Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving

Stapel, Jork; Mullakkal-Babu, Freddy Antony; Happee, Riender

DOI
10.1016/j.trf.2018.11.006

Publication date
2019

Document Version
Final published version

Published in
Transportation Research Part F: Traffic Psychology and Behaviour

Citation (APA)
Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *60*, 590-605. https://doi.org/10.1016/j.trf.2018.11.006

Transportation Research Part F 60 (2019) 590–605

Contents lists available at ScienceDirect

Transportation Research Part F

journal homepage: www.elsevier.com/locate/trf

Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving

Jork Stapel *, Freddy Antony Mullakkal-Babu, Riender Happee

Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, the Netherlands
Transportation and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, the Netherlands

ARTICLE INFO

Article history:
Received 15 February 2018
Received in revised form 1 September 2018
Accepted 10 November 2018
Available online 11 December 2018

Keywords:
Automated driving
On-road
Workload
Experience
Attention

ABSTRACT

Driver mental workload is an important factor in the operational safety of automated driving. In this study, workload was evaluated subjectively (NASA R-TLX) and objectively (auditory detection-response task) on Dutch public highways (~150 km) comparing manual and supervised automated driving in a Tesla Model S with moderators automation experience and traffic complexity. Participants (N = 16) were either automation-inexperienced drivers or automation-experienced Tesla owners. Complexity ranged from an engaging environment with a road geometry stimulating continuous traffic interaction, and a monotonic environment with lower traffic density and a simple road geometry. Perceived and objective workload increased with traffic complexity. When using the automation, automation-experienced drivers perceived a lower workload, while automation-inexperienced drivers perceived their workload to be similar to manual driving. However, the detection-response task indicated an increase in cognitive load with automation, in particular in complex traffic. This indicates that drivers under-estimate the actual task load of attentive monitoring. The findings also highlight the relevance of using system-experienced participants and the importance of incorporating both objective and subjective measures when examining workload.

1. Introduction

Monitoring ability is essential in an increasing number of vehicles offering supervised, or SAE2 automation (SAE International, 2016), which require the driver to monitor the automation and intervene when needed. Driver mental workload is an important factor in the operational safety of supervised automation. When automation relieves the driver from the continuous control tasks, mental underload can occur (De Waard, 1996). Over time, this can lead to a state of drowsiness, inattention and slower reactions (Greenlee, DeLucia, & Newton, 2018; Hirose, Kitabayashi, & Kubota). This has raised concerns regarding the driver's ability to monitor the automation and his/her performance to intervene in critical situations (Kyriakidis et al., 2017).

In order to address these effects, it is important to know how workload is affected by the use of automation, and how this effect varies with driving conditions. This study focuses on two main moderating variables of workload: the complexity of the driving environment and the driver's experience with the automation. Understanding the effect of these moderators can help to predict in which conditions workload is too high or too low. Experience with driving automation can lead to task

* Corresponding author.
E-mail addresses: J.C.J.Stapel@tudelft.nl (J. Stapel), F.A.MullakkalBabu@tudelft.nl (F.A. Mullakkal-Babu), R.Happee@tudelft.nl (R. Happee).

https://doi.org/10.1016/j.trf.2018.11.006
1369-8478/

execution at a lower cognitive level, or reduce the perceived complexity of the traffic situation (Paxion, Galy, & Berthelon, 2014; Young & Stanton, 2007). Automation experience can also lead to better monitoring and improved cognitive readiness for familiar driving situations, resulting in higher control transition performance (Krampell, 2016; Larsson, Kircher, & Andersson Hultgren, 2014; Paxion et al., 2014; Wright, Samuel, Borowsky, Zilberstein, & Fisher, 2016; Young & Stanton, 2007). Moreover, automation experience may reduce task demand, or reduce sensitivity to demand changes, and thus influence workload differently in high and low traffic complexity (Patten, Kircher, Ostlund, Nilsson, & Svenson, 2006; Stanton, Hedge, Brookhuis, Salas, & Hendrik, 2005).

This study investigates how workload changes with monitored automated driving in real-world conditions, and how this change is moderated by traffic complexity and by the driver's prior experience with automated driving. We conducted an on-road experiment on Dutch public highways in a Tesla Model S. The change in workload was assessed subjectively (NASA R-TLX) as well as objectively (auditory detection-response task). Traffic complexity was moderated by driving in a monotonic, low workload and a complex, engaging highway. To moderate automation experience, participants were either automation-inexperienced drivers or automation-experienced Tesla owners. The conditions were driven both manually and with automation. This resulted in a 2 (automation: on vs. off) × 2 (environment: monotonic vs. engaging) × 2 (experience: experienced vs. inexperienced) mixed design as illustrated in Fig. 1.

1.1. Theories of workload

In line with resource theory and the capacity model (Kahneman, 1973), we describe workload as the ratio between task demands and resources available to meet them. (A discussion of alternative definitions can be found in (Cain, 2007)). Task demand depends on the complexity of the driving task and the traffic situation, but also on how the goals are set (i.e. accepting a level of performance), and the strategy chosen to achieve it. To meet these demands, the driver has to allocate physical and mental resources, which are limited in availability. Driving consists of multiple sub-tasks. To model when and how much these tasks interfere, Wickens (1981) proposed the multiple resource theory in which resource pools are available for the different modalities of perception (e.g. visual, auditory, tactile), the codes of processing (spatial or verbal) and response selection and execution (hands, feet, speech). In addition, he proposed a cognitive resource shared across all tasks.

Resources are finite in capacity, but the upper limit is considered elastic (Kahneman, 1973; Young & Stanton, 2002), and closely related to the driver's energetic state. Drivers may exert state related effort to improve their energetic state. Investing computational effort can compensate for increasing demand. Both forms of effort are consciously perceived, and are considered key aspects of perceived workload (De Waard, 1996).

The relation between task demand and workload is u-shaped (De Waard, 1996) and consists of regions of underload, optimal load and overload. In optimal load, performance is generally good and changes in demand have little or no effect on perceived effort or achieved performance. Overload occurs when demands exceed the available resource capacity and performance degrades despite the additional effort invested. Underload occurs when demands are exceptionally low or monotonous in nature. Underload can lead to vigilance decrement, or inattention. However, low task demand can lead to an increase in workload when drivers recognize the development of drowsiness and invest state-related effort to compensate (Warm, Parasuraman, & Matthews, 2008).

Experience can make some demanding tasks impose less or no effort, even when performed concurrently with effortful tasks. These include routine operations and learned skills, executed with a high degree of automaticity. Examples are lane keeping, speed or headway maintenance and event detection. When automatized routines can handle the situation, these driving tasks should be insensitive to changes in cognitive load. According to the cognitive control hypothesis, cognitive load from competing tasks can only emerge for non-automatized tasks or when overruling skill-based behavior (Engström, Markkula, Victor, & Merat, 2017). We thus expect automation-experienced drivers to have a lower workload during automation compared to automation-inexperienced drivers. Conversely, the cognitive control hypothesis predicts that supervised automated driving, which mainly automates skill-based tasks, should not reduce workload for skilled drivers compared to manual driving.

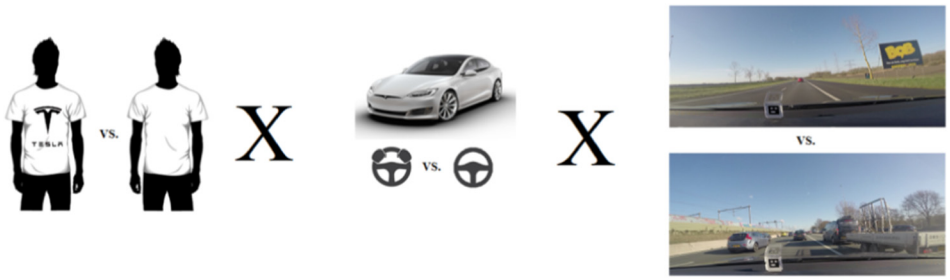

Fig. 1. Illustration of the independent variables: automation experience, automation use, and complexity of the environment.

1.2. Measuring workload

There is an extensive amount of literature reviewing methods to measure workload, e.g. (Cain, 2007; De Waard, 1996; Miller, 2001; Paxion et al., 2014; Stanton et al., 2005, chap. 39; Stanton, Salmon, Rafferty, Walker, Baber, & Jenkins, 2013, chap. 8; Young, Brookhuis, Wickens, & Hancock, 2015). Each measure is sensitive to a different set of resource pools, and in different performance regions (De Waard, 1996). Here we discuss measures used in the present study. The collection of workload measures can be classified into subjective rating (self-report) or objective measures (task-performance and physiological measures).

Subjective rating reflects workload as experienced by the operator (driver) and is thus sensitive to changes in effort. It is the simplest way to measure workload and is considered more reliable than physiological measures (Miller, 2001). The NASA task load index (TLX) (Hart, 2016) is a commonly used subjective measure in aviation and automotive research, and captures operator workload through six dimensions (mental, physical and, temporal demand; effort, frustration and performance) and reduces variability between participants and task contexts by letting participants score the relevance of each of these items. A variant, the Raw TLX (R-TLX), ignores this scoring step and has been found to remain an effective workload measure (Hart, 2016). We adopted the R-TLX to reduce the length of the post-drive questionnaire.

