

# Methodological Approach towards Evaluating the Effects of Non-Driving Related Tasks during Partially Automated Driving

Cornelia Hollander <sup>1,\*</sup>, Nadine Rauh <sup>1</sup>, Frederik Naujoks <sup>2</sup>, Sebastian Hergeth <sup>2</sup>, Josef F. Krems <sup>1</sup> and Andreas Keinath <sup>2</sup>

<sup>1</sup> Research Group Cognitive and Engineering Psychology, Chemnitz University of Technology, 09120 Chemnitz, Germany; Nadine.Rauh@psychologie.tu-chemnitz.de (N.R.); Josef.Krems@psychologie.tu-chemnitz.de (J.F.K.)

<sup>2</sup> BMW Group, 80937 Munich, Germany; Frederik.Naujoks@bmw.de (F.N.); Sebastian.Hergeth@bmw.de (S.H.); Andreas.Keinath@bmw.de (A.K.)

\* Correspondence: cornelia.hollander@psychologie.tu-chemnitz.de; Tel.: +49-371-531-32044

Received: 27 May 2020; Accepted: 26 June 2020; Published: 30 June 2020

**Abstract:** Partially automated driving (PAD, Society of Automotive Engineers (SAE) level 2) features provide steering and brake/acceleration support, while the driver must constantly supervise the support feature and intervene if needed to maintain safety. PAD could potentially increase comfort, road safety, and traffic efficiency. As during manual driving, users might engage in non-driving related tasks (NDRTs). However, studies systematically examining NDRT execution during PAD are rare and most importantly, no established methodologies to systematically evaluate driver distraction during PAD currently exist. The current project's goal was to take the initial steps towards developing a test protocol for systematically evaluating NDRT's effects during PAD. The methodologies used for manual driving were extended to PAD. Two generic take-over situations addressing system limits of a given PAD regarding longitudinal and lateral control were implemented to evaluate drivers' supervisory and take-over capabilities while engaging in different NDRTs (e.g., manual radio tuning task). The test protocol was evaluated and refined across the three studies (two simulator and one test track). The results indicate that the methodology could sensitively detect differences between the NDRTs' influences on drivers' take-over and especially supervisory capabilities. Recommendations were formulated regarding the test protocol's use in future studies examining the effects of NDRTs during PAD.

**Keywords:** partially automated driving; non-driving related tasks; take-over situations; test protocol development; user studies (simulator; closed circuit)

## 1. Introduction

### 1.1. Theoretical Background

In recent years, researchers and practitioners alike have been increasingly motivated to enhance driving assistance and automation resulting in different vehicle automation levels [1], with the overarching goal to improve driving comfort, traffic safety and to reduce traffic congestion [2].

The Society of Automotive Engineers (SAE) [3] defines six automation levels ranging from no automation (level 0) to full automation (level 5). However, only SAE level 1 and level 2 systems are currently available to consumers. Of those, partially automated driving (PAD, SAE level 2) provides continuous steering as well as brake and acceleration support to the driver; however, the driver must

constantly supervise these support features and be prepared to steer, brake, or accelerate as needed to maintain safety [3].

Even though the driver is partially relieved from the driving task and its demands [4], this does not automatically improve driving safety. New problems may arise during PAD when changing the drivers' role from an active to a passive system supervisor [5,6]. For instance, the (partial) relief of driving demands and decreased driving task engagement during automated driving can reduce the drivers' workload and cause cognitive underload [4]. In combination with lengthy supervision periods that are likely to result in both boredom and monotony [7], it can lead to fatigue and therefore inattention towards safety critical aspects of the driving task [8]. Drivers' inattention can be divided into (1) driver restricted attention, which is reflected by the driver's mind wandering for instance [7] and (2) driver diverted attention, emerging through the drivers focusing attention towards driving unrelated tasks (i.e., non-driving related tasks (NDRTs)) [9].

To a certain degree, engaging in NDRTs while driving might not considerably impair driving performance due to drivers' spare attentional capacity. For the manual driving task, which is a predominantly visual task [10], literature proposes that 20–25% [11] or, depending on the driving environment's complexity, even up to 50% [12] of drivers' visual attention is focused on objects unrelated to the driving task (e.g., advertisements, scenery), reflecting drivers' spare attentional capacity during manual driving. This suggests that drivers could execute certain NDRTs during manual driving without performance losses in their gaze and driving behavior as the resources are obtained from the spare attentional capacity [13]. Secondary tasks, identified as suitable for manual driving according to the National Highway Traffic Safety Administration (NHTSA) [14], are likely executed based on these resources. An on-road driving study showed that participants tapped into this spare capacity to increase visual attention towards a newly provided in-vehicle-display during manual driving, indicated by fewer glances towards the speedometer or periphery [15]. Glances towards the street ahead or the mirrors were not affected. Nevertheless, the driving environment's complexity and the workload of the driving task itself e.g., [16,17] might influence how much spare capacity drivers have (e.g., the more complex driving environment, the less free visual attentional capacity available). For PAD, the driving demands are partially reduced since the automated system controls longitudinal and lateral position [4]. Hence, it could be assumed that more spare attentional capacity would be available, for instance, to execute NDRTs without performance losses. However, drivers constantly need to supervise the driving environment as well as the automated system to take over the driving task if necessary, which demands a crucial amount of visual attention. Hence, if the amount of attentional resources needed exceeds the available spare capacity, NDRT execution during PAD will likely negatively impact drivers' supervisory and take-over performance.

Several simulator and real-world environment studies have revealed that drivers tended to engage in NDRTs (e.g., watching a DVD, reading or interacting with a smartphone) to reduce boredom, monotony, and cognitive underload associated with automated driving e.g., [18–20]. However, this behavior might overly distract drivers from vigilantly supervising the system and driving scene, thus reducing their situation awareness if the resources needed to execute the NDRTs are taken from beyond the drivers' spare capacity.

In general, cognitive underload, driver inattention and the associated decrease in system supervision are all related to slower and poorer reactions [7], complete failures to react and poor decision making in the event of a system failure [20,21], which could counteract the potential safety benefits of automated driving [22] and exemplify potential new challenges connected with PAD. Hence, the drivers' new passive role leading to the new problems mentioned, together with the fact that these intermediate automation levels are not consistently reliable yet [23] underline the importance of reasonable usage and implementation of automated driving [14].

### *1.2. Examining the Effects of NDRTs during PAD*

Thus far, many studies have examined the effects of NDRT execution during automated driving. However, most of these studies have focused on automation levels other than PAD. For instance, Carsten et al. [20] observed voluntary NDRT execution (e.g., eating or watching a DVD) during

semiautomated driving, which they defined as automated lateral or longitudinal control, or highly automated driving in a simulator. In contrast, other studies have focused on driver NDRT execution during higher automation levels, such as conditional automation (i.e., SAE level 3) e.g., [18,24].

In addition, many of the PAD studies only included NDRTs as a secondary aspect, while focusing on other main aspects. For example, one simulator study [25] mainly focused on how anticipatory information affected drivers' supervisory behavior during PAD while executing NDRTs (e.g., reading or interacting with a smartphone) on a voluntary basis. A different simulator study mainly centered on how participants' self-regulation during secondary task engagement would affect their supervisory behavior [26]. In contrast, other studies have required participants to execute the NDRTs during PAD instead of leaving it optional. However, the core focus still remained on aspects other than the NDRT execution itself. For instance, Large et al. [27] compared behavioral cues of distraction during NDRT execution (i.e., reading task) across three automation levels: manual driving, PAD, and highly automated driving. Another simulator study examined whether a NDRT could reduce fatigue during PAD [28]. However, concentrating on the systematic evaluation of NDRT execution during PAD is highly important since drivers are likely to engage in these tasks due to their spare attentional capacity available and to reduce the monotony and boredom of the supervisory task. Moreover, it is essential because these tasks might have similar negative and distractive effects on the driver during PAD as they have during manual driving.

In addition to the fact that NDRT execution is often only a secondary aspect, PAD studies often differ considerably regarding their applied methods as well as PAD specifications. For instance, some PAD studies involved take-over requests to redirect the participants' attention towards the driving task e.g., [25,27], whereas others have examined the drivers' ability to detect automation failures during NDRT execution without warning [29]. Moreover, several studies did not include any situations or automation failures requiring participants to regain vehicle control e.g., [20]. Additionally, some studies employed PAD to assist with navigating traffic congestions and managing speeds under 50 km/h e.g., [25], while other studies employed PAD for managing higher speeds, such as 130 km/h e.g., [30].

