filtered THW (moving average over five preceding samples; i.e., 0.5 s of data) was increasing. This suppressing algorithm was implemented in the Graded system only.

Sound feedback was provided only if at the moment of crossing the THW threshold (i.e., ≤ 0.8 s for Stage 1, ≤ 0.5 s for Stage 2, ≤ 0.3 s for Stage 3) the THW was within that threshold at least 0.5 s before. This additional filter suppressed feedback if the threshold was crossed only briefly, causing a maximal time delay of 0.5 s. This additional filter was present in both the Graded system and the Conventional system.

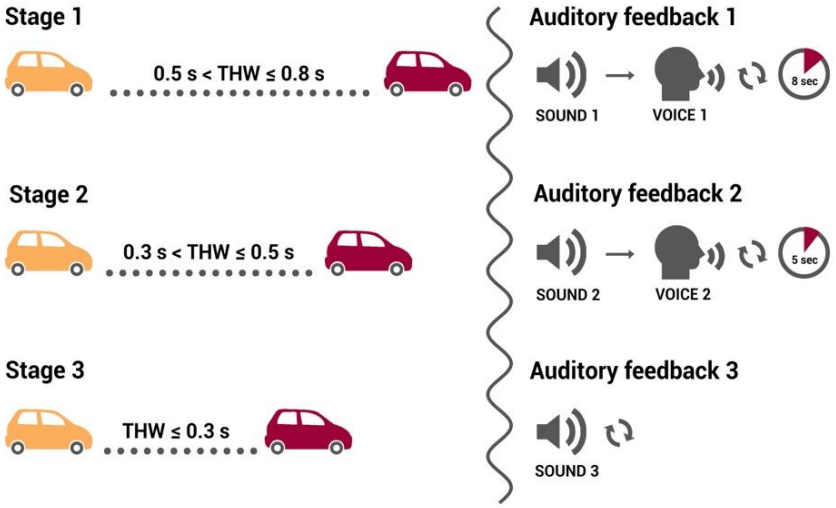


Figure 4. Visualisation of the Graded headway feedback system.

Procedures of the on-road experiment

The auditory feedback was implemented in Python and installed on a Raspberry Pi computer in a Volvo C30. Measurements on driving speed, the status of turning lights, the position of gas and brake pedals, steering angle, THW were obtained through a Mobileye system. All sounds were played through the JBL GO loudspeaker. The leading car was a Nissan Micra.

Twenty people participated in the experiment over the course of three days. The sample consisted of 13 males and 7 females. They had a mean age of 20.5 years (*SD* = 1.6). Five participants indicated to have driven less than 1,000 km in the past 12 months, 12 participants reported 1,001–5,000 km, and 3 participants reported 5,001–15,000 km. Participants provided written informed consent and were informed that the study involved auditory feedback, that the feedback is not necessarily perfect, and that they should remain attentive to the road. Participants were further asked to drive as they normally would. The participants then took place in the test vehicle together with two observers.

The participants drove a total of 14 km on a Dutch highway (A13, from the Molengraaffsingel in Delft to the Schieveensedijk in Rotterdam, and back). When

this road was congested, a track of similar length was driven on the A4 or N470. The driver was not informed about the route but was instructed to follow a car, driven with a normal driving style by one of the authors. Each drive took approximately 15 minutes and was divided into two parts of equal length. Half of the participants started with the Conventional system enabled, and the other half started with the Graded feedback system enabled. Halfway, the feedback system was changed from the Graded feedback system to the Conventional system, or vice versa. The driver was notified of the system change. When the car had returned to its starting location, the driver was asked to complete a questionnaire to measure acceptance of both systems (Van der Laan et al., 1997). The acceptance questionnaire measured two variables, namely the satisfaction and the usefulness of the systems. The participants were also interviewed on how they had experienced the two systems. It consisted of open questions that first identified which differences the participant had noticed between the two systems, and then asked their opinion regarding the used warnings and their timing.

Results

During the experiment, 11 out of the 20 participants (9 males, 2 females, mean age = 20.9, *SD* age = 1.5) drove with a headway close enough to receive feedback from both systems (i.e., at least once in each of the two drives). Only the results of these 11 participants will be considered here. The average time that participants drove faster than 50 km/h was 294 s (*SD* = 123 s, min = 159 s, max = 586 s) for the Conventional system and 286 s (*SD* = 141 s, min = 208 s, max = 682 s) for the Graded system.

Table 3 provides an overview of the number of times that feedback was provided per condition, for the 11 participants combined. It can be seen that the full potential of the Graded feedback system was not tested. That is, participants rarely drove close to the lead vehicle, and therefore Voice 2 was uttered only three times. Sound 3 was provided only once, possibly because of a misdetection or another vehicle cutting in.

Further analysis showed that while driving speed exceeded 50 km/h, the filter of the Graded system suppressed Sound 1 on 7 occasions, Voice 1 on 8 occasions, Sound 2 on 11 occasions and Voice 2 on 3 occasions. In other words, the filter appeared to be effective in *not* providing feedback when the driver was already responding.

Table 3. Number of times that a particular feedback was provided.

	Conventional system	*Graded system*
Sound 1	0 times	52 times
Voice 1	0 times	20 times
Sound 2	49 times	11 times
Voice 2	0 times	3 times
Sound 3	0 times	1 time

Note. The working mechanism of the Graded system is illustrated in Figure 2. The Conventional system provided feedback at a THW of 0.6 s. Only driving speeds greater than 50 km/h were considered.

Figure 5 shows a distribution of the recorded THW for both systems. These results tentatively indicate that the systems affected THW, as THWs higher than 0.6 s were relatively prevalent for the Conventional system whereas THWs higher than 0.8 s were relatively prevalent for the Graded System. In other words, the THW distribution is consistent with the fact that the Conventional system provided feedback at a THW of 0.6 s, whereas the Graded System gave its first beep at a THW of 0.8 s. We refrained from statistical testing due to the relatively small sample size. One issue that we observed was that there were large individual differences in following distance (Figure 6), where some participants received considerably more feedback than others.

Self-reported acceptance

Figure 7 shows the results of the acceptance questionnaire. The magenta markers represent the two systems that were tested herein. The other markers correspond to previous experiments in which participants were provided with a warning (take-over request) indicating that they had to take over control from automated driving (Bazilinskyy et al., 2017). Both the Conventional and Graded Systems received mediocre ratings on the scale from −2 to 2.

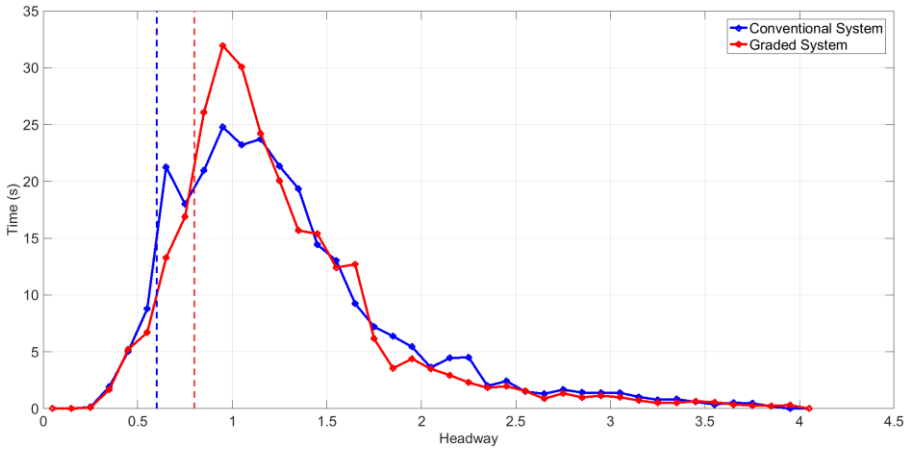


Figure 5. Distribution of time headway (THW). Time headway is defined as the distance headway divided by the own vehicle's speed. A distribution was calculated per participant and then averaged over the 11 participants. Only driving speeds above 50 km/h were considered. The vertical blue and dashed red lines represent the threshold for providing the first feedback in the Conventional and the Graded system, respectively.

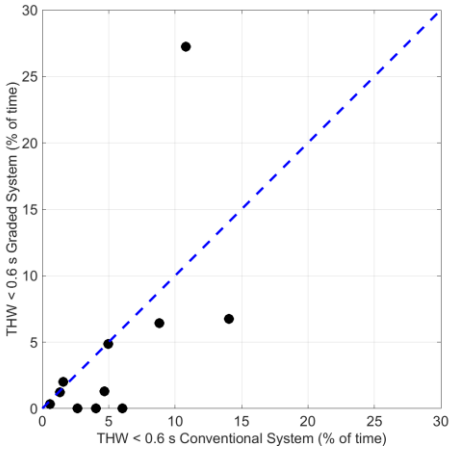


Figure 6. Percentage of time that participants drove at a time headway (THW) smaller than 0.6 s. Only driving speeds above 50 km/h were considered. It can be seen that there were substantial individual differences.

Interviews

The participants were asked questions about how they experienced the two systems, and about their attitude towards the occurrence and selection of sounds. Seven (out of eleven) participants preferred the Graded system over the Conventional system, two preferred the Conventional system, and two accepted neither system. Five participants reported they would like to receive feedback at a shorter THW for the Graded system and one driver would have preferred a shorter THW for the Conventional system. Four participants mentioned that they had experienced a delay in the feedback of the Graded system and regarded this as a negative aspect. Two participants reported trouble in understanding the spoken voice, and one participant reported a negative attitude towards the use of voice for headway warning systems.

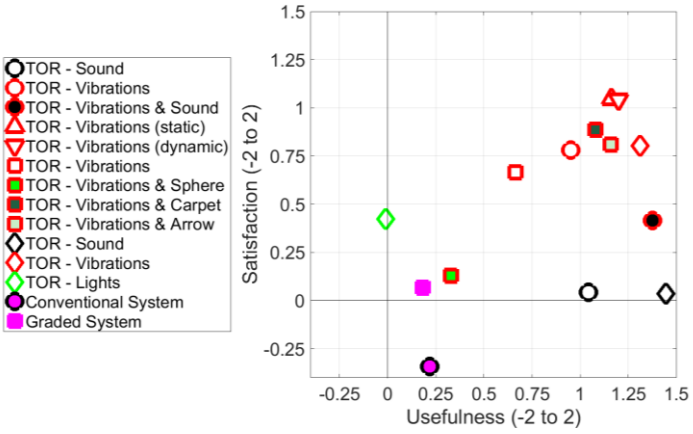


Figure 7. Self-reported usefulness and satisfaction of the tested systems (magenta square and circle) in comparison to previous auditory, visual, and vibrotactile warnings tested in driving simulators. TOR = Take-over request.

Discussion

We designed an auditory feedback system that provided feedback based on THW stage and time spent in a stage (Figure 4). A filter was added to ensure that no feedback was provided when the headway was already increasing. We expected that our design would yield better acceptance than a conventional system that provided binary feedback when a single THW threshold was exceeded.

Our algorithm was pilot-tested on a public road. Results suggest that participants, on average, did respond to the feedback, as shown from the THW distribution (Figure 5). However, we also found that the experimental design was not suitable to properly test the system as participants hardly entered the more dangerous stages. Long-lasting naturalistic driving tests are needed to examine the effect on the THW distribution and the occurrence of hazardous situations (e.g., low time to collision values) (cf. Shinar & Schechtman, 2002). In particular, the topic of individual differences deserves further examination. Some participants may hardly ever receive feedback, whereas others tend to drive at short headways for a significant portion of their driving time.

The filter reduced the number of warnings in the Graded system, especially those following a lane change. However, the interviews revealed that some drivers perceived ‘delayed feedback’ of this system. This delay may have been caused by the filter, which suppresses warnings after a cut-in by another vehicle if the headway already increases, but can trigger a late warning if the headway stops increasing while the THW is still small. The benefits of fewer warnings may, therefore, have caused a reduction in predictability.

The results showed that self-reported acceptance was relatively low as compared to previously tested systems that warn drivers about an impending collision in a driving simulator (Figure 7). It is possible that drivers accept systems that warn them of an imminent threat, but they may be less accepting towards warnings while they are already alert in a regular car following task (as in the present study). Furthermore, it is possible that drivers may be more accepting towards visual or vibrotactile feedback than to auditory feedback, or that simulator-based research yields higher acceptance ratings than on-road research. Future research could be directed towards more refined algorithms that minimize the likelihood of nuisance alarms while retaining a high acceptance.

Acknowledgements

We would like to express our special gratitude to Daria Nikulina for designing the illustration used in the survey.

References

Bazilinskyy, P., & De Winter, J.C.F. (2015). Auditory interfaces in automated driving: an international survey. *PeerJ Computer Science, 1*, e13.

Bazilinskyy, P., & De Winter, J.C.F. (2017). Analyzing crowdsourced ratings of speech-based take-over requests for automated driving. *Applied Ergonomics, 64*, 56-64.

Bazilinskyy, P., Eriksson, A., Petermeijer, S., & De Winter, J. C. F. (2017). Usefulness and satisfaction of take-over requests for highly automated driving. *Proceedings of the Road Safety & Simulation International Conference (RSS)*. LOCATION: PUBLISHER.

Brackstone, M., & McDonald, M. (2007). Driver headway: How close is too close on a motorway? *Ergonomics*, 50, 1183-1195.

Gonzalez, C., Lewis, B.A., Roberts, D.M., Pratt, S.M., & Baldwin, C.L. (2012). Perceived urgency and annoyance of auditory alerts in a driving context. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*, 1684-1687.

Hoogendoorn, S., & Botma, H. (1997). Modeling and estimation of headway distributions. *Transportation Research Record: Journal of the Transportation Research Board, 1591*, 14-22.

NHTSA. (2004). *A comprehensive examination of naturalistic lane-changes*. Washington, D.C., USA: NHTSA National Highway Traffic Safety Administration.

Ohta, H. (1993). Individual differences in driving distance headway. *Vision in Vehicles, 4*, 91-100.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230-253.

Parasuraman, R., Hancock, P.A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics*, *40*, 390-399.

Saffarian, M., De Winter, J. C. F., & Senders, J. W. (2017). The effect of a short occlusion period on subsequent braking behavior: A driving simulator study. Retrieved from https://www.researchgate.net/publication/314658202_The_effect_of_a_short_occlusion_period_on_subsequent_braking_behavior_A_driving_simulator_study

Sarter, N.B. (2005). Graded and multimodal interruption cueing in support of preattentive reference and attention management. *Proceedings of the Human Factors and* Ergonomics *Society Annual Meeting* (Vol. 49, No. 3, pp. 478-481). Sage CA: Los Angeles, CA: SAGE Publications.

Shinar, D., & Schechtman, E. (2002). Headway feedback improves intervehicular distance: A field study. *Human Factors, 44*, 474-481.

Stanton, N.A., & Edworthy, J. (1999). *Human factors in auditory warnings*. Aldershot: Ashgate.

Stipdonk, H. (2017). The impact of changes in the proportion of inexperienced car drivers on the annual numbers of road deaths. *Proceedings of the Road Safety & Simulation International Conference*, The Hague, The Netherlands.

SWOV. (2012). *Headway times and road safety* (Factsheet). Den Haag, The Netherlands: SWOV Institute for Road Traffic Research.

Treiber, M., Kesting, A., & Helbing, D. (2006). Understanding widely scattered traffic flows, the capacity drop, platoons, and times-to-collision as effects of variance-driven time gaps. *Physical Review E, 74*, 016123.

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.

Wang, J., Zhang, L., Zhang, D., & Li, K. (2013). An adaptive longitudinal driving assistance system based on driver characteristics. *IEEE Transactions on Intelligent Transportation* Systems, *14*, 1-12.

Wiese, E.E., & Lee, J.D. (2004). Auditory alerts for in-vehicle information systems: The effects of temporal conflict and sound parameters on driver attitudes and performance. Ergonomics, *47*, 965-986.

Xie, D. F., Gao, Z.Y., & Zhao, X. M. (2008). Stabilization of traffic flow based on the multiple information of preceding cars. *Communications in Computational Physics, 3*, 899-912.

Ye, F., & Zhang, Y. (2009). Vehicle type-specific headway analysis using freeway traffic data. *Transportation Research Record: Journal of the Transportation Research Board, 2124*, 222-230.

User performance for vehicle recognition in three-dimensional point clouds

Patrik Lif, Fredrik Bissmarck, Gustav Tolt, & Per Jonsson
Swedish Defence Research Agency, Linköping
Sweden

Abstract

Unmanned Aerial Vehicles (UAVs) equipped with electro-optical sensors, e.g. visual and infrared cameras, are increasingly used in military, security and search and rescue contexts. Lately, active (laser) sensors have emerged as powerful imaging devices, combining accurate and high-resolution three-dimensional measurements with night-time capabilities. The increasing availability of active sensors raises important human factors' questions, e.g. regarding what spatial resolution is required for users to recognize objects. This paper describes the outcome of a study of the relation between resolution of 3-D data and the possibility for humans to recognize different objects. We designed an experiment where the participants watched video sequences from a simulated UAV-mounted LIDAR (Light detection and ranging) sensor. Participants had to recognize vehicles of different types and point resolution, and to report their confidence level. The main conclusion is that about 100 points on the vehicles are required for users to recognize vehicles with a distinct shape or with no other vehicles of the same type. For recognizing a vehicle among others of similar appearance and size about 1000 points is required. The results show that the recognition ability deteriorates with lower number of points but that the variations between different vehicles are large. The results also show that at low resolutions participants become more precarious (lower confidence estimations) and take longer time to respond.

Background

Recognizing people, vehicles and objects is an important task in many civil and military contexts. Often electro-optical sensors, such as visual and infrared cameras, are used to facilitate the work (Schueler & Woody, 1992). Significant advances have been made recently in automatic object recognition, enabled by break-through in computational and sensor technology and fuelled by applications in robotics, the automotive industry and consumer electronics. Automatic solutions are required when actions have to be taken very quickly or when the amount of data is too much for humans to process. However, in many applications a human operator is better suited to make decisions, e.g. in order to get acceptable performance or for legal reasons.

In this work, we consider the case of using an Unmanned Aerial Vehicle (UAV) to collect sensor data for reconnaissance purposes. UAVs allow for exploring larger areas compared to using ground vehicle-mounted sensors (Fahlström & Gleason,

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

2012; Grönwall, Tolt, Lif, Larsson, Bissmarck, Tulldahl, Henriksson, Wikberg, & Thorstensson, 2015). Without the restrictions of platform movement imposed by ground properties, the UAV offers advantages in terms of viewing certain areas from different angles, thereby providing a more complete data set.

A typical task could be to search for a vehicle (or person) or to ensure that there are no vehicles (or people) within an area. In the first task it is necessary to find a specific vehicle, while in the second task it is sufficient to find any vehicle. Overall, it is not always clear what seeing a person or object means. In research it is necessary to define the situation and task to analyse human performance in different settings. One way to analyse observers' ability to perform visual tasks is to use the Johnson criteria (Donohue, 1991; Johnson & Wolfe, 1985; Sjaardema, Smith, & Birch, 2015), which is often used by scientists who study the capability of sensors. An important differentiation is made between *detection* (an object is present), *orientation* (direction of the present object), *recognition* (type of object can be discerned, e.g. recognize the difference between a human and car) and *identification* (a specific object can be discerned, e.g. type of car or identify a specific person) (Pinsky, Levin, Yaron, & Schuberth, 2016.

The Johnson criteria propose in detail how many pixels (originally line-pairs) an object needs to contain to make the classification possible. Other criteria, such as the National Interpretability Rating Scales (NIIRS) (Irvine, 1997; NIIRS, 2017) could also be used to measure the quality of images and performance of imaging systems that has scales for visible, radar, infrared and multispectral stimuli (NIIRS, 2017). NIIRS is used to assign a number (level 1-9) which indicates interpretability of an image. However, often these measurements (e.g. Johnson criteria and NIIRS) do not provide reliable and valid results since they do not take into account user variability. Therefore, experiments with naïve and expert users are necessary. Also, there are a variety of other factors that must be considered, e.g. contrast between objects and background, atmospheric disturbances, number of objects in the picture, light, contextual clues, colour, and type of optics in the sensor. Moreover, performance is affected by resolution, type of task, the experience of the participants and their level of training for the specific task, motivation, and the relative importance between quick decisions and correct results. In order to assess and evaluate a specific sensor in a specific setting, it is recommended to conduct experiments with users.

In this study a three-dimensional laser sensor that generates point clouds was in focus (Grönwall, Tolt, Lif, Larsson, Bissmarck, Tulldahl, Henriksson, Wikberg & Thorstensson, 2015; Isa & Lazoglu, 2017; Lif, Tolt, Larsson & Lagebrant, 2016). The point clouds consist of enormous amount of data in three-dimensions that are hard to handle for users and to handle these large quantities methods must be developed, preferable automatic methods (Hron & Halounová, 2015; Waldhauser, Hochreiter, Opteka, Pfeifer, Ghuffar, Korzeniowska & Wagner, 2008). If handled correct, these point cloud can help users recognize objects. However, it is hard to recognize objects from static point clouds, but if you twist and turn the cloud or look at a video sequence it is significantly easier to interpret the information. Even if the user cannot see the real object, dynamic point clouds can give some understanding of what is seen. This knowledge originated in the discoveries about biological

motion found by the Swedish researcher Gunnar Johansson in the 1970s (1973). He discovered that by presenting only a few light points placed on the body's joints a person that performed activities could be recognized (Johansson, 1973, 1975). In biological motion, a few points can be used because they are placed on the body's joints, but how many points that are needed to detect, recognize and identify a three-dimensional vehicle when the points are randomly distributed over the vehicle is unclear.

The purpose with the current experiment is to investigate how many points in a point cloud are required for vehicle recognition in a dynamic setting. This information can be used to increase the understanding of which objects that can be detected with a specific three-dimensional laser sensor at a given distance.

Method

A laboratory experiment was conducted to investigate the ability to recognize vehicles presented in the form of point clouds with different resolution. The participants were given the task to watch video sequences and recognize which of ten different vehicles was visualized. The participants also estimated how confident they were at their response on a scale from 0% to 100% in steps of 10%. Their response time (RT) was also registered.

Design

A within-group design with four resolutions × five vehicles was conducted in two experiments, with different vehicles and different resolutions for each experiment. Each vehicle was presented four times for each resolution for each participant, and based on this; a mean value was calculated and used in the subsequent analysis. The experiments were conducted so that the participants did not experience two separate experiments. This meant that the participants recognized each stimulus from ten possible vehicles.

Participants

Twelve participants (three women and nine men) participated in the experiment. All had adequate vision with or without correction.

Apparatus

The video sequences were presented on a 21.5 inch widescreen display (Dell SX 2210) with a resolution of 1920 × 1080 pixels. A PC with Windows 7, Intel® Core™ 2 Duo processor with 3 GHz and 4 GB of RAM was used to register answers and measuring response time.

Stimuli

Synthetic video sequences were generated for ten vehicles (Figure 1) by a sensor simulation system to simulate a real situation captured by a 3-D imaging laser system. Point clouds with different resolutions were generated along a predefined

path in a straight line along the road where the vehicles were located, 50 meter to the right and at 50 meter above the ground (Figure 2). Data were filtered into four different resolutions for each vehicle. There were five civilian vehicles (Volvo V70, Vera Cruz, Toyota pickup, minibus and Isuzu Truck) and five military vehicles (Leopard 122, Combat vehicle 90, Ural truck, Patria Sisu and Galten). To give the reader an understanding of the vehicles presented in the videosequences a high resolution example is depicted in Figure 1.

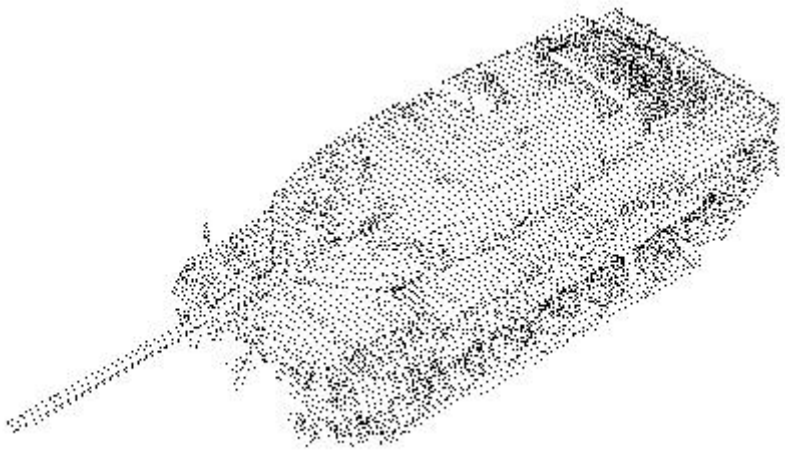


Figure 1. High resolution example of Leopard 122.

Each vehicle was dynamically presented where a simulated unmanned aerial vehicle flew in an arc around the current vehicle for five seconds according to Figure 2. The flight always started by visualizing the front of the vehicle and ending when the position was from the side of the vehicle.

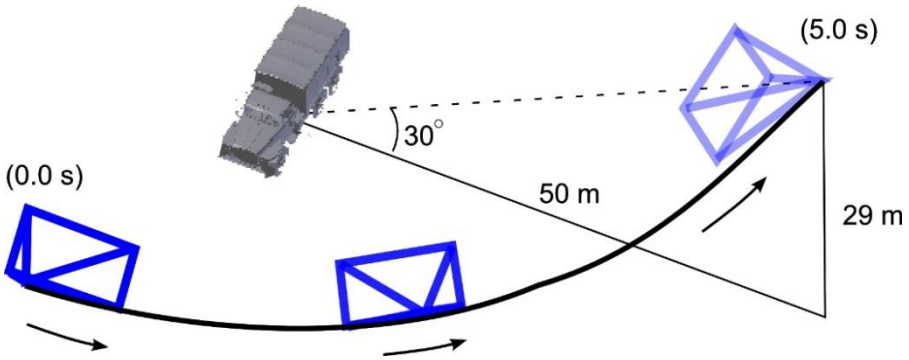


Figure 2. Flightpath and camera orientation from simulated UAV in an arc from vehicle front to vehicle side during five seconds.

The simulations were performed using a combination of tools. First, the laser scans obtained from the UAV motion (see above) was acquired by simulation, by ray casting towards surface models of the vehicles, and a surrounding, quite flat road segment. Then, the obtained, dense 3D point image was down sampled by a filter, resulting in uniform point clouds at desired resolutions. No noise was added in the simulation in order to keep the number of parameters in the experiment tractable.

Five of the ten vehicles were used in experiment 1 while the remaining five vehicles were used in experiment 2. Refer to Table 1 for details about resolution and actual number of points for each vehicle at different resolutions.

Table 1. Vehicles, their belonging to experiment, and resolution indicated in points/m^2 (*) and total number of points (**) for each vehicle for the different resolutions.

Experiment 1				
Vehicle type	Resolution 1	Resolution 2	Resolution 3	Resolution 4
Galten	6.3*/309**	1.6*/84**	0.7*/23**	0.3*/9**
Leopard	6.3*/509**	1.6*/101**	0.7*/58**	0.3*/13**
Patria	6.3*/324**	1.6*/83**	0.7*/36**	0.3*/8**
Toyota Pickup	6.3*/225**	1.6*/62**	0.7*/18**	0.3*/7**
Vera Cruz	6.3*/160**	1.6*/42**	0.7*/15**	0.3*/6**
Experiment 2				
Vehicle type	Resolution 1	Resolution 2	Resolution 3	Resolution 4
Isuzu truck	25.0*/309**	2.8*/144**	1.0*/63**	0.4*/20**
Minibus	25.0*/785**	2.8*/92**	1.0*/23**	0.4*/10**
Combat vehicle 90	25.0*/1196**	2.8*/123**	1.0*/49**	0.4*/17**
Ural truck	25.0*/1385**	2.8*/166**	1.0*/69**	0.4*/21**
Volvo V70	25.0*/564**	2.8*/74**	1.0*/14**	0.4*/8**

Procedure

After welcoming the participants individually and briefing them about the experiment purpose and procedure they received written information and had the opportunity to ask questions to the experiment leader. Then an introduction was given to make sure that the participants were familiar with the situation and test material. They were introduced to the stimuli material during five minutes of training. The participants watched the videos and answered by first pressing the space button whereby the response time (RT) was recorded, then selecting one of the ten vehicles (recognition) and finally stating how confident they were (Figure 3).

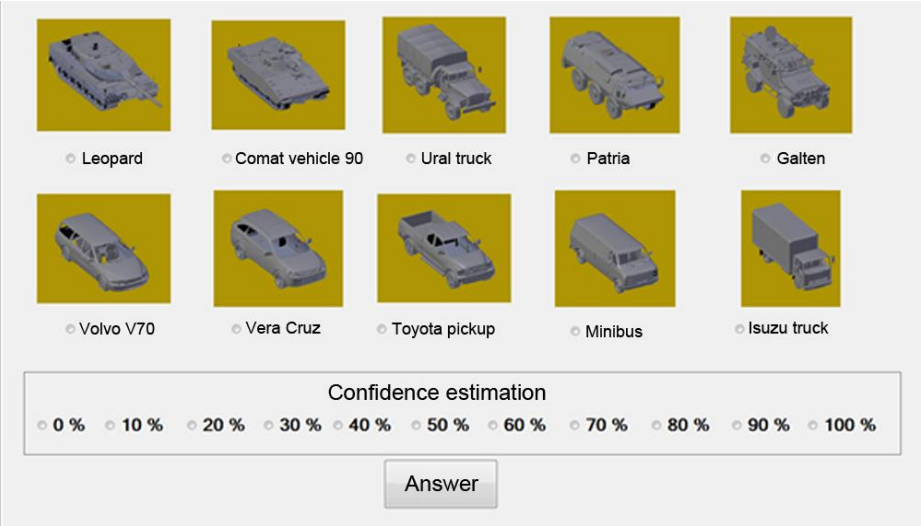


Figure 3. Response window with selection of vehicle and confidence estimation.