Subjective workload ratings have high face validity, but ratings may deviate from the actual workload. Stanton (1995) and Young and Stanton (1997) suggest in the contextual attention theory (CAT) that imbalance between perceived and actual demands and/or resources is one of the mechanisms through which poor performance can emerge, and that such an imbalance is especially likely in automated driving when there is insufficient feedback on the driver's performance (Norman, 1981). In order to capture such a mismatch, it is necessary to also collect objective measures of workload.

Objective workload measures often derive from task-performance, assuming reduced performance with under- and overload. Performance can either be measured on the primary task, or on secondary tasks. Most primary tasks in driving require manual operation of the vehicle (e.g. lane keeping performance), and are not suitable for automated driving. Secondary tasks aim to measure the driver's spare capacity. They tend to have a high reliability and can be designed to target specific resource pools. Consequentially, a variety of secondary tasks can be found throughout literature. One drawback of secondary tasks is that they interfere with the primary task. The detection-response task (DRT) is a secondary task designed to measure driver's cognitive load, and has been verified extensively (NEN-ISO 81748, 2016). Specifically, it measures the driver's ability to shift attention between the primary driving task and the DRT by measuring the delay between stimulus and response. When using a modality not interfering with the driving task (i.e. tactile or auditory), it is regarded as a pure measure of cognitive load. Compared to other secondary tasks, the additional cognitive demand induced by the DRT is generally considered to be low (Martens & van Winsum, 2000), but not effortless and not prone to automaticity (Engström et al., 2017). We selected the auditory DRT since monitoring of automation is centered in the cognitive resource pool, has low interference with the driving task and is not visually distracting. We preferred auditory over tactile stimuli, as this minimizes intrusive instrumentation.

Physiological measures sensitive to changes in workload include cardiovascular activity, galvanic skin response, brain activity and pupilometry. Brookhuis regards physiological measures as "*the most natural type of workload index, since, by definition, work demands physiological activity*" (cited in Stanton et al., 2005, p.17–2). Physiological measures can be recorded continuously and unlike performance measures they do not require any task to be performed, which makes them interesting for driver state monitoring. Cardiac monitoring is one of the most commonly used physiological measures of workload. Mental effort is associated with arousal which increases heart rate, while heart rate variability is found to decrease under high mental effort (Stanton et al., 2005, chap. 20, chap. 39). This relation between heart rate variability and mental effort is related to the sympathovagal balance between the sympathetic (0.02–0.06 Hz) and parasympathetic (0.15–0.40 Hz) nervous system, which is measured in the 0.10 Hz range, or as the ratio between high and low frequency ranges (Though the idea that the LF/HF ratio is a suitable indicator for the sympatho-vagal balance has been challenged; see Billman (2013) for a comprehensive review). However, heart activity (and variability in particular) are not selective measures of workload. They primarily respond to the body's regulatory functions and are hypersensitive to noise from movement, changes in breathing rate and speech (Jorna, 1992; Young, 2000). We recorded heart activity and explored LF/HF ratio and standard deviation of inter-beat intervals, as they are related to mental workload and less affected by artifacts than other variability measures (Stapelberg, Neumann, Shum, McConnell, & Hamilton-Craig, 2017). Eye measures related to workload include blink rate, horizontal gaze dispersion (for highway driving) and pupil diameter (Marquart, Cabrall, & de Winter, 2015). The latter is particularly sensitive to high levels of cognitive load, but requires careful control of light conditions (Kahneman, 1973). In addition, eye tracking can provide further insight into the quality of monitoring (i.e. task performance) by assessing changes in glance frequency and durations to regions of interest (Kircher & Ahlstrom, 2017). We included eye tracking in our study to assess visual load and monitoring quality.

1.3. Empirical workload in automated driving

The theories of workload can help to explain and predict how automation and other moderators affect the driver's workload, but for quantitative effects we need to examine empirical findings. To this end, we selected studies addressing effects of driving automation, traffic complexity and automation experience.

The empirical review from De Winter, Happee, Martens, and Stanton (2014) summarizes workload findings from 32 studies comparing different levels of automation on the NASA TLX and RSME (Rating Scale Mental Effort). TLX responses were converted to a percentage scale for better comparison to RSME, with the lowest item mapped to 0% and the highest to 100%. Studies were mainly performed in simulators, and indicated a workload reduction of 21% on average from manual to automated driving. Six of the reviewed automated driving conditions could be considered SAE2 (Damböck, Weißgerber, Kienle, & Bengler, 2013; McDowell, Nunez, Hutchins, & Metcalfe, 2008; Saxby, Matthews, Warm, & Hitchcock, 2013; Schermers & Malone, 2014). With SAE2 automation workload was only 13.5% lower compared to manual driving. Ratings ranged from 23 to 66% for manual and from 23 to 40% for SAE2 automated driving.

The influence of traffic complexity on workload can be as large as the use of driving automation, with a 35% workload increase from low to high traffic complexity in manual driving (Teh, Jamson, Carsten, & Jamson, 2013). During supervised automated driving, traffic increases demands for the monitoring task (Jamson, Merat, Carsten, & Lai, 2013).

While task complexity increases demand, experience with automation may reduce it. Until recently, the influence of experience with automation could hardly be investigated due to the unavailability of automation-experienced drivers. Simulator studies on workload in automation often include a familiarization period, but the 15–30 min exposure times are too short for the development of experience (Beggiato, Pereira, Petzoldt, & Krems, 2015). Some studies have approximated automated driving experience by using adaptive cruise control (ACC) experienced drivers (Larsson et al., 2014; Naujoks, Purucker, & Neukum, 2016) or developed special procedures to create experience through training (Krampell, 2016). Some effects of experience, such as the perceived risk and trust, may also be hard to study in simulators, which pose limitations on the perceptual fidelity (De Winter, van Leeuwen, & Happee, 2012; Hallvig et al., 2013). However, some recent studies measured mental workload during automated driving on the road.

Solís-Marcos, Ahlström, and Kircher (2018), measured visual secondary task performance in a Volvo S90 equipped with pilot assist (SAE2) and included both automation-inexperienced drivers and vehicle owners who had experienced the automation for 4.5 months on average before participation. In contrast to their expectations, they found that automation use increased the percentage of incorrect responses to the secondary task compared to manual driving, despite similar task completion rates in both conditions and longer glances towards the visual task with automation. TLX ratings of mental effort were high (79% in manual driving and 67% with automation use), which indicates that in supervised automation, secondary visual-motor tasks can be very demanding. Automation-experienced drivers gave shorter glances to the road compared to automation-inexperienced drivers in all conditions. They also gave longer glances at the secondary task, and this behavior was more pronounced during automated driving compared to manual driving, whereas the inexperienced drivers did not change glance time with automation use.

Banks and Stanton (2016) studied the workload of automation-inexperienced drivers during a short but engaging trip in a prototype supervised automated vehicle. In contrast to findings from simulators, the perceived workload was higher during automated driving (42%) compared to manual driving (27%). The participants' lack of prior training with the system, the additional tasks (performing three lateral maneuvers and answering an interview) and reported issues with the automation's behavior may all have contributed to the perceived workload increase.

Heikoop, de Winter, van Arem, and Stanton (2017) performed an on-road test with professional drivers familiar to supercars, but with no prior experience with lateral automation in a Tesla Model S on the highway, following a lead vehicle after 30 min of test-track training. A simple secondary task (counting bridges) was performed during part of the trip. The perceived workload during automated driving was rated very low overall (average of 19%), which is even below findings from simulator literature and reduced over time, suggesting that accustomization occurred during the trip. Accordingly, negative standardized change scores between the pre-drive and post-drive engagement ratings on the Dundee stress-state questionnaire suggest an overall disengagement during the drives.

Eriksson, Banks, and Stanton (2017) investigated the transition time in non-critical control transitions on the road in a Tesla Model S and compared it to a simulator study. Participants in the on-road experiment had prior experience with driving automation while the participants of the simulator study did not. Drivers in the on-road experiment regained control 32% (1.5 s) faster on average compared to the simulator drivers. The workload was perceived as low in both studies and no significant difference was found between the two studies.

Naujoks et al. (2016) performed a field study measuring secondary task uptake, secondary task workload and compensatory behavior in congested traffic while driving manually, with ACC and ACC plus steer assist in a Mercedes-Benz E-Class. They explored the effect of automation experience by comparing drivers with and without prior ACC experience. ACC-experienced drivers performed more secondary tasks in automated driving than in manual driving, in particular when driving at lower speeds, suggesting reduced workload with automation at lower driving speed. The effect however was not present for ACC-inexperienced drivers, suggesting that automation experience is a prerequisite for freeing cognitive resources for secondary tasks.

Based on these preceding works, we formulated the following hypotheses for supervised automation:

H1. Workload will be higher in the engaging condition than in the monotonic condition for both manual and automated driving.

H2. Automation will reduce workload.