Hence, even though these studies often applied similar "[...] paradigms when participants are instructed to undertake a period of automated driving, and additionally given the option to (and are free to when/if comfortable) engage in a range of secondary activities available to them while sitting in the driver seat" [p. 3, 22], the varying methods and specifications these studies used complicate the generalization of the findings. Moreover, the different studies yielded varying results regarding the effects of NDRT execution during PAD. For instance, one simulator study revealed that reaction time during hazardous situations clearly increased when driving with NDRT execution compared to without NDRT execution [25]. Another simulator study permitted participants to freely engage in various smartphone activities during PAD and highly automated driving [30]. On the one hand, results showed that drowsiness and highly motivational NDRTs negatively affected driving performance during PAD in terms of slower reactions. On the other hand, NDRTs with low to moderate visual and mental workloads improved driving performance in a hazardous situation [30].

This brief overview underlines the general need to examine the effects of NDRT execution on drivers during PAD by systematically manipulating various NDRTs. Moreover, the different methods and specifications employed in the studies emphasize the importance of incorporating a standardized methodology that is comparable to, for example, the methodology used by the NHTSA to examine NDRTs during manual driving [14]. Previous efforts in developing standardized methods either focused on higher automation levels, such as the overview of current research questions and relevant methodical approaches in the conditional automated driving field (SAE level 3) [31], or on evaluating PAD system and human-machine-interface (HMI) designs e.g., [32,33]. This clearly emphasizes the need to fill this gap and develop a standardized method to enhance the comparability, reproducibility, and generalization of these studies and their results. The standardization supports continuing examination of NDRT execution effects on drivers' supervisory and take-over capabilities to reach the important, overarching goal of safe PAD usage.

### *1.3. Developing a Standardized Methodology to Evaluate NDRT Execution during PAD*

For manual driving, well-established methodologies and guidelines exist that detail how the effects of NDRTs on driving performance and gaze behavior can be evaluated. For example, NHTSA's well-established methodology focuses on examining different visual-manual NDRTs [14]. In addition to the standardized methodology, NHTSA provides guidelines and cut-off values that clearly regulate whether a NDRT is suitable for execution during manual driving [14]. For instance, to be acceptable, the single gaze durations towards a NDRT should not exceed two seconds. Further, neither driving performance nor gaze behavior during NDRT execution should be poorer than during execution of the manual radio tuning task [34], which is NHTSA's [14] recommended reference task. Employing the methodology and guidelines standardizes the evaluation of visual-manual NDRTs and enhances the comparability and reproducibility between the studies incorporating them.

However, the driving task during PAD is considerably different than manual driving. Hence, the evaluation specifications for manual driving performance (e.g., lane maintenance, speed or distance to another vehicle) no longer apply given that the automated system takes over these tasks in PAD. Instead, the drivers' ability to vigilantly supervise the system during a prolonged PAD period and to take over vehicle control immediately and in a safe manner, if necessary, become more important during PAD. Moreover, since the driving demands during PAD are lower than those of manual driving allowing for more available cognitive resources, the question arises whether the cut-off values for manual driving proposed by the NHTSA [14] are still applicable for PAD.

Therefore, to be useful when examining NDRT execution during PAD, existing methods need to be adapted and fulfill several additional requirements. Firstly, the new method needs to capture and sensitively evaluate drivers' capabilities to perform the new tasks that are important during PAD (i.e., vigilant supervision and taking-over the driving task if needed). To fulfill these requirements, the PAD periods must be interrupted by critical situations in which drivers must recognize the need and be able to take-over vehicle control due to a system failure or limit based on their vigilant system supervision. Secondly, the methodology must be sensitive to different NDRTs with varying distractive potentials and to other aspects relevant to the automobile context (e.g., different (in-vehicle) display locations). Lastly, the methodology must enable the establishment of cut-off values, comparable to those for manual driving, based on an adequate number of testing. A further beneficial characteristic of the method would be the ability to adapt it based on the research questions of interest.

Due to safety considerations, any new methodology for testing PAD should initially be applied in a driving simulator. However, the effects of NDRT execution in a real-world environment are potentially more safety critical than in a driving simulator. Therefore, it is also necessary to examine the external validity of such methods. Accordingly, a high external validity would greatly enhance the methods' generalizability.

#### *1.4. Objectives of the Present Research Project*

Since the methodology assessing manual driving is not applicable for PAD, the current project's overarching goal was to fill this methodological gap and take the initial steps towards developing a test protocol and providing recommendations for the systematic evaluation of drivers' supervisory and take-over capabilities during PAD while engaging in different NDRTs. To achieve this, the well-established methodology for manual driving [14] was extended for PAD based on the formulated requirements (see Section 1.3). The new test protocol was developed, validated and adapted through the course of three studies.

- The first two studies took place in a driving simulator to determine the potential for a new test protocol to assess the effects of NDRT execution during PAD.
- In the second simulator study, the new test protocol was also extended to other relevant aspects, such as (in-vehicle) display locations.
- The third study was conducted in a partially automated vehicle to validate the test protocol in a real vehicle on a closed test track. The main goal was to determine whether the test protocol was applicable to a real driving environment.

The following research questions were addressed within the three studies:

- Research question 1 (RQ1): Can the test protocol sensitively detect differences, as they are expected based on the literature, in the drivers' supervisory and take-over capabilities during PAD depending on various influencing factors?
  - Research question 1a (RQ1a): Is the new test protocol sensitive to the effects of visual-manual NDRTs with varying distractive potentials on the drivers' supervisory and take-over capabilities during PAD?
  - Research question 1b (RQ1b): Is the new test protocol sensitive to the effects of (in-vehicle) display locations with varying proximity to the driving scene on the drivers' supervisory and take-over capabilities during PAD?
- Research question 2 (RQ2): What parameters are minimally necessary and sufficient to sensitively capture and evaluate the take-over and supervisory capabilities of the drivers in light of these aspects?

The focus of the current article and the corresponding research questions solely lie with the examination of the proposed test protocols' suitability to sensitively evaluate the effects of NDRTs on the drivers during PAD. The evaluation of the NDRTs' particular effects on the drivers' take-over and supervisory capabilities using the provided test protocol were not the focus of the current manuscript. The presentation of the specific results will be described in more detail in separate papers.

## 2. Test Protocol Development

The following chapters will describe the relevant aspects of the test protocol's development based on the literature and existing research and specific implementation by the authors within the three studies of the current project, beginning with a presentation of the driving scenario and the implemented take-over situations (Section 2.1), followed by a description of the evaluated independent variables (Section 2.2) and the assessed dependent variables (Section 2.3). In the subsequent chapters, the equipment and materials used are presented (Section 2.4), as well as the detailed experimental design and procedure (Section 2.5). This is preceded by a presentation of the data preparation and analysis (Section 2.6). The final chapter (Section 2.7) describes the participants of all three studies. In addition, two subchapters are integrated into each chapter, providing unique details of the driving simulator studies (Section 2.x.1 Driving Simulator Implementation) and the test track study (Section 2.x.2 Test track implementation).

### 2.1. Driving Scenario and Take-Over Situations

This section will describe the development and the specific implementation of the driving scenario and the take-over situations. To test how visual-manual NDRTs affect drivers during manual driving, the NHTSA methodology recommends incorporating a car-following scenario on a highway road [14]. Effects are evaluated by judging the drivers' gaze behavior and driving performance while executing the NDRTs. More precisely, drivers are evaluated on their ability to maintain distance to the lead vehicle, speed, proper lane maintenance during the car-following scenario, as well as how long, in terms of single and total glance durations, the drivers are glancing towards the NDRTs [14].

To standardize examination, NHTSA [14] prescribes several specifications for the test track and driving scenario. Firstly, NHTSA recommends a car-following scenario where drivers attempt to maintain a certain speed (80 km/h) and distance to the lead vehicle (70 m), which allows the examination of the drivers' ability to fulfill this task during NDRT execution [14]. Moreover, NHTSA advises using a straight highway route with two lanes per direction and a predefined lane width. This reflects a realistic setting and enables examination of the drivers' ability to stay within the lane for instance [14]. Accordingly, straight road segments should be used to examine the drivers' gaze and driving behavior, although curved segments can be included occasionally [14]. Lastly, NHTSA

recommends using a generic driving environment that excludes any external cues (i.e., trees, houses) [14], though they allow occasional (oncoming) traffic during the car-following task.

