Non-Driving-Related Tasks During Level 3 Automated Driving Phases—Measuring What Users Will Be Likely to Do

Elisabeth Shi^{1, 2} and Alexander T. Frey²

¹ Chair of Ergonomics, Technical University of Munich

² Section F4 Automated Driving, Federal Highway Research Institute (BAST), Bergisch Gladbach, Germany

After activation of a Society of Automotive Engineers (SAE) Level 3 automated driving function (SAE International, 2021), the function takes over the driving task and the human user may engage in other, non-driving-related tasks (NDRTs). Meanwhile, the user needs to remain receptive to requests by the function because s/he needs to reengage in the driving task, when the function approaches a system limit and requests the user to take over. Hence, the effects of NDRT engagement on driver state and takeover behavior have been investigated closely. However, concerning relevance to traffic safety, it is important to take a step back and investigate what NDRTs users are likely to engage in, and what methods are suitable to collect respective data. Two experiments were conducted to answer these questions. In Experiment 1, participants experienced Level 3 automated driving in a Wizard-of-Oz vehicle on German motorways. After the ride, participants were asked to name NDRTs that they would engage in, if the function was available in their own vehicle. In Experiment 2, participants were asked to bring their own NDRTs along and experienced Level 3 automated driving in the Wizard-of-Oz vehicle. Comparison of results shows preferences of similar NDRTs (e.g., smartphone usage and reading). Moreover, we found that different methods provide different insights into NDRT engagement (i.e., engagement rate, total duration rate of engagement, and naming rate). Integrating our results in current literature landscape highlights the strong dependence of resulting NDRTs from the investigation method. Results, strengths, and weaknesses of the employed methods are discussed.

Keywords: automated vehicle, level 3, Wizard-of-Oz, non-driving-related task, method

Theoretical Background

The Society of Automotive Engineers (SAE) International Standard J3016 (2021) describes capabilities of automated driving systems operating on a sustained basis by categorizing them into Levels 0–5. Today’s driver assistance systems include functions operating on a sustained basis (Shi et al., 2020) that meet criteria for Society of Automotive Engineers (SAE) Level 2 (SAE International, 2021): They perform the longitudinal and lateral vehicle motion control on a sustained basis. The driver is expected to complete the object and event detection and response (OEDR)

subtask of the dynamic driving task (DDT). This means that the driver permanently needs to monitor the function and the environment and to determine whether the Level 2 function responds adequately to the current driving situation. If the function does not respond adequately, the driver is expected to correct vehicle guidance immediately. Because full OEDR is not a prerequisite of Level 2 functions, a warning does not need to be provided to the driver prior to such an event. Only when sustained driving automation functions meet SAE Level 3 they release and relieve the driver from executing the DDT entirely. However, sustained driving automation of Level 3 takes over the entire driving task with the expectation that the driver is available as a “fallback-ready user,” who intervenes upon request. A Level 3 function is able to perform the entire DDT, including the OEDR subtask, within its operational design domain (ODD). Thus, a Level 3 function recognizes its own limits (e.g., approaching the end of its ODD) and in such cases, the function must provide a timely takeover request to the fallback-ready user. The user is then required to intervene within the given time because the function is not able to perform outside its ODD. This role of a “fallback-ready user” puts several demands on the human.

First, to be able to perceive a takeover request by the function, the driver needs to be awake and conscious. When sleeping, the human is most probably not able to reliably perceive a takeover request and thus to fulfil the role of a “fallback-ready user.” However, doing nothing and staring into space would lead to an increase in passive fatigue, which again may impede the fallback-ready user’s attention and perception (Feldhütter et al., 2019; Frey, 2019; Weinbeer et al., 2019). On the one hand, engaging in non-driving-related tasks (NDRTs) has the potential to keep the fallback-ready user awake and attentive (Wu et al., 2020). On the other hand, NDRTs may also increase task-induced fatigue (Jarosch & Bengler, 2019; Jarosch, Bellem, et al., 2019).

Action Editor: C. Shawn Green was the action editor for this article.

ORCID iD: Elisabeth Shi <https://orcid.org/0000-0001-6742-7023>

Data collection in Experiment 1 was sponsored by the German Research Association for Automotive Technology (FAT). Data Collection in Experiment 2 was sponsored by the German Federal Ministry for Economic Affairs and Energy (BMWi) based on a resolution of the German Bundestag. Data cannot be made publicly available because of the respective data protection concepts. Analytic methods are described in this publication. Study materials can be found in the Appendix. The authors have no conflicts of interest to disclose.

Open Access License: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC-BY-NC-ND). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Contact Information: Correspondence concerning this article should be addressed to Elisabeth Shi, Section F4 Automated Driving, Federal Highway Research Institute (BAST), Brueckerstr. 53, Bergisch Gladbach D-51427, Germany. Email: shi@bast.de

Second, to safely take over the driving task upon request, the driver needs to switch from performing the NDRT to reorienting in the current traffic situation, before deactivating the system and continuing the ride. In this context, the effects of NDRTs on following takeover performance have been investigated extensively (Gold, 2016; Jarosch, 2020; Radlmayr, 2020; Wandtner, 2018; Zeeb, 2016).

In conclusion, NDRTs are of special interest when investigating the human availability in Level 3 driving automation.

Currently Applied Methods to Investigate NDRTs During Level 3 Automated Driving Phases

NDRTs as the Independent Variable

Various methods have been applied to investigate NDRTs. Most prominently, NDRTs have constituted the independent variable of research designs. These studies investigated the effect of manipulated NDRTs on variables such as driver state or following manual driving behavior (for an overview see review by Jarosch, Gold, et al., 2019). A review by Naujoks et al. (2017) summarizes the NDRTs applied in experimental studies on automated driving of different driving automation levels. Naujoks et al. (2017) provide a table of standardized NDRTs (e.g., n-back task, tracking-task, or surrogate reference task) and a table of everyday NDRTs (e.g., listening to music, gaming, or watching a movie) that have been used in studies. Yet, in most cases, the NDRT is a manipulation by the experimenter and therefore not an activity freely chosen by the user of the automation. The question of what users will actually do during Level 3 automated driving phases cannot be answered by manipulating this variable experimentally. Experimental manipulation provides insights into the effects of those NDRTs. But, which NDRTs are actually relevant to the user and thus have an impact on traffic safety?

NDRTs as Multiple-Choice Options Provided by the Experimenter

Current studies on this question often choose multiple-choice formats to investigate what drivers will engage in during Level 3 automated driving periods (Pfleging et al., 2016; Schoettle & Sivak, 2014). However, when using multiple-choice formats, it is necessary for the researcher to draw up a shortlist in advance. This list is then presented to the participant who states how often s/he engages in this task. Therefore, it is inherent to the method that results cannot mirror solely what users of Level 3 automation would choose to do, but always reflect the researcher's expectations, too. The quality of results is dependent on the quality of the shortlist, which in turn is just the subject of the research. In addition, the context of answering the questionnaire may influence participants' responses. Pfleging et al. (2016) conducted a web survey and an in situ survey that yield slightly different results (Table 1). Both surveys based on multiple-choice formats and asked participants to state what activities they engage in as passengers of public transportation.

Most of the categories yielded lower percentages in the in situ survey than in the web survey. This gap suggests that results depend on the situation in which participants answer the survey. From a methodical perspective, it is very interesting that percentages differ already when participants are asked about their current behavior as users of public transportation. When applying the method to Level 3 automated driving and expected user behavior in the future, we

Table 1

Excerpt of Results From Web Survey and In Situ Survey by Pfleging et al. (2016)

Non-driving related tasks	Web survey	In situ survey
Texting	74	58
Talking with passengers	72.3	47
Eating and drinking	54	7
Listening to audio content	72	35

Note. Pfleging et al. (2016) used multiple-choice format to collect the data. Numbers represent percentage of participants stating the listed activity.

would assume even larger differences between today's web surveys as estimates of the future user behavior and future in situ surveys, when Level 3 driving automation systems are available. Furthermore, Pfleging et al. (2016) used multiple-choice format questions which might have served as a basis for participants of the web survey to answer the questions. By providing a shortlist of NDRTs that participants could choose from, participants may have based their answers on plausibility and familiarity, thereby yielding higher percentages when participants answer the web survey compared to the in situ survey. Participants who answered the in situ survey experience the situation that they are asked about. Hence, these participants have more cues in addition to the survey's shortlist of NDRTs.