H3. Workload during automated driving will be higher for automation-inexperienced drivers compared to automation-experienced drivers.

We expect these effects to occur for both objective (auditory DRT) and subjective (R-TLX) workload measures. It should be noted that H2 concurs with a wide range of findings in various tasks, including driving in simulators, but not with the cognitive control hypothesis. Also, opposite effects were reported in two recent on-road studies as reviewed above.

2. Methodology

2.1. Participants

Two groups (N = 8 each) of participants took part in the experiment and were selected through convenience sampling. Automation-experienced Tesla owners were recruited through the Dutch/Belgium section of the Tesla Motors forum (Motors, 2017). Seven reported using a Tesla and its Autopilot on a daily basis. One was an irregular user but reported 10,000 km travelled using Autopilot. One of the experts was the safety instructor, who had observed 8 participants prior to taking part himself.

The automation-inexperienced participants were invited through the universities' employee mailing list and through a list of drivers who had indicated their interest to participate in research regarding automated driving. Inexperienced drivers were required not to have experienced driving automation before. Users of adaptive cruise control were excluded but users of non-adaptive cruise control were included. The demographics of both groups are summarized in Table 1.

2.2. Vehicle and instrumentation

An on-road driving task was performed with a rented Tesla model S 75D equipped with Autopilot (hardware version 1; update 8.0) and the driver's seat on the left side. The vehicle features supervised automation, which combines adaptive cruise control with automated lane keeping. The system supports lane changes (which have to be initiated by the driver) and adapts driving speed to traffic in the adjacent lanes and road curvature. The automation requires the driver to keep the eyes on the road and the hands on the wheel. An overview of the instrumentation can be seen in Fig. 2. Video was recorded with three GoPro cameras observing the traffic in front and behind of the car, as well as the driver. A webcam observed the instrument panel.

An auditory detection response task (DRT) was performed as an objective measure of the driver's cognitive workload. The DRT was implemented in Python on a Raspberry PI 3B running Raspbian Jessie. The implementation and analysis were in line with NEN-ISO 8(1748, 2016), with the following notable exceptions:

– An auditory stimulus was provided randomly with an on-set interval of 3–5 s with a 3.1 kHz tone lasting one second, irrespective of response time.

Table 1
Demographics of the two participant groups, with mean μ, standard deviation σ and [interval].

	Experienced group	Inexperienced group
Age	μ = 43 σ = 14 [27–69]	μ = 41 σ = 14 [21–61]
Years licensed	μ = 22 σ = 15 [4–51]	μ = 21 σ = 15 [3–43]
km driven past 12 months	μ = 26.500 σ = 21.500 [7.500–75.000]	μ = 15.000 σ = 13.000 [3.000–42.500]
Gender	7 male, 1 female	8 male

a: Eye tracker
b: DRT button + heart activity sensor
c: Webcam
d: GPS antenna
e: Experimenters
f: IMU + DRT

Fig. 2. Overview of the instrumentation.

- Stimuli were presented over 5 min at a time (amounting to 72 stimuli per participant per condition).
- The button used to respond to the stimuli was strapped to the participant's right index finger, as the right hand had no driving-related tasks other than steering during the DRT.
- The DRT instruction was phrased as "Press the button as soon as you hear the signal, but keep your attention on the road".

Heart activity was recorded as a psychophysiological measure of arousal and workload. Two variability metrics were analyzed: standard deviation of inter-beat intervals (sdNN), where low variability indicates high workload; and low over high frequency ratio (LF/HF), where a high ratio indicates a high workload. These metrics were calculated every 30 s over 300 s of data.

Heart activity was recorded using an optical sensor mounted to the participant's right middle finger, powered by an Atmel AtMega328P embedded processor board. The sensor was able to obtain a heart rate measure, but occasionally suffered from artifacts (e.g. holding the steering wheel differently changed contact pressure of the sensor or reduced blood circulation in the fingertip). The heart rate and variability metrics were calculated using an open-source Python toolbox (Van Gent, 2017; van Gent, Farah, van Nes, & van Arem). Data were collected at 100 Hz, and low pass filtered with a second-order Butterworth with a cutoff frequency of 5 Hz. The dominant (R-wave) peaks were identified as the maximum sample from any signal section rising above a 1.5 s moving average. Sections of poor data were identified by a variety of error detection and peak rejection algorithms, including the exclusion of heart rates outside the normal range [30–130 bpm] as well as any R-peaks whose associated inter-beat intervals exceeded the [250–300 ms] range.

A pupil labs head mounted eye tracker with the Linux distribution of pupil capture (Pupil-Labs V0.9.1, 2017) was included for the exploratory glance behavior analysis. However, a power outage on the second testing day resulted in software corruption, leading to random crashes of the tracking software. As a consequence, we only obtained full recordings of two participants and dropped the eye tracking from further analysis.

Vehicle motion (6 DOF acceleration, speed and location) was recorded using an MPU6050 IMU and GTPA013 GPS sensor connected to a second Atmel processor.

A safety instructor sat next to the participant and was proficient in the use of the Autopilot and experienced in introducing new drivers to the vehicle. During the drive, his tasks were to inform or warn the driver when needed, to help with the navigation and vehicle settings from the center console and to provide answers to technical questions. He was also allowed to engage in idle conversations except when instructions were given by the experimenter or during the DRT. The participant was allowed to initiate a conversation at any time. We did not inhibit speaking to maximize behavioral validity. By allowing participants the freedom to engage in conversation, the effects of experimenters' presence on behavior became more representative to having any other passenger.

To control for confounders that are inevitable in an on-road study, the DRT data was enriched by annotating events which may influence the response, such as lane changes, uninstructed (dis)use of the automation and verbal interactions. For each stimulus-response pair, the following classifications were made through manual annotation of the video footage:

- Lane change: ego vehicle undergoing a lane change or having indicators activated
- Use of Autopilot (on/off)
- Periods of congested traffic (vehicle or traffic speed slower than 75 km/h)
- Driver speaking (y/n)
- Other occupant speaking (y/n)

A stimulus was classified when these events occurred at any moment between the end of this stimulus and the end of the preceding stimulus.

To obtain an accurate record of the experienced traffic conditions, traffic flow (intensity) and traffic speed were logged from the NDW open data server (Warehouse, 2018) every minute. For each recording that contained both values, traffic density was calculated as lane-averaged intensity divided by lane-averaged traffic speed, where empty lanes were ignored. After pre-processing, the lane-average traffic data was interpolated to the GPS position and time, to obtain a continuous estimate of the traffic condition surrounding the vehicle. This interpolation accounted for differences in information travel in free and congested traffic as described in the traffic-adaptive model of Treiber and Helbing (2002). When GPS data was not available, a single average was computed for the road section over the duration of the condition.

2.3. Subjective measures

Three questions were asked while driving, to which the participant responded verbally on a scale from 1 to 9. The first question covered mental demand, the second regarded alertness and the third question reflected the driver's trust. In each condition, the three questions were asked before and after performing the DRT to verify that this task did not influence the driver's state. The questions were phrased as: *On a scale from 1 to 9, how mentally demanding was the {manual driving, use of autopilot}? On a scale from 1 to 9, how alert were you during the {manual driving, use of autopilot}? On a scale from 1 to 9, how much did you trust {yourself with the driving, the automation}?* For the alertness question, the descriptions of Karolinska sleepiness scale (KSS) (Kaida et al., 2006) were used: 1: very sleepy, great effort to keep awake 2: sleepy, some effort to keep awake 3: sleepy, but no effort to keep awake 4: some signs of sleepiness 5: neither alert nor sleepy 6: rather alert 7: alert 8: very

alert 9: extremely alert. Each time, the participant was reminded of the description of the given response and was permitted to revise the response accordingly. The demand and trust questions were not anchored while driving, but 1 was described as low and 9 as high before departure.

The NASA Raw Task Load Index (R-TLX) was filled out after each driving condition on a 21-point scale. (we report results in percentages, with the lowest possible rating mapped to 0% and the highest possible rating mapped to 100%) Additionally, a confidence questionnaire based on Rendon-Velez et al. (2016) was used, with items (1) *driving manually was easy*, (2) *I felt confident to drive manually*, (3) *I had a feeling of risk*, (4) *using the automation was easy*, (5) *I felt confident to use the automation* and (6) *I had a feeling of risk during automated driving* on a 5-point scale with anchors: *disagree strongly, disagree a little, neither agree nor disagree, agree a little, agree strongly*. Also the 12-item automation trust questionnaire from Jian, Bisantz, and Drury (2000) was adopted on a 7-point scale. This questionnaire was only filled out for the drive as a whole, and not for each condition separately.

2.4. Environment

Two highway sections were selected to represent two levels of driving complexity; an engaging environment with a road geometry stimulating continuous traffic interaction, and a monotonic environment with lower traffic density and a simple road geometry and a low chance for high-demand scenarios to occur (Fig. 3).