Due to the changed driving task during PAD, parameters for manual driving are not applicable anymore and it is necessary to evaluate drivers' supervisory and take-over capabilities when a system limit is reached. Therefore, generic take-over situations had to be implemented that simulate such a system limit. The current studies included two take-over situations: (a) *lead vehicle deceleration* and (b) *drifting of the participant's partially automated vehicle* (i.e., *ego vehicle*). Both situations represented system limitations directly corresponding to the main driving tasks taken over by the partially automated system (lateral and longitudinal vehicle control).

Although these two take-over situations were based on earlier research see [35], several adaptations were made to match the NHTSA scenario specifications more closely. During the take-over situation with *lead vehicle deceleration* (addressing longitudinal vehicle control) the lead vehicle slowed down without brake lights. To mimic realistic braking movement, the vehicle slowed down based on a predefined value. Without any driver intervention (i.e., braking), a collision with the lead vehicle would occur. During the second take-over situation involving *ego vehicle drifting* (addressing lateral vehicle control), the vehicle drifted to the left or right see [35]. To prevent a guardrail collision, the participants had to notice the drifting and steer in the opposite direction. A collision would occur without any driver intervention. To ensure comparability between the two distinct situations, the outcome and time to collision (TTC) were identical: without any driver intervention, a collision (outcome) with the lead vehicle or guardrails would occur after the same predefined TTC.

Following Signal Detection Theory (SDT), the most critical situations are missed warnings, in which errors or events occur without any warnings to the system supervisor [36]. When a system limitation is reached during PAD, the automated system neither gives a warning nor issues any take-over request for the drivers, thus the drivers must vigilantly supervise any system changes [3]. Therefore, any system warning or take-over request for the two take-over situations were excluded. All environmental (i.e., trees or houses) and vehicle (i.e., lead vehicle's brake lights or steering wheel movement) visual cues were excluded to reduce the predictability of the take-over situations.

### 2.1.1. Driving Simulator Implementation

The driving scenario and take-over situations were implemented as follows in the two driving simulator studies. Based on the NHTSA methodology, an identical car-following scenario on a straight, four-lane highway route with two lanes in each direction was included. Moreover, the same specifications for speed (80 km/h) and distance to the lead vehicle (70 m) were applied. The participants drove a partially automated vehicle that controlled the longitudinal and lateral position. The two take-over situations (*lead vehicle deceleration* and *ego vehicle drifting*) were implemented with the general specifications discussed in Section 2.1. The specific braking speed of the decelerating lead vehicle was  $2.3 \text{ m/s}^2$ , which corresponded to an electric vehicle with a regenerative braking movement. Without any driver intervention, a collision with the lead vehicle would have occurred after seven seconds. A collision with the guardrails would have occurred after seven seconds in the ego vehicle drifting scenario if the driver did not intervene in time. The participants were introduced to react by braking or steering, respectively, to regain control from the partially automated system.

The test track was 11 km long in the first and 9 km long in the second simulator study, which was programmed using the Silab 5.0 simulation environment. To reduce the predictability of the take-over situations, the driving environment was as generic as possible, excluding any visual cues (i.e., trees). In contrast to the NHTSA guidelines, the simulation did not include any traffic other than the lead and ego vehicle. This allowed for a controlled execution of the take-over situations without needing to, for instance, check for rear traffic before braking. Further, the aim was to reduce any potential distractions, especially during the reference trial, in which boredom might have encouraged drivers to gaze towards irrelevant vehicles instead of focusing on the system and lead vehicle. Although this aspect is also important in terms of situation awareness, it was not the focus of our studies.



### 2.1.2. Test Track Implementation

To ensure participant safety as well as a standardized data collection free of any interference, the third study occurred on a closed test track in a parking lot. The limited space restricted the precise application of the specifications used in the driving simulator, resulting in several adaptations. These adaptations resulted in differences between the simulator and test track studies regarding, for instance, the execution of the driving scenario and take-over situations. These differences potentially reduced the comparability between the results of the two study types (see Table 1 for a comparative overview).

Firstly, compared to the simulator studies, the driving scenario was downscaled for the test track study in terms of the driving environment (with landmarks), test track (one lane, with curves), speeds (max. 25 km/h), the distance between the two vehicles (speed of ego vehicle/2 + 7 m) and the particular execution of the two take-over situations. Nonetheless, the goal was to mimic the scenario as much as possible by finding a test track with as few curves as possible and with at least one long, straight segment for the *lead vehicle deceleration* take-over situation. The driving scenario and take-over situations relied heavily on non-automated, human execution (i.e., lead vehicle or lateral ego vehicle control maintained via Wizard-of-Oz). Therefore, the take-over situations were always executed on the same track segment to enhance reproducibility and comparability as well as to reduce the chance for human error. Figure 1 (top row, left) shows the final test track with the two segments chosen for the two take-over situations, the execution of lateral vehicle control (top row, right) and the experimental setup within the ego vehicle (bottom row, left) as well as the two vehicles involved in the driving scenario (bottom row, right).



**Figure 1.** Test track depicting the two take-over situation locations (**top row, left**), the execution of lateral vehicle control via Wizard-of-Oz (**top row, right**), experimental setup within the ego vehicle (**bottom row, left**), and the two vehicles involved in the driving scenario (**bottom row, right**).

**Table 1.** Differences and similarities between the simulator and test track.

	Simulator	Test Track	Relevant Differences
Study Environment	Simulator	Test track	Yes, since the experienced risk and situation criticality likely differ between the two study environments. For instance, they are likely higher in a real vehicle (on a test track) than in a simulator, where collisions pose no risk to the participants' safety.
Driving Scenario			

Car-following scenario	Yes	Yes	/
Duration	Approx. 8–9 min	Approx. 12 min	No, since the difference is only minor and therefore unlikely to elicit different levels of monotony or fatigue.
Route	Straight highway	Test track on parking lot with many curves	Yes, because a test track with several curves potentially results in different supervisory behavior (e.g., enhanced supervision of the street) compared to a straight test track. Hence, the supervisory behavior on the simulated straight highway might be underestimated compared to the more realistic curved test track.
Surrounding	Without landmarks	With many landmarks	No, since the landmarks (i.e., trees) possessed low levels of visual attraction and were static. Hence, they were less distractive than, for example, dynamic landmarks.
<b>Take-over Situations</b>			
Types	Lead vehicle deceleration and Ego vehicle drifting	Lead vehicle deceleration and Ego vehicle drifting	/
Number of situations per trial	4	4	/
Collision risk	Yes	No	No, since participants were aware that the theoretical consequence of failing to react in a real vehicle would be a collision, a certain level of situation criticality still existed.
Duration (e.g., time to collision)	7 s	Not precisely realizable	No, since the time to collision of the take-over situations was comparably short in all studies (i.e., only several seconds) and participants needed to react as soon as they noted the take-over situations. The differences in distance and speed would likely result in a reduced comparability of the studies' results (see below).
Occurrence	Defined in simulation	Predefined by test circuit, same track segment	Yes, since it enhances the predictability of the take-over situation (locally).
Predictability (locally)	Low	High	Yes, since it potentially influences participants' (supervisory) behavior during the particular segment (e.g., more gazes towards the street and therefore faster reactions to the take-over situations.)
Predictability (time-wise)	Low	Low	/
<b>Ego vehicle</b>			
Speed	80 km/h	Max. 27 km/h	Yes, both the lower speeds driven and smaller distance to the lead vehicle during the test track study potentially influenced the experienced situation criticality (e.g., reduced criticality through reduced speed; however, partially counteracted through decreased distance between the vehicles). This might have reduced the comparability between the studies' results (e.g., longer gazes towards the NDRTs on the test track due to reduced speeds).
Distance to lead vehicle	70 m	Ego vehicle speed/2 + 7 m	
Longitudinal control	Automated	Automated	/
Lateral control	Automated	Wizard-of-Oz	No, the Wizard-of-Oz execution itself should not influence the results or the comparability of the studies' results, provided participants did not notice the researcher executing lateral control (which was not the case).
<b>Lead Vehicle</b>			
Speed	80 km/h	Max. 25 km/h,	No, the lead vehicle's speed generally is not a separate factor influencing the comparability of the studies' results. It is due to the reduced speeds driven on the test track. Hence, its influence is incorporated



			into the impact of the ego vehicle's speed and the distance to the lead vehicle.
Vehicle control	Simulator (preprogrammed)	Manually driven (with cruise control and motor deceleration/acceleration)	Yes, it reduces the standardization, comparability and reproducibility of the scenario and the take-over situations. Moreover, it influences the driving data (e.g., distance when the take-over situations were triggered). However, the participants' supervising behavior should not be affected.