Regarding inferences from available results of web surveys to future driving behavior, the survey by Schoettle and Sivak (2014) may serve as a starting point, given their large sample size. The researchers addressed the question of what users will do during automated driving phases by using multiple-choice format. Participants from the U.S. ($n = 501$), the U.K. ($n = 527$), and Australia ($n = 505$) were asked about their opinion on autonomous and self-driving vehicles. The questionnaire also included one multiple-choice question on NDRTs during Level 4. Results show that most participants (overall 41.0%) would watch the road even though they would be passengers. Of those who would engage in NDRTs responses slightly varied by country (see Table 2). However, variation does not seem as high as variation between web and in situ survey by Pfleging et al. (2016), again stressing the influential factor of the survey context.

Concluding on multiple-choice formats, we assume that results derived from this method depend on the quality of the shortlist of NDRTs drawn up in advance by the experimenter, and on the context in which participants answer the survey. Furthermore, regarding validity of derived NDRTs, we assume that providing

Table 2

Excerpt of Results From Survey by Schoettle and Sivak (2014)

Non-driving related tasks	Overall	U.S.	U.K.	Australia
Reading	8.3	10.8	7.6	6.5
Texting or talking with friends/family	7.7	9.8	5.5	7.9
Sleeping	7.0	6.8	7.2	7.1
Watching movies/TV	5.3	6.0	4.2	5.7
Working	4.9	4.8	4.9	5.1

Note. Schoettle and Sivak (2014) used multiple-choice format to collect the data. Numbers represent percentage of participants stating the listed activity.

an actual Level 3 automated driving context for participants might yield results that are different from the web surveys, but which might come closer to naturalistic user behavior, analogously to the web survey and in situ survey approach by Pfleging et al. (2016).

NDRTs Derived From Observation in Natura

Researchers investigating NDRTs during Level 0 (SAE International, 2021) used observation methods, naturalistic driving studies (NDSs), or interviews (Dingus et al., 2006; Huemer & Vollrath, 2012; Kathmann et al., 2020; Metz et al., 2013). Because these methods do not require a priori shortlists of NDRT categories, they may reduce the influence of experimenters' expectations on the results. For example, Huemer and Vollrath (2012) conducted the interviews immediately after the participant's ride and chose both urban areas (i.e., parking spaces) as well as motorways (i.e., service stations). Kathmann et al. (2020) observed drivers' behaviors from the roadside. Of course, observers might tend to preferably code conspicuous and salient NDRTs. Therefore, observers need to be trained in advance (Kathmann et al., 2020). However, the described methods of observations, interviews, and NDS are only applicable to today's available systems. Thus, they cannot be used for Level 3 automated driving systems, yet.

NDRTs Derived From Observation in Present Mobility Situations Analogous to Level 3

To apply the method observation in the context of Level 3 automated driving, researchers have looked for situations in today's available means of transportation that come close to the users' role during Level 3 automated driving. For example, Pfleging et al. (2016) also conducted in situ observations in subway trains ("U-Bahn") in and around Munich. Choosing subway trains may fit driver role characteristics of Level 4 and Level 5 automation. Hence, the in situ observations may provide insights into NDRTs executed during active Level 4 or Level 5 automation. However, passengers on those trains do not need to respond to takeover requests, which constitutes the main driver role characteristic of Level 3 automation (SAE International, 2021). Thus, observations in today's available means of transportation seem to be a less suitable test site to examine NDRTs performed during Level 3 automated driving phases.

NDRTs Derived From Observation in a Simulated Level 3 Automated Driving Setting

Another approach to apply the method of observation to Level 3 driving automation is to use the driving simulator setting (Hecht et al., 2020; Large et al., 2017). For example, Large et al. (2017) tested six participants on five consecutive days in a driving simulator study. Each session represented a daily commuting route to work. Large et al. (2017) asked participants to bring their own NDRTs for the automated driving phase and analyzed video recordings of the users' behavior during the automated driving phase. Similarly, Hecht et al. (2020) examined video data of 20 participants from a driving simulator study by Feldhütter et al. (2019). Participants experienced a 60 min automated driving phase and engaged in NDRTs that were freely chosen by themselves. However, in both experiments, participants' choice of NDRTs may be influenced by knowing they are in a physically safe environment, that is, in a

driving simulator. This effect may be even stronger in case of Large et al.'s (2017) study because participants experienced the same ride on five consecutive days. Thus, Large et al. (2017) and Hecht et al. (2020) suggest to use actual driving conditions and more specific setups for future research and to consider effects of vehicle motion and road vibrations on NDRT choice.

Literature Gap and Aim of the Present Study

To our knowledge, there has been no study that uses a Level 3 automated driving setting in a real vehicle and in a real driving situation to answer the question of what users will be likely to engage in during Level 3 automated driving. Furthermore, we apply two methods to examine NDRTs and compare whether they show different results. We thereby extend current findings by (1) providing an actual driving condition to the participants and (2) using methods to collect naturalistic NDRTs that are presumably less influenced by experimenters' expectations than multiple-choice format questions. The two aims of our experiments are (a) to provide insights into what NDRTs users of Level 3 automated driving would be likely to engage in and (b) to examine whether different methods provide different results.

Experiment 1

Introduction

As outlined above, previous studies investigated NDRTs during automated driving. However, there are several limitations to the current literature landscape. First, the level of automation is not always confined to Level 3 (Pfleging et al., 2016; Schoettle & Sivak, 2014). However, the user role in Level 3 automated driving is different from the one in Level 4 and 5 (Shuttleworth, 2019). Second, using multiple-choice formats or observations limit the range of resulting NDRTs inherently. In case of multiple-choice formats, resulting NDRTs are limited to what the experimenter took into account in advance. In case of observations, resulting NDRTs are limited to those being observable. For example, solely by means of observation daydreaming cannot be differentiated from watching the surrounding. However, there may be more NDRTs than those observable or those experimenters have thought of. These remain uncovered because of the applied methods. We conducted Experiment 1 to gain first insights into what users themselves think they will engage in during Level 3 automated driving.

Method

Experiment 1 was part of a broader project (Klamroth et al., 2019). The following sections focus on methods relevant to the research question at hand. For further details on the broader project please refer to the study by Klamroth et al. (2019).

Participants

In total, 39 participants from the general public took part in Experiment 1 (20 female, 19 male; $M_{\text{age}} = 51$ years, $SD_{\text{age}} = 11$ years). Participants were recruited using newspaper announcements. All were in possession of a valid driver's license and reported an annual mileage of $M = 17,570$ km ($SD = 6,958$ km). Moreover, 48.60% stated to use cruise control frequently or very frequently, 10.80% stated to use adaptive cruise control (ACC) frequently or very frequently.

Apparatus and Materials

A Wizard-of-Oz vehicle based on a Volkswagen Caddy was used to simulate Level 3 driving automation. Figure 1 shows the schematic setup of the Wizard-of-Oz vehicle. The participant was seated in the driver's seat. A second driver (the wizard) was seated in the rear of the vehicle separated from the driver's cabin by a tinted window. The wizard took over vehicle motion control when the participant activated the automation during the ride. The wizard then simulated the driving automation function and drove the vehicle from the rear. The wizard could look through the tinted window. To conceal the Wizard-of-Oz principle, the participant in the driver's seat could not see the rear of the vehicle through the window. For safety reasons, there is no front-seat passenger allowed in the Wizard-of-Oz vehicle. The Wizard-of-Oz principle allows investigating human-machine interaction at Level 3 and above without taking the risk of exposing participants to untested technical systems.

The Level 3 automation was available on motorways only. Participants activated and deactivated the driving automation function by pressing a button on the steering wheel. The wizard, who simulated the driving automation function, always adapted the speed according to current traffic conditions and speed limits. If there was no speed limit, the wizard adapted the speed to the advisory speed limit for German motorways (130 km/hr). Approaching a functional limit issued a takeover request (bimodal auditory and visual signal). Functional limits were specified as construction sites, joining a motorway and turning off motorway exits. Takeover requests were issued timely, that is, one kilometer ahead of construction sites and two kilometers ahead of motorway access and motorway exits. Current function status was displayed on an additional human-machine interface (HMI) in the instrument-cluster.