For the engaging environment, the A10 (ring-East of Amsterdam) was selected for its high traffic density throughout the day and the 10–13 on/off-ramps (depending on direction traveled). To maximize the traffic interaction, the driver was instructed to drive in the right lane as much as possible and was allowed to overtake slow moving traffic. On parts of this road it is legal to use the shoulder lane, but we instructed drivers to keep the regular right lane to avoid unpredictable behavior of the Autopilot. The A10 was entered from the A1 and followed down till exit Oud Zuid, either driving manually or using the automation. The highway was then followed in the opposite direction until exit Zeeburg, during which the DRT was performed. The route was then repeated with the remaining mode of automation.

For the monotonic environment the A6 between Almere (exit 7) and Lelystad (exit 10) was selected, which is a straight two-lane highway with low traffic density and no on/off ramps between the two cities. Here the driver was instructed to remain in the right most lane, to not overtake slow traffic and to drive as fast as traffic permits, but not faster than 110 km/h. Drivers got stuck behind a truck or trailer driving approximately 85 km/h in 80% of the monotonic scenarios. The automation was either used on the way towards Lelystad or back towards Almere.

The two driving environments were located 15 min away from one another. The A1 connects the two locations and was used for the familiarization. The A1 was entered from the A9 and first traveled in eastern direction. When the familiarization was to be followed by the engaging environment in Amsterdam, the first available exit was taken before practicing the automated lane change, but no later than Naarden. When the familiarization was to be followed by the monotonic condition, the road was simply continued towards the A6. The order of monotonic/engaging drives after the familiarization was counter balanced. For the inexperienced driver, the total trip lasted for 1.5 h when first driving to the monotonic condition or 1.75 h when first driving to the engaging condition. The experienced driver needed 1.5 h for either route due to the shorter familiarization.

Fig. 3. The driven route. Images were recorded during the drive of participant 5.

The McDonald's Amsterdam Zuidoost was selected as the start/end point of the route, as it was logistically located between the highway entrance and the Tesla supercharger, and provided the facilities needed for welcoming the participants.

2.5. Procedure

The experiment was approved by the human research ethics committee (HREC) of the TU Delft. Upon arrival, the participants were informed of the tasks and risks of the experiment. A pre-drive questionnaire was filled out and the procedures were explained. Prior to departure, the safety instructor informed the participants about the operation and limitations of the vehicle and the automation while the experimenter positioned the eye tracker, heart rate sensor, and DRT button. To explain and demonstrate the basic operation of the vehicle, the safety driver closely followed a checklist covering the controls; all possible automation modes, their functional meaning and methods for activating and disengaging them. Also the automation-related symbols were explained using Fig. 4. Guidelines for safe use were phrased as: *when using autopilot, you should be on the lookout for things that the automation cannot handle correctly, for instance: lane markings that are not well visible or that have to be crossed, situations that require very strong braking or steering, traffic that behaves unexpectedly, when the car does something you would not do, or when it doesn't do something when you would.* The participants were further instructed repeatedly to remain attentive drivers at all times.

Once on the highway, a familiarization drive was performed, during which the participants were introduced to the general operation of the vehicle and the basic behavior of the automation. The performed tasks covered the different methods of activation and deactivation of the automation, the adjustment of the cruise speed setting and the automated lane change. Questions were asked to make sure that the driver understood the instrument panel and operation of the vehicle. The familiarization lasted as long as necessary to let the participant perform each task successfully. The inexperienced drivers needed around 20 min while the experienced drivers required 8 min on average. The participants then drove manually to the engaging or monotonic environments, where they performed 4 rides in manual/automated and engaging/monotonic conditions in a randomized order.

To test for a possible effect of the DRT on subjective workload, within each condition we drove 5 min without DRT and 5 min with DRT. Each condition started with the instructions regarding driving behavior, followed by 5 min of driving without DRT. The three questions regarding mental demand, alertness and trust were asked before and after the DRT.

3. Analysis and results

The drives were performed on workdays between the 3rd and 10th of March 2017. All drives took place between 9:00 and 16:45 to avoid rush hours and congestion. All tests occurred in normal (Dutch) weather conditions, except for two automation-experienced drivers, who drove in heavy rain. The automation operated reliably for all drivers in all experienced conditions. There was one occasion where ACC momentarily braked for no apparent reason (we suspect that the radar caught a guard rail at the end of a highway exit) and a couple of non-critical conflicts with surrounding traffic (e.g. another vehicle cutting in front, Autopilot attempting to undertake another vehicle at the right lane); all were adequately resolved by the participants with no or negligible inconvenience. Two experienced drivers did not follow all instructions during the monotonic condition. One occasionally overtook trucks. The other had not turned off ACC during manual driving in the first half of the monotonic drive.

During the DRT, the 8 automation-inexperienced users collectively made 8 lane changes in the engaging condition during the automated drive (2 with automation turned off) and 5 in the manual drive. The 6 experienced users made 17 in the automated drive (4 with automation turned off) and 15 in the manual drive, suggesting that the experienced users were more comfortable with making lane changes. In the automated engaging drive, the Autopilot was turned off 9.8% and 5.2% of the time by inexperienced and experienced drivers respectively, while in the monotonic drive the Autopilot was turned off 1.9% and 0.7% of the time respectively. The descriptives and effect sizes resulting from a 3-way ANOVA are provided in Tables 4 and 5and will be described in Sections 3.2 and 3.3.

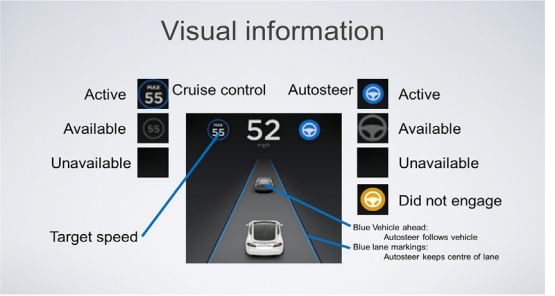

Fig. 4. illustration of the automation-related information on the instrument panel.

3.1. Traffic conditions

Despite a technical malfunction in the power supply to the GPS equipment, we managed to retain the GPS data of 48 5-minute conditions among 11 drives. Since the interpolated traffic estimates require this GPS signal, approximate traffic estimates were made for the conditions where the GPS signal was not available. The two methods were compared to ensure that the approximation is appropriate. Pearson's r and the difference in means between the two calculation methods as well as its significance according to a paired Student's *t*-test are outlined in Table 2. On all three metrics, the two calculation methods correlate well. Although the difference in means of traffic flow and (consequentially) density is statistically significant, we deemed its practical significance small enough to include the approximate traffic data in the analysis.

Table 3 shows the mean driving and traffic conditions among the experimental conditions, as well as the sample sizes for which GPS (and thus traffic speed) is available. Fig. 5 shows traffic density and speed of the individual participants. It demon-

Table 2
Pearson's correlation between interpolated and approximated estimates of traffic speed, flow and density.

	r	$E_{continuous-approx}$	t(1,47)	*p*
Traffic speed	0.836	−0.407 (km/h)	−0.415	0.680
Traffic flow	0.871	−85.7 (veh./h)	−2.858	0.006
Traffic density	0.929	−0.458 (veh./km)	−2.385	0.021

Table 3
Driving speed and traffic properties.

	Monotonic				Engaging			
	Automated		Manual		Automated		Manual	
	μ	σ	μ	σ	μ	σ	μ	σ
Car speed [km/h]	87.3	13.7	88.5	10.5	87.2	16.6	90.0	9.9
Traffic speed [km/h]	111.2	15.6	115.2	12.4	96.0	10.5	96.9	6.5
Traffic flow [veh./h]	751	524	707	530	1303	421	1302	448
Density [veh./km]	9.43	5.70	8.06	4.19	14.14	4.90	14.04	4.29
N_{used}	16		16		16		16	
N_{GPS}	8		8		8		4	
N_{GPS} during DRT	5		5		4		3	
$N_{traffic}$	16		16		14		13	
$N_{traffic}$ during DRT	14		14		13		11	

N_{used} = number of participants driving each condition.
N_{GPS} = number of participants for which GPS (and thus car speed) data is available.
$N_{traffic}$ = number of participants for which traffic data (speed, flow, density) is available.
During DRT = number of samples for which 5 min of data is available while performing the DRT.