On the corresponding test track segment (Figure 1, top row, left), the *lead vehicle deceleration* situation was employed as follows: During this segment, the lead vehicle was driven in activated cruise control mode and only slowed down when the motor decelerated after the cruise control was deactivated. The lead vehicle's brake lights did not activate. During this segment, the ego vehicle was not programmed to maintain distance to the lead vehicle and, therefore, moved closer until the participants intervened. During the *ego vehicle drifting* situation, a researcher sitting in the ego vehicle's passenger seat used a small steering wheel to execute the drifting (Figure 1, top row, right). To reduce human error likelihood and enhance reproducibility, the researcher always drifted the vehicle to the left.

As in the simulator studies, the participants needed to brake or steer in response to the take-over situation, although they could also stop the vehicle by merely touching the steering wheel. Unlike in the simulator studies, the two take-over situations did not result in a collision, even when the participants did not react. For this matter, several fallback solutions were included in case participants failed to intervene, such as programming the ego vehicle to stop automatically if a minimal safety distance is reached and a researcher who could stop the ego vehicle by employing the emergency brake.

## 2.2. Independent Variables

In Section 1.3, the requirement was formulated that the test protocol must be sensitive to the effects of different NDRTs (RQ1a) as well as to other relevant aspects to the PAD context, in this case different (in-vehicle) display locations (RQ1b). In the following two sections (Sections 2.2.1 and 2.2.2) and the corresponding subsections, the theoretical background as well as the specific implementation of the independent variables will be explained.

### 2.2.1. Non-Driving Related Tasks

Several studies have indicated that drivers tend to engage in NDRTs to reduce cognitive underload, boredom, and monotony resulting from the reduced driving demands during PAD e.g., [18–20]. Due to their potentially safety diminishing effects, the new test protocol must sensitively capture the different effects of these NDRTs on the drivers' supervisory and take-over capabilities to evaluate whether a certain NDRT is applicable during PAD.

Amongst other theories and models, the multiple resource model [37] is regularly used to differentiate between NDRTs based on their required modalities as well as between the different visual NDRT effects on drivers' performance and gaze behavior during manual driving e.g., [38,39]. Multiple resource theory, which builds the basis of the model, focuses on the idea that when executing multiple tasks simultaneously, it is necessary to share time and attention between these tasks [37]. Moreover, when these two tasks occupy the same modalities (e.g., both requiring visual attention), these tasks interfere with each other as resources and attention are divided [37,40]. This results in reduced (attentional) resources for both tasks compared to executing only one task at a time [40], thus decreasing performance for one or both tasks [37]. Based on the multiple resource model, it is assumed that visual NDRTs are especially distracting during driving e.g., [38] and cause decreased performance in the driving task, the NDRT, or both since the driving task itself is highly reliant on visual resources [41]. Therefore, visual NDRT execution seems especially problematic during manual driving and are thus given priority by NHTSA. In general, NHTSA focuses on visual tasks with a manual aspect, where the driver must manipulate a device to execute the task [14]. Since

the driving task requires drivers to steer or shift gears, the manual NDRT component would likely interfere with these driving tasks as the resources would overlap. During PAD, the driver's main task is to vigilantly supervise the automated system and driving scene. As it is assumed that visual NDRTs would especially interfere with supervising, the current project focuses on visual tasks as well. In addition, drivers must regain vehicle control during take-over situations and resume steering for instance. Hence, visual NDRTs with a manual component are potentially problematic for PAD as well.

Moreover, the NHTSA guidelines prohibit certain visual tasks, known as *per se lock outs*, due to their distractive characteristics [14]. For instance, displaying photos or watching videos unrelated to the driving task, reading texts from books, the internet or social media as well as automatically scrolling texts or manually entering communication-based texts are prohibited during manual driving [14]. In addition, the guidelines propose that tasks should be interruptible at any time, completed within a maximum of 12 s total gaze time to the task and single gazes to the task should not last longer than 2 s [14]. Congruent with NHTSA, the current project incorporates visual-manual NDRTs. Even though NHTSA excludes the following from any examination, the current project focused on the effects of these *per se lock out* tasks on drivers' supervisory and take-over capabilities during PAD. The goal was to validate the new test protocol by using a broad range of guideline compliant to guideline non-compliant NDRTs. Regarding the latter group, a sensitive test protocol should yield strong effects concerning the drivers' supervisory and take-over capabilities.

Eventually, five NDRTs differing in guideline compliancy as well as similarity to everyday life/artificiality were chosen. Three of these tasks did not comply with the NHTSA guidelines, for instance due to presenting videos unrelated to the driving task. These three tasks included a browsing task, a video watching task and a text reading task, which were all similar to everyday life. The two tasks complying with the NHTSA guidelines included the artificial surrogate reference task (SuRT) [42] and the manual radio tuning task [34]. The latter task, a well-established reference task for manual driving, was designed to reduce the total gaze time of one trial to 20 s [34]. To match these specifications, the trials of the other tasks were designed to last no longer than 20 s as well.

During the browsing task, participants manually entered a departure point, a destination, two flight dates and the number of passengers. Participants received this information from the researchers. During the video task, participants viewed news video segments lasting 20 s and answered a question about the visual or general content of the video. The text reading task presented the participants with 70–100-character texts, which took approximately 20 s to read [43]. The participants had to scroll through the text to read its entirety. After finishing a text, the participants answered a question regarding its content. The SuRT task included finding a target (a bigger circle) amongst many distractors (smaller circles). During the manual radio tuning task, participants needed to set the radio to predefined frequencies.

### 2.2.2. Display Locations

In addition to executing NDRTs during manual driving and PAD, a related trend towards integrating increasing amounts of technology into vehicles has increased the potential of driver distraction and inattention during manual driving [44]. Another trend exists towards using increasing amounts of driving unrelated information [45] as well as smartphones during manual driving [46,47].

The main problem with different (in-vehicle) displays surrounds their proximity to the driving scene. Displays located further away from the windshield and driving scene are associated with enhanced reaction times [48]. For instance, head-up displays (HUD) were associated with significantly shorter reaction times as they are very close to the driving scene or may even overlay it. In contrast, display locations located further away from the driving scene were associated with shorter time to collisions [49]. Additionally, focusing on displays with less vertical proximity to the normal line of sight led to slower reactions than focusing on displays with equivalent horizontal proximity [49]. Moreover, several studies have found that the display location influenced drivers' gaze behavior during manual driving e.g., [45,48,50]. For instance, gazing away from the road

towards a head-down display (HDD) was occurring significantly less often than towards a HUD [50]. In addition, gaze durations during HUD interactions increased compared to HDD (e.g., the instrument cluster or head unit) interaction [45]. However, when focusing on the HUD, driving performance was improved (e.g., fewer lane deviations) because the driving scene was visible peripherally [45].

Hence for manual driving, several studies have shown clear differences in gaze behavior and driving performance depending on the display's proximity to the driving scene. Currently there are no comparable studies examining the effects of different display locations on drivers during (partially) automated driving. Further, no studies exist examining how different NDRTs during PAD affect drivers across display locations. Therefore, the goal of the second study was to incorporate this aspect in the method developed in the first simulator study. The (in-vehicle) display locations were chosen to reflect well-established displays (i.e., instrument cluster, head unit) as well as newer, more innovative technologies (i.e., HUD) and to reflect displays close to the driving scene (i.e., HUD) vs. further away (i.e., instrument cluster, head unit). Moreover, since smartphone usage during manual driving has increased e.g., [46,47], the smartphone was included as a handheld and forbidden display location. In addition, the following three displays were chosen: a head unit, an instrument cluster, and a HUD.

### 2.2.3. Driving Simulator Implementation

In the first simulator study, all five tasks were executed on a touch display. Therefore, the adapted manual radio tuning task for touch-displays [51] was used. This display was situated in the center console, at the same position as the head unit.