During the automated driving phases, participants were instructed to engage in the Surrogate Reference Task (SuRT, ISO/TS 14198, 2012) provided on an Android-based tablet computer mounted on the center console. The tablet computer featured a 10.1" display with 1920 × 1200 px resolution. The SuRT task requires participants to select the largest circle among many distractor circles. Participants

were instructed to select the largest circle by tapping on it. Then, the next screen of circles followed.

The ride took place on German motorways ("Bundesautobahn") within the metropolitan area of Cologne. Participants drove approx. 118 km (approx. 73.3 miles), including seven sections (of up to 15 km length) of Level 3 automated driving.

A questionnaire on NDRTs was provided after the ride, that is, after having experienced Level 3 automated driving. The questionnaire consisted of one open-ended question. Participants should imagine the function was available on their own vehicle. They were asked to state three to five NDRTs they would engage in during the automated driving phases. Each participant was asked to provide three to five NDRTs to uncover a wide range of NDRTs overall. Otherwise it would have been possible that less frequently executed tasks were not assessed.

Procedure

Upon arrival, participants were informed about how to use the vehicle and the Level 3 automation function (including explanations on driver and user roles, activation, deactivation and takeover procedures, and emergency cases). Participants were unaware of the Wizard-of-Oz principle. Next, participants started a training ride on the Bundesautobahn to familiarize with driving the vehicle and using the automation function before starting the experimental ride. During the automated driving phases, participants should engage in the SuRT provided on a mounted tablet. After the ride, participants answered the questionnaire on NDRTs, and finally, were debriefed and informed about the Wizard-of-Oz principle. Participants received 80 Euro for taking part in the study.

Data Analysis

Participants' responses to the questionnaires were listed. Synonymous phrases that describe the same NDRT (e.g., "talking to other passengers" and "making conversation") were unified (e.g., "making conversation"). Naming rates of unified terms were computed and ranked according to size.

Results

The NDRTs stated by each participant were listed and counted. Table 3 shows the top 10 NDRTs. Interestingly, reading is the NDRT mentioned most frequently.

Discussion

In Experiment 1, participants themselves were asked to state the NDRTs they would engage in during a Level 3 automated driving period, after they have experienced the respective driving automation on a motorway route of approx. 118 km in real traffic. This extends current findings as to our knowledge, there is no study on NDRTs that has asked participants from the general public, who have experienced Level 3 automated driving in real traffic, themselves. Rather, previous studies employ the method of observation or multiple-choice questions, thereby inherently limiting the range of NDRTs that may be revealed.

In Experiment 1, we aimed at uncovering a broader range of NDRTs than previous studies. By applying an open-ended question,

Figure 1
Schematic Setup of the Wizard-of-Oz Vehicle

Note. Figure shows the driving positions of the participant in the driver's seat and the wizard driver in the rear. The driver's cabin and the rear are separated by a tinted window, which is only transparent from the back to the front so that the wizard driver is able to look outside, whereas the participant cannot look through to see the rear. After pressing the button to activate the Level 3 driving automation, the wizard driver takes over the driving task and hands over control at a "system limit" by issuing a takeover request.

Table 3

Absolute and Relative Numbers of Naming for Top 10 NDRTs Provided by Participants in Experiment 1 (NDRTs = Non-Driving-Related Tasks)

Type of NDRT	Absolute number	Relative number (in %)
Reading	23	15.54
Making phone calls	15	10.14
Using smartphone	13	8.78
Eating	13	8.78
Reading and writing messages	12	8.11
Taking a look at the environment	11	7.43
Drinking	9	6.08
Making conversation	7	4.73
Sleeping	5	3.38
Relaxing	4	2.07

Note. Relative number of naming given in percent.

participants needed to think of NDRTs themselves and state which NDRTs they would possibly engage in. This strongly reduces influences by preassumptions of the experimenter and presumably results in more realistic representation of NDRT engagement probabilities. We decided to not summarize participants' namings to more global categories (e.g., mobile device usage). Instead, we only unified terms and computed their naming frequencies and rates (e.g., separately for making phone calls, using smartphone, etc.). Based on this procedure, detailed information provided by participants is maintained and statements on the level of singular NDRTs are possible. It is important to be aware of how to aggregate participants' namings when applying this method. Different aggregation strategies may provide different results and conclusions. Hence, it is important to align data analysis methods with the hypotheses. For example, the resulting namings of our participants would have allowed to summarize NDRTs according to the modalities they involve (e.g., visual, auditory, etc.) or according to the devices they include (e.g., mobile device usage), or according to manual load (e.g., hands occupied by NDRT or not), etc. Yet, it is not the aim of the current study to investigate NDRTs according to their involved modalities or the like, but only to investigate what NDRTs users will be likely to engage in. This question is best answered on the level of specific NDRTs.

In the context of previous findings, results from Experiment 1 need to be interpreted in light of the methodical differences. Hecht et al. (2020), for example, report "phone usage" as the most frequent NDRT, followed by reading. Hecht and colleagues' category of "phone usage" results from video observation, where it was not possible to further differentiate what exactly participants engaged in, except of phone calls and listening to music (Hecht et al., 2020). In contrast, the approach of naming allows more detailed information because participants can state mental activities and provide details on mobile device usage. For example, our approach allows differentiating "making phone calls" from "smartphone usage" in general as well, but, in addition naming rates of "reading and writing messages" (8.11%), "watching movies/series" (2.70%), "playing" (2.03%), or "preparing the next appointment" (1.35%) can be calculated as well. It depends on how precise the participants report NDRTs. As shown in Table 3, 8.78% of all namings were "using smartphone" without providing further information on the specific

task, whereas other namings were much more precise and less frequent (e.g., "preparing the next appointment"). When interpreting naming rates, it should be considered that with increasing precision, it becomes less likely that a specific NDRT yields high naming rates.

Our results show that "reading" was the most frequently stated single NDRT, followed by "making phone calls." Similarly, Large et al. (2017) and Hecht et al. (2020) report reading as one of the most executed NDRTs. Replication stresses the practical relevance of this task. In contrast to our findings, Hecht et al. (2020) observed only one participant making a phone call. This could be due to reactivity to the experimental situation, that is, participants are aware of being observed and video recorded and thus do not show their usual behavior (Döring & Bortz, 2016). Our approach of using an open-ended question can be assumed to be less prone to reactivity to the experimental situation. The question specifically addresses a setting in the participant's own car, thereby framing the private setting of interest, which is impossible when observing participants in an experimental setting.

Despite the advantages of the applied method, naming NDRTs involves participants' ability to actively name and declare NDRTs. In Experiment 1, 8.78% of all namings were "smartphone usage." However, no further information was provided on the specific subtask they would execute when using their smartphone (e.g., using social media, listening to music, reading news, writing messages or emails, etc.). In addition, only 7.43% of the participants indicated they would look at the environment, whereas all participants of Hecht et al.'s (2020) study watched the surroundings at some point in time during the automated driving phase. It can be argued that not all participants of our study did assume "watching the surrounding" as a NDRT, thus yielding lower naming rates.

In conclusion, the approach of asking participants on NDRTs may provide more detailed information on specific NDRTs and is less prone to experimental reactivity. However, several aspects should be kept in mind when applying this method: First, when formulating the open-ended question, it is important to frame the respective context of interest (e.g., situation in their own car). Second, depending on the desired specificity of named NDRTs, participants should be explicitly requested to provide specific NDRTs to prevent global categories (e.g., "using smartphone"). Third, when analyzing participants' responses, we strongly recommend staying with the hypotheses. Different analysis methods may emphasize different facets of the collected data. Third, when interpreting the data in the context of other studies, differences in methods should be considered.

Experiment 2

Introduction

Many studies have investigated influences of NDRTs, for example, regarding fatigue or distraction (Jarosch, Bellem, et al., 2019). To align these findings with traffic safety, further insights into what NDRTs users actually engage in are needed. In Experiment 2, participants had the opportunity to truly engage in (almost) every NDRT during a Level 3 automated driving phase in the Wizard-of-Oz vehicle. We conducted Experiment 2 to gain first insights into which NDRTs users actually engage in during Level 3 automated driving.

Method

Experiment 2 was part of a broader project (Ko-HAF final report by Hohm et al., 2018). The following sections focus on methods relevant to the research question at hand. To provide a general overview on the experimental setting, methods influencing the experimental situation are also described. For further details on the broader context of Experiment 2 please refer to the study by Frey (2019).