Participants were instructed to give priority to answering correctly over fast responses. Since the task was mentally demanding the experiment was divided into six separate blocks of five minutes each with the possibility to rest in-between blocks. The order of the conditions for both experiments was randomized to avoid training effects. This means that the participants did not perceive the experiments as separate parts but as one experiment. The whole procedure took about 40 minutes to complete. No feedback was given to the participants during the experiment.

Results

The results include statistical analysis of recognition, confidence estimation and response time. Data were analysed with a two-way ANOVA (Hays, 1994) with resolution × type of vehicle. Tukey HSD were used for post hoc testing (Green & D'Oliviera, 1982). Data from experiment 1 and 2 were analysed separately. Notice the non-linear scale of the x-axis (all figures), which was a necessary compromise to make the figures readable.

Experiment 1

Four resolutions were used (6.3, 1.6, 0.7 & 0.3 points/m^2) for five different vehicles (Galten, Leopard 122, Patria, Toyota pickup and Vera Cruz). Average values were calculated for each participant for each condition.

Recognition

The ability to recognize vehicles was measured by the proportion of correct answers and the analysis was performed by an ANOVA repeated measures. The results show a main effect for type of resolution $F(3, 33) = 24.52$, $p < .001$ and type of vehicle $F(4, 44) = 18.29$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 4.69$, $p < .001$ (Figure 4).

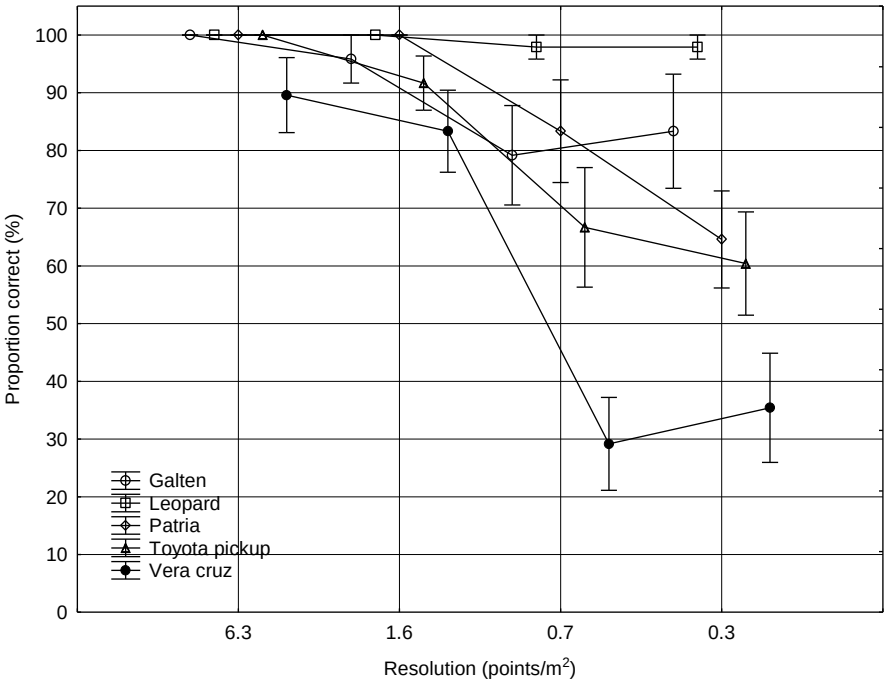


Figure 4. Mean and standard error of mean for proportion recognized vehicles.

Tukey post hoc test shows a significant difference (from the main effect of resolution) between 1.6 and 0.7 points/m^2 ($p < .001$). There is also a significant difference (from the main effect of vehicle) showing that the ability to recognize Vera Cruz was significant lower than the other vehicles ($p < .05$) and the ability to recognize Leopard 122 was significant higher than Toyota pickup and Vera Cruz. The post hoc test from the interaction effect between resolution and vehicle shows that there were no significant differences between vehicles at 6.3 and 1.6 points/m^2. However, at 0.7 and 0.3 points/m^2 the recognition is significant lower for Vera Cruz than for the other vehicles ($p < .001$). At 0.7 points/m^2 the recognition is significant higher for Leopard 122 than for the other vehicles ($p < .01$) and significant higher than for four of the five vehicles (not for Galten) at 0.3 points/m^2 ($p < .01$).

Confidence estimation

The participants' confidence (0-100% confident) was measured and the mean values were used for analysis and performed with ANOVA repeated measures. The results shows a main effect for type of resolution $F(3, 33) = 114.01$, $p < .001$ and type of vehicle $F(4, 44) = 53.31$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 16.78$, $p < .001$ (Figure 5).

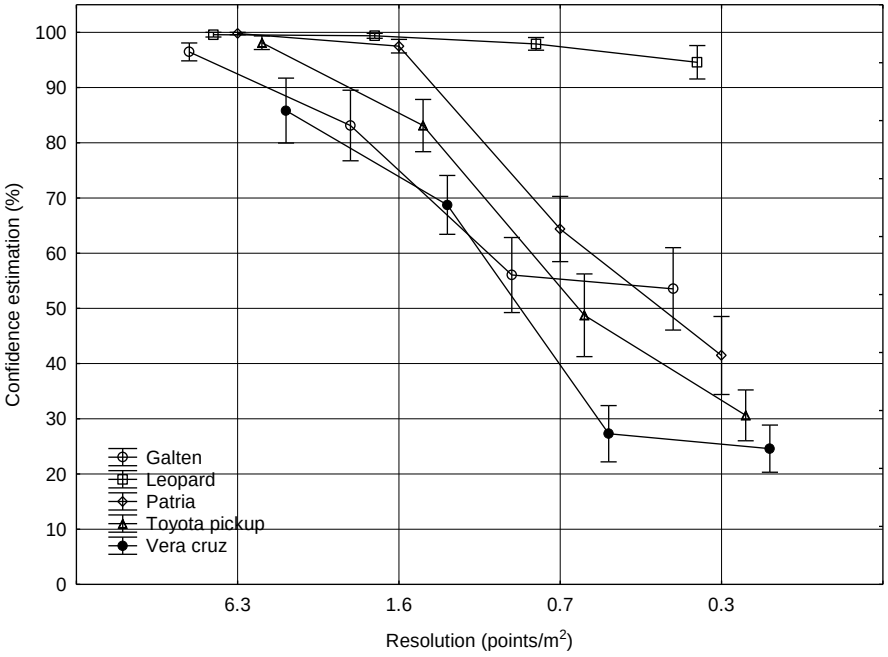


Figure 5. Mean and standard error of mean for confidence estimations.

Tukey post hoc test show a significant difference (from the main effect of resolution) between each resolution where the participants' confidence decreases with decreasing resolution ($p < .05$). The confidence estimation for Leopard 122 is stable for all resolutions (above 90%), but confidence estimations for the other vehicles drastically drops from the highest to the lowest resolution.

Response time

The participants' response times were measured and the mean values used for analysis and performed with ANOVA repeated measures. The results showed a main effect for type of resolution $F(3, 33) = 137{,}25$, $p < .001$ and type of vehicle $F(4, 44) = 113{,}51$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 13.16$, $p < .001$ (Figure 6).

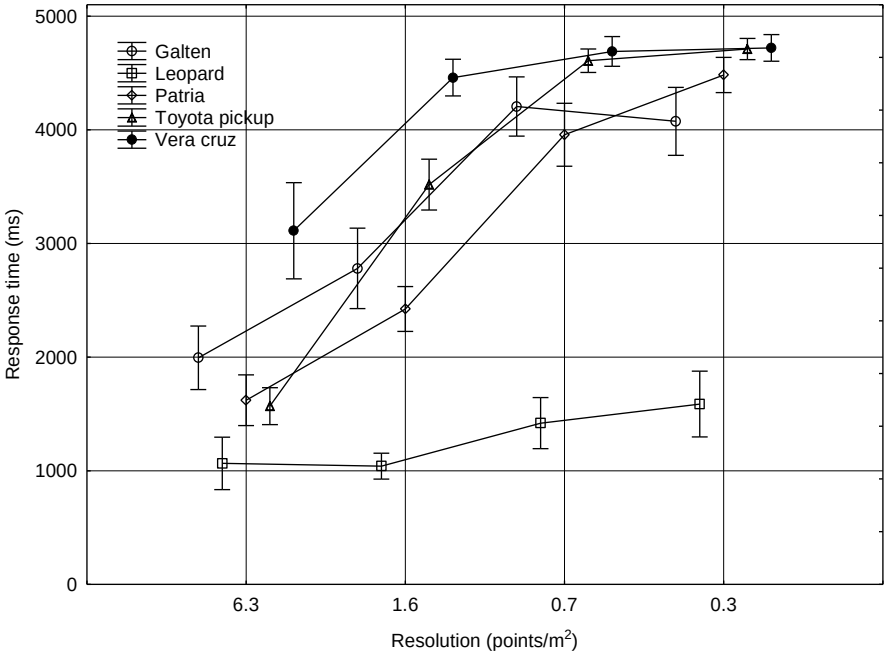


Figure 6. Mean and standard error of mean for response time.

Experiment 2

Four resolutions were used (25.0, 2.8, 1.0 & 0.4 points/m^2) for five vehicles (Isuzu truck, Minibus, Combat vehicle 90, Ural truck & Volvo V70). Average values were calculated for each participant for each condition.

Recognition

The ability to recognize vehicles was measured by proportion correct answers and analysis was performed by ANOVA repeated measures. The results show a main effect for type of resolution $F(3, 33) = 63.0$, $p < .001$ and type of vehicle $F(4, 44) = 17.72$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 5.80$, $p < .001$ (Figure 7).

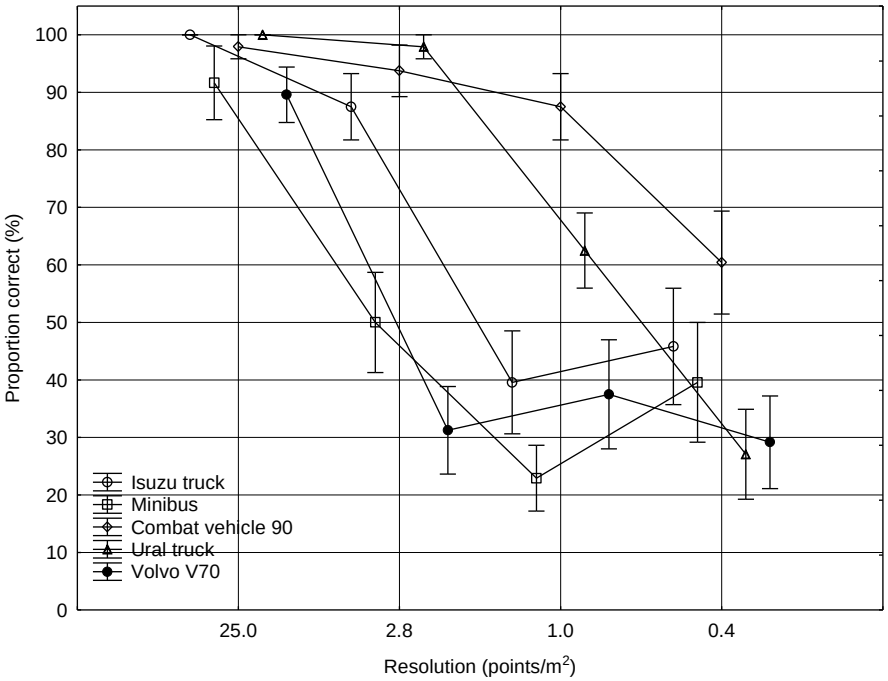


Figure 7. Mean and standard error of mean for proportion recognized vehicles.

Tukey post hoc test showed a statistical difference (from main effect of resolution) between 25.0, 2.8 and 1.0 points/m^2 ($p < .001$). There was also a significant difference (from main effect of vehicle) that showed that the ability to recognize the minibus and Volvo V70 was lower than the other three vehicles ($p < .05$). Also, the ability to recognize Combat vehicle 90 was higher than the other vehicles ($p < .05$). The analysis of the interaction effect shows that the ability to recognize the Ural truck gets significant lower between 2.8 and 1.0 points/m^2, and also between 1.0 and 0.4 points/m^2 ($p < .001$), and for Isuzu truck between 2.8 and 1.0 points/m^2 ($p < .001$). A similar decrease in ability to recognize the minibus and Volvo V70 occurs between 25.0 and 2.8 points/m^2 ($p < .05$).

Confidence estimation

The participants' confidence (0-100% confident) was measured and the mean values were used for analysis and performed with ANOVA repeated measures. The results show a main effect for type of resolution $F(3, 33) = 155.84$, $p < .001$ and type of vehicle $F(4, 44) = 75.99$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 6.53$, $p < .001$ (Figure 8).

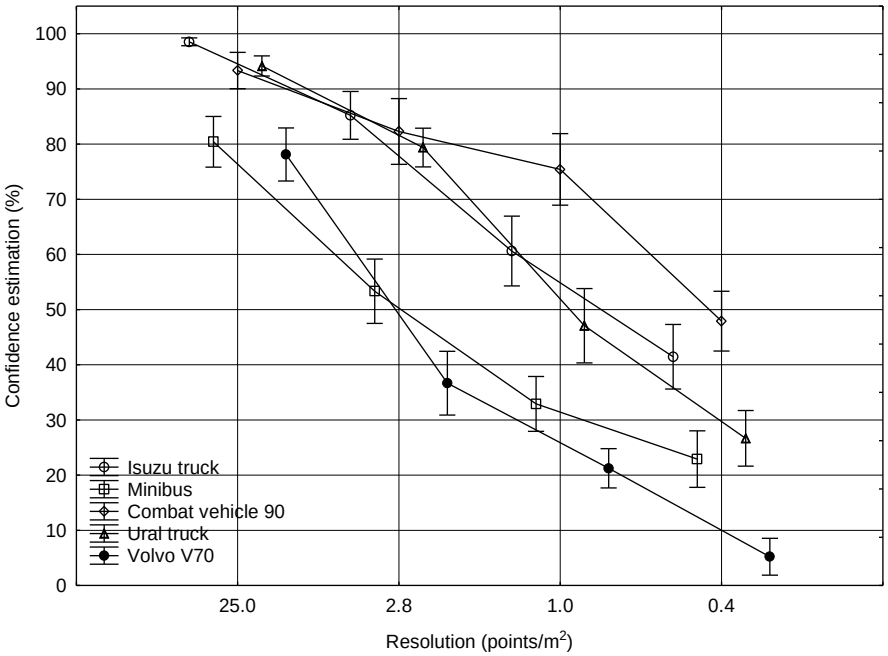


Figure 8. Mean and standard error of mean for confidence estimations.

Tukey post hoc test shows a significant difference (from the main effect of resolution) between each resolution where the participants' confidence decreases with decreasing resolution ($p < .001$). Even though there are differences between vehicles, the overall trend is similar.

Response time

The participants' response time was measured and the mean values used for analysis and performed with ANOVA repeated measures. The results show a main effect for type of resolution $F(3, 33) = 50.83$, $p < .001$ and type of vehicle $F(4, 44) = 28.65$, $p < .001$, and a significant interaction effect between resolution and vehicle $F(12, 132) = 7.24$, $p < .001$ (Figure 9).

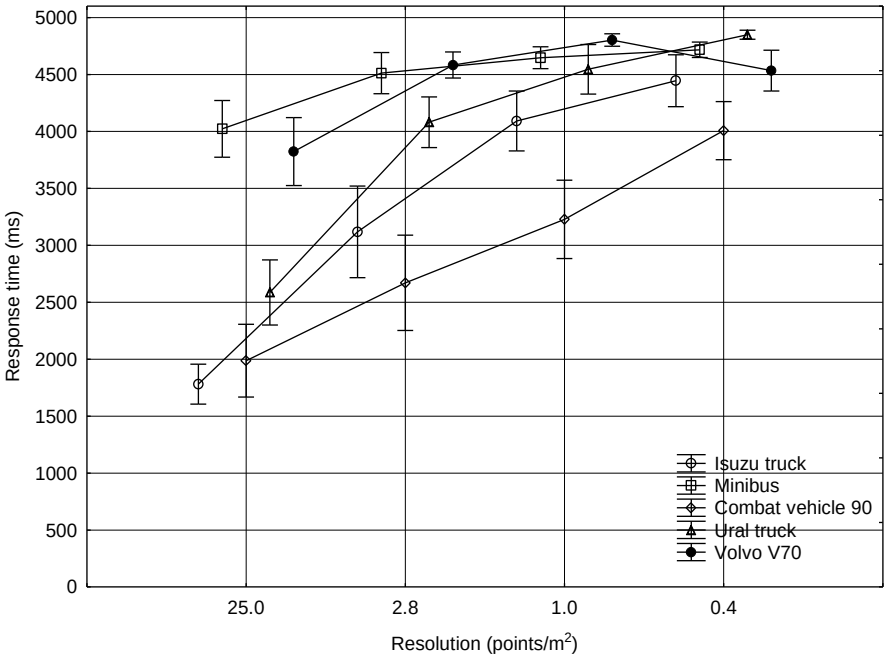


Figure 9. Mean and standard error of mean for response time.

Tukey post hoc test shows a statistical difference between the three highest resolutions ($p < .001$) where response time increases with decreasing resolution. The response time varies between vehicles, but at 25 points/m^2 participants respond faster for Isuzu truck and Combat Vehicle 90 than for the other three vehicles ($p < .005$).

Discussion and conclusions

The purpose of this study was to investigate the ability to recognize vehicles presented in the form of point clouds with different resolution (0.3-25.0 points/m^2). The results clearly show that the ability to recognize vehicles deteriorates when the number of points in the point cloud decreases, but the variation between vehicles is high. The results also show that the participants become more uncertain (lower confidence estimation) and that they take longer time to respond (RT) the lower the resolution is.

The ability to recognize vehicles is depending on whether the vehicle has a distinctive look and how many other vehicles that has similar appearance. Leopard 122 was the easiest vehicle to recognize and even at a resolution of 0.3 points/m^2 participants recognized over 95% of the vehicles with confidence estimations over 95%. This is a remarkably high and probably due to the fact that Leopard 122 is the only vehicle with a distinct gun barrel. This result should therefore be interpreted with caution, because in a context of multiple vehicles with barrels it would probably be much harder to recognize this vehicle. Recognition for 90% correct or

higher requires a resolution of about six points/m^2 when the vehicle has a characteristic appearance or if there are no vehicles of similar appearance. To recognize vehicles with similar appearance (e.g. Volvo V70 and Vera Cruz) 25 points/m^2 are required to get a 90% correct or higher classification. Since the purpose of this experiment was to investigate how many points are required for vehicle recognition, the results were also analysed relative to the number of actual points (instead of points/m^2) for each vehicle and each resolution. By analysing all the mean values of the vehicles where the participants had 90% correct or higher, the following conclusions were drawn for a rule-of-thumb;

1) Approximately 100 points on the vehicle/object/target are required to recognize vehicles that have a characteristic appearance or do not have other vehicles that are similar.
2) Approximately 1000 points on the vehicle/object/target are required to recognize vehicles when there are similar vehicles (such as multiple civil vehicles of similar appearance and size).

In practical operational terms, this means that about 1000 points are needed to recognize vehicles in situations where there are similar vehicles. Especially in civil environments, there are many different similar vehicle models that make recognition and identification problematic. Even though the number of vehicle models is fewer in military environments, many of the vehicles have similar appearance that makes recognition problematic. These conclusions apply provided that information comes from a mobile platform like an unmanned aerial vehicle and that video sequences or other similar dynamic information is presented for the users. Stationary visualisation of vehicles probably requires significantly more points than a dynamic visualization as in this experiment. Numbers given from this experiment should be considered as approximate values (100 and 1000 points for vehicle recognition as stated above) since it will vary depending on the vehicle type and the number of similar vehicles. Also, this experiment should be repeated with other vehicles and possibly different resolutions.

Although this study gives a good idea of how many points that are required to recognize different vehicles, many questions remain unanswered. One such example is to investigate the number of points required for identification rather than recognition and another example is how to visualize point clouds to increase realism and thereby possibly improve participants' performance. Many other factors, such as sensor noise and partial occlusion of the target, should also be taken into account in future experiments. It would also be interesting to investigate how different design tools affect participants' performance, such as a grid on the ground to facilitate the understanding of size.

References

Donohue, J. (1991). *Introductionary review of target discrimination criteria*. Wilmington, MA.

Fahlstrom, P. G., & Gleason, T. J. (2012). *Introduction to UAV Systems*. West Sussex, U.K.: Wiley & Sons, Ltd. Publication.

Greene, J., & D'Oliveira, M. (1982). *Learning To Use Statistical Tests In Psychology*. Philadelphia: Open University Press.

Grönwall, C., Tolt, G., Lif, P., Larsson, H., Bissmarck, F., Tulldahl, M., Henriksson, Wikberg & Thorstensson, M. (2015). 3D sensing and imaging for UAVs. In G. Kamerman, O. Steinvall, K. L. Lewis, & J. D. Gonglewski (Eds.) *Proceeding of SPIE, Volume 9649 Electro-Optical Remote Sensing, Photonic Technologies, and Applications IX; 96490C (2015)*; doi: 10.1117/12.2192834 (p. 96490C). Toulouse, France

Hays, W. (1994). *Statistics* (5th ed.). New York: Ted Buchholz.

Hron, V., & Halounová, L. (2014). Automatic Generation of 3D Building Models from Point Clouds (PDF Download Available). In Automatic Generation of 3D Building Models from Point Clouds. Ostrava, Czech Republic: *Proceedings in Geoinformatics for Intelligent Transportation*.

Irvine, J. M. (1997). National Imagery Interpretability Rating Scales (NIIRS): Overview and Methodology. *In National Imagery Interpretability Rating Scales (NIIRS): Overview and Methodology. Proceedings of SPIE, Volume 3128 (pp. 93-103).*

Isa, M. A., & Lazoglu, I. (2017). Design and analysis of a 3D laser scanner. *Measurement, 111, 122–133.*

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Johansson, G. (1975). Visula motion perception. *Scientific American*, 232 (6), 76-89.

Johnson, R. B., & Wolfe, W. L. (1985). Analysis of Image Forming Systems. In *Analysis of Image Forming Systems* (p. 1016). *Proceeding of SPIE-the International Society for Optical Engineering, 513, part 1 and part 2*.

Lif, P., Tolt, G., Larsson, H., & Lagebrant, A. (2016). Subjective Evaluation for 2D Visualization of Data from a 3D Laser Sensor. *In proceedings of International Conference on Human Interface and the Management of Information (pp. 148-157)*. Springer Link.

NIIRS - National Image Interpretability Rating Scales. (2017). Retrieved August 14, 2017, from https://fas.org/irp/imint/niirs.htm

Pinsky, E., Levin, I., Yaron, O., & Schuberth, W. (2016). Prediction of object detection, recognition, and identification [DRI] ranges at color scene images based on quantifying human color contrast perception. In Image and Signal Processing for Remote Sensing XXII, Lorenzo Bruzzone, Francesca Bovolo (Eds.), *Proceeding of SPIE Vol. 10004, 1000423.* Cambridge University Press.

Schueler, C., & Woody, L. (1992). Digital electro-optical imaging sensors. *International Journal of Imaging Systems and Technology*, *4*(3), 170–200. https://doi.org/10.1002/ima.1850040305.

Sjaardema, T.A., Smith, C.S., & Birch, G.C. (2015). *History and Evolution of the Johnson Criteria*. Albuquerque, New Mexico.

Waldhauser, C., Hochreiter, R., Otepka, J., Pfeifer, N., Ghuffar, S., Korzeniowska, K., & Wagner, G. (2014). Automated classification of airborne laser scanning point clouds. In Solving Computationally Expensive Engineering Problems, (Eds. Koziel, Leifsson & Yang). *Proceedings in Mathematics & Statistics book series (Vol. 97, pp. 269–292)*. Springer New York LLC.

Impulsivity modulates pilot decision making under uncertainty

Julia Behrend[1,2], Frédéric Dehais[2], & Etienne Koechlin[1]
[1]Laboratoire de neurosciences cognitives, Département d'études cognitives, Ecole Normale Supérieure, INSERM, PSL Research University, 75005 Paris,
[2]Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), Université de Toulouse
France

Abstract

Personality has an important influence on the variability in human decision making. Little is known whether intensive training and a highly-procedural environment can alleviate the influence of personality on decision making. Here, we address this issue by investigating the influence of impulsivity as personality factor on decision making among airline pilots. We showed that impulsivity modulated pilots' indecisiveness in uncertain decision scenarios as well as pilots' self-reported compliance to airline guidelines in real life. This result suggests that the personality factor impulsivity is a profound trait that continues to have an influence through intensive training and highly-procedural decision situations.

Introduction

There is a great variability of human behaviour in response to uncertainty. It is well documented that personality influences decision preferences and actions (Byrne, Silasi-Mansat & Worthy, 2015; Sutin & Costa, 2010; Hirsh, Morisano, & Petersen, 2008). In high-risk environments, such as in commercial aviation, individuals often have to make critical decisions under uncertainty and time pressure without compromising safety. For example, a pilot has to decide whether to continue a landing approach - keep action plan - or to discontinue an approach – change action plan. In order to decide, a flight crew, composed of a Captain and a First Officer, should integrate and respect a list of defined airline guidelines, the approach criteria. Approach criteria are technical values such as correct speed, wind, vertical glide path, etc. This particular decision moment is one of the most dynamic and incident-sensible flight phases in aviation (U.S. Department of Transportation, 2015). Here, pilots have to make rapid decisions under time pressure by proving their adaptation skills (Dehais, Behrend, Peysakhovich, Causse, & Wickens, 2017). When approach criteria exceed guidelines, pilots should discontinue the approach by changing the current action plan. Surprisingly, in more than 97% of this type of situation pilots kept their action plan and did not adapt it although it would have been required by airline guidelines (IATA, 2016). Due to the dynamic character and the operational consequences, this type of decision is complex. Much is known about contributing

In D. de Waard, F. di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

factors, such as financial incentives and emotions (Causse, Dehais, Péran, Sabatini, & Pastor, 2013), lack of airline policy and time pressure (IATA, 2016), overconfidence (Goh & Wiegmann, 2001), or safety implications due to the rarity of the event in real life (Dehais et al., 2015; BEA, 2013). However, the understanding of the psychology of non-compliance to airline guidelines lacks.

Pilots are a very homogenous population since they follow a complex selection process that requires a high level of executive functions (O'Hare, 1997) and a stable personality (Childester, Helmreich, Gregorich, & Geis, 2009). An individual's personality could be described as result of constant interactions between inherited genetic influence, epigenetic effects, and social environment (Montiglio, Ferrari, & Réale, 2013). However, flight crews do not compromise the same individuals. A consequence of the worst air disaster in history, the Tenerife airport crash of two airplanes in 1977, was to reduce subjective decisions of the part of pilots (McCreary, Pollard, Stevenson, & Wilson, 1998). This was also the birth of the earlier concept of crew resource management (CRM): a set of mandatory training procedures with a focus on interpersonal communication, leadership and decision-making in the cockpit (Helmreich, Merritt, & Wilhelm, 1999). In this accident, the KLM Captain released the brakes and the airplane crashed into another airplane, even though the First Officer was reading back the ATC clearance to the tower. The KLM Captain made a quick and autocratic decision, although he had seemed to be pace and non-autocratic before. Among other causes, human factors analyses argued that his personal leadership appeared to change – possibly due to his hierarchical status in the cockpit, his responsibility in the company, and the stressful environment under time pressure (McCreary et al., 1998). The question is, *do personality factors persist in highly-trained individuals and in highly-procedural situations, such as in airline pilot decision making?* One hypothesis could be that an intensive training and a highly-procedural environment reduce the influence of personality on decision making. Or alternatively, personality is a profound trait which influence cannot be reduced by intensive training and a highly-procedural environment. We addressed this issue among airline pilots making decisions during landing approach scenarios. The focus was on impulsivity as personality factor.