In the second simulator study, the display location was included as an additional independent variable since the test protocols' ability to differentiate their effects and the opportunity to easily manipulate these displays could be safely validated. Given that the goal was to ensure an economic study design and given that participants performed comparably during the browsing and text reading task, the browsing task was excluded. The remaining tasks ranged from slightly visually distracting (i.e., SuRT and manual radio tuning task) to highly distracting (i.e., text reading task). The video watching task was considered in the middle of this range. With exception of the video watching task, the tasks were not adapted from the first simulator study. The results of the first study led to the assumption that participants were listening to more than looking at the video segments. Therefore, the questions following each video segment were adapted to focus solely on the video's visual content to highlight its importance and enhance the comparability to other, more compelling videos (e.g., blockbuster videos).

The four chosen display locations were implemented as follows. For the head unit, a well-established HDD, the same 9-inch pre-installed display in the driving simulator's fully equipped vehicle mockup, was used as in the first study. For the instrument cluster, also a well-established HDD, a 9-inch display was installed behind the fairing of the vehicle mockup's built-in displays. The installed display thereby covered the tachometer but not the speedometer. Due to the fairing, parts of the 9-inch display were covered, thus the presented information (i.e., NDRTs) had to be downsized. Regarding the head-up display, a glass plate with mirror foil used to retrofit HUDs in vehicles was installed on the dashboard since the vehicle mockup was without a windshield. A 9-inch display was positioned under the glass plate with its presented information reflected onto the mirror foil. For the smartphone, a Huawei P9 with Android was used. Participants needed to hold the smartphone close to the gearstick, simulating the realistic attempt to hide phone usage during manual driving. Therefore, the smartphone condition was considered a part of the HDD category as well. During take-over situations, participants had to put down the smartphone before regaining vehicle control. For an overview of the NDRTs and (in-vehicle) display locations assessed within the three studies, see Table 2.

**Table 2.** Overview of the independent and dependent variables addressed and assessed within the three studies.

		Study 1—Simulator	Study 2—Simulator	Study 3—Test Track
Independent variables	Non-driving related tasks	<ul style="list-style-type: none"> <li>- Manual radio tuning task</li> <li>- Surrogate reference task</li> <li>- Video watching task</li> <li>- Text reading task</li> <li>- Browsing task</li> </ul>	<ul style="list-style-type: none"> <li>- Manual radio tuning task</li> <li>- Surrogate reference task</li> <li>- Video watching task</li> <li>- Text reading task</li> </ul>	<ul style="list-style-type: none"> <li>- Manual radio tuning task</li> <li>- Video watching task</li> <li>- Text reading task</li> </ul>
	(In-vehicle) Display locations	<ul style="list-style-type: none"> <li>- Head-unit</li> </ul>	<ul style="list-style-type: none"> <li>- Head-Up display</li> <li>- Head-unit</li> <li>- Instrument cluster</li> <li>- Smartphone</li> </ul>	<ul style="list-style-type: none"> <li>- Head-unit</li> </ul>
Dependent variables	Supervisory capabilities	<ul style="list-style-type: none"> <li>- Mean Gaze Duration</li> <li>- Total Gaze Duration</li> <li>- Maximum Gaze Duration</li> <li>- Number of Gazes</li> <li>- Number of Transitions</li> </ul>	<ul style="list-style-type: none"> <li>- Mean Gaze Duration</li> <li>- Total Gaze Duration</li> <li>- Number of Transitions</li> </ul>	<ul style="list-style-type: none"> <li>- Mean Gaze Duration</li> <li>- Total Gaze Duration</li> <li>- Number of Transitions</li> </ul>
	Take-over capabilities	<ul style="list-style-type: none"> <li>- Reaction time</li> <li>- Minimal distance to the lead vehicle at initial reaction</li> <li>- Maximal brake pressure</li> <li>- Maximal steering angle</li> <li>- Number of crashes</li> </ul>	<ul style="list-style-type: none"> <li>- Reaction time</li> <li>- Number of crashes</li> </ul>	<ul style="list-style-type: none"> <li>- Reaction time</li> <li>- Minimal distance to the lead vehicle at initial reaction</li> </ul>

#### 2.2.4. Test Track Implementation

Due to limited time resources, only a reduced NDRT selection was used and the display location aspect was excluded entirely. Additional reasons for excluding the latter included safety concerns for participants. The three NDRTs implemented in the study (i.e., manual radio tuning task, reading task, and video watching task) reflected a broad range of distractive potential (as determined during the two simulator studies) and a strong similarity to everyday life. The three tasks were executed on a tablet with touch control, attached at the head unit's position.

#### 2.3. Dependent Variables

The new test protocol was required to be sensitive regarding the effects of diverse NDRTs and display locations on drivers' supervisory and take-over capabilities. For that matter, parameters are necessary that sensitively capture and evaluate the capabilities (RQ2). To meet the requirement and answer the research questions, extensive examinations of different parameters for the supervisory and take-over capabilities were completed.

For manual driving, NHTSA recommends analyzing gaze behavior in terms of mean and total gaze duration towards the NDRTs [14]. Congruent with NHTSA, these parameters were included when analyzing supervisory behavior during PAD. However, since PAD differs from manual driving and the supervisory tasks increased in importance, further parameters were examined to achieve a comprehensive view. The assumption behind the additional parameters, including for instance the number of gazes or transitions between certain areas of interest (AOIs), is that these were assumed to be useful parameters to judge the drivers' compensatory behavior. For example, if long gazes to the NDRTs occur but are accompanied by many transitions between the NDRT and driving scene, the length of these gazes is somewhat compensated. Thereby, the driver will likely know more about current driving events and might react better to system failures compared to a driver executing long gazes to the NDRTs with few transitions. Furthermore, parameters reflecting and examining the

supervisory behavior during PAD based on the NHTSA guidelines and cut-off values (i.e., maximal 2 s per gaze towards the NDRT) were included, such as the maximum gaze duration towards the NDRTs. Hence, the following parameters for supervisory capabilities were of interest:

- the *mean gaze duration* towards the NDRT
- the *total gaze duration* towards the NDRT
- the *maximum gaze duration* towards the NDRT
- the *number of gazes* towards the NDRT
- the *number of transitions* between the driving scene and NDRT AOIs

Regarding the take-over capabilities, new parameters were proposed for the new test protocol given the automated system takes over the driving task during PAD. Firstly, reaction time indicated the criticality of the situation when the initial reaction occurred as well as the quality of the drivers' supervisory behavior. Longer reaction times would indicate reduced or insufficient supervision of the driving scene and system, probably due to the NDRT's greater distractive potential. Moreover, as reaction times increase, the criticality of the situation increases. For instance, the distance to the lead vehicle decreases each second, eventually making collision avoidance impossible. Four additional parameters were included to indicate situation criticality: the number of crashes, the minimal distance to the lead vehicle at initial reaction, the maximal brake pressure and maximal steering angle. For instance, more crashes or a small minimal distance to the lead vehicle would suggest a higher situation criticality. In addition, these variables were assumed to provide context to the reaction time and indications about potential compensatory behavior. For example, strong steering or braking responses might still prevent a collision even with a slow reaction time indicative of a critical situation. The parameters of interest were defined as follows:

- *Reaction time*—The time between the beginning of a take-over situation until the participants' initial reaction (braking or steering).
- *Number of crashes*—The number of collisions with guardrails (lateral) and the lead vehicle (longitudinal).
- *Minimal distance to the lead vehicle at initial reaction*—The distance between the two vehicles when participants initially reacted (braking). Applies only to the *lead vehicle deceleration* take-over situation.
- *Maximal brake pressure*—The highest administered brake pressure during the initial braking interval. Applies only to *lead vehicle deceleration*.
- *Maximal steering angle*—The greatest administered steering angle during the initial steering interval. Applies only to *ego-vehicle drifting*.

### 2.3.1. Driving Simulator and Test Track Implementation

The parameters were comparably assessed across the three studies. In the first study, all described parameters were assessed. However, based on the results of the first study some of the parameters were excluded from the following studies. The rationale behind this will be addressed in more detail within the result and discussion chapters. Table 2 gives a short overview of the parameters assessed within each of the three studies.

### 2.4. Equipment and Materials

In general, both study environments (driving simulator and actual vehicle) had to allow for scenario implementation (i.e., car-following) and independent variable examination. Hence, the following equipment was implemented within the three studies of the current project.

Firstly, at least two displays were required: One providing participants information about system states (i.e., instrument cluster) and one on which participants could execute the NDRTs (e.g., head unit). For the second study, the simulator had to allow to include further, controllable displays to examine the effects of display locations.