Participants

Data were collected from 19 participants from the general public, who were acquired by newspaper advertisements. One participant had to be excluded because he did not comply with the instructions, leaving a total of 18 participants (8 women, 10 men; $M_{\text{age}} = 45$ years, $SD_{\text{age}} = 15$ years). All participants were in possession of a valid driver's license and reported an annual mileage of $M = 22,264.71$ km ($SD = 21,749.37$ km). Seven participants (38.9%) had prior experience with cruise control, three (16.7%) had prior experience with adaptive cruise control, and one (5.6%) had experience with a park assist function, no one had experience with lane keeping functions, hence no one had experience with Level 2 or Level 3 functions.

Apparatus and Materials

Automated driving was simulated by means of a Wizard-of-Oz vehicle, which was the same as in Experiment 1 (see Figure 1). Experiment 2 was run on a test track with an oval course of 2.1 km (approx. 1.3 miles) consisting of three lanes. The Level 3 automated ride took place on the middle lane at a speed of 70 km/hr (approx. 43.5 mph) without surrounding traffic. To keep the risk of physical harm minimal throughout the experiment, the test track was booked for exclusive use for each participant.

Participants were instructed to bring along their own NDRTs. These were not allowed to be too loud ensuring that the participant was always able to hear the sound of the takeover request. In addition, NDRTs were not allowed to take up too much space (e.g., large newspapers) so that the wizard was able to look out the front window from the second row. Furthermore, NDRTs were not allowed that included talking due to concurrent electroencephalogram (EEG) measurement. Magazines were available on the front passenger seat in case the participant did not bring along any NDRTs.

EEG was recorded continuously from 11 scalp sites using an electrode cap. Electrooculogram (EOG) was recorded from four additional sites to account for eye movements. One additional site recorded electrocardiogram (ECG). The EEG method was used to analyze alpha spindles (Simon et al., 2011) which are robust to motoric artifacts. However, speaking may interfere with the measurement. Therefore, NDRTs that include talking were not allowed.

Procedure

In the context of invitation to the experiment, participants were instructed to bring along at least one engrossing NDRT, but no task that would include talking as this would disturb the concurrent EEG measurement.

Upon arrival at the testing center, participants received a mandatory introduction to the testing center, gave written informed consent to participating in the experiment and answered questionnaires. Participants were told that they would use a Level 2 function for approx. 30 min and after that a Level 3 function for approx. 30 min (order counterbalanced across participants). The experimenter explained the vehicle in general (VW Caddy, automatic gearbox, adjusting seat and mirrors, etc.) and both automated driving functions (Level 2 and Level 3).

Instruction on the Level 3 automation includes that the function takes over the driving task, including accelerating, braking, and steering, and that they do not need to supervise the function so that they may engage in NDRTs during the automated driving phases. When approaching a functional limit or in case of a technical defect, the function would provide a timely request to intervene and the participant shall take over the driving task again. Participants were told they have a time budget of max. 20 s to take over the driving task, otherwise, a minimal risk condition will be reached.

Instruction on the Level 2 function includes that the function takes over accelerating, braking, and steering, but the function needs to be supervised. In addition, the function will use the middle lane and will drive at a speed of approx. 70 km/hr. The driver needs to be ready for takeover any time. Participants were instructed that significant deviations regarding speed and lane keeping may occur, that is, speed reduced to below 60 km/hr or increased to above 80 km/hr, or exceedance of lane markings to the left or right. Whenever participants detected such an error, they should press a button that was fixed on one of their fingers (freely chosen by the participant). This task was provided to emphasize the participant's role when the Level 2 function was active, and additionally, for comparison reasons to another experiment. However, no errors occurred during the Level 2 ride.

Following the instructions, participants were prepared for EEG measurement, the button was fixed on one of their fingers, NDRTs were checked regarding size and sound intensity and participants were seated in the vehicle. All participants passed a training phase including two takeover situations at the beginning of the ride. In the experimental phase, each block (Level 2 or Level 3) was announced explicitly so that participants were aware of the current automation level. The Level 3 automated driving phases were interrupted by phases of manual driving (each approx. 2.5 min).

After the ride, participants answered questionnaires and were finally debriefed and informed about the Wizard-of-Oz principle. Participants received 80 Euro for taking part in the study.

Data Analysis

NDRTs were analyzed based on video data. From the camera's perspective, it is only possible to determine what participants held in their hands, but not what they did specifically. For example, when holding their smartphones in their hands, it is not possible to clearly identify whether they were playing, texting, reading, or using social media. This is applicable to other activities as well. Therefore, we decided to base the category labels on what was physically observable from the videos. That means, when participants were holding an ebook, magazine, or smartphone, we used the labels accordingly (i.e., person looking on ebook, magazine, or smartphone) and did not infer any mental activity (e.g., reading). The same applies to the category label "steady gaze to the outside." We treated steady gazes

to the outside as NDRTs, although it could be argued that this is not an active task. Yet, from observing participants looking outside, we cannot know if they were randomly and absently gazing outside or actively observing their surroundings. Furthermore, the camera focused on the participant and not on the outside so that it was not possible to determine what they looked at specifically. For these reasons, steady gazes to the outside were treated as a NDRT and the category label “steady gaze to the outside” was chosen to describe the behavioral information recorded on video without insinuating information on the mental level. This labeling method yielded the following NDRT categories: “steady gaze to the outside,” “smartphone,” “ebook,” “book,” “magazine,” “writing with a pen,” “knitting needles,” “drinking or eating.”

Together with the NDRT categories, respective time stamps of the points of beginning and ending based on the videos of the drivers were coded. NDRTs were coded from the beginning to the end of a Level 3 automated driving period. The resulting data provide information about what drivers did during Level 3 automation, for how long they executed one specific task, and how often they changed tasks during one automated driving phase.

Results

Results show that all participants executed at least two NDRTs. Except one participant, all looked outside first, before executing an active NDRT. The one participant who did not look outside first used his/her smartphone immediately after activating the automation. The second task was much more heterogeneous across participants. However, most participants switched to using their smartphones. Some also read magazines, read books, or wrote something using a pen. No NDRTs were performed simultaneously, and all NDRTs were performed consecutively. Figure 2 shows the number of tasks participants engaged in. As can be seen, only 4 of 18 participants engaged in more than four NDRTs during the approx. 30 min Level 3 automated driving phase.

Table 4 shows that engagement rate and total duration rate of engagements provide different information on NDRT engagement. Regarding total duration rate of engagements, smartphone usages without phone calls ranks highest (total duration rate: 45.13%). Thus, smartphone usage (without phone calls) was the most time

Table 4

Engagement Rate and Total Duration of Engagement Per Observed NDRT in Experiment 2 (NDRT = Non-Driving-Related Task)

NDRT	Engagement rate (in %)	Total duration rate of engagements (in %)
Smartphone (w/o phone calls)	25.00	45.13
Magazines	16.67	23.54
Book	5.56	8.79
Writing using a pen	4.17	7.52
Steady gaze to the outside	38.89	5.90
Knitting needles	1.39	5.59
ebook	6.94	3.37
Eating/drinking	1.39	0.22

Note. NDRTs are derived from observation. For the engagement rate, percentages are relative to the total number of engagements. For the total duration rate of engagements, percentages are relative to the total duration of all engagements.

consuming NDRT. Regarding engagement rate, a steady gaze to the outside ranks highest (engagement rate: 38.89%). However, the total duration rate of steady gazes to the outside was fairly low (5.90%). Assuming that participants were reading when they were holding a magazine, a book, or an ebook in their hands, reading would reach a total duration rate of 35.70% and an engagement rate of 29.16%. Also, these three tasks were performed by a high share of participants (steady gaze to the outside: 18/18, smartphone usage without phone calls: 13/18, assuming reading: 12/18).

Discussion

Experiment 2 adds to the current literature on NDRTs as there are no studies to our knowledge that investigated NDRT engagements in a real driving setting. Our results corroborate current findings indicating that participants are likely to execute a steady gaze to the outside during the automated driving phase, to use their phones (w/o phone calls), and to engage in reading (Hecht et al., 2020; Large et al., 2017; Schoettle & Sivak, 2014).