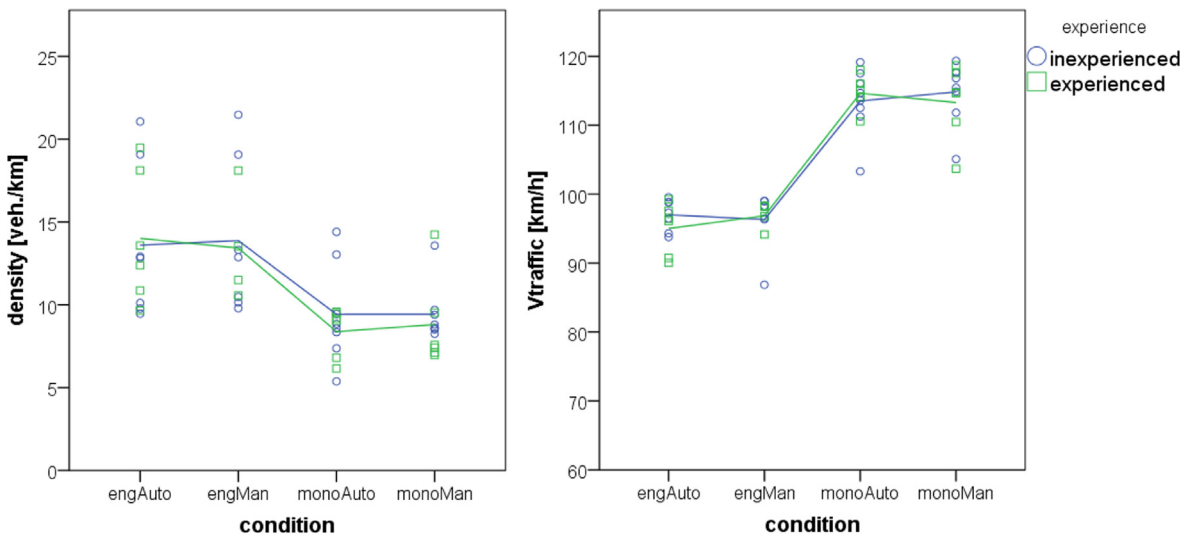

Fig. 5. Traffic density (left) and speed (right) as experienced by the two groups across all test conditions during the DRT. Lines indicate condition means, symbols represent individual participants. engAuto = engaging environment using automation, engMan = engaging environment with manual driving; mono" = monotonic environment with either driving mode.

strates that comparable traffic conditions were obtained between manual and automated driving for both groups, while differentiating between the two driving environments.

3.2. Detection response task (DRT)

The auditory DRT was performed as an objective measure of changes in cognitive load. Due to missing values, the experienced and inexperienced group are represented by six and eight participants respectively.

Lane changes contributed to only 0.10% of the time for the inexperienced group and 0.17% of the time for the experienced group. During lane changes in the engaging condition reaction times were 240 ms ($t(5) = 3.206$, $p = .024$) and 261 ms ($t(7) = 3.797$, $p = .007$) slower for the experienced and inexperienced drivers respectively. These effect sizes are comparable to the 1-back task (adds 232 ms to baseline manual driving according to NEN-ISO 8(1748, 2016, Table E8)) or counting backwards from a 3-digit number (adds 125 ms to baseline manual driving according to Merat, Johansson, Chin, Nathan, and Victor (2006, figure 30)). Of all DRT misses, 30% and 25% occurred during lane changes for the inexperienced and experienced group respectively, resulting in a miss rate of 12% for automation-experienced and 19% for automation-inexperienced drivers during lane changes.

The time spent conversing varied across participants, with 3 out of 14 participants accounting for 65% of all conversations held. When speaking, the participants' reaction time was 240 ms slower compared to when being silent ($t(13) = 3.45$, $p = .004$). No difference in reaction time was found between other occupants being silent or speaking ($t(13) = 1.388$, $p = .188$). The DRT thus showed a higher miss rate and slower response time during lane changes and a slower response time during speaking. This indicates that single DRT responses can uncover additional information when combined with the identification of external events.

To remove some of the confounders for the DRT, we omitted all stimuli that occurred during lane changes, uninstructed automation (dis)use, and during driving or traffic speeds below 75 km/h. This refinement excluded 10.4% of the stimuli (with 17.0% of stimuli removed from the engaging-automated condition, 18.7% from the engaging-manual condition, 1.8% from the monotonic-automated condition and 0.5% from the monotonic-manual condition). Results after this removal are presented in Tables 4 and 5. Stimuli during speaking or listening were not excluded from the analysis, as we consider speaking to be inherent to the driving strategy, but we verified if removing stimuli during speaking would change the trends and effects reported in Tables 4 and 5. All trends with automation, experience and complexity remained. Reaction time received slightly smaller effect sizes, but significant effects remained significant. The effect of environment on miss rate became insignificant ($p = .081$). It is however not surprising that miss rates, which depend on the occurrence of rare events, are sensitive to data removal.

Table 4
DRT reaction time and miss rate, and R-TLX.

	Reaction time (ms)				Miss rate (%)				R-TLX (%)			
	Experienced N = 6		Inexperienced N = 8		Experienced N = 6		Inexperienced N = 8		Experienced N = 7		Inexperienced N = 8	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Engaging												
Automated	548	212	532	161	3.10	2.00	4.19	3.09	24.30	20.3	42.90	20.1
Manual	491	232	460	160	2.55	2.15	3.54	2.99	48.30	19.6	42.60	17.1
Monotonic												
Automated	379	109	451	182	1.70	0.62	1.93	0.70	10.20	7.1	24.70	16.2
Manual	386	114	420	167	1.65	1.09	1.79	1.09	32.00	19.6	29.60	15.4

Table 5
Univariate effects and interactions of 3-way ANOVAs for reaction time, miss rate and overall workload. The effect of automation in the R-TLX contradicts that of RT and MR. Significant effects (confidence $\alpha < .05$) are highlighted. Sub-items indicate which R-TLX sub-scale items were significant (mental, physical, temporal, performance, effort, frustration).

Effects and interactions	Reaction time (ms)			Miss rate (%)			R-TLX (%)			
	F(1,12)	p	η_p^2	F(1,12)	p	η_p^2	F(1,13)	p	η_p^2	Sub-items
Environment	**15.168**	**0.002**	**0.558**	**7.036**	**0.021**	**0.370**	**14.584**	**0.002**	**0.529**	m,ph,t,pe,e
Automation	**11.321**	**0.006**	**0.485**	0.480	0.502	0.038	**8.871**	**0.011**	**0.406**	m,ph,t,e
Experience	0.028	0.870	0.002	0.972	0.344	0.075	0.886	0.364	0.064	–
Environment * experience	2.266	0.158	0.159	0.514	0.487	0.041	0.003	0.957	0.000	–
Automation * experience	1.306	0.275	0.098	0.007	0.934	0.001	**5.945**	**0.030**	**0.314**	m,pe,f
Environment * automation	4.111	0.065	0.255	0.360	0.559	0.029	0.175	0.682	0.013	–
Environment * automation * experience	0.218	0.649	0.018	0.000	0.999	0.000	1.119	0.309	0.079	–

Reaction times and miss rates show similar trends (see Table 5). Both reaction time and miss rate have a main effect of environment (engaging environment has 99 ms slower reaction time and 1.6% higher miss rate compared to the monotonic environment), indicating that the engaging environment is more demanding than the monotonic environment. This effect size is similar to adding the 0-back task to baseline manual driving, which increases reaction time with 99 ms according to NEN-ISO 8(1748, 2016, Table E8). Reaction time also shows a main effect of automation (38 ms slower reaction during automated driving compared to manual driving), indicating that automation use resulted in less spare cognitive capacity compared to manual driving. Miss rate did not increase significantly with automation. However this can be attributed to a capping effect from the measurement resolution; with 72 stimuli, only miss rate increments of 1.4% can be distinguished with each measurement. No main effect for experience was found. None of the interactions was statistically significant.

Although the interaction automation * environment is not statistically significant ($p = .065$), the increase in reaction time during automated driving is more pronounced in the engaging environment ($E_{auto\text{-}man} = 64$ ms, SE = 18) compared to the monotonic environment ($E_{auto\text{-}man} = 13$ ms, SE = 16). Thus, automation seems to increase cognitive workload particularly in the engaging condition. These effect sizes are similar (but opposite) to comparing DRT performance with and without manual driving (baseline driving increases DRT reaction time with 52 ms on average according to NEN-ISO 8(1748, 2016, Table E8)).

3.3. NASA R-TLX

Subjective workload ratings were collected using the NASA R-TLX. Due to missing values, the experienced and inexperienced group were represented by seven and eight participants respectively. The subjective workload is given in Table 4 and Fig. 6, which also includes the six contributing items. The main effects and interactions are given in Table 5. For each effect and interaction, we indicated the sub-scale items that were statistically significant. The familiarization condition was excluded from further analysis, as it differs in road type.

Overall, workload was perceived 15.4% higher in the engaging than in the monotonic environment, and 12.6% lower during automated driving compared to manual driving. The interaction automation * experience however indicates that only the experienced group ($E_{man\text{-}auto} = 22.9\%$, p = .003) perceived a workload reduction with automation use, whereas the inexperienced group did not experience a workload difference between manual and automated driving ($E_{man\text{-}auto} = 2.29\%$, p = .699). The perceived workload reduction with automation for the experienced group was consistent for both traffic conditions and for the six contributing items (Fig. 6 left). No main effect of experience and no further interactions were observed.

3.4. Heart activity

Individual heart rate traces are shown in Fig. 7. An ANOVA on the mean-adjusted heart rate did not reveal any main effects or interactions on environment or automation use. Similarly, no effects or interactions were found for sdNN or LF/HF. Heart rate varied over time without apparent relation to the conditions or observed events, We suggest that the inherent variability of on-road driving along with artifacts from speaking, gripping of the steering wheel and other confounders overshadowed any possible effects resulting from automation or complexity of the environment.