As discussed in Section 2.3, it was essential that the take-over capabilities and gaze behavior could be captured. Regarding the former, it was necessary to record participants' driving or take-over behavior. For this matter, the simulator software had to be programmed to record all relevant variables (see Section 2.3). The real vehicle had to contain data recording devices and the necessary sensors as well (e.g., LiDAR) to record data and compute relevant parameters.

Concerning the supervisory behavior, several methods to capture gaze behavior were employed of which the general (dis-)advantages will be discussed in the following section before describing the specific implementation within the three studies in the respective subchapters. Head-mounted eye trackers are a common tool to assess gaze behavior (e.g., Tobii Pro Glasses). Advantages of head-mounted eye trackers are, amongst others, the opportunity to analyze gaze data across different levels of detail (e.g., level of fixations or gazes). Moreover, AOIs are seen from the participants' perspective and their gazes are directly projected on to these AOIs. This allows for easy and reliable manual mapping of gazes towards AOIs, even for relatively small AOIs. In addition, the included eye tracking analysis software often provides the opportunity to automatically map raw gaze data on to relevant AOIs. However, it is still necessary to check the accuracy of the automatic mapping, and often manual remapping is required. An important disadvantage of head-mounted eye trackers is that most do not allow participants to wear glasses, thus these participants cannot take part in the study. This is especially problematic when examining older age groups as they are more likely to wear them. For instance, in 2014, 63.5% of all German citizens wore glasses and 92% of those older than 60 wore them compared to only 32% for those aged 20 to 29 and 38% for 30 to 44-year-old citizens [52]. Moreover, participants are highly aware of wearing these head-mounted eye-trackers and wearing them for prolonged times can be very uncomfortable.

Another method to assess supervisory behavior is using video annotations, whereby gaze behavior is annotated manually using multiple, synchronized videos facing the participants. This method is a non-invasive alternative to eye trackers since participants do not have to wear anything extra. This also allows participants wearing glasses to take part. Based on detailed annotation schemes, including descriptions of the AOIs that should be mapped and instructions on how to detect gazes to these AOIs, as well as the inclusion of training annotations with detailed feedback, it is possible to reliably annotate gaze behavior even across multiple researchers. However, in contrast to head-mounted eye tracking, where AOIs are seen from the participants' perspective, the videos are facing the participants. Therefore, only gaze *directions* towards a certain region representing AOIs (e.g., instrument cluster or street) can be annotated and differentiation between smaller AOIs closer together is difficult. Nevertheless, when relatively large AOIs (e.g., instrument cluster, head unit, mirrors or street) are of interest, this is less problematic. Another disadvantage is that video annotation does not allow the annotation of fixations. However, when focusing predominantly on gaze levels, as often done in this type of research e.g., [14], this disadvantage is less relevant.

In addition to the technical aspects, several formal aspects were necessary, such as information regarding the study, an informed consent and a data privacy statement. Furthermore, to standardize the information, participants received all instructions (e.g., concerning the partially automated system's activation and deactivation, the NDRT's system failures and execution) in written form.

To supplement the performance data, demographics such as participant age, gender and prior system experience were assessed. This allowed for an even distribution of gender and age as recommended by NHTSA [14], and this information could function as control variables during the analyses. Additionally, the participants' subjective experience regarding, for instance, the PAD or NDRT executions during PAD were assessed by questionnaires (e.g., Van-der-Laan-Acceptance-Scale, NASA TLX, and Trust in Automation). These subjective evaluations enriched the objective results or clarified aspects such as the participants' willingness to execute certain NDRTs during PAD before and after the study was completed.

To run the study as smoothly as possible, at least two researchers were deployed. One researcher focused on technical aspects (i.e., starting the simulator or driving the lead vehicle) and the other focused solely on supporting and supervising participants, including answering their questions or monitoring for simulator sickness.

### 2.4.1. Driving Simulator Implementation

To employ the developed test protocol in a simulated environment, the driving simulator used in the study included simulation software that presented the test track and scenario. The current project utilized a fixed-base driving simulator that consisted of a fully equipped mockup of the front of a vehicle (up to the B-pillar) with side- and rearview mirrors. Three connected screens presented a 180° horizontal field of view. In both studies, the driving simulator contained several cameras focused on the driving scene, the pedals and the driver (from two different angles). In the first study, drivers' gaze behavior was analyzed based on video annotations. The main reason for choosing this method was that no reliable eye tracker was available. Nevertheless, using non-invasive video annotations to examine gaze behavior enhanced participant comfort and allowed those with glasses to participate as well, thereby increasing the potential participant pool. This method could also reliably assess the relevant AOIs (e.g., the street, head unit or the instrument cluster). However, the second study incorporated the head-mounted Tobii Pro Glasses eye tracker [53] to record gaze behavior. Even though this method excluded glasses-wearing participants, it was very useful to assess more refined AOIs (e.g., handheld smartphone) and differentiate AOIs (e.g., differentiation within the instrument cluster between one part presenting system-related information and the other presenting the NDRT).

Moreover, during both studies, the instrument cluster presented the various states of the partially automated system (e.g., active, inactive, and deactivated) using a very minimal design. The second simulator study implemented a self-turning steering wheel inside the vehicle mockup. This reflected the actual PAD experience more closely since an actual vehicle's steering wheel moves during curved segments as well. The slight steering wheel movement during the *ego vehicle drifting* take-over situation could, however, lead to faster recognition of the situation and hence, faster reactions. Lastly, two researchers were present for both studies, one focusing on technical aspects and the other on participants. Participants received written instructions to enhance standardization and also received several questionnaires.

### 2.4.2. Test Track Implementation

To reproduce the test protocol in a real driving environment, it was necessary to include two vehicles: (a) A partially automated ego vehicle that was programmable to deliberately trigger system failures and enable the capturing of driving and take-over data, and (b) a lead vehicle with (if possible) advanced driving assistance systems (ADAS) such as cruise control (see Figure 1 (bottom row, right) in Section 2.1.2).

The lead vehicle's cruise control started at 25 km/h and was driven manually by a researcher. To ensure maximum comparability, the researchers received detailed instructions and absolved several training runs. With exception of the *lead vehicle deceleration* take-over situation, the researcher always drove the vehicle in second gear with the motor executing all necessary accelerations or decelerations to maintain a constant speed as much as possible (approximately 15 km/h).

A second, programmable vehicle served as the ego vehicle, equipped with various measurement technologies (e.g., Denso LiDAR, Novatel DGPS). The partial automation was achieved through combining genuine vehicle automation and Wizard-of-Oz techniques. The automation controlled longitudinal movement, vehicle speed (max. 27 km/h), and held constant the distance to the lead vehicle relative to speed except during the *lead vehicle deceleration* take-over situation. The distance was set to half of the ego vehicle's speed with an additional buffer of seven meters ( $\text{ego vehicle speed}/2 + 7 \text{ m}$ ). The Wizard-of-Oz techniques controlled lateral movement. The researcher in charge of programing the partial automation sat in the passenger's seat during the entire study to secure safety and execute lateral control (steering) using a small steering wheel unseen by participants (see Figure 1 (top row, right) in Section 2.1.2). Although participants did not notice the researcher steering the vehicle, this led to a major disadvantage in that the steering movements and lane keeping were not completely identical during each drive.

As in the simulator studies, the ego vehicle contained several cameras focused on the participants (from two angles), driving scenario, and vehicle interior. The recordings of the



participants were used to analyze drivers' gaze behavior towards the AOIs (e.g., the street, head unit, and instrument cluster) based on video annotations.

Three researchers were present in this study: one focused on technical aspects and ego vehicle steering, one focused solely on driving the lead vehicle and one focused on supervising and supporting participants in between trials. Again, participants received written instructions and questionnaires with the discussed contents (See Section 2.4).

## 2.5. Experimental Design and Procedure

A within-subjects design was used to test the NDRTs, take-over situations, and other independent variables such as the different display locations. This approach allowed to reduce the number of participants necessary for high statistical power by directly comparing each participant to themselves and excluding any influences from interindividual confounding variables.

The participants experienced both take-over situations, all NDRTs and a reference trial without NDRT execution. During trials with NDRT execution, participants needed to continuously execute the task whenever the partial automated system was active and to only cease task execution during take-over situations. In the reference trial, participants drove partially automated and experienced both take-over situations. The trials with and without NDRTs were randomized and counterbalanced to reduce order effects.