Just as in other studies investigating natural user behavior by observation, the specific experimental setting must be considered. Multiple cameras and concurrent EEG measurement always reminded participants on the fact that they were in an experimental situation and being observed. Hence, any conclusions suggesting that natural behavior was observed must be made with caution because reactivity to the experimental situations must be assumed. The experimental setting itself could have altered participants' behavior (Döring & Bortz, 2016). Another limitation of our study is that for several reasons, restrictions were put on participants' choice of NDRTs, that is, not too loud, not too large, no talking, and no front-seat passenger allowed. Therefore, all NDRTs that would have involved one of these characteristics could not be observed, which may bias proportions of the observed NDRTs. Furthermore, magazines were provided on the front passenger seat in case participants did not bring any NDRTs with them. This was only the case for one participant. Yet, the availability of magazines may have led to an increase in reading compared to engagements in other NDRTs. However, comparing frequencies between Experiments 1 and 2 and taking other studies into account (Hecht et al., 2020;

Figure 2

Bar Graph Depicting Number of Participants Per Total Number of Non-Driving-Related Tasks (NDRTs) Performed in Experiment 2

Note. Figure shows how many NDRTs were performed by how many participants.

Schoettle & Sivak, 2014), it can be assumed that this effect was rather small.

General Discussion

To our knowledge, we have conducted the first experiments that used a Level 3 automated driving setting in a real vehicle and in a real driving situation to address the question of what users will likely engage in during Level 3 automated driving. We applied two methods to examine NDRTs. In Experiment 1, we asked participants to name NDRTs, whereas in Experiment 2, we let participants bring along their own NDRTs and observed their actual behavior during Level 3 automated driving phases. The two aims of the study were (a) to provide insights into what NDRTs users of Level 3 automated driving would be likely to engage in and (b) to examine whether different methods provide different results.

Experiment 1 provides insights into what participants think they would engage in during Level 3 automated driving. Based on Experiment 1, smartphone usage (w/o phone calls), eating/drinking, and phone calls were mentioned most frequently. Phone calls and using the smartphone were deemed as separate NDRTs and were not condensed to one NDRT. This was done because a smartphone may be used in many ways, for example, playing, reading news, social media consumption. Making phone calls constitutes one way of using a smartphone. Thus, it can be argued that phone calls and using a smartphone could be summarized to one general category of "using smartphone." However, we decided not to do so because this would have resulted in loss of information.

Experiment 2 provides insights into what participants actually do, when they use a Level 3 automated driving function. Most participants engage in at least two NDRTs. Smartphone usage (w/o phone calls) and reading are the two most time-consuming NDRTs, although another NDRT has a higher engagement rate. Participants spent 45.13% of their time on smartphone usage (engagement rate: 25.00%), and 35.70% of their time on reading (engagement rate: 29.16%). A steady gaze to the outside was the NDRT that participants most frequently engaged in (engagement rate: 38.89%). Interestingly, the total duration rate of NDRT engagement was only 5.90%. This leads to the assumption that different NDRTs seem to be associated with different forms of engagement. Each time participants attend to NDRTs like reading or smartphone usage, they spend more time on the NDRT compared to NDRTs like looking outside. A steady gaze to the outside is performed more frequently. However, each engagement seems to be associated with a shorter period of time.

Hence, we could show that different methods provide different results with regard to detailed information on NDRT engagement. Overall, both asking participants an open-ended question after they have experienced Level 3 automated driving in real traffic in a real vehicle and observing participants while they use Level 3 automated driving in a real vehicle seem adequate to examine what NDRTs are relevant to users. It depends on the exact research question at hand, what method/s should be used.

Comparison and Comparability of Results From Experiments 1 and 2

Both experiments employed different methods to answer the question of what users would likely do during the Level 3 automated

driving phase. The first method (asking an open-ended question after participants experienced Level 3 automated driving in a real vehicle and in real motorway traffic) may provide information on NDRTs that are chosen deliberately and that users probably consciously engage in. For example, 14.86% of all stated NDRTs refer to eating/drinking. However, when executing the NDRT, participants in Experiment 2 spent only 0.22% of the total time of NDRT engagement on eating/drinking and 1.39% of all engagement were engagements in eating/drinking. The relative number of naming eating/drinking (14.86%) is comparable to the relative number of naming those two NDRTs that were executed most of the time (reading: 15.54%; smartphone usage w/o phone calls: 16.89%). Participants in Experiment 1 were explicitly asked to provide NDRTs. Thus, information on frequency of engagement and time spent on the NDRT is missing. All stated NDRTs have the same weight. But, naming of NDRTs requires that participants are aware of engaging in this task. Therefore, the resulting list of NDRTs may provide information on what NDRTs users engage in rather deliberately and attentively.

The second method (observation in real vehicle) is not based on participants' declaration and may uncover NDRTs that are executed in a real ride. In comparison to the first method, observation may on top uncover NDRTs that are executed less attentively. For example, the NDRT "steady gaze to the outside" is associated with a naming rate of 7.43% in Experiment 1. At the same time, this NDRT makes up 38.89% of all NDRT engagements in Experiment 2. Hence, naming rate and engagement rate cannot be equaled, but rather seem to describe different facets of NDRT engagement. Observation not only provides information on number but also on the duration of engagements. For example, eating/drinking reaches a naming rate of 14.86% in Experiment 1, but was executed only 0.22% of the total time of NDRT engagement in Experiment 2. Total duration rate of NDRT engagement and naming rate may also diverge, as well as engagement rate and total duration rate of engagements. The latter two may provide distinct information on NDRT engagement. For example, taking the engagement rate of 38.89% together with the total duration rate of 5.90%, it can be inferred that gazes to the outside were performed rather frequently compared to other NDRTs, while at the same time engagement into gazes was relatively short compared to other NDRTs. When adding information about the naming rate of "steady gazes to the outside" (7.43%), it seems this NDRT might not be executed very attentively or on purpose like other NDRTs, such as reading (naming rate: 15.54%, engagement rate: 29.16%, total duration rate: 35.70%).

Hence, it may be concluded that measures derived from Experiments 1 and 2 describe different facets of NDRT engagement and are comparable in the sense that they highlight similar NDRTs as relevant. Still, the measures provide different information on NDRT engagement and therefore, are not comparable on a numeric level.

Some methodological limitations need to be discussed in the context of comparing the measures. We aimed at providing a realistic driving setting to our participants. However, it was still an experimental setting. Especially when using the method of observation, it is possible that rare NDRTs remain undetected. For example, in Experiment 2, relaxing and sleeping were not performed, although stated (rarely) in Experiment 1 (sleeping: 3.38%; relaxing: 2.07%). The same applies to eating/drinking (naming rate: 14.86%, engagement rate: 1.39%, total duration rate of engagements: 0.22%). Reasons for not performing a

NDRT may be found in the experimental setting. For example, in Experiment 2, the trip duration of approx. 30 min may have been too short to allow participants to sleep, relax, or become hungry/thirsty. Moreover, in Experiment 2, it was participants' first time using a Level 3 driving automation, which might have been experienced as a thrilling event, thereby, inhibiting tediousness, drowsiness, or fatigue. Furthermore, the experimental setting included EEG measurement and multiple cameras in the driver's cabin that possibly hamper participants to sleep or relax. This stresses that reducing reactivity to the experiment and creating a natural environment is crucial for detecting natural NDRT engagement. In addition, it emphasizes the potential of our first method (i.e., asking participants what they would engage in, after they have experienced Level 3 automated driving in a 118 km test ride in a real vehicle in real motorway traffic). Asking experienced participants still uncovered these NDRTs.

Representativity of Our Participant Sample

Our participants were neither drawn from university students nor from employees of a specific company. Instead, our samples were drawn from the general public and did not need to fulfill any criteria, besides having a driving license. Because participants were recruited via newspaper advertisement, they may have a higher preference of the reading NDRT compared to the general public. However, as both of our samples were acquired via newspaper advertisement, this characteristic arguably did not systematically influence comparison of resulting NDRTs between Experiment 1 and Experiment 2. Our samples' reading rates are discussed in the context of current literature in the next section.