Impulsivity is a multi-dimensional personality construct that is frequently described as "a predisposition toward rapid, unplanned reactions to internal or external stimuli without regard to the negative consequences of these reactions to the impulsive individuals or to others" (Moeller, Barratt, Dougherty, Schmitz, & Swann, 2001). For example, impulsive individuals are more likely to choose immediate-smaller over larger-delayed rewards; demonstrated via decision preference (Bialaszek, Gaik, McGoun & Zielonka, 2015), physiological activity (Korponay, Dentico, Kral, et al. 2017), and brain activity (Garavan, Ross, Murphy, Roche, & Stein, 2002). One area of significant importance to the measurement of impulsivity is executive function and decision making (Stanford, Mathias, Dougherty, Lake, Anderson, & Patton, 2009). Executive control is characterized as the capacity to coordinate thoughts and to perform non-automatic actions for the purpose of adaptation to stimuli (Koechlin, 2016). Individuals with executive deficits, e.g. cognitive impairment, tend to score higher on impulsivity (Stanford et al., 2009). Garavan et al. (2002) found a positive correlation between cognitive impairment and anterior cingulate activation in the

"Go/noGo task"; which measures impulsive control behaviourally. Importantly, cingulate activation is crucial to inhibition tasks, where deliberative responses are more appropriate than automatic responses. The "Wisconsin Card Sorting Test" (WCST) assesses cognitive flexibility, which is part of executive functioning and can be described as the ability to switch between different task sets and to decide flexibly. Cheung, Mitsis, and Halperin (2004) used this test demonstrating that motor impulsivity explained significant parts of the performance variance of cognitive flexibility. Two studies among general aviation pilots showed that perseverative errors on the WCST (Causse, Dehais, Arexis, Pastor, 2011a) as well as flight experience, motor impulsivity, and updating capacity could predict landing decision relevance (Causse, Dehais, & Pastor, 2011b). Indeed, in the second study the pilot's ability to detect meteorology degradation during the decision making process was measured. It was found that general aviation pilots with a higher motor impulsivity score showed less adaptation skills by continuing the current action plan. Although impulsivity is often characterized as a negative and dysfunctional state, it has been shown that being impulsive can be positive and more adaptive in simple decision tasks (Dickmann, 1990). Importantly, the decision context plays a crucial role to an individual's response behaviour (Maule, Hockey, & Bdzola, 2000).

Analysing a pilot's individual decision in a questionnaire – a non-dynamic context - can be useful to improve the understanding of the decision-relevant information and the interpretation of airline guidelines. In this study, we investigated the influence of impulsivity along with other factors such as flight hours, hierarchy, and prior airline career on individual pilot decision-making in a questionnaire.

Material and methods

Participants and demographic information

Forty randomly-selected airline pilots (age-range 32-65 years) from the same airline participated in this study. The planning department of an airline randomly chose these pilots from the pilot pool. Afterwards, we contacted these pilots by e-mail in order to ask for their agreement. Nationalities represented in our sample included the following: France (n = 38), and Belgium (n = 2). French was their native language. Table 1 resumes the demographic characteristics of this sample size. Captains were significantly older ($t(38) = 4.46$, $p < 0.001$) and had more flight hours ($t(38) = 4.69$, $p < 0.001$) than First Officers in this sample size. Half of the airline pilots reported having worked for at least another airline prior to their current employment. The percentage of pilots with a military career was 10%. All participants were paid for their participation by their airline and gave written consent prior to the experiment. Confidentiality was guaranteed.

Table 1. Demographic characteristics of this sample size

Participants (n)	*Gender, Male % (n)*	*Age, years M (SD)*	*Flight experience, hours M (SD)*
All (40)	93 (37)	47.9 (7.4)	11613 (4142)
Captain (24)	96 (23)	51.4 (5.6)	13633 (3355)
First Officer (16)	88 (14)	42.7 (6.8)	8581 (3317)

Barratt Impulsiveness Scale (BIS-11)

The BIS-11 (Patton, Stanford, & Barratt, 1995) is a self-report measurement of impulsivity with three sub traits: attentional impulsivity (e.g. "I don't pay attention"), motor impulsivity (e.g. "I act on the spur of the moment") and non-planning impulsivity (e.g. "I say things without thinking"). The questionnaire's instructions ask subjects to indicate how often description of impulsive behaviour pertain to themselves on a 4-point-Likert scale. Lower questionnaire scores indicate lower levels of impulsivity. The BIS has good internal consistency (Cronbach's Alpha = .83) and test-retest reliability (Spearman's rho = .83) (Stanford et al., 2009). In this sample size, internal consistency was computed and considered acceptable (Cronbach's Alpha = .73).

Landing questionnaire

Participants considered eighteen decision scenarios. The order of these scenarios was randomized across participants. For each scenario, participants were asked: "Based on the following information, would you continue the approach?" They could reply "Yes", "No" or "I don't know". We manipulated the presence of uncertainty in the landing decisions (uncertain vs. certain continue approach or discontinue approach). All decision scenarios were chosen from an airline's real event database. Prior to the experiment, we asked five experts - all flight instructors - to evaluate the chosen landing scenarios. All flight instructors agreed on 12 certain (8 continue, 4 discontinue the approach). The remaining 6 scenarios were labelled as uncertain (at least two instructors chose the opposite of the three others). Information complexity was reduced to three main approach criteria (localizer deviation, glide slope deviation, and airspeed) and two additional decision criteria (wind, weather conditions). For each scenario, the type of approach and the airport were identical. The information relevant for landing decisions was either within the airline guidelines (certain/continue the approach), out of the airline guidelines (certain/discontinue the approach) or at thresholds of airline guidelines (uncertain/continue or discontinue the approach). Certain decisions to continue required all criteria to be within airline guidelines. Certain decisions to discontinue occurred when at least one criterion was out of airline guidelines. Uncertain decisions (to continue or discontinue) occurred when at least one criterion was at threshold. After the 18 landing decisions, pilots were asked in an open question if they had ever taken a decision that was not in line with airline procedures (non-compliance).

Experimental design

The experiment was performed within a period of 30 days. All participants replied to the questionnaires after a full-flight simulator training. Each participant was seated separately in a room with paper and pencil. Pilots were told that the experiment was part of a research project aiming to better understand their evaluation of approach criteria. Afterwards, they were asked to complete the paper-and-pencil version of the BIS-11. Finally, they gave demographic information (Figure 1). They had no time restriction to complete the questions. The experiment duration was between 30 and 80 minutes. Figure 1 shows the protocol timeline.

Figure 1. Protocol timeline.

Results

Statistical analysis

Normality of variables was evaluated using Kolmogorov-Smirnov-Test. Normally-distributed variables were: impulsivity and flight hours. Descriptive statistics summarized pilots' approach decisions of all 18 scenarios. If sample sizes were small, Fisher's exact test for categorical variables - instead of chi-square statistics – was used. T-tests compared the pilots' level of impulsivity to normative data of other studies. Linear regression was used to describe pilots' indecisiveness during uncertain approach scenarios. Logistic regression was performed in order to encode pilots' self-reported compliance to airline guidelines in real life. A p-value .05 was considered significant. Statistical tests were performed two-tailed.

Uncertainty rating in approach decisions

We first analysed whether pilots rated the approach scenarios in the same way as pilot experts. Table 2 shows that 93% of the participants agreed on the decision to continue the approach in the certain/continue scenario. In the certain/discontinue scenario, 90% of participants made the decision to discontinue the approach. In the uncertain scenarios, 54% of the participants decided to continue, whereas 35% decided to discontinue the approach. 11% of the participants expressed their indecisiveness. Fisher's exact test confirmed significant differences between both certain/continue scenarios and uncertain scenarios ($p < .001$), certain/discontinue scenarios and uncertain scenarios ($p < .001$) as well as certain/continue and certain/discontinue ($p < .001$).

Table 2. Mean of decision agreement with expert judgement for the three types of scenarios

Decision agreement in %	*Certain/continue*	*Certain/discontinue*	*Uncertain*
Continue	93	4	54
Discontinue	3	90	35
Indecisiveness	4	6	11

BIS-11 impulsivity scores

Next, the pilot's mean impulsivity score ($M = 51.9$, $SD = 5.4$) was compared to other studies. Therefore, we calculated the impulsivity *t*-value of different studies by comparing it to the impulsivity *t*-value of this experiment. Table 2 shows that pilots scored significantly lower on the BIS-11 than healthy controls of two studies (Patton

et al., 1995; Spinella, 2007). There were no significant differences between the adult sample size of Stanford et al.'s study (2009) and this study.

Table 2. Comparison of BIS-11 impulsivity scores between the reference study and other studies

Study authors	*Reference study*	*Patton et al. 1995*	*Spinella . 2007*	*Stanford et al. 2009*
Sample type	Pilots	Male undergraduates	Adults	Adults
N	40	130	700	1577
M (SD)	51.9 (5.4)	64.9 (10.1)	64.2 (10.7)	62.3 (10.3)
|t|, p < .05	2.02	> 2.41 *	> 2.27 *	< 1.93

Linear regression: encoding indecisiveness during uncertain approach scenarios

More than half of the participants (52.5%) expressed at least once their indecisiveness during all 18 landing scenarios. They were named indecisive pilots in the analysis. Pilots, who never expressed their indecisiveness during all scenarios, were labelled decisive pilots. Chi-square test revealed that the percentage of expressed indecisiveness (indecisive vs. decisive pilots) did not significantly differ by hierarchy (Captain vs. First Officer) (χ^2 (2) = 0.1, p < .69**,** φ = .01, n = 40). An independent *t*-test was conducted to compare the level of impulsivity and the number of flight hours in decisive vs. indecisive pilots. There was a significant difference in the impulsivity scores between decisive pilots (*M* = 53.61, *SD* = 5.6) and indecisive pilots (*M* = 50.10, *SD* = 4.5); *t*(38) = -2.1, p < .03, Cohen's d = .69). There were no significant differences regarding the flight hours between the decisive (*M* = 11481, *SD* = 3644) and indecisive pilots *(M* = 11758, *SD* = 4730); *t*(38) = -.08, *p* = .83, Cohen's d = .06). In order to gain a more precise understanding of the level of indecisiveness, we then calculated an indecisiveness score that was defined as the number of times a pilot expressed indecisiveness during all uncertain approach scenarios. Next, a linear regression analysis was conducted to predict this indecisiveness score using flight hours, impulsivity, hierarchy and prior airline experience. Together, these measures explained 27 % of the variance in the individuals' indecisiveness score (*F*(4,35) = 3.2, *p* < .02). Individually, impulsivity (*t* = -2.02, p < .05) and flight hours (*t* = 2.00, *p* < .04) were significant (see Figure 1). These results suggest that the number of flight hours influenced positively (*β* = .40) the level of indecisiveness, whereas the level of impulsivity (*β* = - .31) influenced negatively the level of indecisiveness.

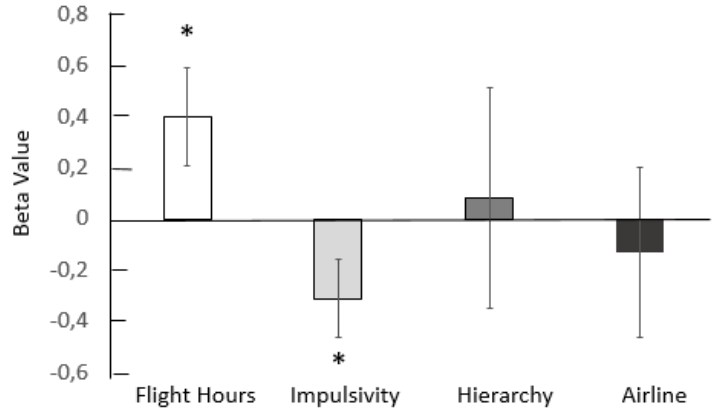


Figure 1. Standardized betas and standard errors for all factors of the model

Logistic regression: encoding self-reported compliance to airline guidelines in real life

A logistic regression was conducted to encode pilots' self-reported compliance to airline guidelines in real life (compliers vs. non-compliers) for 40 airline pilots using flight hours, impulsivity, hierarchy, and prior airline experience as predictors. A test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguished between compliers and non-compliers of airline procedures in real life (χ^2 (3) = 11.47, p < .001, n = 40). Nagelkerke' R square was .364. Prediction success was 64.9 %. The Wald criterion demonstrated that impulsivity (p < .02) made significant contributions to prediction (see Figure 2).

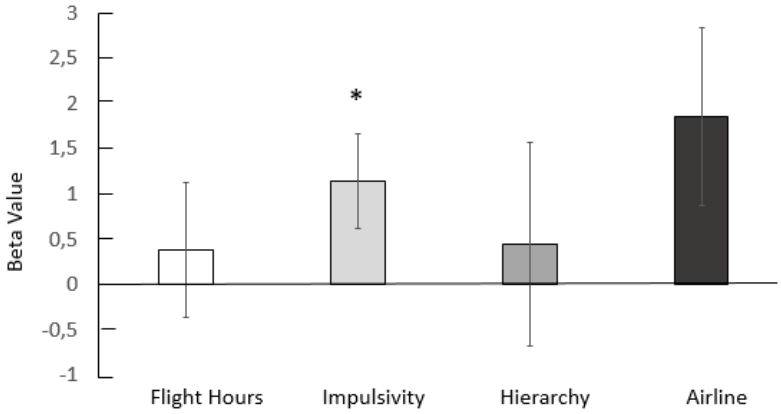


Figure 2. Standardized betas and standard errors for all factors of the model

Discussion

General discussion

The aim of this experiment was to investigate the influence of impulsivity among other factors - flight experience, hierarchy, and prior airline experience - on airline pilot decision making. In line with expert ratings, participants strongly agreed on decisions that were well-defined by airline guidelines. Nevertheless, we explored the existence of response uncertainty in the questionnaire when airline guidelines allowed interpretation: half of the pilots expressed at least once indecisiveness despite the existence of airline guidelines. Pilot experts reported that airline guidelines were theoretically applicable in the questionnaire scenarios. Indecision may be an indicator of (a) evaluation difficulty of the situation and decision complexity due to outcome uncertainty, (b) a lack of information or (c) non-familiarity with decisions (Anderson, 2003; Rassin, 2007). Further; pilot experts emphasized that indecisiveness in a dynamic situation could be described as momentary persistence in the current action plan.

It was pertinent to study the influence of impulsivity as personality factor on pilot decision making. Impulsivity predicted decisions in real life (self-reported compliance to airline guidelines) and decisions in this static questionnaire (uncertain approach scenarios).

Self-reported compliers of airline guidelines in real life were less impulsive than non-compliers. Previous research has shown a link between impulsivity, punishment and reinforcement sensitivity (Gray, 1987; Martin & Potts, 2004). Potts, George, Martin, and Barratt (2005) measured sensitivity to punishment among individuals with low and high impulsivity BIS-11 scores. They found reduced behavioural inhibition among participants with higher impulsivity scores. Martin and Potts (2009) demonstrated in a risky choice paradigm with electroencephalography that low impulsive individuals – in contrast to high impulsive individuals – were more sensitive (i.e. larger error-related negativity) to the consequences of high-risk choices. This is in line with the findings of this experiment. It is possible that self-reported non-compliers of airline guidelines in real life are less sensible to possible punishments of the airline. Qualitative data suggested that non-compliers of airline guidelines in real life reported having taken a decision that was not within guidelines for a positive reason, i.e. in order to avoid a worst-case scenario. The question arises if, in this case, a little bit more impulsivity may be functional. Dickmann (1990) describes functional impulsivity as behaving rapidly with positive outcomes.

The exploratory variables of indecisiveness in the approach scenarios were flight hours and impulsivity. Both factors are independent. This means that impulsivity persists despite intensive training and a highly-procedural environment, whereas flight hours can be acquired. More experienced pilots expressed more indecisiveness than less experienced pilots. Previous research has shown that experience improves performance in aviation studies (Harkey, 1996; Taylor, Kennedy, Noda, & Yesavage, 2007), especially when decision making is concerned (Wiegmann, Goh, & O'Hare, 2002). More experienced pilots recognize the uncertain character of the decision situation and its complexity by delaying their decision. They might aim to

acquire more information in order to make a more appropriate decision. In addition, less impulsive pilots expressed more often their indecisiveness than decisive pilots. Delaying action options is the opposite of making rapid, unplanned decisions, which is positively correlated with self-reported motor impulsivity on the BIS-11 (Baumann & Odum, 2012).

The randomly-chosen airline pilots represent a low impulsive population in comparison to normative data. *Can training and environment modify the influence of personality on decisions?* In a literature review, Baumeister, Gailliot, DeWall, and Oaeten (2006) argue that ego depletion moderates the effect of personality traits on choice behaviour. If an individual's ability to self-regulate behaviour is depleted, desires may have a stronger impact on actions. Therefore, the ability of self-regulation may suppress individual differences in behaviour. Montiglio et al. (2013) emphasize the link between the social context and the prevalence of certain personality traits by the term behavioural flexibility.

Limitations and future research

One limitation is that participants were instructed to make their decisions in a non-dynamic environment. In real life, decision parameters are dynamic and may evolve over time since they depend on pilots' technical skills and actual weather conditions. Importantly, deviation detection of parameters (context updating) is therefore another challenge prior to the actual decision. Thus, pilots had no time restriction for responses and approach decisions were reversible, contrary to dynamic situations. Under time pressure in the real world, potential consequences of their actions may be valued differently as in a questionnaire. Next, this experiment focused on individual decision making under uncertainty. Although this type of decision has a low-procedural interdependence character, i.e. each pilot in the cockpit is allowed to make the decision; at least two pilots are physically present in a cockpit: Both pilots exchange information concerning the decision. Future field studies in a full-flight simulator might confirm the static results by investigating the influence of hierarchy and personality factors on uncertain and dynamic decisions.

Conclusion

Despite the existence of guidelines, the complex selection process of an airline pilot, the intensive training and the highly-procedural environment, a personality factor – impulsivity- mainly accounted for decision making differences among individuals. Impulsivity modulated pilots' indecisiveness in the questionnaire scenarios and pilots' self-reported compliance to airline guidelines in real life. Results emphasize that personality is a profound trait which influence on decision making cannot be removed by intensive training and a highly-procedural environment.

Acknowledgements

This paper originates from an interdisciplinary project that was supported by Air France. We thank Jérôme Rodriguez for technical support. We would also like to express our sincere gratitude to all airline pilots who participated in this research.

References

Anderson, C.J. (2003). The Psychology of Doing Nothing: Forms of Decision Avoidance Result From Reason and Emotion. *Psychological Bulletin, 129*, 139–167.

Baumann, A.A. & Odum, A.L. (2012). Impulsivity, risk taking, and timing. *Behavioral Processes, 90*, 408-14.

Baumeister, R.F., Gailliot, M., DeWall, C.N., & Oaten, M. (2006). Self-regulation and personality: how interventions increase regulatory success, and how depletion moderates the effects of traits on behavior. *Journal of Personality, 74*, 1773-801.

BEA. (2013). *Study of aeroplane state awareness during go-around* (Report: No. FRAN-2013-023). Paris, France: Author.

Bialaszek, W., Gaik, M., McGoun, E., & Zielonka, P. (2015). Impulsive people have a compulsion for immediate gratification – certain or uncertain. *Frontiers in Psychology, 5*, 515.

Byrne, K., Silasi-Mansat, C., & Worthy, D.A. (2015). Who chokes under pressure? The big five personality traits and decision-making under pressure. *Personality and Individual Differences, 74*, 22-28.

Causse, M., Dehais, F., Arexis, M., & Pastor, J. (2011a). Cognitive aging and flight performances in general aviation pilots. *Aging, Neuropsychology, and Cognition, 18*, 544-561.

Causse, M., Dehais, F., & Pastor, J. (2011b). Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *The International Journal of Aviation Psychology, 21*, 217-234.

Causse, M., Dehais, F., Péran, P., Sabatini, U. & Pastor, J. (2013). The effects of emotion on pilot decision-making: A neuroergonomic approach to aviation safety. *Transportation Research Part C: Emerging Technologies 33*, 272-281.

Cheung, A. M., Mitsis, E. M., & Halperin, J. M. (2004). The relationship of behavioral inhibition to executive functions in young adults. *Journal of Clinical and Experimental Neuropsychology, 26*, 393–404.

Childester, T.R., Helmreich, R.L., Gregorich, S.E., & Geis, C.E. (2009). Pilot personality and crew coordination: implications for training and selection. *International Journal of Aviation Psychology, 1*, 25-44.

Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., & Wickens, C.D. (2017). Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: a behavioural and eye-tracking study. *The International Journal of Aerospace Psychology, 27*, 15-28.

Dickmann, S.J. (1990). Functional and dysfunctional impulsivity: personality and cognitive correlates. *Journal of Personality and Social Psychology, 58*, 95-102.

Garavan, H., Ross, T.J., Murphy, K., Roche, R.A.P., & Stein, E.A. (2002). Dissociable executive functions in the dynamic control of behavior: Inhibition, error detection and correction. *Neuroimage, 17*, 1820-1829.

Goh, J. & Wiegmann, D.A. (2001). *An investigation of the factors that contribute pilots' decisions to continue visual flight rules flight into adverse weather.* Proceedings of the Human Factors and Ergonomics Society 45th Annual (pp. 26-29). Santa Monica, CA: Human Factors and Ergonomics Societytime p.

Gray, J. A. (1987). Perspectives on anxiety and impulsivity: A commentary. Journal of Research in Personality, 21, 493–509.

Harkey, J. A. Y. (1996). Age-related changes in selected status variables in general aviation pilots. Transportation Research Record: Journal of the Transportation Research Board, 1517(-1), 37-43.

Helmreich R.L., Merritt A.C., & Wilhelm J.A. (199). The evolution of crew resource management in commercial aviation. *International Journal of Aviation Psychology*, 9, 19–32.

Hirsh J.B., Morisano D., & Peterson J.B. (2008). Delay discounting: Interactions between personality and cognitive ability. *Journal of Research in Personality, 42*, 1646–1650.

IATA (2016). *Unstable Approaches: Risk Mitigation Policies, Procedures and Best Practices (Report* ISBN 978-92-9229-317-8, No. 2). Montreal-Geneva*:* International Air Transport Association.

Koechlin, E. (2016). Prefrontal cortex function and adaptive behavior in complex environments. *Current Opinions in Neurobiology, 37*, 1-6.

Korponay, C., Dentico, D., Kral, T., Ly, M., Kruis, A., Goldman, R., Lutz, A., & Davidson, R.J. (2017). Neurobiological correlates of impulsivity in healthy adults: Lower prefrontal grey matter volume and spontaneous eye-blink rate but greater resting-state functional connectivity in basal ganglia-thalamocortical circuitry. *Neuroimage, 157*, 288-296.

Martin, L & Potts, G. F. (2004). Reward sensitivity in impulsivity. *Cognitive Neuroscience and Neuropsychology, 15*, 1519–1522.

Martin, L. & Potts, G. (2009). Impulsivity in decision-making: An event-related potential investigation. *Personality and Individual Differences, 46*, 303-308.

Maule, A.J. Hockey, G.R., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica, 104*, 283-301.

McCreary, J., Pollard, M., Stevenson, K. & Wilson, M. B. (1998). Human factors: Tenerife revisited. *Journal of Air Transportation World Wide, 3*, 23-32.

Moeller, F.G., Barratt, E.S., Dougherty, D.M., Schmitz, J.M., & Swann, A.C. (2001). Psychiatric aspects of impulsivity. *American Journal of Psychiatry, 158*, 1783-1793.

Montiglio, P.O., Ferrari, C., & Réale, D. (2013). Social niche specialization under constraints: personality, social interactions and environmental heterogeneity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 8*, 20120343.

O'Hare, D. (1997). Cognitive ability determinants of elite pilot performance. *Human Factors, 39*, 540-52.

Patton, J.H., Stanford, M.S., & Barratt, E.S. (1995). Factor structure of the Barratt Impulsiveness scale. *Journal of Clinical Psychology,* 51, 768–764.

Potts, G. F., George, M. R., Martin, L. E., & Barratt, E.S. (2005). Reduced punishment sensitivity in neural systems of behavior monitoring in impulsive individuals. Neuroscience Letters, 397, 130–134.

Rassin, E. (2007). A psychological model of indecisiveness. *The Netherlands Journal of Psychology, 63*, 2–13.

Spinella, M. (2007). Normative data and a short form of the Barratt Impulsiveness Scale. *International Journal of Neuroscience, 117*, 359-368.

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences, 5*, 385-395.

Sutin, A.R. & Costa, P.T. (2010). Reciprocal influences of personaliy and job characteristics across middle adulthood. *Journal of Personality, 78*, 257-288.

Taylor, J., Kennedy, Q., Noda, A., & Yesavage, J. (2007). Pilot age and expertise predict flight simulator performance: A 3-year longitudinal study. Neurology, 68(9), 648.

U.S. Department of Transportation. (2015). *Safety alert for operators, roles and responsibility for PF and PM* (Report No.15011). Washington, DC: Flight Standards Service.

Wiegmann, D., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human Factors, 44*, 189.

Innovative cockpit touch screen HMI design using Direct Manipulation

Marieke Suijkerbuijk, Wilfred Rouwhorst, Ronald Verhoeven, & Roy Arents
Netherlands Aerospace Centre (NLR)
Amsterdam, the Netherlands

Abstract

As a widely-used and proven technology, touchscreens are entering the cockpits of civil aircraft. As part of the project ACROSS (Advanced Cockpit for Reduction Of StreSs and workload), NLR designed an innovative cockpit display with touch interaction for Tactical Flight Control; changing the aircraft's (vertical) speed, heading and/or altitude. In current cockpit configurations, the controls for this auto-pilot (AP) functionality are spatially separated from the visualization of the parameters they adjust, introducing aspects of physical and mental workload. In this paper, the Human Machine Interface (HMI) design process of eliminating this physical gap and creating an intuitive interaction by means of Direct Manipulation (DM) is described. DM is characterized by manipulating graphical objects directly on the position where they are visualized in a manner that at least loosely corresponds to manipulating physical objects. It has the potential to be highly intuitive, and less prone to error. Therefore, the HMI design was hypothesized to reduce pilot's workload and simultaneously increase Situational Awareness (SA). The concept is evaluated using NLR's flight simulators. Experiment results showed that the Tatical Flight Control design concept has great potential, but the interaction implementation needs further improvement, since it increased the pilot's workload, especially under turbulent conditions.

Introduction

This project aimed to research novel technologies to reduce the workload levels for flight crews in civil aircraft. Amongst other things a way to reduce crew workload during tactical flight control is researched.

Tactical flight control

Two methods are discerned for flying an aircraft from point A to point B; *strategic* and *tactical* flight control. In a *strategic approach*, the aircraft is automatically guided along a predefined trajectory between A and B. In a *tactical approach*, the flight crew sets speed, heading, altitude and/or vertical speed to accomplish a desired flight manoeuvre, in general using the auto-pilot (AP) system.

In D. de Waard, F. di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Direct manipulation & touch screens

The term Direct Manipulation (DM) was introduced by Shneiderman (1982). Since then it has been widely adopted as a successful Human Machine Interface (HMI) design style. The idea behind DM is to create a direct intuitive interaction with visually presented objects in a manner that at least loosely corresponds to manipulating physical objects. A well-known example is drag-and-drop functionality in file systems.

The introduction of touch screens enabled a very high sense of DM; a user is able to manipulate visual objects in a way they recognize from the physical world, like moving, resizing and rotating an image with the use of their fingers. Touch screens are not novel technology in everyday's live, but they are in a cockpit; especially for use in main piloting tasks. According to Avsar et al. (2016a, 2016b), Boeing (2016), Gauci et al. (2015) and Gulfstream (n.d.), touch screens are gradually entering the cockpit of business and civil transport aircraft.

In *tactical* flight control the input devices for setting the heading, speed, altitude or vertical speed are spatially separated from the visual representation of the chosen values. An example is shown in the cockpit of an A320, shown in Figure 1 (flight deck picture taken from Meriweather (n.d.)). The pilots use knobs (pushable, pullable and rotatable) on the centre of the glare shield to set a desired speed, heading, altitude or vertical speed. The chosen values are numerically displayed above the knobs. Besides, they are also graphically presented (within the orange indicatory circles) on the two displays in front of the pilot's eyes; the Primary Flight Display (PFD) and the Navigation Display (ND). These are two of the main displays during flight; a pilot continuously scans these displays as they indicate the most important variables for safe flight. Since crew procedures mandate that values inputted using the knobs on the glare shield are visually verified on the PFD and/or ND, the input and output of the AP system have become spatially separated. This creates an additional aspect in mental and physical workload. Using rotary buttons to set target values for the (vertical) speed, heading and altitude has no correspondence to manipulation of physical objects and is therefore not necessarily intuitive. Using DM for this task is hypothesized to reduce workload and simultaneously increase Situational Awareness (SA). NLR has designed and evaluated a touch screen HMI for control of the auto-pilot using DM.

Design phase 1

As a first step in the design, a single pilot crew experiment was set up in NLR's fixed based flight simulator APERO, hosting an Airbus A320-alike aircraft model. With the DM philosophy in mind a solution was searched for manipulating the AP variables at the place where they are graphically presented; the PFD and the ND. Manipulating the variables at the scales leads to complications since the range of the scales is limited. To prevent these complications, and to stay close to the use of the rotary knobs on the glare shield, it has been decided to use interaction wheels.