Each trial included four take-over situations and started with a short familiarization segment (see Figure 2). The four take-over situations within each trial were to avoid predictability and, hence, a change in gaze behavior. With only two take-over situations, participants could easily predict the second take-over situation after experiencing the first one. It would also be problematic having three take-over situations if the first two were the same, the third would have been easily predictable. With four or more take-over situations within one trial, it was possible to make the order of the situations unpredictable. The two take-over situations were sequentially counterbalanced across the four occurrences to reduce predictability as well as order and learning effects. The first and third occurrences always included the two take-over situations: *lead vehicle deceleration* and *ego vehicle drifting*. The order of the two situations was alternated. For the other two occurrences, the two take-over situations were randomly assigned to reduce predictability. However, identical situations would not follow each other more than twice.

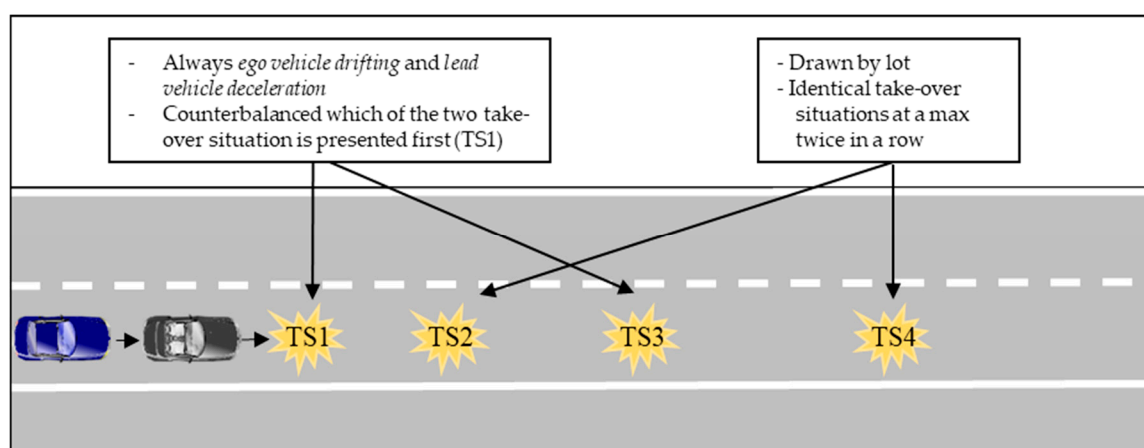


Figure 2. Overview of the four take-over situations presented to the participants per trial.

Participants began the studies receiving information regarding the goal of the study and an informed consent. They experienced manual and partially automated familiarization drives to get accustomed to the driving simulator or ego vehicle and partial automated system. Participants received written instructions for the partially automated system, the take-over situations and task priorities (i.e., giving the safe driving task the highest priority). Moreover, before each trial, the corresponding NDRT was introduced and explained to the participants and, if applicable, each display location as well. Congruent with NHTSA's guidelines, participants practiced the tasks during

vehicle standstill to ensure a comparable level of understanding before the trial with data recording started.

### 2.5.1. Driving Simulator Implementation

In the first simulator study, the NDRTs were included as within-subjects factors, resulting in every participant executing all five NDRTs. In addition, all participants experienced the reference trial, leading to a total of six trials. The second study contained only four of the five NDRTs (see Table 2). The reference trial was excluded to ensure an economic study design. For the same reason, participants only experienced two out of four display locations. Three of the four NDRTs (i.e., the SuRT, text reading, and video watching tasks) were executed on these display locations. The manual radio tuning task was executed on an additional display representing the typical location for in-vehicle sound systems/radios. This resulted in seven trials per participant. The procedure is described in Section 2.5. Study participation took 2–2.5 h.

### 2.5.2. Test Track Implementation

In the test track study, three of the five NDRTs (see Table 2) were included as within-subjects factors. In addition, all participants experienced a reference trial, resulting in four trials total. The procedure was identical to the simulator studies (see Section 2.5) with one small exception: an additional manual familiarization drive used to accustom participants to the test track. This drive did not include take-over situations. The study lasted approximately two hours for each participant.

## 2.6. Data Preparation and Analysis

Regarding the supervisory behavior, either the data from the head-mounted eye tracker or the video annotations were used for further analysis. The data were prepared and analyzed using the specification of a gaze towards an AOI, following the ISO standard (EN ISO 15007-1) [54] definition of glance duration. This can be defined as the time from when a gaze initially moved towards an AOI to when it moved away towards another AOI, which would include all consecutive fixations towards this AOI during that time. This includes all saccades occurring within this time as well [54]. The three studies focused on gazes towards the following AOIs:

1. Driving scene—Gazes through the windshield, directed towards the driving scene
2. NDRT—Gazes inside the vehicle, towards where NDRTs were executed (i.e., towards the head unit in the first and third study or to different locations in the second study)
3. Instrument cluster + steering wheel—Gazes inside the vehicle, towards the instrument cluster and steering wheel
4. Vehicle interior—Gazes inside the vehicle that were not directed to the NDRT or other relevant locations (e.g., gazes to the researcher in the passengers' seat during the third study)

To analyze participants' supervisory capability during PAD, a predefined segment prior to a system failure and take-over situation occurring was examined (i.e., several seconds before the take-over situation occurred).

Data recorded by the simulator or vehicle were gathered for further analyses of the take-over capabilities.

Across the three studies, only the first and third take-over situations including both take-over situation types were further analyzed for each trial. This was done to reduce predictability of the upcoming situation and to ensure a consistent number of analyzed events for each participant and trial. As the first take-over situation was chosen randomly, it was completely unpredictable and thus participants' gaze behavior was assumed to be as natural as possible (i.e., checking for both possibilities, lane deviations or reduced distance to the lead vehicle). The second take-over situation was chosen randomly as well. Hence, the second situation could be the same take-over situation as the first one or it could be the other one. It was believed that participants would likely expect the other take-over situation they had not experienced to occur and hence adjust their gaze behavior accordingly (e.g., just checking for lane deviations). The third situation was always the take-over

situation participants had not experienced in the beginning. It was thought that participants' will once again scan for both possible take-over situations (i.e., show natural gaze behavior) after realizing that there is no systematic presentation of the take-over situations (e.g., in an alternating manner). The fourth situation was again chosen randomly.

For take-over capabilities, the performance from when a take-over situation begins to when the participants' make their initial response was analyzed. If participants did not react, their performance from when the take-over situation commenced until a crash or an intervention of the vehicle or researcher occurred was analyzed. The take-over and supervisory capabilities during the NDRT execution trials were compared to each other, to the reference trial without NDRT and, additionally, to the manual radio tuning task.

### 2.6.1. Driving Simulator Implementation

In both simulator studies, a 1-km segment equal to a duration of approximately 45 s prior to the beginning of the first and third take-over situation for each trial was used to analyze the supervisory capabilities based on the relevant parameters (see Section 2.3). It was assumed that the gaze behavior prior to the take-over situations did not differ depending on the following situation due to the study's design implemented to reduce predictability (see Section 2.6). As the data supported this assumption, the supervisory behavior prior to the situations was averaged across both events. The take-over capability was analyzed from the start of the take-over situation until participants' initial reaction or a collision (see Section 2.6).

To examine the test protocol's ability of differentiating the effects across different NDRTs and display locations, as well as whether these differences are as expected (RQ1), repeated measures ANOVAs (rmANOVAs) were used that are highly robust, even with slight deviations from the assumption of normality [55].

### 2.6.2. Test Track Implementation

In the test track study, a 10-s segment prior to the beginning of a system failure and take-over situation was examined to determine participants' supervisory capabilities. The segment length was chosen to ensure that the previous take-over situation would not interfere with the analyzed segment. Hence, this required the previous take-over situation to be completed and the partially automated system to be active again so that the 10-s segment included only actual supervisory behavior during PAD. This was necessary because the two take-over situations could have occurred within one round with relatively little time in between.

The take-over capabilities were analyzed identically to those in the simulator studies with one exception: The analyzed interval would end with the researcher's intervention in the ego vehicle or with the ego vehicle itself in case participants did not react.

Due to the within-subjects design, robust repeated measures ANOVAs (rmANOVAs) were computed.