Referring to Current Literature

Reading as a NDRT

Compared to results by Schoettle and Sivak (2014), we found lower percentages for sleeping and, interestingly, also for reading. Interestingly, 41.0% of the total sample of Schoettle and Sivak would read during a "ride in a completely self-driving vehicle (Level 4)" (Schoettle & Sivak, 2014, p. 18). However, percentages vary strongly between U.S. (10.8%) and U.K. and Australia (44.0% and 43.4%). In our Experiment 1, 15.54% of our participants stated to engage in reading during the automated driving phases. In Experiment 2, we observed engagement rates of 5.56%, 6.94%, and 16.67% and total duration rates of 8.79%, 3.37%, and 23.54% for the categories "book," "ebook," and "magazine," respectively. As outlined above, our rates resulting from two Wizard-of-studies are much lower than the rates obtained by Schoettle and Sivak's (2014) using a web survey. This pattern can also be found in the study by Pfleging et al. (2016) who observed higher percentages for NDRTs in the web survey compared to the in situ survey. We argue that this gap indicates that results depend on the context in which participants answered the survey. Participants in our studies experienced Level 3 automated driving in a real vehicle and in a real ride. Thereby, in Experiment 1, our participants were more experienced regarding Level 3 automated driving when they answered our question, compared to participants by Schoettle and Sivak (2014). The latter ones only read about "completely self-driving vehicle (Level 4)" (Schoettle & Sivak, 2014, p. 18) in the web

survey. In Experiment 2, we even observed our participants actually performing NDRTs in the vehicle during Level 3 automated driving. Based on this comparison, we assume that our sample's possible preference for reading (as assumed because of recruitment by newspaper advertisement) does not affect results strongly, if at all.

Sleeping as a NDRT

Of note, 7.0% of the total sample of Schoettle and Sivak would sleep during the ride. In our Experiment 1, 3.38% stated to engage in sleeping. It should be noted that the method, sample size, and Level of automation differ here. Schoettle and Sivak (2014) constructed a situation with a "completely self-driving vehicle (Level 4)" and asked participants to select from a list of NDRTs. In contrast, we let participants use a Level 3 motorway automation and then asked participants to state those NDRTs they would engage in, if the automation was available in their own vehicle. The difference in Level of automation may contribute to the resulting difference in percentage for sleeping (7.0% vs. 3.38%). Nevertheless, 3.38% of our participants seem to be prone to behavior that is not in accordance with their role as a fallback-ready user. Interestingly, our participants have received information on their role and have used a Level 3 automation function in real-traffic on German motorways. Still, for 3.38%, sleeping seems to be an acceptable NDRT.

Comparison to Multiple-Choice Format

In general, in Experiment 1, we find lower percentage numbers for each NDRT than studies using multiple-choice formats (Pfleging et al., 2016; Schoettle & Sivak, 2014). First, in Experiment 1, participants provided own answers to an open-ended question, thereby the number of possible responses was open. In contrast, when using multiple-choice formats, the number of response options is usually set. Second, when applying multiple-choice questions, interpretation of participants' responses should consider the context of response options. By providing response options, participants' responses might reflect what NDRT they would most likely engage in, given the provided response options. Third, response options may limit and influence participants' own ideas of NDRTs. Hence, if the goal is to reach a picture of what users are likely to do during Level 3 automated driving phases, users themselves will be able to give a more realistic result. Furthermore, to answer this question, users do not need preselected response options by experimenters, but are able to think of NDRTs themselves, just as they would need to when truly using Level 3 driving automation.

Comparison to Driving Simulator Setting

Hecht et al. (2020) and Large et al. (2017) conducted driving simulator studies to come closer to a realistic driving setting than studies before. Furthermore, they both applied the method of observation to detect NDRTs and identified reading as one of the most executed NDRTs, which our two experiments confirm in the Wizard-of-Oz setting. It should be noted that dependent variables differ between our experiments and studies by Hecht et al. (2020) and Large et al. (2017). Yet, in line with Hecht et al. (2020), we also found a high engagement rate and naming rate for "steady gazes to the outside," and similar to Hecht et al. (2020), all our participants in Experiment 2 executed a steady gaze to the outside sometime during

the automated driving phase. As Large et al. (2017) observed only six participants, they found substantial interindividual variance regarding NDRTs. However, the authors found all NDRTs to be visually and cognitively demanding, which also applies to our findings. In contrast to the studies by Hecht et al. (2020) and Large et al. (2017), we found naming rates of 2.7% for watching movies/series in Experiment 1, which is rather low compared to other NDRTs. In Experiment 2, we did not observe anyone watching movies/series/videos. On the one hand, Hecht et al. (2020) implemented a Level 3 automated driving phase of 60 min, whereas ours was approx. 30 min. This may have resulted in less participants engaging in movies or series. On the other hand, it is possible that our participants may have engaged in these tasks when they used their smartphone. From the video recordings, we could not identify the specific NDRT when participants were using their smartphones.

Methodological Discussion on Our Wizard-of-Oz Approach

We applied the method of Wizard-of-Oz to provide a realistic driving setting. According to the study by Müller, Weinbeer and Bengler (2019), a Wizard-of-Oz vehicle study can be interpreted as a psychological test in which “the test manual defines the study-specific testing procedure for examiners and the automation behavior simulated by the driving wizard” (Müller et al., 2019, p. 183). Based on this, the main test quality criteria of psychological tests (i.e., objectivity, reliability, and validity) can be applied to the Wizard-of-Oz method.

First, objectivity in the context of Wizard-of-Oz methods refers to the extent to which results from the study are independent from the driving wizard. In each Experiment, one trained driving wizard simulated Level 3 driving automation for all participants. Experiment 1 took place in real traffic on German motorways. The wizard was instructed to follow the German traffic rules. For motorway sections without speed limit, the driving wizard adapted the advisory speed limit of 130 km/hr. Participants were told that the system was able to conduct lane changes and overtake other vehicles taking traffic density and other traffic participants into consideration. Overall, the wizard should apply a defensive driving style. Hence, overtaking maneuvers were not conducted frequently. All takeover requests were issued by the wizard at predefined positions (for details see Klamroth et al., 2019, p. 31). This instruction allows for a standardized and replicable driving style throughout Experiment 1. The experimental setting (i.e., real traffic), however, is dependent on other traffic participants and further environmental factors (e.g., nonconsiderate or aggressive driving behavior of surrounding traffic participants, weather conditions), which cannot be controlled by the experimenter or the wizard. Experiment 2 took place on an oval course of a test track, which was booked for exclusive use, ensuring that there was no surrounding traffic. The driving wizard drove on the middle lane and kept a constant speed of 70 km/hr by using cruise control. The wizard did not change lanes. Takeover requests were issued after 15 rounds on the oval course. The rounds were counted automatically to reduce the risk of miscounting. This instruction allows for a standardized and replicable driving style throughout Experiment 2.

Second, reliability in the context of Wizard-of-Oz vehicle studies describes the extent to which the driving style is replicable. In addition to the instructions on driving style and issuance of takeover

requests, the driving wizard took sufficient breaks to avoid errors due to fatigue and to maintain a comparable driving style for all participants.

Third, validity in the context of Wizard-of-Oz vehicle studies addresses whether participants actually believed in the cover story (i.e., vehicle being equipped with a technical Level 3 driving automation system) and whether the simulated Level 3 driving automation truly behaves like a technical Level 3 driving automation system. Müller et al. (2019) suggest asking participants “Do you trust the vehicle?” after the ride. In our experiments, participants answered trust questionnaires. In Experiment 1, we used the Trust in Automation (TIA) questionnaire by Körber (2018). Participants reported an average value of 4.26 ($SD = .58$) on the subscale “Trust in Automation” (consisting of the items “I trust the system.” and “I can rely on the system.”, scale ranges from 1 = *low trust* to 5 = *high trust*). This indicates relatively high trust in automation. In Experiment 2, we used the Automation Trust Scale by Jian et al. (2000) in a German translation, which largely based on the studies by Beggiato (2015) and Vogelpohl et al. (2016). We observed a numerical increase in trust after the experimental ride ($M_{\text{trustL3_before}} = 4.33$, $SD_{\text{trustL3_before}} = .20$, $M_{\text{trustL3_after}} = 5.57$, $SD_{\text{trustL3_after}} = .23$; scale ranges from 1 to 7; after recoding reversely scaled items higher values indicate higher trust). As trust values increased numerically after experiencing the simulated Level 3 driving automation, it can be assumed that participants trusted the driving automation system. Following the study by Müller et al. (2019), we can infer that participants believed in the cover story. In addition, after debriefing participants, their reactions revealed great surprise. For example, one participant of Experiment 1 compared it to the situation when a child learns that Santa Claus is not real. In summary, the first validity aspect by Müller et al. (2019) can be assumed to be given. The second aspect cannot be estimated, yet. To compare whether the simulated Level 3 driving automation behaves like a technical Level 3 driving automation system, it needs a Level 3 driving automation system that operates within the investigated use cases. However, such systems are not available yet.