A linear regression on the mean adjusted heart rate shows a time on task effect for the inexperienced group ($b = -0.034$ bpm/min, $F(1,169) = 51.71$, $p < .001$), but not for the experienced group ($b = -0.002$ bpm/min, $F(1, 138) = 0.024$, $p = .878$). This indicates that the inexperienced drivers may have been acclimatizing to the vehicle and automation use during the experiment. Apart from this trend, heart rate proved ineffective to disclose significant effects of automation and traffic complexity.

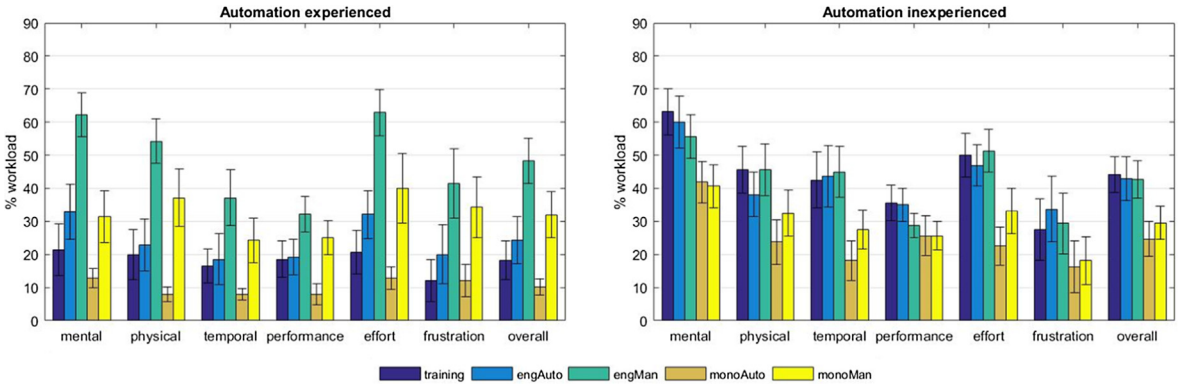

Fig. 6. R-TLX for the experienced (left) and inexperienced (right) group, converted to percentage. Whiskers indicate standard errors.

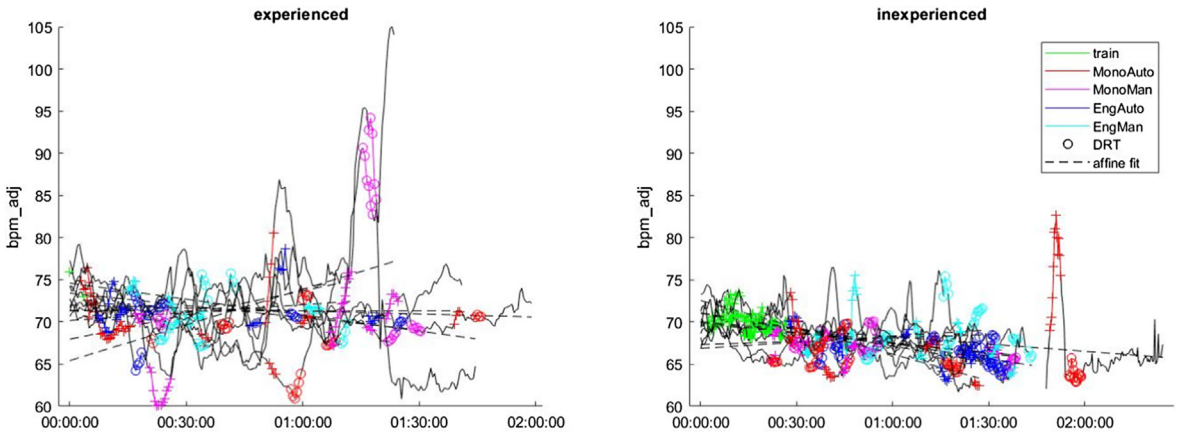

Fig. 7. Mean-adjusted heart rate (bpm) over time of experienced (left) and inexperienced (right) drivers for each individual driver. Each sample is calculated over a 5 min sliding window. Data collected during the conditions are highlighted.

3.5. Questionnaires

To assess the impact of the DRT, the participants rated mental demand, sleepiness and trust both before and after performing the DRT. Overall, mental demand was 32.4% without the DRT and 36.9% with DRT, which is a measurable yet small increase in mental demand ($F(1,9) = 3.361$; $p = .027$). No difference between driving with and without DRT was found for the KSS ($F(1,9) = 0.941$; $p = .357$) or for trust ($F(1,9) = 0.764$, $p = .405$).

Drivers reported lower sleepiness (KSS) in the engaging condition compared to the monotonic condition ($E_{mono-eng} = 0.875$ points; $F(1,14) = 18.08$, $p = .001$), but KSS was not affected by automation ($F(1,14) = 1.577$, $p = .230$), experience ($F(1,14) = 0.140$, $p = .714$) or on any of the interactions.

An overall confidence rating was computed over the items of the confidence questionnaire. Ratings ranged from 63% to 95% as can be seen in Table 6. The confidence questionnaire only showed a main effect of environment, with the engaging condition providing 13.4% less confidence than the monotonic condition ($F(1,12) = 13.38$; $p = .003$). The ratings suggest that experienced drivers felt more confident during automated driving than in manual driving, while the inexperienced drivers felt more confident during manual driving than in automated driving, but the interaction automation * experience was not statistically significant ($F(1,12) = 4.37$; $p = .059$).

In the 12-item automation trust questionnaire, automation-experienced drivers reported a higher trust in the automation than the automation-inexperienced drivers. (experienced: $\mu = 84.9\%$, $\sigma = 9.42$; inexperienced: $\mu = 67.7\%$, $\sigma = 14.0$; $t(1,13) = 2.82$; $p = .014$)

3.6. Post-hoc analysis

To complement the ANOVAs presented in Table 5, we checked post-hoc for any relations between traffic conditions, heart rate (BPM, sdNN, LF/HF) and DRT response time within the experimental conditions. Pearson's correlation was used to explore relations between the participant averages in each condition. Traffic density correlated inversely with traffic flow on both environments (Pearson's $r = -0.590$ for engaging; $r = -0.624$ for monotonic), which is in agreement with traffic flow models. Within the conditions, none of the traffic metrics correlated with neither RT nor with any of the heart rate metrics.

We further checked if the difference in mileage between participant groups could confound the results by observing the correlations between the DRT and R-TLX measures for both groups and all conditions. As shown in Table 7, 19 out of 24 cor-

Table 6
Confidence ratings among conditions and participant groups.

	Confidence (%)			
	Experienced N = 7		Inexperienced N = 7	
	μ	σ	μ	σ
Engaging				
Automated	91.9	11.7	63.0	30.4
Manual	67.9	20.0	73.7	28.4
Monotonic				
Automated	95.2	12.5	80.9	19.8
Manual	84.6	24.2	89.3	14.1

Table 7
Correlations between mileage and reaction time, heart rate and R-TLX. engAuto = engaging environment using automation, engMan = engaging environment with manual driving; mono" = monotonic environment with either driving mode. The *p*-values are not corrected for the 8 independent comparisons.

	Reaction time		Miss rate		R-TLX	
	Pearson's r	*p*	Spearman's ρ	*p*	Pearson's r	*p*
Inexperienced						
engAuto	−0.153	0.717	−0.621	0.100	−0.348	0.399
engMan	0.039	0.927	−0.185	0.660	−0.451	0.262
monoAuto	0.233	0.578	−0.113	0.791	−0.241	0.565
monoMan	0.202	0.631	0.296	0.476	−0.403	0.323
Experienced						
engAuto	−0.577	0.231	0.174	0.742	−0.351	0.440
engMan	−0.441	0.382	−0.696	0.125	−0.078	0.868
monoAuto	−0.292	0.574	−0.893	0.016	−0.323	0.480
monoMan	−0.395	0.439	−0.585	0.222	0.583	0.170

relations were of negative sign, which suggests a higher mileage is associated with better performance at the secondary task. However, the sample sizes are too small to make any conclusive statements regarding the effect of mileage. Although the miss rate correlation of the experienced group in the monotonic condition during automation was statistically significant (p = .016), this may be attributed to an inflated type I error from the multiple comparisons being made. A Bonferroni correction for the 8 comparisons would dictate that the probability is to be tested at a confidence of $\alpha' = 0.0064$ instead of $\alpha = 0.05$, in which case also this correlation is not statistically significant.

Because the dynamics of mental demand vary at a smaller time scale than a five-minute average can reveal, we further explored relations among the measurements at a shorter time scale by looking for correlations between individual DRT responses, heart activity and traffic conditions, as well as time on task effects. Since the raw response times are not normally distributed, Spearman's non-parametric correlation was used. No correlations were found between heart rate and the individual responses. Similarly, relations between the traffic conditions and individual responses did not reveal further relations within the driving conditions. No time on task effects were found for DRT reaction time using linear regression (b = 0.1 ms/stimulus, $p = .645$).

4. Discussion

In this study, we investigated how workload changes with attentively monitored automated driving in real-world conditions, and how this change is moderated by traffic complexity and by the driver's prior experience with automated driving.