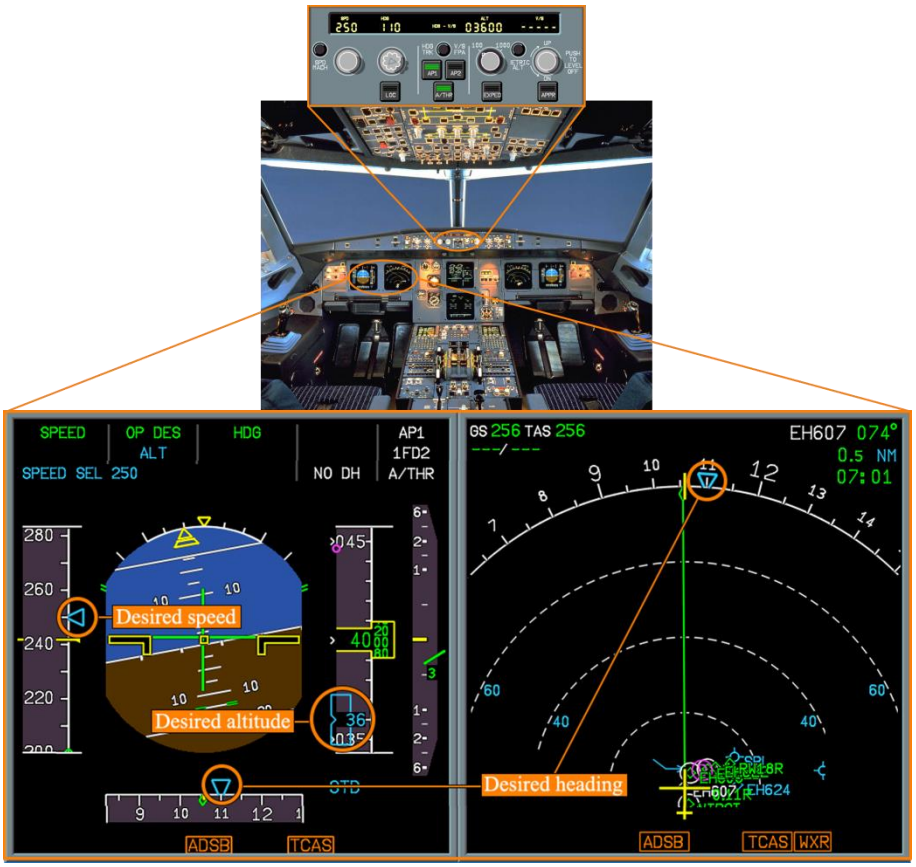


Figure 1. Location of auto-pilot input device (top, on glare shield) and graphical presentation (bottom, in front of pilots)

For every AP input variable to be adjusted, a wheel was created. These wheels were placed next to the graphical presentation of the variable at hand. By dragging the wheels one could adjust the target value; the graphical indicator next to the wheel changed position and value so the user directly got feedback. The wheels were developed in such a way that a gentle dragging resulted in a small adjustment of the value and a swipe resulted in a large adjustment of the value. In this first phase the vertical speed was left out of the design for reasons of simplicity. In Figure 2 the final design of phase 1 is presented. The HMI was displayed on a 10” tablet positioned in front of the pilot, fixated on a stand as shown in Figure 3. As can be seen, the PFD and the ND are also presented just as in the normal cockpit. The AP panel on the glare shield however, was covered. The tablet solely functioned as the input device to the AP system for adjustment of speed, heading and altitude.

Figure 2. HMI design phase 1 using touch wheels

Figure 3. APERO flight simulator with the HMI presented on a tablet

Within the project, besides adjusting the auto-pilot input variables also some other cockpit controls were transferred to the touch HMI. The pilot was able to extend/retract the landing gear, adjust the flaps/slats setting and select an ND-range (in ARC-mode only) via a pinch-zoom gesture. Furthermore a novel interaction function was developed for choosing a new runway on the destination airport (see Rouwhorst et al., 2017). Since on a 10" tablet the amount of pixels is limited, it was decided to let the user decide which display would get emphasis and was magnified in the centre of the tablet. This can be seen in Figure 4; at the bottom of the tablet there are miniatures of the four optional central displays: the PFD, the ND, the gear indicator and the flaps/slats indicator. By touching one of the miniatures, the

selected display was magnified in the centre of the screen. The chosen miniature was highlighted with a green border.

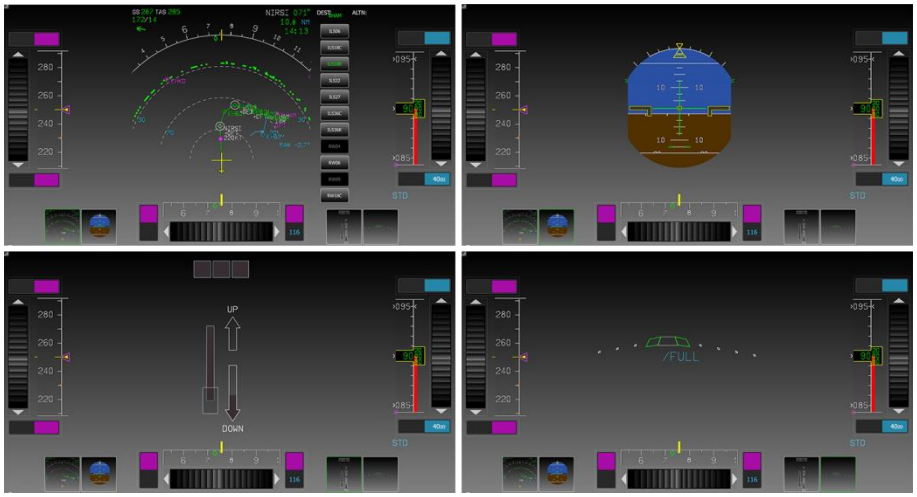


Figure 4. The user decides which display gets emphasized

As can be seen from Figure 4, the three scales for speed, heading, and altitude, which are normally attached around the PFD, were permanently present, no matter what display was chosen to be centrally emphasized. The idea behind this design decision was that the pilot should always be able to adjust the target speed, heading and altitude values, regardless of the active centre display. The three scales were placed near the edges of the touch screen. In this way the operator received some physical support from the tablet hardware; for example he could grab the tablet on the left side and use his thumb to change the target speed value.

With today's use of the knobs on the AP panel on the glare shield not only can the target value of the selected variable be adjusted, the pilot can also decide to change the flight mode from *strategic* to *tactical* and vice versa. In Airbus terminology the two modes are called *managed* (strategic) and *selected* (tactical) mode. In the graphical presentation on the PFD and the ND these two modes are distinguished by colour: *magenta* for managed mode and *cyan* for selected mode. This mode switching functionality had to be transferred to the tablet HMI as well. It was decided to use toggle switches for that purpose: for every wheel two switches were presented (see Figure 2). When the user adjusted the target values in such a way that it fell outside the range of the scale, the value was numerically presented on such a switch above (if higher than the maximum value on the scale) or under (if lower than the minimum value on the scale) the scale. The switches are coloured in correspondence of the active mode. To toggle between modes, the switch could be dragged towards the inactive side (resulting in a colour swap). In the toggle switch design an important rationale was admitted; sliding a toggle towards the centre of the display resulted in a switch towards *strategic* mode. Sliding a toggle outwards resulted in a switch towards *tactical* mode. The idea behind this was copied from the AP-panel of the A320, where a knob push results in *strategic* mode and a pull results

in *tactical* mode. Pushing the knob can be interpreted as giving control to the aircraft (*strategic*) and pulling the knob can be seen as taking control in your own hands (*tactical*). On the tablet HMI sliding the toggles towards the centre display area gave control to the aircraft and sliding it outwards gave control to the user.

Evaluation phase 1

The HMI design was evaluated both subjectively and objectively by ten airline pilots during various descent & approach scenarios. Detailed results on all designed items can be found in Rouwhorst et al. (2017). Here only the main results concerning the new auto-pilot HMI design are presented.

In general, the touch screen as input device was well received. Positive about the design was that building up and maintaining SA appeared to be just as effortless as in a conventional cockpit, and for some pilots even less effortless. This AP tablet HMI design did not lead to a faster, more efficient operation and subjective workload increased with 3 points on a 0-150 Rating Scale Mental Effort (RSME) scale (Zijlstra (1993)). This was caused by the fact that setting a specific value turned out to be rather difficult; swiping the wheel appeared to be too sensitive and correcting this and fine-tuning the value time consuming. The ND-range pinch-zoom interaction felt intuitive and was highly appreciated. The toggle switches for changing the flight control mode were well received.

Design phase 2

To both asses the influence of turbulence on touch screen operation and that of multi-crew operation procedures, the evaluation platform was changed to NLR's full motion simulator GRACE. In Figure 5 the new set-up can be seen, with a seat for both a pilot-flying (PF) and a pilot monitoring (PM). With future expansion of pilot tasks using touch technology in mind, it was decided to switch from a separate tablet to fully integrated touch technology. This was achieved by using three 20" touch screens replacing the cockpit LCDs.

In this phase the interaction design is shifted from the scales of the PFD towards the ND. A trend is seen towards *strategic* flight control. This implies that the role of the ND in the cockpit will become increasingly important. Moreover, the ND would ease the understanding of the flight crew about the consequences of the tactical intervention in terms of SA, since information about terrain, traffic and weather is presented. Finally, with the use of the ND, also a correspondence to manipulation of physical objects can be obtained, which gives DM its intuitive character. To this end, the physics nature of the AP variables was considered. Speed, heading, altitude and vertical speed can be contained in just one physical object: a vector in 3D space, originated at the aircraft (see Figure 6 for a top view and a side view representation of this vector).

Figure 5. Touchscreen HMIs integrated in full-motion flight simulator GRACE

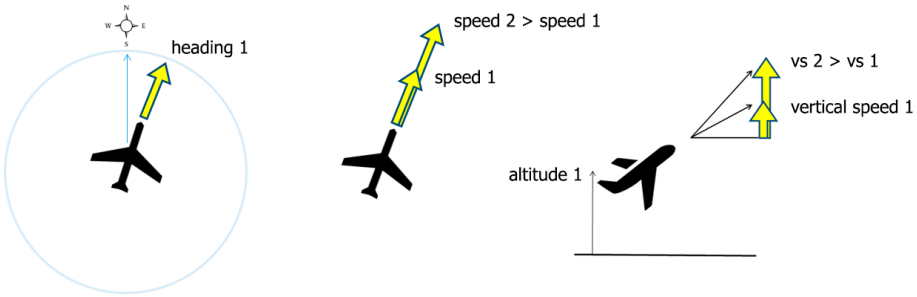


Figure 6. AP variables physics nature as a vector in 3D space

An aircraft has a position and an altitude which determine the origin of this vector. Its speed and heading give the vector length and direction. One can imagine that it is possible to control the behaviour of an aircraft by manipulating this vector; rotate it to change heading, extend/shorten its length to increase/reduce speed, tilt it to adjust its vertical speed to climb or descent towards a new target altitude. This formed the basis of the new HMI design. Since it is difficult to adjust a vector in a 3D space on a 2D screen, it was decided to split it into two views: a top view presented on the ND and a side view presented on a so-called Vertical Situation Display (VSD). This VSD has a similar nature as the ND with information about altitude, vertical speed, distance, terrain, weather and other traffic.

In Figure 7 the final HMI design of this phase is shown including this VSD. Since the mode switches were positively assessed, they are preserved and positioned in between the ND and VSD.

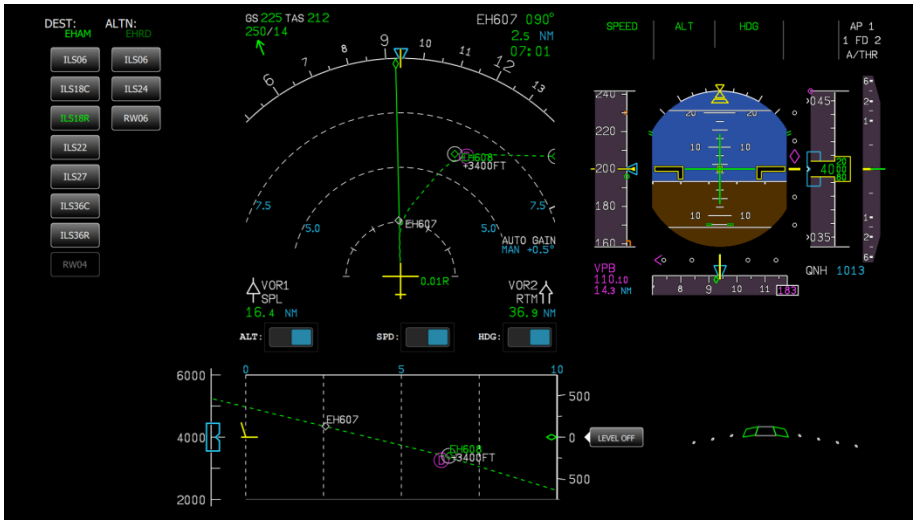


Figure 7. HMI design phase 2 including a VSD below the ND

Heading and speed adjustments

The heading and speed could be adjusted using the ND display. When touching the aircraft symbol on the ND, an interaction screen appeared on top of the ND, while dimming the rest of the display, see Figure 8. The heading rose was extended to a full 360 degrees circle and a speed scale was presented covering all selectable airspeeds. The blue vector originated in the aircraft symbol represented the current target heading and speed physics of the aircraft (in the figure 91 deg and 220 kts (11.32 m/s) respectively).

With this HMI design it was possible to simultaneously adjust the speed and the heading. When touching and holding the tip of the vector (at the position of the yellow arrows indicating the possible interaction directions), the user could drag it around within the heading rose, thereby adjusting both the length (i.e. aircraft speed) and the direction (i.e. aircraft heading) of the vector (see Figure 9). Since Air Traffic Control (ATC) commands can contain a combination of heading and speed, this feature was hypothesized to increase efficiency. When one stayed within the grey circular band, merely the heading was changed. On contrary, when one stayed within the speed scale, merely the speed was changed. For ease of operation and due to the fact that most adjustments do not request a higher accuracy than 5 units, both the speed and heading values were snapped at a multiple of 5 kts (2.57 m/s) or 5 degrees. For fine-tuning purposes both for the speed and the heading values, two buttons were added next to the numerical indication of the adjusted value. Touching these buttons increased/decreased the heading or speed values by 1 degree or 1 knot (0.5144 m/s).

Figure 8.Interaction screen for speed and heading adjustments

Figure 9. Drag the tip of the vector to adjust the aircraft speed and heading

Altitude and vertical speed adjustments

The altitude and vertical speed could be adjusted using the VSD display. When touching the aircraft symbol in the VSD, an interaction screen appeared on top of the VSD, while dimming the rest of the display, see Figure 10. The yellow vector originating from the aircraft symbol represented the physics nature of both the current altitude and vertical speed. The aircraft symbol was positioned at half height of the VSD. The altitude scale on the left is a moving tape, indicating the actual altitude of the aircraft at the aircraft symbol. The blue icon on this altitude scale represented the target altitude; after descending/climbing towards this target altitude, the aircraft will level off when reaching the target value. A pilot could decide to control the climb/descent towards this new altitude by setting a target vertical speed. When flying with a certain vertical speed towards a certain altitude, it can be calculated what distance will be covered before reaching this altitude. These three variables were all presented simultaneously in the VSD. Vertical dotted lines were plotted in the VSD at fixed distance from the aircraft (in this view at every 10 nautical miles (NM) (18.52 km)). Imaginary sloped lines from the aircraft symbol towards the right side of the VSD scale indicated how much altitude could be covered in 40 NM (74.08 km) and therewith represented the vertical speed (indicated in small sloped lines on the right side of the VSD scale).

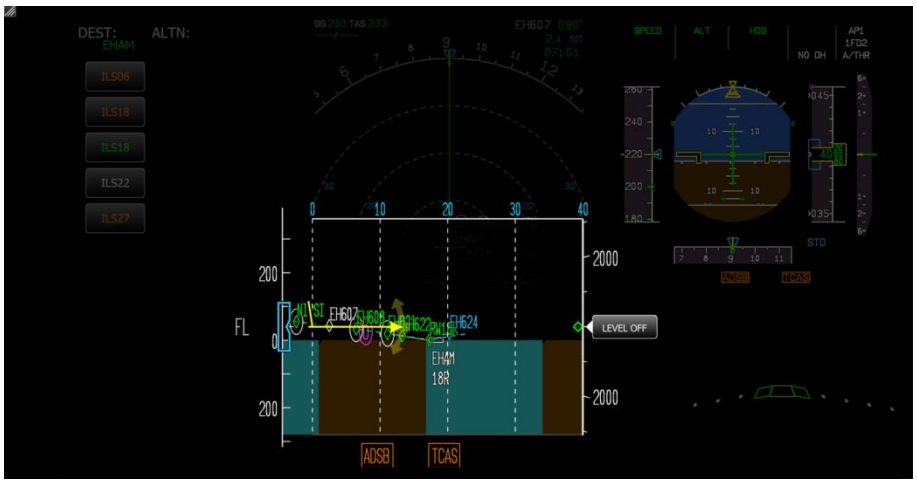


Figure 10. Interaction screen for altitude and vertical speed adjustments

This made the VSD an ideal interaction display for simultaneously adjusting the altitude and the vertical speed. When touching and holding the tip of the vector the user could drag it around within the VSD scale, thereby adjusting the desired distance at which to achieve a certain altitude. For example in Figure 11, where a pilot wanted to climb to an altitude of 15000 ft (4572 m), while climbing at 1500 ft/min (7.62 m/s), he immediately received feedback about the distance the aircraft would need to travel to reach the target altitude at this vertical speed (i.e. almost 30 NM (55.56 km)). If he wanted to reach this altitude earlier, he needed to drag his finger to the left, resulting in a higher vertical speed. This was hypothesized to be a

very intuitive feature, which was not yet available using the AP-panel on the glare shield.

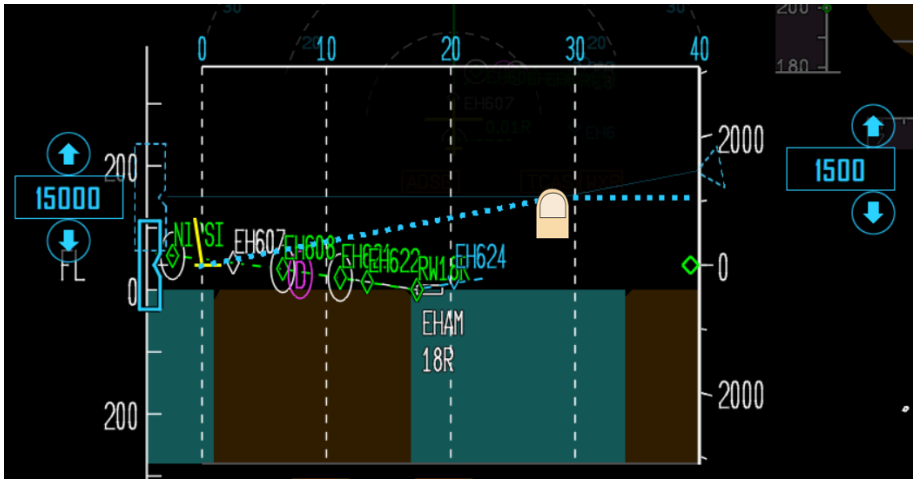


Figure 11. Simultaneously adjusting the target altitude and vertical speed

When a pilot merely wanted to set a target altitude, he could drag the tip of the vector on the left side outside the VSD scale, thereby releasing the vertical speed. The altitude value then could be adjusted by dragging the indicator along the altitude scale or by touching the fine-tuning buttons (with increments of 100/1000 ft (30.48/304.8 m)) depending on the flight phase), see Figure 12.

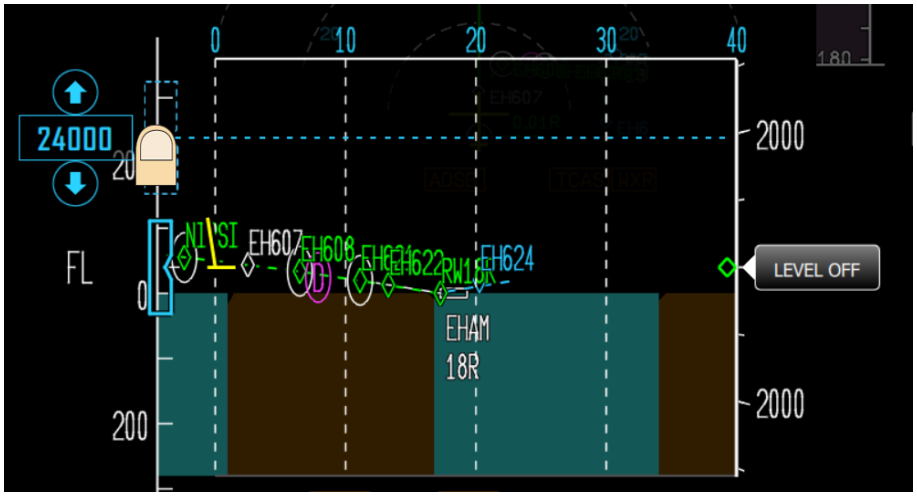


Figure 12. Merely adjusting the altitude using the VSD interaction scale

When a pilot merely wanted to set a target vertical speed, he could drag the tip of the vector on the right side outside the VSD scale, thereby releasing the altitude. The vertical speed value could be adjusted by dragging the indicator along the vertical

speed scale or by touching the fine-tuning buttons (with increments of ±100 ft/min (0.508 m/s)), see Figure 13. To instantly reset the target vertical speed to 0 ft/min (0 m/s) the *Level off* button could be touched (see Figure 12).

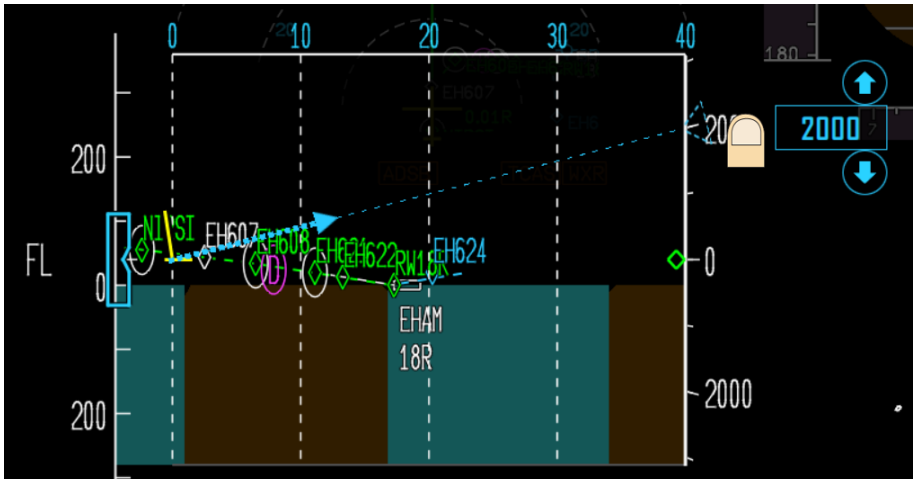


Figure 13. Merely adjusting the vertical speed using the VSD interaction scale

Evaluation phase 2

It has to be noted here, that during the experiment, the auto-pilot functionality was just one of several new things to be evaluated. In Rouwhorst et al. (2017) more detail is presented on some other touch functionalities.

In general, when receiving the briefing and the design thoughts of the new HMI, all pilots supported the idea of setting auto-pilot values in this intuitive way. The idea of not having to switch focus to the glare shield display, the possibility of combined inputs of heading and speed and setting a level-off point graphically were considered promising.

All pilots experienced a steep learning curve. When comparing the touch HMI design to using the rotary knobs of the AP-panel, the pilots reported it as more demanding; it took more effort (the number of actions), time and mental workload. Although beforehand they expected the combined functionality of the speed and heading inputs to be intuitive, most of the pilots advised to decouple them, since it was too hard and time consuming to accurately set them both simultaneously. Pilots had difficulty operating the system under high stress levels, such as turbulence and complex ATC commands, like those that included speed, heading and altitude changes. Such a plural request would require interaction on the ND at first, followed by another interaction on the VSD. This took too much time, and number of operations was too high; pilots tended to forget the actual instruction provided to them. Comments were received on the grey circle band. Since its placement depended on the actual speed, the radius of the band became too small for adjusting the heading when flying at low airspeeds. Pilots liked the graphical representation of the point where the aircraft will level-off to a new set altitude, however controlling

this point with a finger, thereby choosing a value for the altitude and the vertical speed simultaneously appeared to be troublesome. With the use of the VSD, they predicted an expansion of use of the vertical speed mode, since this was received as very intuitive. In terms of multi-crew operation the design appeared to be inadequate; the PM had trouble staying in the loop of what the PF was doing and could not easily verify whether for example instructions received from ATC were properly addressed by the PF. During the experiment therefore a master-slave construction was developed in which the actions of the PF were passively visible on the screen of the PM.

Design Phase 3

Unfortunately, there was not enough time to do a complete third design and piloted evaluation session, but based on the pilot comments and outcome of the experiments, the project allowed final improvements to be made. The most important improvement was the decoupling of the heading & speed and altitude & vertical speed input. Since the grey circular band appeared to be too small to set accurate heading values at low speeds, the heading was decided to be set along the outer ring of the ND arc (see Figure 14).

Figure 14. Interaction screen for decoupled speed and heading adjustments

Another improvement was the addition of the current target speed value as a reference, presented by the yellow line and cyan triangle in the speed tape on the ND. In the previous design, the interaction overlay disappeared automatically after you had stopped adjusting your input. The pilots got confused by this; they lost track of what they were actually doing, had to wait a short while before their inputs were taken over by the aircraft and missed the possibility to reset the entire action. This has been improved by given them control; an "*acknowledge*"- and "*cancel*"-icon are added. When the pilot felt confident about his actions, he could acknowledge them by touching the green check mark.

The same idea is adopted for the VSD, see Figure 15. Also on the VSD the altitude and vertical speed settings were decoupled; adjusting the altitude could be done by dragging the indicator along the altitude scale on the left and adjusting the vertical speed could be done by dragging the indicator along the vertical speed scale on the right. As the pilots had trouble finetuning the target altitude on the small scale it was decided to fixate the value indicator at the vertical centre of the VSD. In the previous design the fine-tuning increments depended on the flight phase; to give the pilots more sense of control, additional fine-tuning buttons were added, for achieving an accuracy of 100ft (30.48 m) as well as 1000ft (304.8 m). Because the pilots liked the feature of knowing where the level-off point is situated, this is preserved as a dashed bold line (so in the example in Figure 15, when climbing at 2400 ft/min (12.19 m/s) to a target altitude of 13200 ft (4023.4 m), the level off point was situated at a range of 15 NM (27.28 km)).

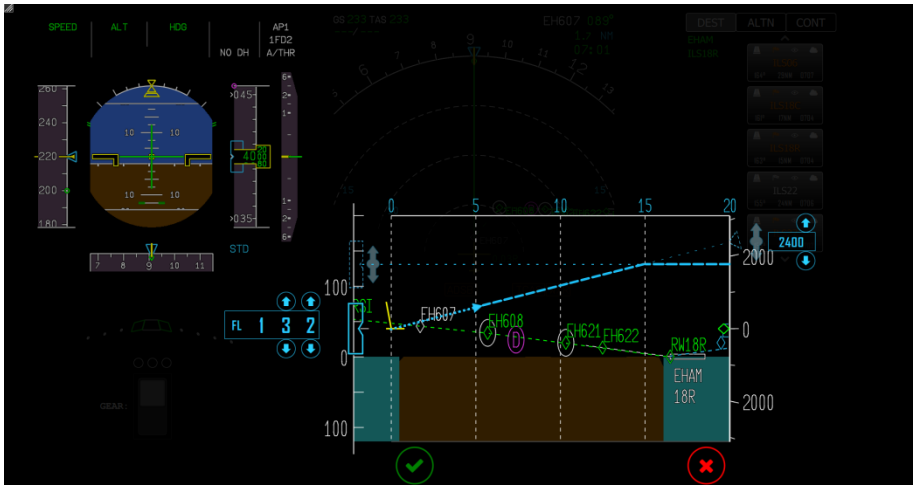


Figure 15. Interaction screen for decoupled speed and heading adjustments

Discussion

It can be concluded that the touch screen itself has great potential, and other functionalities evaluated in the project were already very well received (see also Rouwhorst et al. (2017)). For the task of Tactical Flight Control it can be concluded that the design concept was well received and has the potential to increase SA, but there is room for improvement of the interaction implementation on the HMI. Only with extensive iterative testing and evaluating a complex HMI such as the present design can be fine-tuned to be an impeccable system. As a first step for further research the HMI design of phase 3 could be evaluated in a pilot-in-the-loop experiment. A solution should be found for dealing with turbulence when using a touch screen. It is unlikely that the HMI design concept will reduce workload when solely comparing the Tactial Flight Control task with the conventional AP knobs functionality. It has however great potential to increase SA and be part of a full blown touch cockpit. Such integrated touch cockpit has the potential to reduce overall workload levels. This research can be seen an important step towards this

future touch cockpit, but more iteration cycles are needed on the HMI interaction design.

Acknowledgment

The project was co-funded by the EU in the 7th Frame Work Program under contract number ACP2-GA-2012-314501. The authors would like to thank all the participating pilots and involved ACROSS partners.

References

Avsar, H., Fischer, J.E., Rodden, T. (2016a). Designing Touch Screen User Interfaces for Future Flight Deck Operations. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th* (pp. 1-9). New York, USA: IEEE Institute of Electrical and Electronics Engineers.

Avsar, H., Fischer, J.E., Rodden, T. (2016b). Mixed method approach in designing flight decks with touch screens: A framework. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th* (pp. 1-10). New York, USA: IEEE Institute of Electrical and Electronics Engineers.

Boeing (2016), *Touchscreen come to the 777X Flight Deck, bringing today's technology in the hands of pilots*, http://www.boeing.com/features/2016/07/777x-touchscreen-07-16.page

Gauci, J., Cauchi, N., Theuma, K., Zammit-Mangion, D. and Muscat, A. (2015). Design and evaluation of a touch screen concept for pilot interaction with avionics systems. In *Digital Avionics Systems Conference (DASC), 2015 IEEE/AIAA 34th* (pp. 3C2-1 - 3C2-19). New York, USA: IEEE Institute of Electrical and Electronics Engineers.