In total, our methodological approach can be evaluated as objective, reliable, and valid from today’s perspective.

Future Research

For future studies on the question what NDRTs users will likely engage in, especially when using the method of observation, reducing factors that support reactivity to the experimental setting is highly recommended. Furthermore, when applying the method of Wizard-of-Oz, it would be interesting to investigate larger participant samples.

NDRTs During Monotonous Versus Discontinuous Automated Driving Phases

We investigated NDRTs in the context of Level 3 automated driving on motorways. In Experiment 1, the automation was available outside construction sites, motorway access, and motorway exits. In Experiment 2, the ride took place on a test track with a constant speed of 70 km/hr (approx. 43.5 mph). In both cases, participants experienced a Level 3 driving automation function that took over vehicle motion control within a long-distance ODD, where the ride was relatively monotonous. It would be

interesting for future work to investigate whether NDRT performance is different during discontinuous Level 3 automated rides, for example, in case of a traffic jam with mostly stop-and-go traffic.

Differences Between NDRTs Identified by Different Methods

We discussed that asking participants for NDRTs probably yields NDRTs that users engage in rather deliberately and attentively compared to NDRTs resulting from observation. The question arises whether these NDRTs (chosen deliberately, e.g., reading) differ from other NDRTs (e.g., looking outside) with regard to traffic safety. For example, it would be interesting to examine whether the former ones differ from the latter ones in terms of higher potential to occupy the users' mental capacity (e.g., attention) or additional hurdles (e.g., motivational) on disengaging from the NDRT at takeover.

Significance of Prior Experience With Driving Automation

From the experiments conducted, we assume that prior experience with automated driving is very important for our research question. All our participants did not have any prior experience with automated driving before participating in our study. In Experiment 1, participants first experienced automated driving on German motorways. After the ride, they were asked to imagine the function was available on their own vehicle. Participants then stated three to five NDRTs they would engage in during the automated driving phases. To answer this question, participants may have used the strategy to base their responses on the comparison of what they are allowed to do today while driving manually to what they would like to do while driving. Please keep in mind, that participants were only asked one question ("What NDRTs would you engage in during automated driving phases?"). We assume that the provided answers strongly result from the difference between the set of answers to the questions "What would I like to do while driving?" and "What am I allowed to do while driving today?" This comparison may dominate the selection of NDRTs and therefore bias responses in Experiment 1 and choices in Experiment 2. For example, in Experiment 2, participants were allowed to bring along their own NDRTs for the Level 3 automated driving phases. One participant brought along a textbook to read during the automated driving phases. From the video recordings, this textbook was about DIN A3 in size when opened and seemed to be heavy. During the Level 3 automated driving phase, the participant leaned the book against the steering wheel to read. However, the situation seemed uncomfortable to the participant because relative to the available space in the driver's cabin, the textbook was too large and heavy. Moreover, the steering wheel would move during the automated driving phase. From the video recordings, we could detect that these circumstances made it difficult for the participant to read. Therefore, we assume that next time this participant has to decide what NDRT she brings along, she would consider size and weight of the NDRT relative to the available space inside the driver's cabin more strongly than the first time. In contrast, the first time (i.e., participation in Experiment 2) she may have based her decision more strongly on what she would like to do and what is not possible today while driving. From this assumption, it follows that the prevalence of NDRTs smaller in size and light in weight may increase over time, that is, with drivers' increasing experience with Level 3 automated driving. Therefore,

the exact NDRTs resulting from this study may be limited in their predictive value with regard to NDRTs performed in the far future. Rather, NDRTs resulting from this study have more predictive value for NDRTs performed in the near future after release of Level 3 automated driving functions. Future research may investigate the described bias for choosing NDRTs more closely and examine whether this bias will reduce with increasing experience with Level 3 automated driving. For example, future studies could invite participants a second time and investigate whether there are differences between the NDRTs chosen in the first and second trial (e.g., with regard to size, weight, etc.).

Misuse of Level 3

We detected a naming rate of 3.38% for sleeping as a NDRT during a Level 3 automated driving phase. As outlined earlier and given today's available HMIs, sleeping is not in line with the role of a fallback-ready user who needs to intervene when the function requires it. Research on strategies to reduce misuse of Level 3 driving automation functions could provide a substantial contribution to traffic safety.

Conclusion

In conclusion, we could show that similar NDRTs were highlighted as relevant when using different methods, for example, reading, steady gazes to the outside, and smartphone usage. In Experiment 1, we let participants experience Level 3 automated driving in real motorway traffic in a real vehicle, and afterwards, asked participants on NDRTs they would engage in during a Level 3 automated driving phase. Naming rates were calculated for stated NDRTs. In Experiment 2, we invited participants to bring along NDRTs they want to execute during an approx. 30 min Level 3 automated driving phase, and observed what NDRTs participants actually engaged in. Engagement rates and total duration rate of engagements were calculated. The two methods provided different details regarding NDRT engagement. Open-ended questions seem adequate to generate a list of NDRTs that users choose deliberately and probably engage in consciously. In addition, this method may assess NDRTs that are performed rarely. However, as each NDRT has the same weight, one cannot differentiate between NDRTs with regard to their relevance for traffic safety, that is, how much time do they spent on a NDRT. In contrast, in situ observations offer the possibility to measure duration and number of engagements in NDRTs, which allows to describe NDRT engagement in more detail. Furthermore, NDRTs are directly derived from participants' behavior during Level 3 automation, whereas a questionnaire requires the bypass of participants' mental imagination and theoretical assumptions. Any bypass is prone to errors and confounding effects. We highlight the importance of and recommend providing realistic settings to derive natural NDRTs for Level 3 automated driving phases.

References

- Beggiato, M. (2015). *Changes in motivational and higher level cognitive processes when interacting with in-vehicle automation* [Doctoral dissertation]. University of Technology Chemnitz.
- Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., & Knipling, R. R. (2006). *The 100-car naturalistic driving study, phase*