The engaging traffic environment resulted both in a higher overall subjective (R-TLX) and objective (DRT reaction time and miss rate) workload compared to the monotonic environment. Additionally, the drivers remained as sensitive to changes in driving complexity while using automation as they were while driving manually. This supports hypothesis H1 and demonstrates that monitoring automation imposes a considerable task demand.

Hypothesis 2, reduced workload with automation, is only supported for the perceived overall workload (R-TLX) for automation-experienced drivers but not for automation-inexperienced drivers. Furthermore in both driver groups the objective cognitive load (DRT) increased with automation. These results were unexpected and show opposing effects on subjective and objective workload in the experienced drivers.

The R-TLX ratings suggest that automation experience is a prerequisite for a reduction in perceived workload. The automation-experienced user may have developed a less demanding (or automated) strategy for monitoring the automation, while the inexperienced driver may stay closer to strategies from manual driving. This view is also supported by Solís-Marcos et al. (2018), who showed that automation-inexperienced and -experienced drivers have different glance behavior and that only the automation-experienced group changes glance behavior with automation use. In our study, during automated driving, automation-experienced drivers did not perform better on the DRT task compared to automation-inexperienced drivers. Although we believe our per-group sample size is rather small to formally test hypothesis H3, we would like to point out that this observation also aligns with Solís-Marcos et al. (2018), who despite the longer glances of experienced drivers to the secondary task did not find a difference in task performance between the two groups. In contrast, Naujoks et al. (2016) found that automation use resulted in a higher secondary task completion rate compared to manual driving for ACC experienced drivers. A possible explanation could be the difference in driving speed, since higher performance in Naujoks et al. (2016) only occurred in slow-moving (<60 km/h) congested traffic.

It should be emphasized that our findings for automation-inexperienced drivers are in conflict with results from simulator studies as reviewed in De Winter et al. (2014), where automation-inexperienced drivers reported workload reduction due to automation of a magnitude similar to the reduction we found for experienced drivers. Since the automation-inexperienced group reported lower trust and confidence in the automation compared to the automation-experienced group, we suspect that this difference between real-world and simulator findings relates to the low validity of risk perception in driving simulators. This further suggests that driver trust accounts for a large difference in perceived workload reduction

by automation. We should however remark that on-road studies provide mixed results. Automation-inexperienced drivers perceived low workload during automation in Heikoop et al. (2017), while Solís-Marcos et al. (2018) found high workload ratings for both inexperienced and experienced users of automation. With the emerging on-road studies addressing workload in automated driving with automation-experienced users, we believe that a new meta-analysis on the effects of automated driving may be in order.

The dissociation between perceived overall workload (R-TLX) and objective cognitive load (DRT) for the effects of automation deserves further examination. Although overall workload incorporates more than cognitive load, we believe that a direct comparison between the two measures is fair, because the monitoring sub-task, which is centered in the cognitive resource pool, forms a large part of the driving task, and because the mental demand sub-items of the R-TLX show similar patterns as the overall load. The increase in reaction time and miss rates could theoretically be attributed to mental underload since such performance reduction is an indicator of vigilance decrement (Greenlee et al., 2018). We have however several indications that this is not the case. The Karolinska sleepiness scale did not indicate any development of drowsiness, and when drowsiness had been compensated with state-related effort, we should have seen this reflected in the effort or mental demand sub-scales of the R-TLX. Furthermore, the periods of automated driving were relatively short (10 min of automated driving at a time, interluded with verbal ratings after 5 min). Although vigilance decrement can develop in such time span, we should have been able to see such decrement as a time-on-task effect, which we did not in our regression analysis. Finally, the reaction time increase with automation use was larger in the engaging condition compared to the monotonic condition for both driver groups, which contradicts the hypothesis that longer reaction times of this study signify underload.

The increase in DRT reaction time can also not be fully explained by the malleable attentional resource theory (Young & Stanton, 2002, 2007), which suggests that total capacity reduces when task demands are low. In order to explain the *reduction* in DRT performance, the capacity reduction should have been larger than the reduction in primary task demand, whereas Young and Stanton (2002) propose that spare mental capacity should still improve, but disproportionally to the reduction in primary task demand. The cognitive control hypothesis cannot explain the increase in reaction time, but provides an explanation why supervised automation did not reduce objective workload.

The increase in objective cognitive load (DRT) suggests that the ratio between task demand and allocated resource increases with automation, whereas the reduction in subjective workload suggests that this increase is not perceived as such. Assuming that the TLX ratings are not confounded by confirmation bias or attribution error, we believe that this is caused by a mismatch between perceived and actual workload as suggested in contextual attention theory. Stanton (1995) proposed there can be a mismatch between the perceived and actual demands, between perceived and actual resources, or between the perceived demands and perceived resources. Under-estimation of task demand should result in too few resources being mobilized for both primary and secondary task. Such underestimation would reduce monitoring performance but may still lead to improved secondary task performance, unless the primary task demand is met with a resource allocation that is higher than the perceived requirement. An over-estimation of allocated resource is particularly likely in low-effort conditions, and could explain the reduced DRT performance under lower perceived workload. We can further expect workload to be rated lower than actual load, when a surplus demand investment (either perceived, or allocated in response to the instruction to monitor attentively) is ignored or weighted less in the overall workload rating. This suggests that either (1) automation increases demand while automation-experienced drivers perceive less effort and lower demands, or that (2) automation does not reduce demand as much as we think, and we allocate a larger fraction of our resource to monitoring than we made available for it. The first suggestion would be in conflict with hypothesis H2 whereas the second is not. Both however indicate a difference between perceived and actual workload. While drivers remain sensitive to changes in task demand (i.e. changes in traffic), they appear to over-estimate their resource allocation. The idea that experienced users under-estimate the actual task load has implications for safe usage of SAE2 automation. It indicates that supervised automation does not increase spare capacity as much for secondary tasks or interaction with in-vehicle information systems as the driver believes. These perceptual differences should be incorporated in the design and user-education for these systems. The results also reinforce the importance of measuring workload both objectively and subjectively. We recommend to incorporate objective measures of both primary task performance (i.e. monitoring) and spare capacity when studying mismatches between perceived and actual workload.

Our DRT findings suggest that attentively supervised automation results in a healthy workload (i.e. a little higher than manual driving) and thereby do not support the concern that supervised automated driving causes mental underload. In contrast, in particular the automation-experienced drivers perceived a reduced workload with automation. This means that, when drivers supervise attentively, they can maintain a healthy workload while perceiving a meaningful comfort benefit. However the mismatch between objective and perceived load may be a point of concern when it motivates users to pay less attention than is required, which in turn could mediate underload and may compromise safety.

5. Limitations

Eye tracking could not be assessed due to technical malfunction. The number of participants was limited, which in particular makes the between-subject comparisons sensitive to individual differences. However, the within-subject effects were very consistent among participants, and persisted when correcting for speaking during the experiment. The instructions and presence of a safety instructor and experimenters motivated attentive supervision of the automation. The results should

therefore be regarded as workload under intended use, which may differ from every-day use. The idle conversations held may have reduced the sensation of being in an experiment, raised energetic state and increased workload. Although these aspects are representative for a drive with other occupants, results may differ when driving alone, with no one to talk to. The effects of supervised automation with longer periods of automated driving in a naturalistic setting without additional motivations to supervise remain to be investigated. Convenience sampling balanced years licensed, age and sex. Mileage was not balanced between groups, but sample size was too low to correlate this between-group difference. Furthermore, the Tesla users were sampled from a forum which actively discusses the limitations and abilities of the vehicle. The owner's disposition towards the vehicle may have resulted in confirmation bias or attribution error (i.e. general satisfaction being expressed on the workload scale). As an approach to compensate for such rater bias in future experiments, we propose to assign automation users to different vehicle brands, or to group participants based on their disposition regarding the automation.

Acknowledgment

We kindly thank Paul van Gent for providing us with the heart rate and GPS equipment, and his support in pre-processing the cardiovascular recordings.

Funding

This work was supported by the NWO-TTW Foundation, the Netherlands, under the project "From Individual Automated Vehicles to Cooperative Traffic Management - Predicting the benefits of automated driving through on-road human behavior assessment and traffic flow models (IAVTRM)"-STW#13712.

References

Banks, V. A., & Stanton, N. A. (2016). Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53*(Pt B), 389–395. https://doi.org/10.1016/j.apergo.2015.06.020.

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour, 35*, 75–84. https://doi.org/10.1016/j.trf.2015.10.005.

Billman, G. E. (2013). The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in Physiology, 4*(26), 1–5. https://doi.org/10.3389/fphys.2013.00026.

Cain, B. (2007). A review of the mental workload literature. *Defence Research and Development Canada Toronto Human System Integration Section*, 1–35.

Damböck, D., Weißgerber, T., Kienle, M., & Bengler, K. (2013). Requirements for cooperative vehicle guidance. *IEEE Annual Conference on Intelligent Transportation Systems (ITSC), 16*, 1656–1661.