Gulfstream (n.d.), *Gulfstream symmetry flight deck, Piloting Perfected*, http://www.gulfstream.com/technology/symmetry-flight-deck

Meriweather (n.d.), A320 Flight Deck by Jerome Meriweather, http://meriweather.com/flightdeck/320/fd-320.html

Rouwhorst, W.F.J.A, Verhoeven, R, Suijkerbuijk, H.C.H., & Arents, R. (2017). Use of Touch Screen Display Applications for Aircraft Flight Control. In *Digital Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th Sept. 2017.* New York, USA: IEEE Institute of Electrical and Electronics Engineers

Shneiderman, B. (1983). Direct manipulation: A step beyond programming languages. In *Computer, Volume: 16, Issue: 8* (pp. 57 – 69). Los Alamitos, CA, USA: IEEE Computer Society Press.

Zijlstra, F.R.H. (1993). *Efficiency in work behavior. A design approach for modern tools.* PhD-thesis, University of Delft, The Netherlands, University of Delft: Delft University Press

Assessment of stress sources and moderators among analysts in a cyber-attack simulation context

Stéphane Deline, Laurent Guillet, Clément Guérin, & Philippe Rauffet
University of South Brittany
France

Abstract

With the prominence of cybersecurity questions, the role of analysts in managing cyber-attacks is crucial. Studies investigating human factors in cyber defence context generally focus on analyst training, situation awareness or cognitive biases (e.g. Gutzwiller et al., 2015) in order to reduce analyst errors. Champion and collaborators (2012) showed that social factors such as team communication influence the cyber teamwork. In this present study, we have examined elements contributing to the analyst's stress level. More precisely, we have studied the effects of cyber threats and the moderator effects of social support on analyst stress. We venture the hypothesis that 1) cyber-threats have an impact on stress levels and 2) social support reduce individual stress levels. This study has taken place in a cyber-security centre where cyber-attacks on a Vital Organisation have been simulated with engineer-students as cyber-defenders. Stress levels have been measured according to their heart frequency, and social communications have been coded from the video. Results show that threats do not directly affect stress, whereas obtaining -informational - social support is associated with a decrease of stress level.

Introduction

Many organisations (industrial firms, financial institutions, public administrations, companies operating in the fields of defence and energy or more generally organisations depending on the use of computer data or internet) have a critical need to protect against cyber-attacks. In order to ensure the security of their information system, these organisations need to get protecting against data theft and alteration. With the increasing number of cyber-crimes, it is essential to identify factors improving effectiveness and efficiency of cyber defenders. These operators have to assess how serious the situation is swiftly, identify priorities and make relevant decisions. The strong pressure (time pressure, high risk) felt by these operators can generate significant stress and therefore impact their performances. In this context, we will observe how the cyber team operates and is managed during cyber-attacks.

The aim of this exploratory study is to examine the effects of cyber events and the moderator effects of social support on stress level in cyber-attack simulation. This study takes place in a cyber-security centre where cyber-attacks on a Vital Organisation have been simulated with engineer-students as cyber-defenders.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Theoretical framework

In the literature, Champion, Rajivan, Cooke, & Jariwala (2012) suggest that a cybersecurity analyst team can be characterized as a group of individuals working independently with few communication or collaborative efforts among team members. They identified three major factors impacting teamwork: overall organisation of the team, team communication and information overload. Some authors (Gutzwiller, Fugate, Sawyer, & Hancock, 2015; Champion & al., 2012) focused on situation awareness in cyber defence context but not on stress processes.

With regard to stress, several models exist. This study uses the Lazarus (1984, 1999) transactional model of stress. Stress occurs when person/environment transactions lead the individual to perceive a discrepancy between the situational demands and her/his resources or abilities to cope with those demands. The nature and type of coping generated by a person will be determined by the coping resources in the personal environment. The model identifies four types of coping resources: individual resources, social support, beliefs, and problem solving skills. Granåsen & Anderson (2015) explore the within-team communication in a cyber-attack situation to understand and get knowledge on team effectiveness in cyber defence exercises without taking social support into account. Our study aims to assess the relationship of social support on stress level. The social support will be measured through the communications. We focus on social support which is considered as a major moderator. Indeed, Kaufmann & Beehr (1986) suggest that positive communications might buffer individual occupational stress, while negative communications might have a reverse buffering effect.

According to House (1981), social support is defined as a positive resource that a person can use to cope with stressful situation. House (1981) distinguished four types of social support:

- Emotional support consists in expressing to a person the positive affect that one feels towards her (friendship, love, comfort, sympathy), and generating feelings of reassurance, protection or comfort.
- Appreciation support is about reassuring a person in terms of skills and values. This encouragement will allow her to strengthen her self-confidence in times of doubt when she is concerned that the demands of a situation will exceed her resources and capacities (overwork, role conflict, burnout ...).
- Informative support involves advice, suggestions, knowledge about a problem, proposals for solving a new problem, for example.
- Instrumental support involves effective assistance such as lending or giving of money or tangible goods or providing services in difficult times. It also characterizes assistance in the form of donating time or work.

Frese (1999) shows that social support buffers the effect of stressors on health. Buffer effects in the relationships between stressors and psychological or psychosomatic dysfunctioning are higher when social support is high and lower when social support is low. Malviya, Fink, Sego, & Endicott-Popovsky (2011) aim to determine whether situational awareness of team members participating in a cyber-competition could predict the overall team’s score. Various data were

collected (e-mail, machine logs, video and audio sources), and they suggest supplementary data sources such as physiological stress measurements should be introduced in order to complete their research. The stress can be measured with heart frequency in dynamic situation (e.g. in driving, Healey & Picard, 2005). To our knowledge, no studies have been conducted on the assessment of stress in cyber defence situations, and none involves the study of heart frequency.

The objective of this research is to explore the effects of cyber events and the moderator effects of social support on analyst stress. We venture the hypothesis that (H1) cyber-threat have an impact on stress levels, and that (H2) social support reduces individual stress levels. We have designed a methodology to record all communications during a simulated cyber defence exercise, focusing on social support communications. The different situations of cyber-threats are then studied with regard to the potential stress generated. This stress is measured through heart rate and matched with social support.

Method

The cyber context

The study was carried out in the Cyber Security Centre (CSC) of a Higher National Engineering School in France. Cyber-attacks on a VO (Vital Organisation, e.g. an energy company) information system can be simulated. In order to be more realistic, a scenario of a hospital attack is worked out: following a series of triggering events (planned by the author's scenario), the repercussions of these events on the hospital and its environment were simulated (e.g., social conflicts) associated with a series of cyberattacks (e.g. DDOS, Defacement). These attacks could occur at any time during the day. The operators had no information on the development of the scenario and they had to resolve the situation using their defence skills. Several cells were constituted for the exercise management. The Cyber cell, called the Blue Team, constitutes the SOC (Security Operational Centre) in charge of the organisational security. There is a Management Team making choices and confirming the decisions, the Red Team launching the cyber-attacks, an Animation Team regulating depending on the sequence of events and a White Cell for interacting with the media and providing potential reinforcement.

Participants

The sample is composed of 29 graduate students from the engineering school of the University of Southern Brittany in France. They are aged from 21 years to 32 years (mean age = 23.,93 years, standard deviation = 2.62). The sample is composed of 1 female student and 28 male students. The participants' anonymity is guaranteed and a request for consent was signed by the participants fifteen days before the experiments.

Data collection

Observations and measurements of activity were made during an exercise of cyber defence simulation training. The Cyber Crisis Exercise took place over five days in

February 2017. Spread out in four teams, each team was placed alternately in the Cyber Security Centre in a cyber-crisis situation. In this study, the team focussed on is the Blue Team. It consisted in 6 operators and a Real-Time Coordinator (CTR) responsible for coordinating crisis management operations. They had to deal with threats and attacks that could damage the system of the simulated VO. A measurement of the heart rate (HR) of the team members was performed using a BioHarness3 ™ heart rate monitor, three days before the exercise for HR baseline calibration, and continuously, throughout each exercise day for physiological stress measure. Two cameras recorded participants continuously during the cyber-crisis exercise. Communication were recorded using microphones and dictaphones all along. All events, communication and activities (e.g. movements) were coded according to a coding scheme. The coding scheme was designed to identify cyber events and social support verbalisation.

The coding procedure was focused on taking the oral communications of the Blue Team into account. The coding scheme on social support was carried out in four steps: 1) Identification Sender (CTR or operator); 2) Purpose of social support (contribution, expectation or proposal); 3) Identification of the recipient (CTR or operator); 4) Qualification of the type of support (instrumental, informational, emotional, appreciation). For example, in the case of a social support contribution (SSC) from the CTR, SSC could take one of the four types of support cited above.

Proposition, expectation and contribution social support behaviours had been coded but for the illustration, the focus was on contribution. An example of each kind of contribution is presented in Table 1.

Table 1. Illustration of types of social support contribution (SSC) during the cyber exercise.

CONTEXT DESCRIPTION	COMMUNICATION (Op = Operator / CTR = Coordinator in real-time)	SOCIAL SUPPORT CONTRIBUTION TYPES
Op1 uses a tool that he doesn't know very well and asks Op2 what he should do.	Op1 "What should I do in the software **Op2 "you click here** (showing directly with the mouse of the Op1's PC)".	Instrumental
Op1 doesn't know how to do an analysis task. Op2 responds.	**Op2 "For the UFW you do a 'Ptinstall', you look for the configurations Apt in each machine.**	Informational
The CTR has made a request for reinforcement. The reinforcement arrives. Here is the reaction of the CTR when he understands who the reinforcement is.	**CTR "Oh that's a great gift"** [to have this person]	Emotional
CTR follows an Op task and encourages him in his task.	CTR "how is it going?" Op1 "I have made back-ups on my PC, in case the machine gets attacked." **CTR "It's good okay".**	Appreciation

Heart rate analysis and Controlled variables

In the analysis of heart rate (HR) variations, we decided to ignore a behaviour interval of 20s around the behaviour in order to exclude from the data, HR variation caused by the communications induced by the behaviour. When a studied behaviour occurred, we compared mean HR during 20s before (Interval 1) and 20s after the SSC behaviour interval (Interval 2) (see Figure 1).

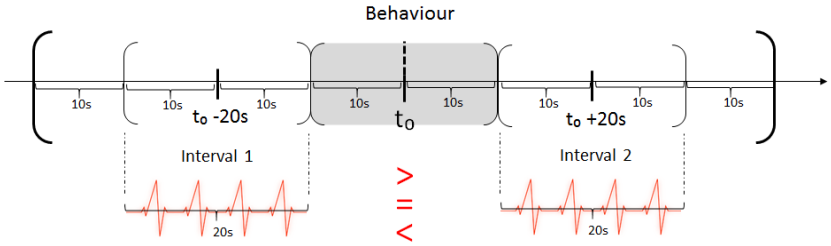


Figure 1. Heart rate comparison method.

Physical activity (standing or sitting position) was controlled around interval 1 & 2 to limit ecological context bias. The HR is known to be sensitive and slowly decreasing, so we controlled the activity of the individual in order to limit activity influence on HR. If the operator was not seated during a fixed interval (40 seconds before and after a studied behaviour), the behaviour was excluded from the analysis.

Results

The result section presents the effects of threatening cyber events on individuals and the effects of social support on individuals HR.

Threatening cyber events

In this part, the communications on threatening cyber events that contribute to collective representation among the cyber team is analysed. These events are: detection of threats, detection of attacks, or more generally additional information contributions on such events. It was found that 66 occurrences of these threatening cyber events were coded. After activity control (standing versus sitting position), only 30 occurrences of threatening cyber events were taken into account. The boxplot depicted in Figure 2 shows the variation of HR during these occurrences.

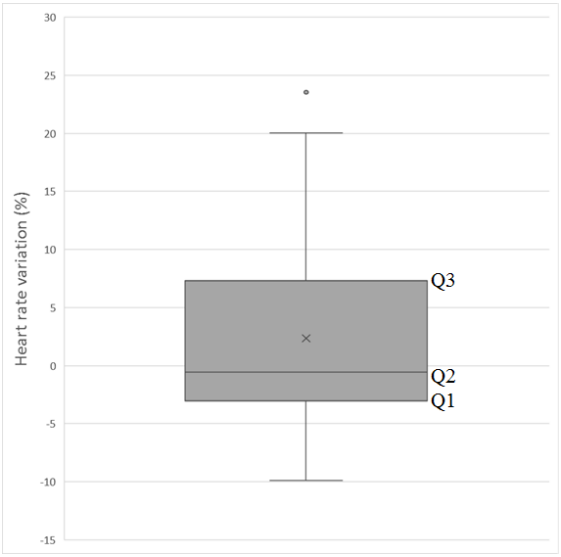


Figure 2. Variation of heart rate with threatening cyber events. Q1 = Quartile 1, Q2 = Median, Q3 = Quartile 3, × = Mean.

When comparing before and after the onset of threatening cyber events, no effect was found on HR (Mean = + 2.34 %; NS with a Wilcoxon paired test, n = 30). To illustrate, Figure 2 presents each individual HR pattern variation between interval 1 & 2 (see), depending on the criticality level of the cyber-attack (level 1: low hazard & low recovery, level 2: medium hazard & medium recovery, level 3: high hazard & high recovery; from a cyber-subject matter expert) in the simulation context.

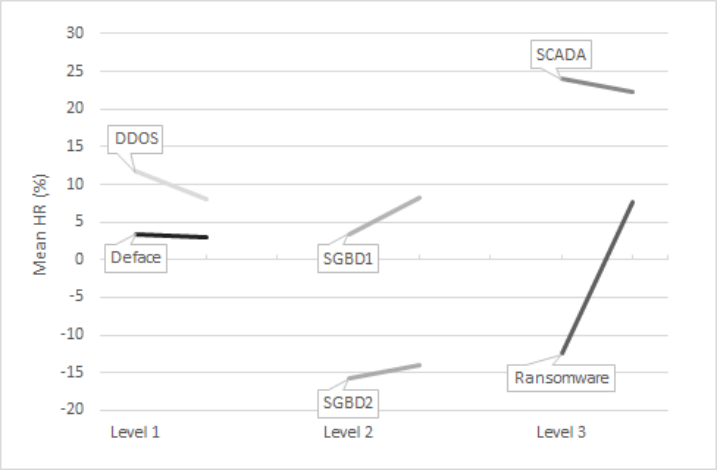


Figure 3. Heart rate variations depending on the criticality level of cyber-attacks (level 1 to level 3; n = 6).

The different individual patterns observed (Figure 3) suggested that attack criticality influences the individual reaction. When the attack criticality level was low (level 1), the 2 individual patterns did not indicate an increase of HR. However, when the attack criticality was higher (Level 2 or 3), pattern show an increase of HR except for the SCADA attack. An explanation of the decrease of HR for SCADA attack is that the individual already had a high HR before the attack, and so an increase of HR was less likely.

Social support

In this part, the social support behaviours were investigated. In total, 320 occurrences of social support behaviours were coded from communication. The most prevalent social support behaviours coded were contributions (n = 211) then expectations (n = 89) and finally social support propositions (n = 20). Among the social support contributions (SSC), the most frequent SSC was informational social support contribution (n = 144), then instrumental (n = 32), appreciation (n = 20) and finally emotional (n = 15).

In the following, we focused on the SSC, and analysed the variation of HR with its occurrences. In accordance with our second hypothesis, the analysis indicated a decrease of HR following a SSC (Mean = -3.410 %; W = 4.367; $p < 0.001$, with a Wilcoxon paired test, n = 117) compared to before the SSC.

In the boxplot depicted in Figure 4, the HR variation depending on SSC types are presented.

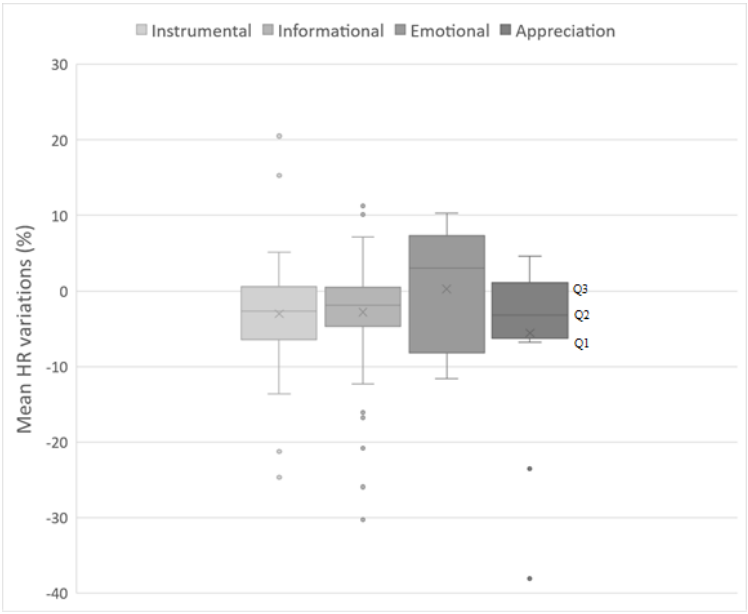


Figure 4. HR variations depending on types of social support contribution; Q1 = Quartile 1, Q2 = Median, Q3 = Quartile 3, × = Mean.

We differentiated SSC depending on the central tendency and the homogeneity of HR variations after these SSC. All SSC -except emotional- were associated with a HR median variation from -1 to -5% and an interquartile interval from 1 to -6 %. These distributions indicated that HR was stable or decreased for 75 % of SSC occurrences. The analysis of HR following a SSC showed a significant difference for informational SSC (mean = -2.886 %; $t = 3.634$, $p < 0.003$, $n = 78$) but not for instrumental SSC (Mean = -3.146 %; NS; $n = 19$), emotional SSC (Mean = 0.250 %; NS; $n = 5$) and appreciation SSC (Mean = -5.583 %; $W = 2.166$; $p < 0.07$; $n = 15$). These results are in accordance with our hypothesis of an effect of SSC on HR, depending on the type of SSC.

Discussion

The current study was conducted to investigate stressors (cyber-threats or cyber-attacks) and stress moderators in cyber context.

Cyber stressors

Our hypothesis that threatening cyber events increase stress among cyber operators is not verified. As Figure 3 suggests, the increase tendency of stress with cyber-attacks seems to be influenced by individual features and potential level of criticality of cyber events. First, as the Lazarus stress model (1984, 1999) proposes, it can depend on the individual personality which can be identified with appropriate questionnaires. Operators can be differently affected depending on their anxiety or current stress level. On one hand, if they are not anxious or engaged in their task, their stress level will be more stable. On the other hand, if they are already stressed, an increase of stress level is less likely. Such questionnaires have been administered and constitute the next phase of our research. Secondly, the specificity of the context which is ecological but not a real one, could affect their stress level. The fact that the exercise is a simulation can reduce impact of attacks compared to real live attacks, even if the exercise was also an evaluation of individual cyber defence abilities. Thirdly, the operators are trained to defend information systems from attacks, and with expertise, they have to manage stress during cyber events. So probably, threatening events are not the most important stressor, other factors, as validating countermeasure or making management decisions (e.g. to disconnect website) could be more stressful and constitute an interesting perspective for study.

Heart rate variation with social support contribution

In a general manner, social support contribution is associated with a decrease of HR. It provides an additional argument to the Lazarus model of stress (1984, 1999) that SSC is a relevant moderator of stress. It is also in accordance with the hypothesis of positive or negative communication buffering effect (Kaufmann & Beehr, 1986) suggesting that positive communication might buffer individual occupational stress. However, social support's contribution has different effects on stress depending on social support contribution types. Informational social support contribution reduces stress whereas the other types do not. More surprising is that instrumental social support contribution does not influence heart rate despite its tangible feature. One explanation is that a relevant tangible social support contribution can be very useful

for the operator's task, inducing new cognitive tasks and so potentially additional mental workload. It could be interesting to insert a combination of measures to control the impact of mental workload on heart rate. Moreover, appreciation social support contribution tendentially reduces heart rate which means that when an operator is stressed, discouraged or submerged, encouraging him could have a positive effect on stress. Under stressful conditions like a cyber-crisis context, encouraging or helping collaborators in need, could have an important effect on stress and so contribute to team cohesion. In a performance perspective, such a vector of team cohesion and stress could be an interesting way to optimize team functioning.

Conclusion

The study shows that in a cyber-attack context, threatening events may not be sources of stress but the attack criticality could be. Moreover, the social support contribution -and more specifically informational contribution- seems to moderate the stress level. It would be interesting to continue the investigation on stressors and moderators in a cyber defence context with a combination of stress and mental workload tools, in order to dissociate their respective influence on heart rate.

Acknowledgments

Thanks to the Cyber Centre of Excellence, the Brittany region and the University of South Brittany.

References

Champion, M.A., Rajivan, P., Cooke, N.J., & Jariwala, S. (2012, March). Team-based cyber defense analysis. *In Cognitive Methods in Situation Awareness and Decision Support (CogSIMA),* 2012 IEEE International Multi-Disciplinary Conference (pp. 218-221).

Frese, M. (1999). Social Support as a Moderator of the Relationship Between Work stressors and Psychological Dysfunctioning: A Longitudinal Study With Objective Measures. *Journal of occupational health psychology, 4,* 179-192.

Granåsen, M. & Andersson D. (2015). Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. *Cognition, Technology & Work, 18*, 121-143.

Gutzwiller, R.S., Fugate, S., Sawyer,B.D., Hancock. P.A. (2015). The Human Factors of Cyber Network Defense. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59, 1, 322-326.

Healey, J.A. & Picard, R.W. (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. in IEEE Transactions on Intelligent Transportation Systems, vol 6, 2, 156-165.

House, J.S. (1981). *Work and stress and social support*. Reading, Mass: Addison-Wesley.

Kaufmann, G.M., & Beehr, T.A. (1986). Interactions between job stressors and social support: Some counterintuitive results. *Journal of Applied Psychology*, *71*, 522-526.

Lazarus, R.S., Folkman, S. (1984). Stress, appraisal, and coping. New York, Springer Publishing Company.

Lazarus, R. S. (1999). Stress and emotion: a new synthesis. New York, US: Springer Publishing Co.

Malviya, A., Fink, G.A., Sego, L., & Endicott-Popovsky, B. (2011). Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work. Proceedings - 2011 8th International Conference on Information Technology: New Generations, ITNG 2011. 937-942

Potential of wearable devices for mental workload detection in different physiological activity conditions

Franziska Schmalfuß, Sebastian Mach, Kim Klüber, Bettina Habelt, Matthias Beggiato, André Körner, & Josef F. Krems
Chemnitz University of Technology
Germany

Abstract

Wearable devices have gained high popularity in the last years, especially for health monitoring. Some devices aim at identifying mental states, but scientific studies on the potential of wearable devices for identifying mental states are rather sparse. Heart rate parameters proved to be valuable indicators for increasing mental workload and growing levels of physical activity. The question arises, if wearable devices can be used to identify high mental workload in different physiological activity conditions. Thirty-two participants (18 female) participated in an experiment with a 2 (mental workload) x 4 (physiological activity) factorial within-subject design. Participants sat, stood, stepped or cycled while they fulfilled either no secondary task (5 minutes) or a counting backwards task (5 minutes). Heart Rate was measured via a wrist-worn mobile device and a stationary device. Results showed that measurements of the two devices did not correlate consistently. Heart Rate and Inter-Beat Intervals, measured via the stationary device differed significantly with varying levels of physical activity and mental workload. Data from the wearable device showed only the physical activity effect. Findings indicate that wearable devices are not fully capable of identifying mental workload. Still, wearable devices have potential for identifying and fostering reduction of high physical load in everyday usage.

Introduction

The market share of wrist-worn wearable devices is on the rise (IDC, 2017). This shows the high popularity this new technology has gained in the last years. Their potential for health monitoring and health support has been intensively investigated and discussed (e.g., Marakhimov & Joo, 2017). They allow consumers to continuously monitor physiological parameters and manage their health and well-being on a personal basis. Additionally, they can help physicians to get access to their clients' health data to offer personalized medical care (e.g., Kim & Kim, 2016). Some devices even aim to identify mental states, stress or emotions, but scientific studies on the potential of wearable devices for identifying different mental states in different situations are rather sparse.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

The potential identification of high workload or stress opens new opportunities for daily life usage. Wrist-worn wearable devices might be applied in the driving or working context in order to identify situations in which users need support. For instance, workers' health parameters could be tracked in order to implement solutions that respond to the observed health status and reduce the physical and cognitive burden at work (Lavallière et al., 2016). Through technical solutions using wearable devices, personal recommendations can be made about the sequence of the pending work tasks, exercise (e.g., daily step count), nutrition, or practices to reduce stress and optimize job-related (mental) workload (Swan, 2012).

Mental Workload and Heart Rate Variability (HRV)

Mental workload can be described as the relationship between the cognitive resources that are necessary to fulfil a specific task and the operator's cognitive resources that are available (e.g., Wickens, 2008). Valuable physiological indicators for increased mental workload and growing levels of physical activity are an increase in Heart Rate (HR) and decrease in Heart Rate Variability (HRV; Mulder, 1992; De Waard & Brookhuis, 1991). HR presents the number of heart beats per minute and Heart Rate Variability is defined as the variability of the intervals between two heart beats, the Inter-Beat Interval (IBI; for a comprehensive overview see, Shaffer et al., 2014). HR and HRV can be measured via electrocardiogram (ECG) by recording the electrical activity from the heart or wearable devices that often use optical heart rate monitors. HRV parameters can either be time-domain or frequency-domain parameters. Typical time-domain parameters that are supposed to be indicators for parasympathetic nervous system (PNS) activity are the standard deviation of the RR-intervals (SDNN), root mean square of sequential deviations (RMSSD), and the number of adjacent pairs of IBIs differing more than 50 ms divided by the total number of Inter-Beat Intervals (pNN50). Commonly used frequency-domain parameters are power of the high-frequency band (HF: 0.15-0.40 Hz) and the low-frequency band (LF: 0.04-0.15 Hz) and the LF/HF ratio. Sympathetic nervous system activities as reaction to physical activity or stress reactions should reflect in LF (Shaffer et al., 2014). The LF/HF ratio is considered as a marker for shifts in sympathetic or parasympathetic dominance. The 0.10 Hz component that corresponds to the LF component is supposed to be especially sensitive to changes in mental demand (De Waard, 1996), but often all above mentioned parameters are analysed (Hsu et al., 2015).

Potential of wearable devices for heart rate monitoring

Wearable devices represent an easy-to-use alternative to measure HR parameters in daily context. Stahl et al. (2016) showed that measures of different wearable devices such as Mio Alpha, Microsoft Band and Fitbit Charge HR correlate highly with the criterion measure (Polar RS400) and with each other, even when people walk or run. Another study showed relatively high error rates for walking, but more acceptable error rates for cycling and running (Shcherbina et al., 2017). Furthermore, wearable devices proved satisfying HRV measurements in order to differentiate between high and low demanding cognitive tasks (Barber et al., 2017), although other studies showed that HRV parameters could be too inaccurate (Reinerman-Jones et al., 2017).

On the basis of the reviewed literature, we expect that HR increases and HRV parameter decreases when mental demand is higher (H1). Furthermore, higher physical demand should reflect in higher HR and lower HRV (H2). It is further assumed that these effects can be detected using an ECG and a wearable device. Sun et al. (2012) showed that higher mental workload can be identified using HR parameters in different physiological activity conditions, but it has not been investigated whether this is replicable with using wearable devices.

Methods

Participants

Thirty-two healthy participants finished the experiment. One data set could not be used due to technical problems. The remaining n = 31 participants (18 female, 13 male) were on average 25 years old (SD = 5.5), 87% were right-handed and none of them had diagnosed diabetes, cardiovascular complaints or diseases or other health issues that would constitute a risk for participants in the study. Students (n = 29) received course credits for participation.

Design

In an experiment with a 2 (mental workload) x 4 (activity) factorial within-subject design, participants' HR and HRV was assessed in each condition. They either sat, stood, stepped, or cycled while solving an arithmetic task in parallel or doing nothing additional. The sequence of activities was varied using the latin square. The study procedure was approved by the ethical committee of the Chemnitz University of Technology (no. V-163-BM-FS-Factory-24112016).

Apparatus and material

HR and HRV were measured 1) with the Microsoft Band 2 (MB2) on the non-dominant hand and 2) with a 1-channel ECG, the SUEmpathy® (SUE) with disposable adhesive electrodes (Dahlhausen type 405, Ag/AgC1; 45 mm diameter) positioned on the abdomen and chest area. The SUEmpathy100 is a measuring device for the functional diagnostics of the autonomous nervous system of the company SUESS Medizintechnik ECG 1303, SUEmpathy® Vitalbox, SUESS Medizin-Technik Aue). The data from the ECG sensor were recorded at 512 Hz with 12-bit resolution, and the Microsoft Band 2 data at 1 Hz. The ECG recordings were pre-analysed with the associated software SUEmpathy100, version SUE1-4.36j Scientific (SUESS Medizin-Technik Aue, 2009). The Windows software development kit (SDK) coming with the MB2 allows for real-time data streaming via Bluetooth between the device and a computer. Therefore, a self-developed logging application was installed on a Lenovo notebook. For analyses, the collected HR and IBIs were of interest.