- II-results of the 100-car field experiment (No. NTIS-PB2006110756). United States: Department of Transportation, National Highway Traffic Safety Administration. https://rosap.ntl.bts.gov/view/dot/37370/dot_37370_DS1.pdf
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>
- Feldhütter, A., Hecht, T., Kalb, L., & Bengler, K. (2019). Effect of prolonged periods of conditionally automated driving on the development of fatigue: With and without non-driving-related activities. *Cognition Technology and Work*, 12(1), 309–318. <https://doi.org/10.1007/s10111-018-0524-9>
- Frey, A. (2019). Müdigkeit und Vigilanz in einer automatisierten Realfahrt. In VDI Wissensforum GmbH (Ed.), *Mensch-Maschine-Mobilität 2019: Der (Mit-)Fahrer im 21. Jahrhundert!* (pp. 121–132). VDI-Verlag.
- Gold, C. (2016). *Modeling of take-over performance in highly automated vehicle guidance* [Doctoral dissertation]. TU Munich. <http://nbn-resolving.org/urn:nbn:de:hbz:5:1-63882-p0011-7>
- Hecht, T., Feldhütter, A., Draeger, K., & Bengler, K. (2020). Non-driving related activities in automated driving—An online survey investigating user needs. In T. Ahram, W. Karwowski, S. Pickl, & R. Taiar (Eds.), *Advances in intelligent systems and computing. Human systems engineering and design II* (Vol. 1026, pp. 182–188). Springer. https://doi.org/10.1007/978-3-030-27928-8_28
- Hohm, A., Klejnowski, L., Skibinski, S., Bengler, K., Berger, S., Vetter, J., Krug, S., & Stürmer, T. (2018). *KO-HAF—Kooperatives Hochautomatisiertes Fahren: (Projektübergreifender Schlussbericht)*. https://www.ko-haf.de/fileadmin/user_upload/projekt/19S14002_Ko-HAF_partnertübergreifender-Schlussbericht_final.pdf
- Huemer, A. K., & Vollrath, M. (2012). Ablenkung durch fahrfremde Tätigkeiten—Machbarkeitsstudie: Bericht zum Forschungsprojekt FE 82.376/2009. In *Berichte der Bundesanstalt für Straßenwesen M, Mensch und Sicherheit: Vol. 225*. Wirtschaftsverl. NW Verl. für neue Wiss. <http://bast.opus.hbz-nrw.de/volltexte/2012/382/pdf/M225b.pdf>
- ISO/TS 14198: 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. (2012). <https://www.iso.org/standard/54496.html>
- Jarosch, O. (2020). *Non-driving-related tasks in conditional driving automation: Occurrence and effects of passive task-related fatigue on take-over performance in conditional driving automation* [Dissertation]. TU München. <https://mediatum.ub.tum.de/doc/1524061/1524061.pdf>
- Jarosch, O., Bellem, H., & Bengler, K. (2019). Effects of task-induced fatigue in prolonged conditional automated driving. *Human Factors*, 61(7), 1186–1199. <https://doi.org/10.1177/0018720818816226>
- Jarosch, O., & Bengler, K. (2019). Is it the duration of the ride or the non-driving related task? What affects take-over performance in conditional automated driving? In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Advances in intelligent systems and computing. Proceedings of the 20th congress of the international ergonomics association (IEA 2018)* (Vol. 823, pp. 512–523). Springer. https://doi.org/10.1007/978-3-319-96074-6_54
- Jarosch, O., Gold, C., Naujoks, F., Wandtner, B., Marberger, C., Weidl, G., & Schrauf, M. (2019). The impact of non-driving related tasks on take-over performance in conditionally automated driving—A review of the empirical evidence. In TUM Lehrstuhl für Fahrzeugtechnik mit TÜV SÜD Akademie (Chair), 9. *Tagung Automatisiertes Fahren*. Symposium conducted at the meeting of Lehrstuhl für Fahrzeugtechnik mit TÜV SÜD Akademie. <https://mediatum.ub.tum.de/doc/1535156/1535156.pdf>
- Jian, J. Y., Bisanz, 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
- Kathmann, T., Johannsen, M., von Heel, E., Hermes, T., Vollrath, M., & Huemer, A. K. (2020). Nutzungshäufigkeit von smartphones durch pkw-fahrer: Erhebung 2019 [Frequency of smartphone use by car drivers: Survey 2019]. *Berichte der Bundesanstalt für Strassenwesen M, Mensch und Sicherheit*. Fachverlag NW in Carl Ed. Schünemann KG.
- Klamroth, A., Zerbe, A., & Marx, T. (2019, October 29). Transitionen bei level-3-automation: Einfluss der verkehrsumgebung auf die bewältigung-leistung des fahrers während realfahrten [Transitions at level 3 automation: Influence of traffic environment on drivers' accomplishment in real rides]. *FAT-Schriftenreihe* 323. https://www.vda.de/dam/vda/publications/2019/FAT/FAT-Schriftenreihe_323.pdf
- Körper, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Advances in Intelligent Systems and Computing. Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (Vol. 823, pp. 13–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2
- Large, D. R., Burnett, G., Morris, A., Muthumani, A., & Matthias, R. (2017). *Design implications of drivers' engagement with secondary activities during highly automated driving—A longitudinal simulator study*. <https://nottingham-repository.worktribe.com/output/1214099/design-implications-of-drivers-engagement-with-secondary-activities-during-highly-automated-driving-a-longitudinal-simulator-study>
- Metz, B., Landau, A., Hargutt, V., & Just, M. (2013). Womit beschäftigt sich der fahrer während der fahrt? Ergebnisse zur häufigkeit von ablenkung beim fahren aus einer Naturalistic Driving Study [What do drivers engage in during the ride? Results from a Naturalistic Driving Study on distraction frequency while driving]. In VDI-Berichte: Vol. 2205, *Der Fahrer im 21. Jahrhundert: Fahrer, Fahrerunterstützung und Bedienbarkeit; 7. Vdi-Tagung, Braunschweig, 5. Und 6. November 2013*. VDI-Verl.
- Müller, A. I., Weinbeer, V., & Bengler, K. (2019). Using the wizard of Oz paradigm to prototype automated vehicles: methodological challenges. In C. P. Janssen, S. F. Donker, L. L. Chuang, & W. Ju (Eds.), *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications: adjunct proceedings* (pp. 181–186). Association for Computing Machinery. <https://doi.org/10.1145/3349263.3351526>
- Naujoks, F., Befelein, D., Wiedemann, K., & Neukum, A. (2017). A review of non-driving-related tasks used in studies on automated driving. In N. Stanton (Ed.), *Advances in Intelligent Systems and Computing: Vol. 597. Advances in Human Aspects of Transportation*. AHFE 2017 (pp. 525–537). Springer. https://doi.org/10.1007/978-3-319-60441-1_52
- Pfleging, B., Rang, M., & Broy, N. (2016). Investigating user needs for non-driving-related activities during automated driving. In J. Häkila & T. Ojala (Eds.), *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia – MUM '16* (pp. 91–99). ACM Press. <https://doi.org/10.1145/3012709.3012735>
- Radlmayr, J. (2020). *Take-over performance in conditionally automated driving: effects of the driver state and the human-machine-interface* [Dissertation]. Technische Universität München.
- Society of Automotive Engineers (SAE) International. (2021). *SAE international standard J3016: (R). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. https://www.sae.org/standards/content/j3016_202104/
- Schoettle, B., & Sivak, M. (2014). *A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia*. University of Michigan. <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/108384/103024.pdf?sequence=1&isAllowed=y>
- Shi, E., Gasser, T., Seeck, A., & Auerswald, R. (2020). The principles of operation framework: A comprehensive classification concept for automated driving functions. *SAE International Journal of Connected and Automated Vehicles*, 3(1), 27–37. <https://doi.org/10.4271/12-03-01-0003>
- Shuttleworth, J. (2019). *SAE Standards News: J3016 automated-driving graphic update*. <https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic>

- Simon, M., Schmidt, E. A., Kincses, W. E., Fritzsche, M., Bruns, A., Aufmuth, C., Bogdan, M., Rosenstiel, W., & Schrauf, M. (2011). EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clinical Neurophysiology*, 122(6), 1168–1178. <https://doi.org/10.1016/j.clinph.2010.10.044>
- Vogelpohl, T., Vollrath, M., Kühn, M., Hummel, T., & Gehlert, T. (2016). *Übergabe von hochautomatisiertem fahren zu manueller steuerung (Forschungsbericht Nr. 39)* [Handover from conditional automated driving to manual driving (Report No. 39)]. Gesamtverband der Deutschen Versicherungswirtschaft e. V. [German Insurance Association (GDV)]. https://udv.de/sites/default/files/tx_udvpublications/fobe_39_hochautomfahren.pdf
- Wandtner, B. (2018). *Non-driving related tasks in highly automated driving: Effects of task characteristics and drivers' self-regulation on take-over performance* (Inaugural-dissertation). Julius-Maximilians-Universität Würzburg. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/17395/file/Wandtner_Bernhard_AutomatedDriving.pdf
- Weinbeer, V., Muhr, T., & Bengler, K. (2019). Automated driving: The potential of non-driving-related tasks to manage driver drowsiness. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. IEA 2018. *Advances in intelligent systems and computing*, Vol 823. Springer. https://doi.org/10.1007/978-3-319-96074-6_19
- Wu, Y., Kihara, K., Hasegawa, K., Takeda, Y., Sato, T., Akamatsu, M., & Kitazaki, S. (2020). Age-related differences in effects of non-driving related tasks on takeover performance in automated driving. *Journal of Safety Research*, 72, 231–238. <https://doi.org/10.1016/j.jsr.2019.12.019>
- Zeeb, K. (2016). *Der einfluss fahrfremder tätigkeiten auf die fahrerübernahme während des hochautomatisierten fahrens* [The impact of non-driving related activities on drivers' takeover in the context of conditionally automated driving] [Doctoral dissertation, Heinrich-Heine-Universität Düsseldorf]. HHU Repository. https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-43957/Diss_Zeeb_ver%C3%B6ffentlicht-1.pdf

Appendix

Fahrfremde Tätigkeiten

Stellen Sie sich vor, Sie könnten die automatisierte Fahrfunktion auch in ihrem privaten PKW verwenden. Welchen Nebentätigkeiten würden Sie während der automatisierten Fahrtstrecken nachgehen? Bitte geben Sie mind. 3 – 5 Nebentätigkeiten an.

1. _____
2. _____
3. _____
4. _____
5. _____

Received April 14, 2020
 Revision received February 6, 2021
 Accepted February 10, 2021 ■