De Waard, D. (1996). *The measurement of drivers' mental workload*. Traffic Research Centre VSC.

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

De Winter, J. C. F., van Leeuwen, P. M., & Happee, R. (2012). Advantages and disadvantages of driving simulators: A discussion. *Proceedings of Measuring Behavior*, 47–50.

Engström, J., Markkula, G., Victor, T., & Merat, N. (2017). Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors, 59*(5), 734–764. https://doi.org/10.1177/0018720817690639.

Eriksson, A., Banks, V. A., & Stanton, N. A. (2017). Transition to manual: Comparing simulator with on-road control transitions. *Accident; Analysis and Prevention, 102*, 227–234. https://doi.org/10.1016/j.aap.2017.03.011.

Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2018). Driver vigilance in automated vehicles: Hazard detection failures are a matter of time. *Human Factors and Ergonomics Society, 18720818761711*. https://doi.org/10.1177/0018720818761711.

Hallvig, D., Anund, A., Fors, C., Kecklund, G., Karlsson, J. G., Wahde, M., & Akerstedt, T. (2013). Sleepy driving on the real road and in the simulator–A comparison. *Accident; Analysis and Prevention, 50*, 44–50. https://doi.org/10.1016/j.aap.2012.09.033.

Hart, S. G. (2016). Nasa-Task Load Index (NASA-TLX); 20 years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), 904–908. https://doi.org/10.1177/154193120605000909.

Heikoop, D., De Winter, J. C. F., Van Arem, B., & Stanton, N. A. (2018). *Acclimatizing to automation: Driver workload and stress during partially automated car following in real traffic*. Manuscript submitted for publication. https://doi.org/10.13140/RG.2.2.26822.16964.

Hirose, T., Kitabayashi, D., & Kubota, H. Driving characteristics of drivers in a state of low alertness when an autonomous system changes from autonomous driving to manual driving. SAE Technical Paper 2015-01-1407. Advance online publication. doi: 10.4271/2015-01-1407.

Jamson, A. H., Merat, N., Carsten, O. M. J., & Lai, F. C. H. (2013). Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transportation Research Part C: Emerging Technologies, 30*, 116–125. https://doi.org/10.1016/j.trc.2013.02.008.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology., 34*, 237–257.

Kahneman, D. (1973). Attention and effort. New Jersey: Englewood Cliffs.

Kaida, K., Takahashi, M., Akerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., & Fukasawa, K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 117*(7), 1574–1581. https://doi.org/10.1016/j.clinph.2006.03.011.

Kircher, K., & Ahlstrom, C. (2017). Evaluation of methods for the assessment of attention while driving. Accident; Analysis and Prevention. Advance online publication. doi: 10.1016/j.aap.2017.03.013.

Krampell, M. A. H. (2016). Accelerated behavioural adaptation through targeted training programs – the case of highly automated driving (Master's thesis). Department of Computer and Information Science (IDA) at Linköping University.

Kyriakidis, M., de Winter, J. C. F., Stanton, N., Bellet, T., van Arem, B., Brookhuis, K., & Happee, R. (2017). A human factors perspective on automated driving. *Theoretical Issues in Ergonomics Science, 53*, 1–27. https://doi.org/10.1080/1463922X.2017.1293187.

Larsson, A. F. L., Kircher, K., & Andersson Hultgren, J. (2014). Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour, 27*, 229–237. https://doi.org/10.1016/j.trf.2014.05.008.

Marquart, G., Cabrall, C., & de Winter, J. C. F. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing, 3*, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783.
Martens, M., & van Winsum, W. (2000). *Measuring distraction: The peripheral detection task*. TNO Human Factors.
McDowell, K., Nunez, P., Hutchins, S., & Metcalfe, J. S. (2008). Secure mobility and the autonomous driver. *IEEE Transactions on Robotics, 24*(3), 688–697. https://doi.org/10.1109/TRO.2008.924261.
Merat, N., Johansson, E., Chin, E., Nathan, F., & Victor, T. (2006). Specification of a secondary task to be used in safety assessment of IVIS.
Miller, S. (2001). Literature review: Workload measures. National Advanced Driving Simulator. Iowa City, United States.
National Data Warehouse for Traffic Information (2018). NDW open actuele dataservice. http://opendata.ndw.nu/.
Naujoks, F., Purucker, C., & Neukum, A. (2016). Secondary task engagement and vehicle automation – Comparing the effects of different automation levels in an on-road experiment. *Transportation Research Part F: Traffic Psychology and Behaviour, 38*, 67–82. https://doi.org/10.1016/j.trf.2016.01.011.
NEN-ISO 17488 (2016). ISO 17488:2016: DRT for assessing attentional effects of cognitive load. Nederlands Normalisatie Instituut.
Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*(1).
Patten, C. J. D., Kircher, A., Ostlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident; Analysis and Prevention, 38*(5), 887–894. https://doi.org/10.1016/j.aap.2006.02.014.
Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology, 5*, 1344. https://doi.org/10.3389/fpsyg.2014.01344.
Pupil-labs v0.9.1. (2017). Pupil Capture, Player, and Service release v0.9.1. https://github.com/pupil-labs/pupil/releases/tag/v0.9.1.
Rendon-Velez, E., van Leeuwen, P. M., Happee, R., Horváth, I., van der Vegte, W. F., & de Winter, J. C. F. (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour, 41*(41), 150–169. https://doi.org/10.1016/j.trf.2016.06.013.
SAE International (2016). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016 SEP2016).
Saxby, D. J., Matthews, G., Warm, J. S., & Hitchcock, E. M. (2013). Active and passive fatigue in simulated driving: Discriminateing styles of workload regulation and their safety impacts. *Journal of Experimental Psychology: Applied.*.
Schermers, G., & Malone, K. M. (2014). Dutch evaluation of Chauffeur Assistant (DECA): Traffic flow effects of implementation in the heavy goods vehicles sector.
Solís-Marcos, I., Ahlström, C., & Kircher, K. (2018). Performance of an additional task during level 2 automated driving: An on-road study comparing drivers with and without experience with partial automation. *Human Factors: the Journal of the Human Factors and Ergonomics Society, 001872081877363*. https://doi.org/10.1177/0018720818773636.
Stanton, N. A. (1995). Ecological ergonomics: Understanding human action in context. *Contemporary Ergonomics*, 62–67.
Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2013). Human factors methods: practical guide for engineering and design (2nd edition). Wey Court East, Union Road, Farnham, Surrey, GU9 7PT England: Ashgate Publishing Limited.
Stanton, N. A., Hedge, A., Brookhuis, K., Salas, E., & Hendrik, H. (2005). *Handbook of human factors and ergonomics methods*. Boca Raton: CRC Press.
Stapelberg, N. J. C., Neumann, D. L., Shum, D. H. K., McConnell, H., & Hamilton-Craig, I. (2017). The sensitivity of 38 heart rate variability measures to the addition of artifact in human and artificial 24-hr cardiac recordings. *Annals of Noninvasive Electrocardiology: The Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc, 23*(1). https://doi.org/10.1111/anec.12483.
Teh, E., Jamson, S., Carsten, O., & Jamson, H. (2013). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. *Transportation Research Part F: Traffic Psychology and Behaviour, 22*, 207–217. https://doi.org/10.1016/j.trf.2013.12.005.
Tesla Motors (2017). Tesla Motors Club Foum: Belgium and the Netherlands. https://teslamotorsclub.com/tmc/forums/belgium-and-the-netherlands.118/.
Treiber, M., & Helbing, D. (2002). Reconstructing the spatio-temporal traffic dynamics from stationary detector data. *Cooperative Transportation Dynamics, 3* (1-3), 24.
Van Gent, P., Farah, H., van Nes, N., & van Arem, B. Analysing noisy driver physiology real-time using off-the-shelf sensors: heart rate analysis software from the taking the fast lane project. (Submitted for Publication to the Journal of Open Research Software).
Van Gent, P. (2017). Python heart rate analysis toolkit. https://github.com/paulvangentcom/heartrate_analysis_python.
Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50*(3), 433–441. https://doi.org/10.1518/001872008X312152.
Wickens, C. D. (1981). Processing resources in attention, dual task performance and workload assessment.
Wright, T. J., Samuel, S., Borowsky, A., Zilberstein, S., & Fisher, D. L. (2016). Experienced drivers are quicker to achieve situation awareness than inexperienced drivers in situations of transfer of control within a Level 3 autonomous environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 270–273. https://doi.org/10.1177/1541931213601062.
Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics, 58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151.
Young, M. S., & Stanton, N. A. (1997). Automotive automation: Investigating the impact on driver mental workload. *International Journal of Cognitive Ergonomics, 1*(4), 325–336.
Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors, 44*(3), 365–375. https://doi.org/10.1518/0018720024497709.
Young, M. S., & Stanton, N. A. (2007). What's skill got to do with it? Vehicle automation and driver mental workload. *Ergonomics, 50*(8), 1324–1339. https://doi.org/10.1080/00140130701318855.
Young, M. S. (2000). Attention, automaticity, and automation: Perspectives on mental underload and perfrmance.