All instructions for physical and arithmetic tasks were presented via LabView (version 2014). In the arithmetic task, that was used to increase the mental workload, participants were asked to count backwards from 5,200 by, for instance, 13 (similar to Meinel, 2013). In the stepping condition, participants stepped with both legs on

and off a step board (height: 20.5 cm x length: 89 cm x width: 35.5 cm) using one leg at a time and following an 80 bpm beat given by a metronome (Yixiang, 2015). In the cycling condition, participants cycled on a bike fixated with a roller fix frame (In'Ride 300, 550 Watt B'TWIN) and followed a 90 bpm beat. The whole experiment was video-recorded with a Sony Digital HD-video recorder to assure data matching in case of system failure.

Subjective workload was assessed using the NASA-TLX scale (Hart & Staveland, 1988) which includes 6 items covering mental, physical and temporal demand, effort, frustration level and performance. Items were rated on 20-point bipolar scales (from 0 = *low* to 20 = *high*; for performance scale: 0 = *success* to 20 = *failure*). Demographic data as well as self-judgments regarding math skills were collected via a questionnaire.

Procedure

First, participants read the instruction including the information that disqualifying criteria were cardiovascular complaints or diseases, diabetes, etc. and filled in the socio-demographic questionnaire. The experimenter checked for exclusion criteria, positioned the MB2 on participant's non-dominant arm and started the video recording. Participants signed the confirmation agreement, equipped him-/herself with the electrodes of the SUEmpathy® following detailed instructions and with potential assistance of the experimenter. When the participant stated to be ready, the experimenter started the LabView presentation and physiological data recording. The procedure (see Figure 1) was in accordance to Sun et al. (2012) with the only exceptions that each condition lasted 5 minutes and no mediation music was presented in baseline and recovery phases. In the stepping and cycling condition, participants could shortly test the physical activity (max. 1 minute).

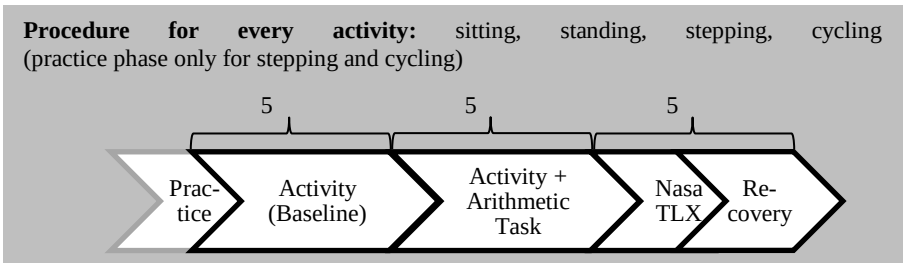


Figure 1. Experimental sequence.

After each condition with increased mental workload, the NASA-TLX (Hart & Staveland, 1988) was filled in by the participants. In sum, the experiment lasted 90 to 120 minutes.

Results

The HR data for both devices were transferred to Kubios (Version 3.0.2; Tarvainen et al., 2014) in order to calculate further heart parameters. The MB2 data for the first 10 participants had to be excluded from analysis, because data were too unreliable

due to the position of the MB2 (sensors on the upside of the wrist). Outlier analyses (Grubbs, 1969) identified 0 to 4 outliers for the varying HR parameters for each device that were excluded from further analyses. To investigate measurement validity, inter-class correlation coefficients (ICC) were calculated between MB2, SUE and NASA-TLX data.

Hypotheses were tested using ANOVAs for repeated measurements as well as post hoc tests with Bonferroni correction and paired *t*-tests. In case of violating the assumption of sphericity, the Greenhouse-Geißer Correction was used (Field, 2013). When data for the different conditions were not normally distributed, log-transformation (ln) was applied.

Stationary apparatus versus wearable device

As an example, Figure 2 shows raw HR data for one participant.

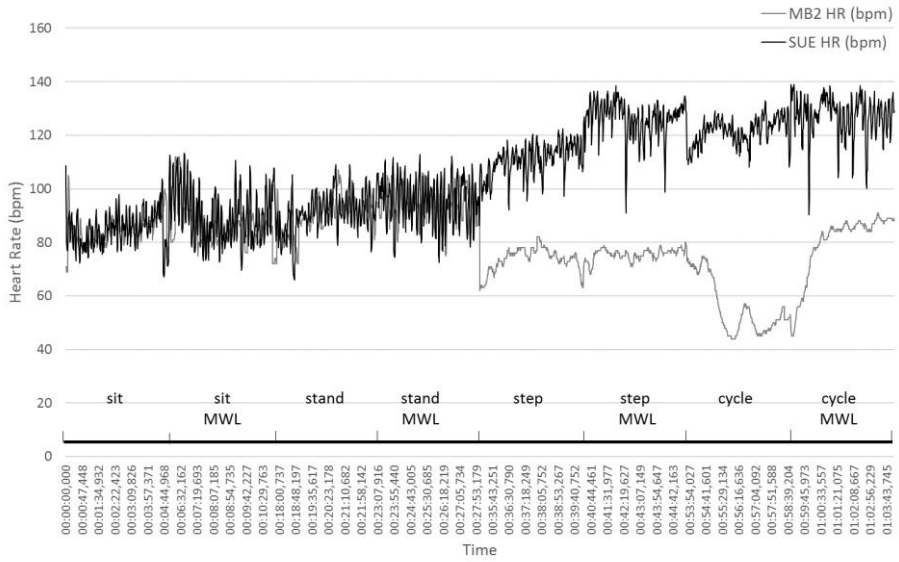


Figure 2. HR data for one participant (No. 21) for each condition measured via stationary (SUE) and wearable device (MB2).

Note. MWL = conditions with higher mental workload (MWL).

The ICCs (Table 1) show that the two devices' measures correlated only significantly in the sitting and standing conditions.

Table 1. ICCs and 95% Confidence intervals for the HR and IBI mean measured via stationary and wearable device (p < .05; ** p < .01).*

Parameter	*Condition*	*Sit*	*Stand*	*Cycle*	*Step*
Mean HR (bpm)	activity	.83** [0.58; 0.93]	.86** [0.65; 0.95]	-.26 [-2.17; 0.50]	.02 [-1.48; 0.61]
	activity + MWL	.70** [0.24; 0.88]	.66* [0.13; 0.86]	-.65 [-3.16; 0.35]	-1.02 [-4.11; 0.20]
Mean IBI (ms)	activity	.84** [0.59; 0.94]	.74** [0.35; 0.90]	.07 [-1.36; 0.63]	.11 [-1.26; 0.67]
	activity + MWL	.57* [-0.10; 0.83]	.49 [-0.28; 0.80]	-.39 [-2.51; 0.45]	-.14 [-1.89; 0.55]

Note. $N = 20$.

Influence of Mental Workload and Physical Activity on HR parameters

Means of several HR parameters for each condition are displayed in Table 2. In accordance to the ICC results, values of the two devices differ especially in the cycling and stepping condition. Still some differences between the conditions can be detected.

Regarding hypothesis H1, SUE data confirmed that participants in conditions with higher mental workload show higher HR and lower mean IBI (Table 2 and 3). Contrary to H1, an increase was found for SDNN and pNN50 when mental workload was raised (Table 2). RMSSD result showed no clear direction of change. For the stationary device, significant main effects with large effect sizes were found for HR, IBI and SDNN (Table 3). For the MB2, significant differences were found in RMSSD, SDNN and pNN50, but not in HR or mean IBI. Average RMSSD, SDNN and pNN50 values were higher when mental effort was higher compared to when mental effort was lower. Frequency-domain components LF and HF showed also an increase when extra mental effort was needed. The changes were significant (Table 2 and 3). This is contrary to the hypothesised direction of the mental workload effect and does not support our hypothesis (H1). Additionally, no significant effect was found for the LF/HF ratio. ANOVAs for the MB2 data showed comparable effects in the frequency-domain parameters.

For the SUE, all interactions *mental workload* x *physical activity* were significant (Table 3). For the MB2, HR, RMSSD, SDNN and LF parameters showed the significant interactions *mental workload* x *physical activity*. These results point out that the main effects of mental workload might not be always apparent. For the main effects that were in the hypothesised direction, we then tested each activity condition for significant differences between conditions with increased mental load and without by using one-tailed *t*-tests for paired samples. For interpreting the *p*-values, we applied a Bonferroni-Holm correction (Holm, 1979). For the SUE data, significant effects of the arithmetic task were found in the HR and mean IBI data for each of the activity conditions. As expected HR increased (sit: $t(28) = -5.53$, $p < .001$, $d = -1.04$; stand: $t(28) = -2.40$, $p = .012$, $d = -0.45$; cycle: $t(28) = -8.66$, $p < .001$, $d = 1.63$; step: $t(28) = -7.92$, $p < .001$, $d = -1.50$) and IBI decreased significantly (sit: $t(28) = 5.53$, $p < .001$, $d = 1.04$; stand: $t(28) = 2.40$, $p = .012$, $d = 0.45$; cycle: $t(28) = 8.66$, $p < .001$, $d = 1.64$; step: $t(28) = 11.28$, $p < .001$, $d =$

2.13) with more demanding physical activity. For the MB2 data, none of the significant mental workload effects retrieved from the ANOVAs pointed in the hypothesised direction. However, in one-tailed *t*-tests for paired samples with HR and mean IBI values, a medium effect was found when comparing the varying levels of mental workload in the sitting condition ($t(19) = 2.34$, $p = .016$, $d = 0.54$). After the Bonferroni-Holm correction it was found to no longer be significant. All other differences were non-significant and with small effects. Overall, the mental effort effect (H1) could be confirmed by HR and IBI data collected via SUE, but not by MB2 data.

Table 2. Results of HR and various HRV parameters measured via SUEmpathy® (SUE) and Microsoft Band 2 (MB2).

Para-meter	De-vice	*n*	Sit	Sit MWL	Stand	Stand MWL	Cycle	Cycle MWL	Step	Step MWL
HR (bpm)	SUE	29	73.8	79.4	87.4	90.8	107.5	116.9	112.2	122.1
	MB2	20	76.3	80.2	88.2	86.9	79.7	75.5	90.4	94.6
Mean IBI (ms)	SUE	29	828.3	766.7	703.4	673.9	567.1	524.9	542.9	495.3
	MB2	20	783.8	738.5	674.5	681.8	766.4	805.4	663.0	638.6
RMSSD (ms)	SUE	29	45.3	44.4	26.6	34.0	14.0	14.1	18.0	15.5
	MB2	19	114.0	161.6	125.3	160.7	253.3	253.2	220.1	213.1
SDNN (ms)	SUE	28	51.9	61.7	39.7	53.0	17.2	20.9	19.0	18.1
	MB2	19	90.0	126.9	98.2	124.1	188.3	203.0	175.0	173.5
pNN50 (%)	SUE	28	21.5	21.8	8.4	10.9	2.2	2.8	1.5	0.9
	MB2	19	30.5	53.3	38.5	54.6	72.9	71.9	77.2	72.2
LF (ms^2)	SUE	30	1576	2711	1206	2520	208	420	212	245
	MB2	18	2475	2957	2526	4238	24725	14440	7489	6549
HF (ms^2)	SUE	29	1086	1037	399	674	83	304	198	269
	MB2	18	10178	13942	5121	9426	34010	74639	22000	16229
LF/HF	SUE	27	2.85	3.88	7.82	5.75	5.67	3.91	4.36	3.57
	MB2	21	1.01	0.75	1.05	0.67	0.48	0.43	0.45	0.46

Note. MWL = conditions with higher mental workload (MWL).

Table 3. Results of ANOVAs with repeated measurements for various HR parameters measured via SUEmpathy® (SUE) and Microsoft Band 2 (MB2).

Parameter	Device (*n*)	*n*	Mental workload			Activity			Interaction		
			F	*(df)*	η^2_p	*F*	*(df)*	η^2_p	*F*	*(df)*	η^2_p
HR (bpm)	SUE	29	**68.9*** **	**(1, 28)**	**.71**	**368.9*** **	**(3, 84)**	**.93**	**3.3***	**(3, 84)**	**.11**
	MB2	20	0.2	(1, 19)	.01	**11.6*** **	**(2.1, 39.9)**	**.38**	**3.5***	**(3, 57)**	**.15**
Mean IBI (ms)	SUE	29	**68.8*** **	**(1, 28)**	**.71**	368.8***	(3, 84)	**.93**	**4.9****	**(2.4, 66.8)**	**.15**
	MB2	20	4.0	(1, 19)	.17	**17.9*** **	**(1.3, 23.7)**	**.49**	2.1	(2, 38)	.10
RMSSD (ms)	SUE	29	0.6	(1, 28)	.02	**51.0*** **	**(2.0, 56.2)**	**.65**	**18.7*** **	**(2.3, 63.2)**	**.40**
	MB2	19	**6.6***	**(1, 18)**	**.27**	**26.7*** **	**(2.3, 41.4)**	**.58**	**6.3****	**(3, 54)**	**.26**
SDNN (ms)	SUE	28	**16.9*** **	**(1, 27)**	**.39**	**71.2*** **	**(1.9, 50.3)**	**.73**	**12.9*** **	**(2.2, 59.8)**	**.32**
	MB2	19	**10.1****	**(1, 18)**	**.36**	**32.5*** **	**(3, 54)**	**.64**	**5.1****	**(3, 54)**	**.22**
pNN50 (%)	SUE	28	3.4	(1, 27)	.11	**54.3*** **	**(2.2, 60.4)**	**.67**	**6.9*** **	**(3, 81)**	**.20**
	MB2	19	**13.4****	**(1, 18)**	**.43**	**35.0*** **	**(2.2, 40,1)**	**.66**	**17.8*** **	**(2.0, 35.9)**	**.50**
LF (ms^2)	SUE	30	**33.2*** **	**(1, 29)**	**.53**	**85.6*** **	**(2.2, 64.5)**	**.75**	**10.7*** **	**(2.2, 62.6)**	**.27**
	MB2	18	**6.0***	**(1, 17)**	**.26**	**31.7*** **	**(3, 51)**	**.65**	**3.3***	**(3, 51)**	**.16**
HF (ms^2)	SUE	29	**15.6*** **	**(1, 28)**	**.36**	**47.1*** **	**(1.8, 51.0)**	**.63**	**7.1*** **	**(3, 84)**	**.20**
	MB2	18	**9.0****	**(1, 17)**	**.35**	**19.1*** **	**(3, 51)**	**.53**	2.8	(3, 51)	.14
LF/HF	SUE	27	0.1	(1, 26)	.01	**10.3*** **	**(3, 78)**	**.28**	**6.9*** **	**(2.3, 59.4)**	**.21**
	MB2	21	2.4	(1, 20)	.11	**5.8****	**(3, 60)**	**.23**	0.5	(3, 60)	.03

Note. ***$p < .001$, **$p < .01$, *$p < .05$, significant effects are bold written.

Regarding the effect of higher physical activity (hypothesis H2), a general trend in SUE data can be retrieved from Table 2. With more demanding physical activity, HR increases and HRV parameters values decrease. As one example, mean HR was the lowest in the sitting condition, followed by the standing condition and cycling. Stepping had the highest average HR. ANOVA results for all parameters showed that the main effect for activity was significant and very large (Table 3). The post hoc tests revealed significant differences between all activity conditions for HR and mean IBI, ($p \leq .015$). Post hoc tests between cycling and stepping were not significant for RMSSD, SDNN, pNN50, LF and HF. However, all other pairwise comparisons proved significance ($p < .001$); means were highest in the sitting condition, followed by standing, then stepping and cycling on third rank.

For the MB2, HR parameters in the cycling and stepping conditions differed extremely from SUE data (see also Table 1). Still, main effects of activity were found for each parameter, but post hoc tests showed quite inconsistent patterns. For HR and IBI significant differences occurred between sitting and standing, sitting and stepping as well as cycling and stepping ($p \leq .002$). Additionally, IBI data showed a

difference between standing and cycling (p = .034). For RMSSD, SDNN, LF, and HF, sitting and stepping as well as cycling and stepping did not vary significantly, but all other post hoc tests revealed statistically significant results ($p \leq .009$). LF/HF mean values only differed significantly between standing and cycling as well as stepping ($.009 \leq p \leq .013$). Based on the low reliability of MB2 data, SUE data get more weight in drawing conclusions regarding our hypothesis H2.

Results of NASA-TLX as Manipulation Check

Regarding the NASA-TLX, the raw task load index (RTLX) was analysed (Hart & Staveland, 1988) and confirmed our manipulation attempts (Table 4). The overall score showed that workload differed significantly between most conditions ($p \leq .002$) except for standing and cycling as well as for cycling and stepping. Results of the subscale *Physical demand* showed that sitting was the least demanding and cycling the most demanding activity. The post hoc tests revealed significant differences between almost all activities ($p < .001$), except for sitting and standing as well as for cycling and stepping. The *Mental demand* subscale indicated that the conditions with higher workload were comparable between the different physical activity conditions, but differences between sitting and standing (p = .003) as well as sitting and stepping (p = .046) were significant.

Table 4. Selected scores and ANOVA results for the NASA-TLX data.

	Overall score		*Mental demand*		*Physical demand*	
	M	*SD*	*M*	*SD*	*M*	*SD*
sit	39.7	27.3	57.3	19.5	8.2	7.5
stand	46.4	27.9	68.6	18.8	13.6	9.3
cycle	53.7	23.0	65.2	22.6	52.9	19.9
step	55.9	23.3	68.1	19.5	50.5	22.9
ANOVA results	$F(1.99, 59.71) = 18.67$, $p < .001$, $\eta_p^2 = .384$		$F(2.34, 70.22) = 4.93$, $p = .007$, $\eta_p^2 = .141$		$F(2.00, 60.13) = 82.72$, $p < .001$, $\eta_p^2 = .734$	

Note. N = 31, M = Mean, SD = Standard deviation.

In order to compare the results of the NASA-TLX and the devices, ICCs were calculated between the mean IBI values of the different activity conditions with mental workload task and the overall score values. However, no significant correlation was found for SUE (ICC(119) = -.186, 95%-KI[-.702, .173]) and MB2 (ICC(79) = -.135, 95%-KI[-.770, .272]).

Discussion

The present experimental study aimed at identifying increased mental workload in the course of different physical activities comparing a wearable device with a stationary device. Data of the two devices showed no correlation when participants moved. The low accuracy of wearable device data is contrary to findings from other studies (e.g., Stahl et al., 2016). Worth mentioning is that the real-time data assessment using the Microsoft SDK is only developed for reliable measurements when resting. However, even in the less active conditions reliability was not as high as in other studies (Barber et al., 2017). One possible explanation might be that the device has the function to manually switch to another activity mode when recording heart rate while, for instance, cycling. This was not used in the study, because it would not be realistic to regularly switch modes of the wearable device while working and standing up or starting to walk. Although it was not explicitly stated, other researchers might have used such switches in modes. Future devices automatically correcting HR recordings according to the physical activity might be of higher potential for identifying increased mental workload or even overload at work.

Still, there was a tendency in MB2 data that a higher level of mental workload came along with smaller mean IBI values. Only in the sitting condition, the effect reached a medium size. Contrary to findings of Barber et al. (2017), detecting increased mental workload using the Microsoft Band 2 and HR parameters while sitting, did not work satisfyingly in the current study. For the SUE data, the effects of increased mental workload for all HR parameters was strong, but the hypothesised direction of the effect was only found for HR and IBI. Most of the other analysed HRV parameters behaved in the opposite direction. Higher mental workload resulted in higher SDNN, LF, and HF values. This opposite direction of the mental workload effect was also found and discussed by Schubert et al. (2009). One explanation is that naming numbers orally, as part of the arithmetic task, influences the HRV parameters too much, so that the mental workload effect detection is difficult when speaking.

Results regarding the physical workload revealed more consistent effects; for all activities, a significant decrease in mean IBI durations was found when mental workload was increased. Additionally, both devices revealed significant effects for the activity; higher levels of physical workload led to lower levels in mean IBI durations. Still, the MB2 showed completely different results in the cycling condition which leads us to the conclusion that measurement is too much biased in order to draw any conclusions from the data. Overall, results are in line with previous findings on the effects of physical workload on heart rate variability (e.g., Sun et al., 2012). The potential of a wearable device for detecting higher physical activity on the basis of HR parameters, as also tested by Hwang and Lee (2017), was proven again.

In line with findings of Matthews et al. (2015), NASA-TLX scores did not correlate with physiological data. The RTLX index score showed that mental demand was quite comparable for the different conditions with the arithmetic task. Only the sitting condition was somewhat less demanding. This can be partly explained by the

additional task to follow the beat while cycling and stepping. This could have additionally raised the mental demand. Additionally, arithmetic tests might have varied slightly regarding their difficulty. Future research might address this and use even more equally difficult tasks and/or another cognitive demanding task as well as other means for ensuring similar physical demand.

Overall, the limited reliability of the used wearable device regarding HR measures in varying activity conditions reduces the potential of (comparable) wearable devices for a fine-grained monitoring of physical and mental effort. Thus, short-term adaptation of workload on the basis of comparable, easy-to-use devices that measure HR does not seem reasonable right now. Future research might concentrate on identifying rather long-term changes that indicate stress and/or develop algorithms that address the reduced reliability of wearable devices, especially when moving and/or considering more variables for identifying changes in workload (e.g., step count, galvanic skin response).

Acknowledgements

The research leading to these results has received funding from Horizon 2020, the European Union's Framework Programme for Research and Innovation (H2020/2014-2020) under grant agreement no 723277.

References

Barber, D., Carter, A., Harris, J., & Reinerman-Jones, L. (2017). Feasibility of wearable fitness trackers for adapting multimodal communication. In S. Yamamoto (Ed.). *International Conference on Human Interface and the Management of Information* (pp. 504-516). Cham: Springer.

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

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

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*, 1-21.

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. doi:10.1016/S0166-4115(08)62386-9

Holm, S. 1979. A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hsu, B.W., Wang, M.J.J., Chen, C.Y., & Chen, F. (2015). Effective indices for monitoring mental workload while performing multiple tasks. *Perceptual and Motor Skills, 121*, 94-117.

Hwang, S., & Lee, S. (2017). Wristband-type wearable health devices to measure construction workers' physical demands. *Automation in Construction*. doi:10.1016/j.autcon.2017.06.003

IDC (2017). Worldwide wearables market to nearly double by 2021, According to IDC. https://www.idc.com/getdoc.jsp?containerId=prUS42818517.

Kim, S., & Kim, S. (2016). A multi-criteria approach toward discovering killer IoT application in Korea. *Technological Forecasting and Social Change*, *102*, 143-155.

Lavallière, M., Burstein, A.A., Arezes, P., & Coughlin, J.F. (2016). Tackling the challenges of an aging workforce with the use of wearable technologies and the quantified-self. *Dyna*, *83*(197), 38. doi:10.15446/dyna.v83n197.57588

Marakhimov, A., & Joo, J. (2017). Consumer adaptation and infusion of wearable devices for healthcare. *Computers in Human Behavior*, *76*, 135-148. doi:10.1016/j.chb.2017.07.016

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, *57*, 125-143.

Meinel, J. (2013). *Spezifische Effekte visueller und kognitiver Ablenkung bei der Kraftfahrzeugführung (Specific effects of visual and cognitive distraction while driving a motor vehicle)* (PhD thesis). Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II. https://doi.org/http://dx.doi.org/10.18452/16678

Mulder, L.J.M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology*, *34*, 205-236.

Reinerman-Jones, L., Harris, J., & Watson, A. (2017, July). Considerations for using fitness trackers in Psychophysiology research. In S. Yamamoto (Ed.). *International Conference on Human Interface and the Management of Information* (pp. 598-606). Cham: Springer.

Schubert, C., Lambertz, M., Nelesen, R.A., Bardwell, W., Choi, J.B., & Dimsdale, J.E. (2009). Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological Psychology*, *80*, 325-332.

Shaffer, F., McCraty, R., & Zerr, C.L. (2014). A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.01040

Shcherbina, A., Mattsson, C.M., Waggott, D., Salisbury, H., Christle, J.W., Hastie, T., Wheeler, M.T., & Ashley, E.A. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, *7*(3), 1-12.

Stahl, S.E., An, H.S., Dinkel, D.M., Noble, J.M., & Lee, J.M. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine*, *2*, doi:10.1136/bmjsem-2015-000106

Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., & Griss, M. (2012). Activity-aware mental stress detection using physiological sensors. In M. Gris and G. Yang (Eds.), *Mobile Computing, Applications, and Services: Second International ICST Conference, MobiCASE 2010, Santa Clara, CA, USA, October 25-28, 2010, Revised Selected Papers* (pp. 282–301). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-29336-8_16

Swan, M. (2012). Sensor mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, *1*, 217-253.

Tarvainen, M.P., Niskanen, J.P., Lipponen, J.A., Ranta-Aho, P.O., & Karjalainen, P.A. (2014). Kubios HRV–heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, *113*, 210-220.

Wickens, C.D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449-455.

Yixiang, X. (2015). Metronom Pro – Das Profi Metronom (Version 3.13.2) [Mobile application software]. Retrieved from: http://appsto.re/de/FNrFC.i

Ocular-based automatic summarization of documents: is re-reading informative about the importance of a sentence?

Orlando Ricciardi, Giovanni Serra, Federica De Falco, Piero Maggi, & Francesco Di Nocera
Sapienza University of Rome
Italy

Abstract

Automatic document summarization (ADS) has been introduced as a viable solution for reducing the time and the effort needed to read the ever-increasing textual content that is disseminated. However, a successful universal ADS algorithm has not yet been developed. Also, despite progress in the field, many ADS techniques do not take into account the needs of different readers, providing a summary without internal consistency and the consequent need to re-read the original document. The present study was aimed at investigating the usefulness of using eye tracking for increasing the quality of ADS. The general idea was of that of finding ocular behavioural indicators that could be easily implemented in ADS algorithms. For instance, the time spent in re-reading a sentence might reflect the relative importance of that sentence, thus providing a hint for the selection of text contributing to the summary. We have tested this hypothesis by comparing metrics based on the analysis of eye movements of 30 readers with the highlights they made afterward. Results showed that the time spent reading a sentence was not significantly related to its subjective value, thus frustrating our attempt. Results also showed that the length of a sentence is an unavoidable confounding because longer sentences have both the highest probability of containing units of text judged as important, and receive more fixations and re-fixations.

Introduction

Summarization is a strategy used to understand and store knowledge (Anderson & Armbruster, 2000). The goal of a summary is to produce a document shorter than the original by eliminating unnecessary information, allowing the readers to optimize their use of time and cognitive effort (Renkl & Atkinson, 2007) and to organize the text in a structure that facilitates comprehension (Leopold et al., 2013). The activity of rewriting text is the last phase in the process of summarization. Indeed, when an individual reads text to study it, s/he proceeds with a quick first reading, then determines the main contents and, finally, rewrites them into a new, shorter document (Flower & Hayes, 1980; Taylor & Beach, 1984; Wittrock & Alesandrini, 1990). An essential value of a summary is that it reflects precisely what the reader wants to learn about a topic. On the other hand, it is evident that the activity of providing summaries requires time and cognitive effort, especially if the text is

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

lengthy. Such problems are made worse by the increasing amount of electronic information available online and the consequent need to manage it quickly (for a review see Eppler & Mengis, 2004). ADS represents a partial solution to this problem by allowing the creation of summaries in a few seconds, by selecting the essential contents of a text (Gupta & Lehal, 2010).

Research in this field started with the interest in the production of abstracts for technical documentation (Saggion & Poibeau, 2013). Particularly, the first attempt to use ADS was in 1958, when Luhn proposed an algorithm that employs the frequency of a word to measure sentence relevance, leading to the first and most straightforward approach for creating a summary. Subsequently, more complex strategies have considered syntactic analysis of the text (Climenson et al., 1961), grammatical rules of discourse construction, and the semantic relationships among words and sentences (Mani & Maybury, 1999). In general, these approaches have used statistical techniques to extract one or more phrases to provide a summary (Paice, 1990) and are usually divided into two categories: abstractive and extractive methods (Hahn & Mani, 2000). Abstractive methods use linguistic approaches to identify the central concepts and produce a shorter text that may include new sentences, not stated in the original version (Erkan & Radev, 2004). Alternatively, extractive methods focus on statistical analysis of the text's features such as the unit's location in the source text, how often it occurs, the appearance of cue phrases, and statistical significance metrics (Hahn & Mani, 2000). This class of techniques attributes a weight or a score to each different word and sentence and uses statistical analysis to integrate linguistic features (word/phrase frequency, location of cue words) into a shorter document (Kyoomarsi et al., 2008). The fundamental principle is that the most frequent or the better-positioned content in a sentence is considered the most important. Even if these approaches are easy to implement, several issues limit their efficiency. The main problem is that this approach avoids analysing the text`s meaning, providing a summary that may be incomplete or without internal coherence (Hahn & Mani, 2000). For example, if two successive sentences explain different aspects of the same concept, and if only one of them is extracted, a re-analysis of the text might be required. It is worth noting that many studies have investigated this topic, but rarely have the human factors of summarization behaviour been examined (Xu et al., 2009).

Human Factors in summary evaluation

Summary evaluations have been discussed since the late 1990s (Jones & Endres-Niggemeyer, 1995; Mani et al., 2002). Automatic summary evaluation methods can be divided into "intrinsic" and "extrinsic" (Jones & Galliers 1996; Hahn & Mani, 2000). Intrinsic evaluations methods are based on the characteristics of the summary and do not consider the final user. These assessments focus on the consistency between different parts of the text, on the correspondence between the weights assigned to original sentences and extracted sentences and on the information reported into the summary compared to that in the original version. Alternatively, in extrinsic evaluation methods, the final user is the centre of the evaluation process. This technique evaluates how much the summary responds to the user's needs, considering readability, relevance and efficiency of the review based on a query.

Both intrinsic and extrinsic techniques can be automated or manual. In automated approaches, the evaluation consists of a comparison of the summary with one or more reference summaries (Saggion & Poibeau, 2013). Automatic procedures are better suited to extractive methods, whereas evaluations of summaries generated with abstract methods can be made only with a manual approach due to the difficulty of interpreting the meanings of new sentences (Saggion & Poibeau, 2013). In manual procedures, a team of different users evaluates a summary considering various features, such as style, grammar, content, readability, etc. These types of evaluations are often required as benchmarks, even in automated evaluation methods, but they are vulnerable to user subjectivity.

In summary, all ADS methods include the measurement of two fundamental properties: the Compression Ratio and the Retention Ratio. The first refers to the length of the summary relative to that of the original text, whereas the retention ratio indicates how much information from the original version has been retained in the summary (Mani & Maybury, 1999). Even though a significant amount of literature has been provided on this topic, the problem is still far from being solved. Thus, it is possible that a human-cantered perspective could help in addressing the issue. Our idea is to try to improve the quality of automatic summarization techniques by integrating a subjective behavioural indicator of importance into the extractive methods. Particularly, eye movements made during initial reading activity could reflect the reading strategy involved in detecting essential aspects of a text.

Eye behaviour in reading activity

The relationship between eye movements and attention has been widely studied, and several approaches have been proposed to describe it. The Premotor Theory (e.g., Rizzolatti et al., 1987; Rizzolatti, Riggio, & Sheliga, 1994) for example, suggests that attention and eye movements rely on the same brain structures. Furthermore, the Eye-Mind hypothesis (Just & Carpenter, 1984) advises a strong correlation between gazes and cognitive processes. Other studies have found that attention and saccades depend on the same mechanisms involved in spatial attention and in saccade orientation (Shepherd et al., 1986; Kowler et al., 1995; Kowler, 1996). In line with these results, Hoffman and Subramaniam (1995) found that subjects have difficulty in moving their eyes to one location and attending to another, even when instructed to do so and, in contrast, that making a saccade to an area improves the detectability of information presented in that location. In this framework, eye movements could be used to detect reading behaviour indicators that could be used as weights to select the information that will be included into the summary.

The availability of new eye-tracking technology allows us to gain a deep understanding of the eye movement behaviour during reading (Clifton et al., 2016; Radach & Kennedy, 2013; Rayner, 1975, 1978; Rayner & Pollatsek, 1987). Notably, several studies have reported the details of saccades, fixations, skipping and re-fixations (Liversedge et al., 2011; O'Regan & Ltvy-Schoen, 1987; O'Regan et al., 1984; Pynte, 1996; Pollatsek & Rayner, 1990; Reichle et al. 2003). For instance, there is a consensus on the variability of saccades (20-50 milliseconds) and fixation durations (200-500 milliseconds) due to the relation between oculomotor system behaviour and comprehension processing difficulties (Reichle et al. 2003). The most

important factors affecting fixation durations are word length, frequency, age of acquisition, predictability (how predictable a word is from the context of a sentence) and similarity with other words (Ehrlich & Rayner, 1981; O'Regan, Levy-Schoen, Pynte & Brugaillère, 1984; Balota et al., 1985; Inhoff & Rayner, 1986; Kliegl et al., 2004; Hyönä , 2011; Rayner, 2009). Moreover, the duration of fixations may include the encoding of the antecedent word ("spillover effect"; Reichle et al., 2003), or the encoding of the successive word ("preview benefit"; Inhoff et al., 2000; Schroyens et al., 1999). About 10-15% of saccades are called "regressions" because the eyes move back to a part of a text that has been already inspected (see Rayner, 2009). These movements may be due to several factors such as: the correction of oculomotor errors (see Bicknell & Levy, 2011; O'Regan, 1990) for searching for the "optimal viewing position" (O'Regan, 1990; Brysbaert & Nazir, 2005), or difficulties in linguistic processing (Reichle et al., 2003).

Although a large body of research has been conducted with eye tracking during reading, few studies have tried to use eye-movement related metrics in ADS. Xu et al. (2009) for example, assume that the amount of time that a reader spends on a word is related to its importance in the comprehension process of the entire text. Following this reasoning, they inserted the "duration of fixations" into automatic summarization software as a criterion to determine which sentences to include in the summary. Although their results showed some superiority over other automatic summarization software, some issues remain. One issue is the relationship between attention and fixation duration. Several studies have indicated that fixation time on a word does not necessarily reflect its importance or the depth of cognitive processing. Indeed, long fixation time might also reflect difficulty in processing both the word fixated and the information derived from words in parafoveal vision (Kennedy & Pynte, 2005; Kliegl et al., 2006). The lack of attention paid by Xu and co-workers to the variety cognitive processes potentially affecting fixation times has been criticized by Buscher et al. (2012) in a more recent survey. In their study, the authors investigated the relationships among the following variables: "coherently read text length", considered as the length of text in characters that has been read consistently without skipping any text; the "thorough reading ratio", computed as the amount of text that has been detected as having been read divided by the amount of reading or skimmed text; the "regression ratio", i.e. the ratio between the number of regressions made on a single Area of Interest (AOI) divided by the total number of saccades received from that AOI, and the "mean forward saccade length", calculated as the average length of progressive saccades. The authors found that "coherently read text length," "thorough reading ratio" and "regression ratio" increased with perceived relevance of the text, but "mean forward saccade length" decreased with perceived relevance. These results suggest that essential sentences are the recipients of more accurate reading. Also, more important paragraphs and phrases are more frequently inspected by the reader. Finally, Buscher et al. (2012) found that fixation duration, as predicted in the literature, was not related to the importance of a sentence or paragraph. Despite the scientific contribution of this research, the results are not yet conclusive and satisfactory to produce more consistent automatic summarization algorithms.

Study

In the present study, we used eye tracking with the aim to improve the quality of ADS techniques. Thanks to the evolution in eye tracking technology and data analysis methods, we aimed to collect information during a reading task to be used to provide an index of importance attributed by the reader to a sentence. This measure could then be used to improve the quality of automatic summaries by tailoring the summary to the reader's goals. Specifically, we suggested that the study of eye movements during a natural reading activity could allow identification of the reading strategy used to create a summary. As indicators, we have considered first fixation duration as the time spent on the first reading of a sentence and re-fixation duration as the time spent on the second reading of a sentence (explained in more detail in the method section). We used the highlights made on a printed version of the same text as a measure of the subjectively perceived importance (Nist & Hogrebe, 1987; Peterson, 1991). The research hypothesis was that the time spent in re-reading a sentence reflects the subjectively perceived importance (SPI) of that sentence. Along with this assumption, we expected that a sentence receiving longer re-fixations also should receive more highlights. To test this hypothesis, eye fixation behaviour recorded during a screen reading task was compared with the importance attributed to specific sentences by observers who underlined parts of its printed version (Nist & Hogrebe, 1987; Peterson, 1991).

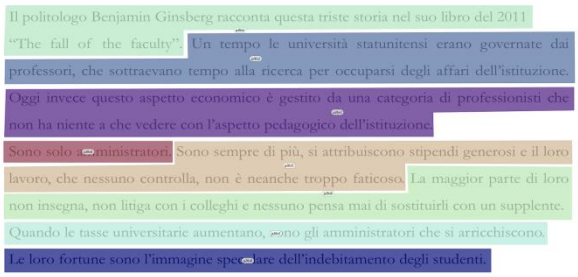
Participants

Thirty university students (25 females, mean age = 26.4; sd = 4.5) volunteered to participate. All had normal or corrected-to-normal vision.

Materials and Method

We used a magazine article for the study, with the aim to involve the participants in the reading activity. The text chosen is the official Italian translation of the article "Academy Fight Song" by Thomas Frank (available at the web address https://thebaffler.com/salvos/academy-fight-song). The text was divided into 41 pages and presented as slideshow. Each sentence was attributed to an AOI for collecting the fixations with the Tobii Studio software, allowing counting the fixations for each AOI. For example, Figure 1 shows the editing of different AOIs in the Tobii Studio software *(a)* and the version read by the participants during the experimental session *(b)*.

a)



b)

Il politologo Benjamin Ginsberg racconta questa triste storia nel suo libro del 2011 "The fall of the faculty". Un tempo le università statunitensi erano governate dai professori, che sottraevano tempo alla ricerca per occuparsi degli affari dell'istituzione. Oggi invece questo aspetto economico è gestito da una categoria di professionisti che non ha niente a che vedere con l'aspetto pedagogico dell'istituzione.
Sono solo amministratori. Sono sempre di più, si attribuiscono stipendi generosi e il loro lavoro, che nessuno controlla, non è neanche troppo faticoso. La maggior parte di loro non insegna, non litiga con i colleghi e nessuno pensa mai di sostituirli con un supplente.
Quando le tasse universitarie aumentano, sono gli amministratori che si arricchiscono. Le loro fortune sono l'immagine speculare dell'indebitamento degli studenti.

Figure 1. a) A specific Area of Interest was assigned to each sentence in the text; b) The version of the text read by the participants during the experimental session.

First fixations (FF) were defined as the early exploration on the "n" AOI until the eyes moved on to the "n + 1" AOI. Then, each backward movement on the "n" AOI was considered a re-fixation (re-reading, RR; Figure 2). The total number and durations of FFs and RRs were weighted according to the length of the AOI, to avoid biasing the data by the number of characters present in the sentences. The X2-30 eye tracker system (Tobii, Sweden) was used to record eye movements during the reading activity.

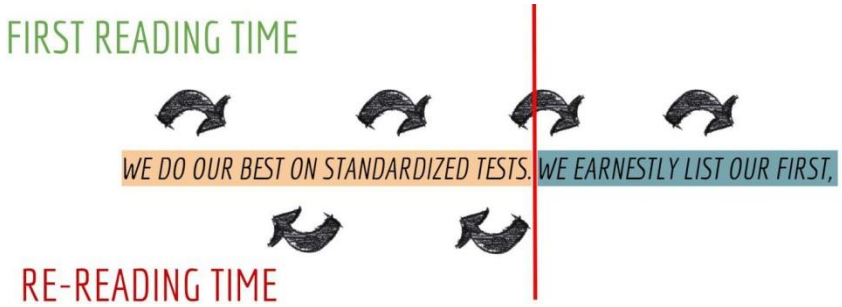


Figure 2. First reading was considered until a backward movement happened from the "n + 1" AOI to the "n" AOI. All successive fixations on the "n" AOI were considered as re-reading.

Procedure

The experiment consisted of two phases: in the first, participants were asked to read a magazine article on a 17" screen while they were positioned at about 60 cm from

the display, and the text was displayed in a full-screen mode, to facilitate a comfortable reading. Before the experiment, a 9-point calibration was performed. Subjects were instructed in using the spacebar for moving to the next slide. The second phase of the experiment took place after a week, with the same subjects. The task consisted of reading the same version of the magazine article in a printed version. We asked them to highlight the most important concepts contained in the text with the objective of collecting an indicator of the subjective importance attributed to the sentences. In both phases, reading comprehension was assessed with a brief structured interview (i.e., "What problem is discussed in the article that you have just read?"; "Why are American students willing to apply for a loan to attend a college or a university?").

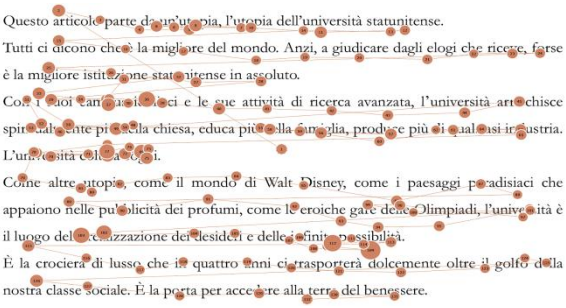


Figure 3. Scan path: saccades (segments) and fixations (spheres) recorded during the reading task.

Data analysis and results

The data used for the analyses were the numbers and durations of fixations and re-fixations directed to each sentence (AOI). The highlights collected on the printed version of the article were used as a subjective measure of perceived importance in the analyses. AOIs were classified into four categories (“very low,” “low,” “high,” “very high”), according to the sum of highlights received from all the subjects (quartiles were considered for classifying the sentences).

Data analyses showed a high positive correlation between the number and the duration of fixation both when the subject read for the first time (r = .99) both in re-reading (r = .97). Due to this high correlation, we have decided to further analyse only duration. The correlation between the number of highlights for each sentence and the re-reading time was significant but low (r = .21). Also, the correlation with the first reading time was found to be significant (r = .31). Moreover, a significant positive correlation was found between the highlights and the sentence’s length (r = .33).

Due to this correlation, this variable has been included as a covariate in an ANCOVA design, where first reading time and re-reading time were used as dependent variables and the number of highlights (category) as a factor. The analysis resulted in the absence of the primary effect of both the variables, showing no difference in reading or reading time depending on the highlights marked on each

group of sentences [$F_{3,205}$ = .45, p > .05]. Indeed, a positive relation was found for the covariate (Figure 4 and 5).

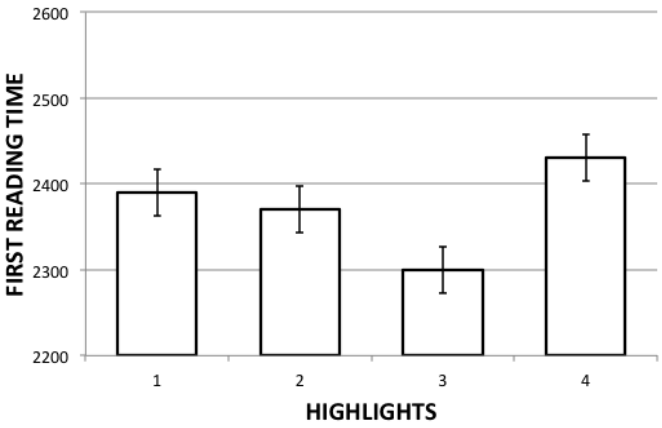


Figure 4. First Reading time (in milliseconds) by highlighting (quartiles indicate highlighting increment and therefore the importance of the sentence); bars represent the 95% confidence interval.

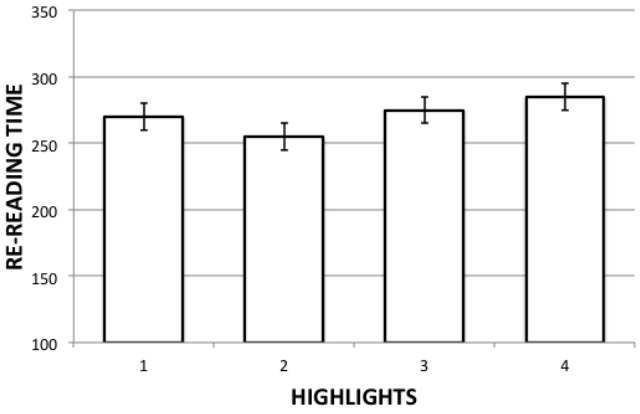


Figure 5. Re-Reading time (in milliseconds) by highlighting (quartiles indicate highlighting increment and therefore the importance of the sentence); bars represent the 95% confidence interval.

Therefore, the sentence's length was used as a dependent variable in a one-way ANOVA design, using the highlights as a factor. The analysis confirmed the significant relationship of this variable [$F_{3,206}$ = 10.61, p < .0001; Figure 6).

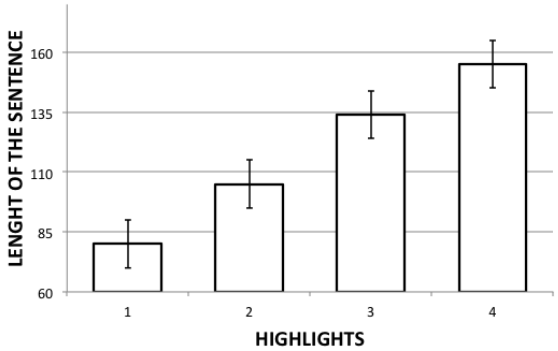


Figure 6. Length of the sentence (in characters) by highlighting (quartiles indicate highlighting increment and therefore the importance of the sentence); bars represent the 95% confidence interval.

Discussion

The objective of the present study was to improve ADS techniques by devising an indicator of the subjective importance of text based on the analysis of eye movement behaviour. Therefore, we compared eye movements and highlighting strategies to find a relation between a behavioural and an individual attribution of salience across the text. Indeed, each sentence has been considered as an AOI and the time spent on it was correlated with the number of highlights received. Highlights were considered an indicator of personal importance (Nist & Hogrebe, 1987; Peterson, 1991), as each reader was free to select the most critical concepts based on his or her previous knowledge and the reader`s objectives. At the same time, eye movement behaviour is deeply involved in cognitive aspects of reading activity (Liversedge et al., 2011; O'Regan & Ltvy-Schoen, 1987; O'Regan et al., 1984; Pollatsek & Rayner, 1990; Rayner et al., 1998; Reichle, 2003).

The idea of integrating a behavioural indicator into an extractive summarization technique is due to the current problems with these classes of methods, since they are usually unable to provide a summary that can satisfy the reader's goals (Hahn & Mani, 2000). Indeed, these algorithms are easy to implement because they select sentences with higher scores, depending on some indicators such as word frequency or word location (Kyoomarsi et al., 2008). However, they do not analyse text meaning and often provide a summary with poor internal coherence because of the loss of crucial concepts (Hahn & Mani, 2000). Eye movement data could offer a subjective weight that can be used to tailor the summary according to the reader's goal, advancing the utility of the extractive methods of summarization. If this relation were confirmed, the applicative rate would be enormous.

Summarizing a text is a common strategy to study and store information (Anderson & Armbruster, 2000), and automatic summarization is used to reduce time and cognitive efforts needed to produce a summary (Renkl & Atkinson, 2007). This is even more important considering the enormous amount of information available online and the need to manage it quickly (Eppler & Mengis, 2004).

The primary hypothesis of our study was that the duration of the regressions made during reading reflects the subjectively perceived importance to the reader. More specifically, we hypothesized that higher re-reading times should correspond with more highlighted sentences, whereas lower re-reading times should be associated with sentences that received fewer highlights.

The results obtained did not confirm our initial hypothesis. The correlation matrix showed a very high correlation between time and number of fixation (both in FR and in RR), so we decided to use the average time of fixation as dependent variable for the analyses. Only a weak correlation was found between RR and highlights. At the same time, the relation between FR measures and the highlights was also low. However, a significant correlation was found between the length of sentences and the number of highlights. It is worth noting that the number of characters in each phrase was balanced, as the FR and RR on each AOI were divided by the number of characters contained in that AOI. FR and RR were used as dependent variables in an ANCOVA design, using the proportion of highlights (Very Low, Low, High and Very High categories) as a factor and the sentence`s length as a covariate. The main effects were nonsignificant, and no interaction effect was found, suggesting that the time spent on reading or rereading the text was not related to the highlighting strategy. Indeed, a significant effect of the covariate was found in both analyses, confirming the relationship between the sentence's length and the eye movement behaviour.

Although the results of this study are far from being conclusive, they seem to disconfirm the hypothesis that the time spent on a sentence reflects its relative importance. This effect was found by Buscher et al. (2012), for example. In their study, the authors noted that essential contents induce more precise eye movements and a higher probability of re-fixation, while contents perceived as not relevant are more related to "skimming" behaviour; that is, a higher likelihood of scanning the text very quickly, skipping many words and without re-fixation. The difference between our results and those from Buscher and colleagues (2012) could be due to several factors as, for instance, the text's language. Italian and English writing styles have different linguistic structures, and one of the main aspects is the sentence`s length, usually shorter in the English language. Moreover, in our study participants did not have to provide a summary of the text, but only to perform a verbal assessment of text comprehension. It is possible that the goal of the task determined a lower level of effort in the second task, and that the subjects gave more importance only to the more extended sentences.

Limitation and further research

The study has some limitations, including the assumption that eye movement regressions and re-reading time reflects the importance given to a sentence. Indeed, re-reading should also reveal the reader's difficulties in understanding a sentence (Rayner et al., 2006). This hypothesis was not explored in the present study but, considering our results, needs to be addressed in further research on this topic.

An issue lies in the instrumentation used to detect eye movements; i.e., a low-cost eye-tracker (sampling rate of 30Hz) that might have led to some errors in measuring

the eye movement behaviour during reading. It is possible that some fixations and saccades were not detected, or they had been assigned to an incorrect AOI. Of course, a higher sampling rate would make easier to collect these kinds of data (Holmqvist et al., 2011), but we observed that our effort for improving ADS techniques by making use of eye-tracking measures is one of the first studies of its kind in this area. Another limitation could be attributed to the experimental design, in which several factors might have altered ocular behaviour. The absence of pauses during the task can be considered a problem, as the subjects reported being tired at the end of the recording session (that lasted from 30 to 40 minutes). Indeed, we had chosen not to allow pauses during the reading activity to avoid the need to re-calibrate the eye tracker. That limitation might have led to a loss in ecological validity. Regarding the data analysis, even though we had tried to weigh and normalize the collected measures, some features that may change the subjective reading behaviour have not been considered. We did not control the word frequency and the word predictability, for example. We were aware that these variables are considered significant in studies that focus on eye movements and reading. However, we should emphasize that our intent was not to study reading activity itself but to analyse the relationship between ocular behaviour during reading and a subsequent highlighting strategy during the study of a document. For this objective, we thought that the best way to investigate this relationship was to use a natural text, accepting loss of control of its structure. This choice was considered trivial to maintain the ecological validity of the study and to enable application of our results to future research and developing ADS technology.

Overall, additional research is required to better understand the relationship between re-reading times and perceived importance of text segments. Future experimental designs should include some modifications to reduce the impact of the described limitations. For example, splitting the reading task into shorter sessions and presenting the subjects with different types of documents, could help to control effects due to fatigue and document type. Additionally, all the features that influence reading times should be limited or controlled.

References

Anderson, T.H., & Armbruster, B.B. (2000). Studying. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, and R. Barr (Eds.), *Handbook of reading research, Volume III* (pp. 657 - 679). Mahwah, NJ: Erlbaum.

Balota, D.A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17*, 364-390.

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In L. Carlson, C. Holscher, and T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 931–936). Austin, TX: Cognitive Science Society.

Brysbaert, M., & Nazir, T. (2005). Visual constraints in written word recognition: evidence from the optimal viewing position effect. *Journal of Research in Reading, 28*, 216-228.

Buscher, G., Dengel, A., Biedert, R. & Elst. L.V. (2012). Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems, 1*, 9.

Clifton, C., Ferreira, F., Henderson, J.M., Inhoff, A.W., Liversedge, S.P., Reichle, E.D., & Schotter, E.R. (2016). Eye movements in reading and information processing: Keith Rayner's 40year legacy. *Journal of Memory and Language, 86,* 1-19.

Climenson, W.D., Hardwick, N.H., & Jacobson, S.N. (1961). Automatic syntax analysis in machine indexing and abstracting. *American Documentation*, 178-183.

Ehrlich, S.E, & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20,* 641-655.

Eppler, M.J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society, 20,* 325-344.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22,* 457-479.

Flower, L., & Hayes, J. (1980). The dynamics of composing: making plans and juggling constraints. In L. Gregg, and E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gupta, V., & Lehal, G.S. (2010). A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence, 2,* 258-268.

Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. *Computer, 33,* 29-36.

Hoffman, J.E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Attention, Perception, & Psychophysics, 57*, 787-795.

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

Hyönä, J. (2011). Foveal and parafoveal processing during reading. In S.P. Liversedge, I.D. Gilchrist, and S. Everling (Eds.), *The Oxford Handbook of Eye Movements* (pp. 819 – 837). Oxford, England: Oxford University Press.

Inhoff, A.W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40,* 431–439.

Inhoff, A.W., Starr, M., & Shindler, K. (2000). Is the processing of consecutive words during eye fixations in reading strictly serial? *Perception & Psychophysics, 62,* 1474-1484.

Jones, K.S., & Endres-Niggemeyer, B. (1995). Automatic summarizing. *Information Processing & Management, 31*, 625-630.

Jones, K.S., & Galliers, J. (1996). Evaluating natural language processing systems: An analysis and review. *Natural Language Engineering, 4*, 175-190.

Just, M.A., Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research, 45,* 153–168.

Kyoomarsi, F., Khosravi, H., Eslami, E., Dehkordy, P.K., & Tajoddin, A. (2008). Optimizing text summarization based on fuzzy logic. In *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on* (pp. 347-352). IEEE.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1-2), 262-284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General, 135*, 12.

Kowler, E. (1996). Cogito ergo moveo: cognitive control of eye movement. In *Exploratory vision* (pp. 51 - 77). Springer New York.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision research, 35*, 1897-1916.

Khosravi, H., Eslami, E., Kyoomarsi, F., & Dehkordy, P. (2008). Optimizing text summarization based on fuzzy logic. *Computer and Information Science,* 121-130.

Leopold, C., Sumfleth, E., & Leutner, D. (2013). Learning with summaries: Effects of representation mode and type of learning activity on comprehension and transfer. *Learning and Instruction, 27,* 40-49.

Liversedge, S.P., Gilchrist. I.D. & Everling, S. (2011). *The Hoxford Handbook Of Eye Movements*. New York: Oxford University Press Inc.

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization, Volume 293*. Cambridge, MA: MIT press.

Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., & Sundheim, B. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering, 8*, 43-68.

Nist, S.L., & Hogrebe, M.C. (1987). The role of underlining and annotating in remembering textual information. *Literacy Research and Instruction, 27*, 12-25.

O'Regan, J.K. (1990). Eye movements and reading. *Reviews of oculomotor research, 4*, 395.

O'Regan, J.K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B.É. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 250.

O'Regan, J.K., & Ltvy-Schoen, A. (1987). Eye movement strategy and tactics in word recognition and reading. In M. Coltheart (Ed.), *Attention and Performance: Vol.12. The psychology of reading* (pp. 363-383). Hillsdale, NJ: Erlbaum.

Paice, C.D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management, 26*, 171-186.

Peterson, S.E. (1991). The cognitive functions of underlining as a study technique. *Literacy Research and Instruction, 31*, 49-56.

Pollatsek, A., & Rayner, K. (1990). Eye movements and lexical access in reading. In D.A. Balota, G.B. Flores d'Arcais, and K. Rayner (Eds.), *Comprehension processes in reading* (pp. 143 - 164). Hillsdale, NJ: Erlbaum.

Pynte, J. (1996). Lexical control of within-word eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 958-969.

Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology, 66*, 429-452.

Rayner, K. (1975). Parafoveal identification during a fixation in reading. *Acta Psychologica, 39*, 271-281.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*, 618-660.

Rayner, K. (2009). The Thirty Fifth Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62,* 1457–1506.

Rayner, K. & Pollatsek, A. (1987) Eye movements in reading: A tutorial review. *Attention and performance, 12I,* 327–362.

Rayner, K., Chace, K.H., Slattery, T.J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading, 10*, 241-255.

Rayner, K., Reichle, E.D., & Pollatsek, A. (1998). Eye movement control in reading: An overview and model. *Eye guidance in reading and scene perception*, 243-268.

Reichle, E.D., Rayner, K., Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading Comparison to other models. *Brain and Behavioral Sciences, 26,* 445–476.

Renkl, A., & Atkinson, R.K. (2007). Interactive learning environments: Contemporary issues and trends. An introduction to the special issue. *Educational Psychology Review, 19*, 235-238.

Rizzolatti, G., Riggio, L., & Sheliga, B.M. (1994). Space and selective attention. *Attention and performance XV, 15*, 231-265.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*, 31-40.

Saggion, H., & Poibeau, T. (2013). Automatic text summarization: Past, present and future. In *Multisource, multilingual information extraction and summarization* (pp. 3 - 21). Springer Berlin Heidelberg.

Schroyens, W., Vitu, F., Brysbaert, M., & d'Ydewalle, G. (1999). Eye movement control during reading: Foveal load and parafoveal processing. *Quarterly Journal of Experimental Psychology, 52*, 1021-1046.

Shepherd, M., Findlay, J.M., & Hockey, R.J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology, 38,* 475-491.

Taylor, B.M., & Beach, R.W. (1984). The effects of text structure instruction on middle-grade students comprehension and production of expository text. *Reading Research Quarterly, 19,* 134-146.

Wittrock, M.C., & Alesandrini, K. (1990). Generation of summaries and analogies and analytic and holistic abilities. *American Educational Research Journal, 27*, 489-502.

Xu, S., Jiang, H. & Lau, F.C. (2009). User-oriented document summarization through vision-based eyetracking. In *Proceedings of the 13th International Conference on Intelligent User Interfaces,* (pp. 7-16). New York: ACM.

If Nostradamus were an Ergonomist: a review of ergonomics methods for their ability to predict accidents

Eryn Grant, Paul M. Salmon, & Nicholas J. Stevens
Centre for Human Factors and Sociotechnical Systems
University of the Sunshine Coast
Australia

Abstract

Forecasting loss events before they occur is the biggest challenge facing safety science. Typically, improving safety has been underpinned by retrospective accident analysis. While this approach has been valuable, many domains have reached a safety plateau where incident rates are not decreasing as they once were (indeed some are increasing). A proactive strategy for monitoring system performance with the aim of predicting adverse events provides a means to redress this. This article describes the first step in the development of a new accident prediction method, which included a review and evaluation of ergonomics methods for their ability to be used in a predictive manner. Six systems ergonomics methods were evaluated for the extent to which they could identify a series of core accident causation tenets derived from integrating contemporary accident causation models. The findings suggest that Cognitive Work Analysis and Event Analysis of Systemic Teamwork are the most suited for development into a formal accident prediction methodology. Implications for practice and future research steps are discussed.

Introduction

Safety traditionally relies on the experience of adverse events, where retrospective analysis is used to learn from the past and prevent future accidents. While this is a valued method for enhancing safety, many domains employing such analysis are finding themselves in a safety plateau, where incident rates are not decreasing as they once were (Dekker & Pizer, 2016; Salmon et al., 2017; Walker et al., 2017). A proactive approach to system safety is the next logical step. Previous analysis of the most widely used systems thinking based accident models has shown that there may be value in integrating their principle tenets of accident causation. Further, the identified tenets may support the development of a new structured approach to accident prediction (Grant et al., 2018). Grant et al. (2018) identified fifteen core systems thinking tenets, which describe the system properties underpinning accidents. The study concluded that the tenets could be combined with an appropriate systems analysis methodology to provide a framework for accident prediction.

In D. de Waard, F. Di Nocera, D. Coelho, J. Edworthy, K. Brookhuis, F. Ferlazzo, T. Franke, and A. Toffetti (Eds.) (2018). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In line with the contemporary knowledge on accident causation, any prediction method should be underpinned by systems thinking. This requires that methods account for the complexity of systems and analyse the interactions between the social and technical domains, while not reducing the system to its constituent parts (Underwood & Waterson, 2014, Salmon et al., 2012). Various candidate systems analysis methods are available (Salmon et al., 2011; 2017), including accident analysis methods such as AcciMap (Rasmussen, 1997), Functional Resonance Analysis Method (FRAM; Hollnagel 2012), and the Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004) and systems analysis and design methods such as the Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2008), Cognitive Work Analysis (CWA; Vicente, 1999) and Hierarchical Task Analysis (Stanton, 2006). All methods, accept for AcciMap, have been applied in some form of predictive context. Risk assessment investigations have been applied using EAST (Stanton et al., 2017); STAMP (STPA; Leveson et al., 2015); FRAM (Jensen and Avin, 2015) and HTA (NET HARMS; Dallat et al., 2017). CWA, has previously been applied predictively in a transportation context (Salmon et al., 2014). The aim of this paper is to communicate the findings of a method assessment to determine, which of the above methods would be most suited to prediction. To assess each methods utility for accident prediction a pre-defined criterion was developed based on Stanton et al. (2013) criteria for assessing Human Factors methods. Second the methods were evaluated for their capacity to identify fifteen systems thinking tenets.

Systems thinking tenets

The systems thinking tenets were identified as part of a wider program of research pursuing a predictive ergonomics method for accidents in complex sociotechnical systems. As part of a literature review of the most cited systems based accident analysis methods it became apparent that many accident causation models exist with wide-ranging contributions to safety (Grant et al., 2018). The review identified fifteen common accident tenets that represent the shared principles of accident causation. However, an outcome of this analysis identified the tenets were also ‘key values’ of system safety with the capacity to reveal a dichotomy between safe and unsafe systems (see Table 1). That is, the tenets can apply to both safe and unsafe operations of a system at any one time. This is important as a proactive approach to safety requires both the diagnosis of unsafe conditions and a return to safe operation.

The tenets themselves are not a method for prediction. To fully measure their potential a suitable ergonomics method that can be used in conjunction with the tenets is required. The aim of this paper is to communicate the findings an assessment to determine which method is best suited to this task.

Table 1. Systems Thinking Tenets (Grant et al., 2018)

	Definition	Safe	Unsafe
Vertical Integration	Interaction between levels in the system hierarchy	Decisions and actions at the higher levels filter down to lower levels and impact behavior. Information regarding the status of the system filters	Decisions and actions do not filter through the system and impact behavior on the front line. Information on the current status of the system is not used when making

		back up the hierarchy and influences higher level decisions and actions	process decisions
Constraints	Influences that limit the behaviours available to components within a system.	Specific constraints introduced to control hazardous processes	Restricts appropriate performance variability
Functional dependencies	The necessary relationships between components in a system.	Relationships between functions are expected and sustained	Dependencies that are not wanted or expected
Emergence	An outcome or property that is a result of the interactions between components in the system that cannot be fully explained by examining the components alone.	Emergent behaviours that support the goals of the system	Behaviours that undermine the goals of the system
Normal performance	The way that activities are actually performed within a system, regardless of formal rules and procedures	Behaviour is flexible enough to cope with adverse conditions	Behaviours cannot cope with the unfolding situation
Coupling	An interaction between components that influences their behaviour; both tight and loose interactions	Tight: connections between components are evident Loose: recovery from disturbances in the system is possible	Tight: Cascading failures when one component breaks down Loose: Loss of control regulating behaviours. Too much independence. Duplication of functions leading to inefficiencies.
Non-Linear interactions	Interactions are complex relationships between components where the outcome is not predictable	Allows for adaptations in the system.	Inconsequential events have large effects, cannot predict the effect of changes
Linear interactions	Direct cause effect relationships between components where the outcome is predictable.	Predicable and dependent	Interactions are predefined and fixed with no allowances for adaptations
Modularity	The organisation of a system where sub systems and components interact but are designed and operate largely independently of each other.	The system is resilient to breakdowns, replacement or substitutions of components and organisation of sub systems can be easily made	The system is tightly integrated and complex, substitutions cannot be made
Feedback Loops	Communication structure and information flow to evaluate control requirements of hazardous processes	Feedback is received on system breakdowns allowing control of hazards	Communication structures are not in place to provide or receive system feedback.
Decrementalism	Small changes in normal performance that gradually result in large changes.	Complex systems need to adapt, small adaptations are required to maintain optimisation	Constant small organisational changes create conflicts and pressure
Sensitive dependence on initial conditions	Characteristics of the original state of the system that are amplified throughout and alters the way the system operates (interconnected webs of relationships).	Mechanisms for monitoring changes are available	No understanding of initial conditions and their influence on the system
Unruly technologies	Unforeseen behaviours or consequences of technologies.	Technology that supports adaptation through a mechanism that is beyond the	Technology that introduces and sustains uncertainties about how and when things

		scope of what is was designed for affording flexibility.	may fail
Performance variability	Systems and components change performance and behaviour to meet the conditions in the world and environment in which the system must operate.	Performance varies to meet the needs of changing conditions	Performance does not change when conditions change
Contribution of the protective structure	The organised structure and system control that are intended to optimise the system, instead they do the opposite.	Protective structures are effective, flexible and adaptable in maintaining controls	Protective structure inhibits performance variability. Introduces new tasks that do not contribute to the goal. Unnecessary controls

Selected methods for review

A review of the literature identified six candidate methodologies selected on the basis that they were systems ergonomic methods that had previously been applied to examine accident causation or system properties. The methods identified from this process were; AcciMap (Rasmussen, 1997), the Functional Resonance Analysis Method (Hollnagel, 2012), Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004), Hierarchical Task Analysis (Stanton, 2006), Cognitive Work Analysis (CWA; Vicente, 1999) and the Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2008). A brief overview of each method is presented below.

AcciMap

The AcciMap method was developed as a technique for depicting the causal web underlying accidents in line with Rasmussen's risk management framework. Rasmussen's (1997) model represents complex sociotechnical systems as a hierarchy, accounting for the dynamic context in which systems operate, which is characterised by rapid change, high dependence on information and communication technologies and often volatile economic and political landscapes (Vicente & Christoffersen, 2006). To complete an AcciMap the analyst identifies contributory factors and relationships between them and places these onto the hierarchy. Links are then made between the nodes at various levels creating a descriptive diagram of the system. AcciMap is primarily used as an accident analysis method to describe how dynamic sociotechnical systems are subject to a fast pace of change, and how accidents occur because actors within the system adapt to change in unpredictable ways.

FRAM

Functional Resonance Analysis Method explains a system in terms of the mutually coupled or dependant functions relative to the whole system focusing on what a system does rather than what it is (Hollnagel, 2012). The system is described by the functions required to complete its tasks and possible variability that may occur in those functions (Lundberg et al., 2009). A FRAM analysis begins by identifying system functions using six basic characteristics (or aspects) these being; input, output, precondition, resource, control, and time (Hollnagel, 2012). Various

functions are linked by these characteristics to show how different functions are coupled. Functional resonance is the detectable signal when a variable combination of system functions causes one functions' variance to be unusually high and is explained as the 'unintended interaction of normal variability' (Hollnagel, 2012). To complete a FRAM analysis the analyst will describe all the functions the system requires (inputs and outputs) for success and how work in that system is done (as opposed to how it is imagined). It will then 'model' the expected or potential variance in the system using possible scenarios.

STAMP

Leveson's (2004) model of system behaviour, System Theoretic Accident Model and Processes (STAMP) uses functional abstraction to model the structure of a system and describe the interrelated functions. In comparison to other accident analysis methods STAMP's aim is to identify control and feedback loops and where they failed. To do this STAMP utilises a hierarchical control structure, which is a model explaining the regulation of a sociotechnical system. A taxonomy of control failures is found both in STAMP and STPA (Leveson, 2015) which is a hazard analysis technique based on STAMP. This taxonomy includes: inadequate control actions, inadequate execution of control actions and missing or inadequate feedback (Leveson, 2004; 2015, Salmon et al., 2012). The control structure is divided into two models, one for system development and one for operations. STAMP employs the use of constraints to maintain safe operation in systems.

HTA

Hierarchical Task Analysis (HTA) is used to describe the system under analysis in terms of goals, operations, and plans. HTA has a long history within Human Factors and ergonomics and was developed in the late 1960's to analyse complex non-repetitive tasks (Annett & Duncan, 1967). HTA is used to decompose goals and sub-goals to reveal the operations required to achieve them. It does this by focussing on observed behaviour to describe task goals and sub goals in a hierarchical form. In an HTA a goal is broken down to its component parts to show the top-level goal of the system and is then accompanied by a description of the necessary task step in a hierarchy. A novel application of HTA and its associated task network, is NET-HARMS which has revealed positive results as a risk assessment method (Dallat et al., 2017).

CWA

Cognitive Work Analysis is a method used to support the analysis, development and evaluation of sociotechnical systems (Jenkins et al., 2009). It was originally developed by Rasmussen et al. (1994) in response to the need for designers to consider non-routine situations when designing process control rooms. It is a composite of five phases designed to consider types of constraints on a system. The phases are used to model possibilities for the different types of behaviours available rather than how a behaviour may actually be done (Read et al., 2015a). CWA is an optimising tool, often used in design contexts because of this (Read et al., 2015b). CWA has undergone development as others in the field have further contributed to

its application most notably Vincente (1999) and Naikar et al. (2005). CWA has been used in multiple domains including but not limited to, command and control, interface design, transportation safety, pedestrian safety and health care.

EAST

The Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2008; Walker et al., 2010) provides an integrated suite of methods for analysing the performance of human-technical systems. It is underpinned by a 'networks of networks' approach in which three interlinked network-based representations are used to describe and analyse activity. The networks represent the tasks, social organisation and information the system requires to operate successfully. Task networks are used to describe the goals and subsequent tasks being performed within a system. Social networks are used to analyse the organisation of the system and the communications taking place. Information networks show how information and knowledge is distributed across different agents within the system. An important contribution of EAST is its integration of both human and non-human agents into the three networks. Task, social and information networks are finally combined revealing the complexity of the system under analysis and demonstrating a deeper understanding of behaviour in human-technical systems.

Method

To assess the methods the authors first applied a pre-defined checklist to the systems ergonomics methodologies under review and second identified what tenets could be identified as an output or result analysis of each method. The pre-defined criteria based on Stanton et al.'s (2013) procedure determined the qualities of each method. While the criteria did not directly assess qualities for prediction, it was an overview of affordances that assisted in selecting the most appropriate method. A description of the criteria can be found in Table 2. To assess the extent that methods could identify the tenets, the authors independently evaluated the application stages of each method using the descriptive tenet definitions found in Table 1. Separately the authors applied these definitions to evaluate if a tenet could be identified as an output or result of analysis for each method. If it was believed a tenet could be identified a further evaluation of "explicit" or "implicit" was recorded. If rated explicitly this referred to a clear and obvious identification of the tenet as an output of the method, if rated as implicit the tenet could not be directly associated as an analysis output, but may be present and could not be fully excluded.

Once the authors had independently reviewed the predefined criteria and tenet identification for each method under review the results were compared and discussion of any discrepancy was undertaken. If consensus was not met a decision was based on a majority rule. Once the assessments had been completed the methods where then weighted based on their scores from the pre-defined checklist and the number of tenets they could identify (including the degree that they could do so.)

Table 2. Description of assessment criteria based on Stanton et al. (2013)

Criteria	**Description of the criteria assessed**
Application	A general description of the method's application and uses. The criteria provide a contextual background.
Safety related published applications (August 2017 Google Scholar)	A count of the safety publications using the method. This criterion provides a useful measure to evaluate the past associations within safety contexts.
Used Predictively	Examples of previous predictive uses of the method. This criterion provides useful insight into predictive qualities of the method.
Tailorable	An assessment of the flexibility of the method to be used in different ways. This criterion is useful to indicate if a method can be altered for the purposes of prediction.
Approximate training and application times	An assessment of the complexity of the method to use learn and apply. This is included to assess the ease of use for future applications of the method in practice.
Related methods	This criterion shows similarities between methods, and underpinning methods which are related and/or integrated into its application.
Reliability and validity (has reliability and/or validity been tested and if so with what result)	Has the method undergone reliability and validity testing? This criterion is useful to assess if the method has proved reliable and effective for its purpose.
Tools needed	What tools are required to perform the analysis? This criterion is useful to understand the necessary requirements to complete an analysis.
Systems thinking tenets identified overall	A list tenet available as an output of the method (either explicit or implicit). This criterion shows if a method can identify the systems thinking tenets which informs how capable the method could be when used with the tenets and its potential for future uses.

Results

The results of the criteria and tenet identification are presented in Table 3. The top six rows of the table explain the method criteria assessment. The bottom row indicates in bold the tenets that were explicitly identified and tenets in plain text are implicitly identified for each method.

The findings suggest that AcciMap, CWA and EAST are the most suited to accident prediction; the criteria evaluation were sound and the methods were deemed to be capable of identifying fourteen (14) of the fifteen (15) systems thinking tenets. The main differences between the three methods arose from the criteria results, which identified that AcciMap had low reliability and validity scores (Branford et al. 2011; Waterson et al. 2017). However, it must also be noted that both CWA and EAST have not undergone full reliability validity testing. AcciMap was notably easier to learn and apply that other methods. While CWA and EAST shared the disadvantages of high complexity and application times they both had been previously applied predictively (see Salmon et al., 2014; Stanton et al., 2017).

The remaining methods FRAM, STAMP and HTA did not perform as well. FRAM and STAMP were deemed capable of identifying thirteen (13) and twelve (12) tenets respectively and HTA nine (9). It was noted in the criteria evaluation that FRAM and STAMP had not undergone formal reliability and validity testing and were complex with high application times. However, STAMP has been used predictively via the STPA method (see Leveson 2015). HTA scored the lowest on paper, however it should be noted that it is arguably the most flexible method and can be used with other of techniques such as human error identification (see Baber and Stanton 1996). While HTA has not been used predictively the SHERPA method has, which uses HTA as a stage of analysis (Embrey 1986). NET-HARMs a new risk assessment method is underpinned by HTA's task network. HTA is easy to learn, however it does have high application times. While HTA has not undergone formal reliability and validity on its own, Human Error Identification techniques have, which are underpinned by HTA (Stanton & Stevenage 1998).

Table 3. Results of method criteria and assessment of systems thinking tenets

	AcciMap	**FRAM**	**STAMP**	**HTA**	**CWA**	**EAST**
Application	**Accident causation method** Generic used in multiple domains. It is a graphical representation of factors and their causal relationships to the occurrence of an accident represented across multiple levels of a system.	**Accident causation method** Generic and can be applied to complex systems. The aim of FRAM is to identify potential variability within the functions of a system. This is represented by combinations of relationships between causal factors.	**Accident causation method** Generic and can be applied to complex systems. A graphical representation of a systems structure showing multiple levels and how they interact. Controls are enforced to prevent unsafe behaviours.	**Task Analysis method** Generic method. Describes activity under analysis in terms of hierarchy of goals sub-goals, operations and plans.	**Cognitive task analysis method**. Generic Method. Models complex socio technical systems. Functional properties, nature of activities It is used to describe constraints in a domain.	**Descriptive Method** Integrates several methods. Its aim is to adequately describe all the degrees of freedom inherent in complex socio technical systems
Safety related published applications (May 2016) (Google Scholar)	54	20	26	20	42	14
Used Predictively	No	Yes (Jenson & Aven 2017)	Yes (Leveson et al. 2015)	Yes (SHERPA; Embrey 1986)	Yes (Salmon et al. 2014)	Yes (Stanton & Harvey 2017)
Tailorable	Yes	No	Yes	Yes	Yes	Yes
Approximate training and application times	Low training time however considerable application time	It is proposed that FRAM is easy to learn however the analyst is required to have in-depth knowledge of the system under investigation (it is plausible that training and application times would be higher).	Low training time however considerable application time	Low training times but application times may be high depending on the system under analysis	Method is complex. Training times are high. High application times	Moderate training the method is complex and requires a lengthy application time.

Related methods	Actor map Rasmussen risk management framework			Used as the first step in many other HF methods (HEI, HRA and mental workload assessment). Best used alongside other methods.	Abstraction hierarchy, decision ladder, Contextual activity template, Strategies Analysis Diagram, information flow maps, SRK framework, Cognitive Work Analysis Design Toolkint (Read et al. 2016)	HTA, task networks, social network analysis, situation awareness networks
Reliability and validity (has reliability and/or validity been tested and if so with what result)	Yes (Waterson et al., 2017; Branford, 2011)	No	Yes (Underwood et al., 2016)	Yes (Stanton & Young, 1999)	No	No
Tools needed	Pen and paper. Software drawing packages are required to produce outputs	Pen & paper Software tool is available to draw visual output	Pen & paper Software drawing packages are required to produce outputs	Pen & paper HTA software tool	Pen and paper. Video and audio recording equipment. Software drawing packages are required to produce outputs	Pen and Paper. Software drawing packages and applications to draw visual output
Systems thinking tenets identified by method Bold = explicitly identified Normal text = implicitly identified	**vertical integration, functional dependencies, emergence, normal performance non-linear and linear interactions, modularity, decrementalism and unruly technologies,** constraints, coupling, and sensitive dependence on initial conditions, performance variability and contribution of the protective structure	**constraints, functional dependencies, emergence, normal performance, coupling, non-linear and linear interactions, modularity, feedback loops, decrementalism, unruly technologies and performance variability,** sensitive dependence on initial conditions	**vertical integration, constraints, functional dependencies, linear interactions and feedback loops,** emergence, modularity, unruly technologies, performance variability, coupling, non-linear interactions and contribution of the protective structure	**normal performance, linear interactions and feedback loops,** vertical integration, constraints, coupling, non-linear interactions and contribution of the protective structure	**constraints, functional dependence, emergence, normal performance, performance variability, contribution of the protective structure, non-linear interactions and linear interactions** vertical integration, coupling, sensitive dependence on initial conditions, unruly technologies, modularity and feedback loops	**vertical integration, functional dependence, emergence, normal performance, decrementalism, unruly technologies, and feedback loops performance variability, contribution of the protective structure, non-linear interactions, linear interactions** constraints, coupling and modularity

Discussion

This methods review aimed to determine which of a series of systems ergonomics methods best met the criteria to be used in a predictive capacity. The results show that AcciMap, CWA and EAST appear to be the most appropriate for use in

conjunction with the tenets to predict accidents in complex sociotechnical systems. Of the remaining methods FRAM and STAMP performed well, however both were shown to be overly complex methods and they failed to identify as many systems thinking tenets as other methods. HTA appears to be the least favourable on paper, however it is the most flexible, has been used predictively and is easy to learn. FRAM, STAMP and HTA were excluded due to their respective scores.

The suitability of AcciMap as predictive method is questionable as it has not been applied predictively and the method itself relies on retrospective information. Its main emphasis is the analysis of systems that have already been subject to incidents. AcciMap does have a high publication record, attesting to its value as an accident analysis method. CWA and EAST differ in comparison as they represent an analysis of systems as they are (and as they are intended to be) and both have been applied predictively. While CWA has been applied extensively in the literature, its safety related publications were relatively low in comparison. This also applied to the EAST method. While this could be interpreted as unfavourable, it provides an opportunity for their possible extension into safety contexts and future predictive applications. For this reason, CWA and EAST are the most likely candidate systems ergonomic methods that may be applied in a predictive context.

A general evaluation of the tenets indicates they are well represented across the six methods, however there may also be room to improve methods that do not identify all tenets as analysis outputs. For example, Sensitive dependence on initial conditions and Decrementalism were the least identified tenets in this review. This may mean that the systems ergonomic methods are better at identifying some tenets over others. Considering that the tenets are important to understanding (and possibly predicting) safety performance, an opportunity exists to extend ergonomic methods and improve toolkits where they may be deficient.

Limitations

The most functional means to assess each of the selected methods for their suitability in a predictive context would have been a practical test of each one. However, given the time limitations of the research this was not available. To overcome this the authors have endeavoured to be as thorough as possible in the method criteria assessment as outlined in the method section of this paper.

Research agenda

It is the authors' belief that both methods require further analysis to determine which performs best in a safety related context specifically their execution of the systems thinking tenets. Therefore, the next phase will test both CWA and EAST on a safe and unsafe scenario and further test their ability to use the systems thinking tenets as a diagnostic tool.

Conclusions

An essential next step to move beyond the safety plateau experienced by many safety critical domains is to predict accidents before they occur. While several

systems ergonomics methods have been applied predictively, there is yet no structured approach to accident prediction. This paper presented the findings of a method assessment of six systems ergonomics methods to determine the most appropriate to be used in a predictive context using a method criteria assessment and application of a set of tenets believed to be key properties of both safe and unsafe systems states. The following methods were assessed; AcciMap (Rasmussen, 1997), Functional Resonance Analysis Method (FRAM; Hollnagel, 2012), Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004), Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2008), Cognitive Work Analysis (CWA; Vicente, 1999) and Hierarchical Task Analysis (Stanton, 2006). Results show that CWA and EAST are equally favourable; both achieved sound results in the method criteria and showed a high capability to identify the systems thinking tenets. Further testing is needed and will require an application of the methods to existing accidents in both safe and unsafe states. This will test whether CWA or EAST is more efficient at identifying the systems thinking tenets and ultimately the most suitable for prediction.

References

Annett, J., & Duncan. K.D. (1967). *Task analysis and training design*. (ED 019 566). Dept. of Psycology, Hull University UK.

Baber, C., & Stanton, N.A. (1996). Human error identification techniques applied to public technology: predictions compared with observed use. *Applied Ergonomics, 27*, 119-131.

Branford, K. (2011). Seeing the big picture of mishaps: Applying the AcciMap approach to analyze system accidents. *Aviation Psychology and Applied Human Factors*, *1*, 31.

Dallat, C., Salmon, P.M., & Goode, N. (2017). Identifying risks and emergent risks across sociotechnical systems: the NETworked hazard analysis and risk management system (NET-HARMS). *Theoretical Issues in Ergonomics Science.* https://doi.org/10.1080/1463922X.2017.1381197.

Dekker, S., & Pitzer, C. (2016). Examining the asymptote in safety progress: a literature review. *International Journal of Occupational Safety and Ergonomics*, *22*, 57-65.

Embrey, D.E. (1986). SHERPA: A Systematic Human Error Reduction and Prediction Approach. In *Proceedings of the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems* , Knoxville, Tennessee American Nuclear Society La Grange Park, Illinois 60525.

Grant, E., Salmon, P.M., Stevens, N.J., Goode, N., & Read, G.J.M. (2018). Back to the future: What do accident causation models tell us about accident prediction? S*afety Science, 104*, 99-109.

Hollnagel, E., (2012). *FRAM: the functional resonance analysis method: modelling complex socio-technical systems*: Ashgate Publishing, Ltd.

Jenkins, D.P. (2009). *Cognitive work analysis: coping with complexity*: Ashgate Publishing, Ltd.

Jensen, A., & Aven. T. (2017). Hazard/threat identification: Using functional resonance analysis method in conjunction with the Anticipatory Failure

Determination method. In *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*:1748006X17698067.

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety science, 42*, 237-270.

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. *Reliability Engineering & System Safety* 136:17-34.

Lundberg, J., Rollenhagen, C. and Hollnagel. E. (2009). What-You-Look-For-Is-What-You-Find – The consequences of underlying accident models in eight accident investigation manuals. *Safety Science 47* 1297-1311. DOI: http://dx.doi.org/10.1016/j.ssci.2009.01.004.

Naikar, N. (2005). Theoretical concepts for work domain analysis, the first phase of cognitive work analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 3, pp. 312-316). Sage CA: Los Angeles, CA: SAGE Publications.

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science, 27*, 183-213.

Rasmussen, J., Pejtersen, A.M., & Goodstein, L.P. (1994). *Cognitive systems engineering*: Wiley.

Read, G.J.M., Salmon, P.M., & Lenné, M.G. (2015a). Cognitive work analysis and design: current practice and future practitioner requirements. *Theoretical Issues in Ergonomics Science* 16 (2):154-173. doi: 10.1080/1463922X.2014.930935.

Read, G.J.M., Salmon, P.M., Lenné, M.G., & Stanton, N.A. (2015b). Designing sociotechnical systems with cognitive work analysis: putting theory back into practice. *Ergonomics* 58, 822-851. DOI: 10.1080/00140139.2014.980335.

Read, G. J. M., Salmon, P. M., & Lenné, M. G. (2016). When paradigms collide at the road rail interface: evaluation of a sociotechnical systems theory design toolkit for cognitive work analysis. *Ergonomics, 59*, 1135-1157. doi:10.1080/00140139.2015.1134816

Salmon, P.M., Stanton, N.A., Lenné, M., Jenkins, D.P., Rafferty, L. and Walker, G.H.. (2011). *Human Factors Methods and Accident Analysis Practical Guidance and Case Study Applications*. U.K.: Ashgate.

Salmon, P.M., Cornelissen, M. and Trotter. M.J. (2012). Systems-based accident analysis methods: A comparison of Accimap, HFACS, and STAMP. *Safety Science, 50*, 1158-1170.

Salmon, P.M., Walker, G.H., Read, G J. M., Goode, N., & Stanton., N.A. (2017). Fitting methods to paradigms: are ergonomics methods fit for systems thinking? *Ergonomics, 60*, 194-205. doi: 10.1080/00140139.2015.1103385.

Salmon, P.M., Lenné, M., Read, G.J.M., Walker, G., & Stanton. N.A. (2014). Pathways to failure? Using work domain analysis to predict accidents in complex systems. *Advances in Human Aspects of Transportation: Part II*: 258-266.

Stanton, N.A., and Stevenage, S.V. (1998). Learning to predict human error: issues of acceptability, reliability and validity. *Ergonomics, 41*, 1737-1756.

Stanton, N.A. (2006). Hierarchical task analysis: Developments, applications, and extensions. *Applied Ergonomics, 37*, 55-79. DOI: http://dx.doi.org/10.1016/j.apergo.2005.06.003.

Stanton, N. Baber, C., & Harris, D. (2008). *Modelling command and control : event analysis of systemic teamwork*. Aldershot, Hampshire, England ; Burlington VT: Ashgate.

Stanton, N.A., & Harvey, C. (2017). Beyond human error taxonomies in assessment of risk in sociotechnical systems: a new paradigm with the EAST 'broken-links' approach. *Ergonomics, 60,* 221-233. DOI: 10.1080/00140139.2016.1232841.

Stanton, N.A., Salmon, P.M., Rafferty, L.A., Walker, G.H., Baber, C., & Perkins, D.P., (2013). *Human factors methods: a practical guide for engineering and design*: Ashgate Publishing, Ltd.

Underwood, P. & Waterson, P. (2014). Systems thinking, the Swiss Cheese Model and accident analysis: A comparative systemic analysis of the Grayrigg train derailment using the ATSB, AcciMap and STAMP models. *Accident Analysis & Prevention, 68,* 75-94. DOI: http://dx.doi.org/10.1016/j.aap.2013.07.027.

Underwood, P., Waterson, P. and Braithwaite, G. (2016). Accident investigation in the wild' – A small-scale, field-based evaluation of the STAMP method for accident analysis. *Safety Science, 82,* 129-143.

Vicente, K.J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*: CRC Press.

Vincente, K.J. and Christoffersen, K. (2006). The Walkerton E. coli outbreak: a test of Rasmussen's framework for risk management in a dynamic society. *Theoretical Issues in Ergonomics Science,* 7, 93–112.

Walker, G.H., Stanton, N.A., Baber, C., Wells, L., Gibson, H., Salmon, P.M., & Jenkins, D. (2010). From ethnography to the EAST method: A tractable approach for representing distributed cognition in Air Traffic Control. *Ergonomics, 53,* 184-197.

Walker, G.H., Salmon, P.M., Bedinger, M., & Stanton. N.A. (2017). Quantum ergonomics: shifting the paradigm of the systems agenda. *Ergonomics, 60,* 57-166. DOI: 10.1080/00140139.2016.1231840.

Waterson, P., Jenkins, D. P., Salmon, P. M., & Underwood, P. (2017). 'Remixing Rasmussen': The evolution of Accimaps within systemic accident analysis. *Applied ergonomics, 59,* 483-503.