## 2.7. Participants

Following NHTSA's methodology, the goal of the three studies was to balance participants' gender and age across the four age groups described in the methodology [14]. This was to achieve a heterogeneous participant group, which allows for controlling and assessing any gender or age effects. Table 3 shows the distribution across age groups, gender, and the total number of participants in the studies and the total number participants actually analyzed.

**Table 3.** Number of participants, age, and gender distributions of participants per study.

Studies	N (Total)	N (Analyzed) <sup>1</sup>	Age Groups				Gender Distribution
			18–24	25–39	40–54	Older than 55	
Study 1—Simulator	57	47	12	15	12	8	27 male, 20 female
Study 2—Simulator	58	50	15	11	13	11	27 male,

							23 female
Study 3—Test track	39	36	8	11	9	8	20 male, 16 female

<sup>1</sup> Exclusion of several participants due to, e.g., simulator sickness or technical problems, resulted in the reduced number of analyzed cases.

### 3. Results

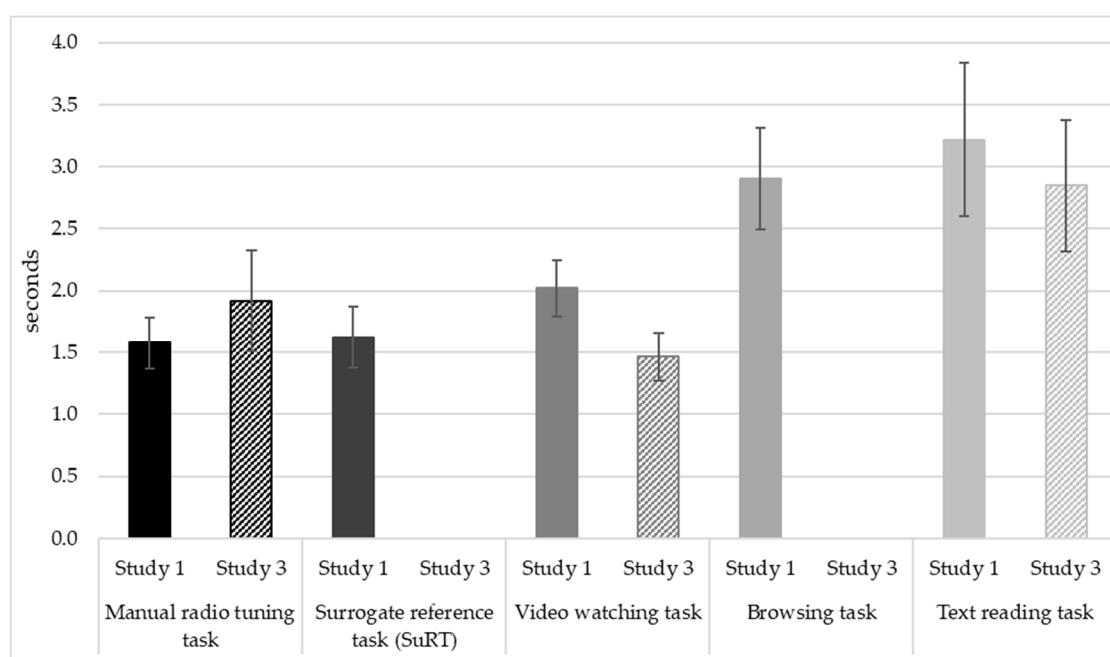
This section shortly covers the results of the formulated research questions that examine the sensitivity of the test protocol and necessary parameters to sensitively evaluate participant take-over and supervisory capabilities. The current article focused on the examination of the proposed test protocol's suitability to sensitively evaluate the effects of NDRTs on the drivers during PAD and not on the NDRTs' particular effects on the drivers' take-over and supervisory capabilities. For a better understanding, an exemplary description of the supervisory capabilities (i.e., mean gaze duration) will be presented. The specific results of the supervisory and take-over capabilities across the NDRTs and display locations within the three studies will be described in more detail in separate papers e.g., [56].

Regarding the supervisory capabilities, several parameters were analyzed (see Section 2.3). Table 4 shows the effect sizes for the main effects of NDRTs and display locations across these parameters. Following the convention of Cohen [57], the effect sizes were categorized into weak ( $\eta^2 < 0.06$ ), medium ( $\eta^2$  between 0.06 and 0.14), and strong effects ( $\eta^2 > 0.14$ ). All three studies revealed predominantly strong effects for the assessed parameters regarding the NDRTs. The results also corroborated the expectations. Figure 3 exemplary shows the effects of NDRT execution on the drivers' supervisory capabilities in terms of mean gaze duration towards the executed NDRT (in seconds) for the first and third study. In line with the expectations, the non-compliant browsing and text reading task resulted in considerably longer mean gaze durations towards the NDRTs than the manual radio tuning task. The guideline conform SuRT resulted in comparably long mean gaze durations towards the task as the reference task (i.e., manual radio tuning task). Less expectedly, the mean gaze duration towards the video watching task in the first study was only slightly longer and in the third study even shorter compared to the manual radio tuning task. The main effect of the display locations yielded strong effects as well. Again, the results were congruent with expectations. For instance, executing NDRTs with the smartphone was more captivating and resulted in less supervision of the driving scene than execution on the instrument cluster. Hence, with regard to the supervisory capabilities, the strong effects for both independent variables indicated the test protocols' ability to sensitively differentiate between NDRTs with different visually distractive potentials (RQ1a) as well as to sensitively detect differences between various (in-vehicle) display locations (RQ1b). Moreover, with exception of the video watching task, the results were in line with the expectations based on the literature (RQ1).

**Table 4.** Effects of the parameters used to examine supervisory capabilities across the three studies.

	Study 1—Simulator	Study 2—Simulator		Study 3—Test Track
	Non-Driving Related Tasks (NDRTs)	NDRTs	Display Locations	NDRTs
Mean Gaze Duration	Strong effect <sup>1</sup>	Medium effect	Strong effect	Strong effect
Total Gaze Duration	Strong effect	Strong effect	Strong effect	Strong effect
Maximum Gaze Duration	Strong effect	/	/	/
Number of Gazes	Strong effect	/	/	/
Number of Transitions	Strong effect	Strong effect	Strong effect	Strong effect

<sup>1</sup> Following Cohen [57], weak effects are defined as  $\eta^2 = 0.00$ – $0.06$ , medium effects as  $\eta^2 = 0.06$ – $0.14$  and strong effects as  $\eta^2 > 0.14$ .



**Figure 3.** Mean gaze duration towards the NDRTs (in seconds) depending on the executed non-driving related task for study 1 in the simulator and study 3 on the test track.

Concerning the take-over capabilities, several parameters were analyzed. Table 5 shows effect sizes for the main effects of NDRTs and display locations. The first simulator study revealed predominantly strong effects regarding the main effects of NDRTs concerning reaction time. In the other two studies, NDRT effects regarding reaction time ranged from weak to strong depending on the particular display location on which the tasks were executed (or weak in the case of the third study). The results were also congruent with expectations. For instance, more distractive NDRTs (e.g., browsing task) resulted in impaired take-over capabilities including longer reaction times. Medium sized effects existed for the reaction time dependent on the display location. For example, the smartphone resulted in longer reaction times as was expected, indicating the test protocol's ability to sensitively detect these differences. The effect sizes regarding the other parameter of the first study were predominantly strong and weak regarding the third study. Hence, specific parameters of the test protocol (e.g., reaction time) were sensitive to NDRT effects with varying visually distractive potentials (RQ1a) and to some extent sensitive to display location effects (RQ1b). Further, the results were in line with the expectations based on the literature (RQ1).

In sum, regarding the first research question (RQ1) and the sub questions (RQ1a and RQ1b), the results showed that the test protocol was sensitive to the effects of different NDRTs and (in-vehicle) display locations. Especially, the supervisory capabilities were proven very sensitive to these effects.

**Table 5.** Effects of the parameters used to examine take-over capabilities across the three studies.

	Study 1—Simulator	Study 2—Simulator		Study 3—Test Track
	NDRTs	NDRTs	Display Locations	NDRTs
Reaction time	Strong effect <sup>1</sup>	Weak—strong effect	Medium effect	Weak effect
Minimal distance to the lead vehicle at initial reaction	Strong effect	/	/	Weak effect
Maximal brake pressure	Strong effect	/	/	/
Maximal steering angle	Weak effect	/	/	/
Number of crashes	Only descriptive analyses.			/

<sup>1</sup> Following Cohen [57], weak effects are defined as  $\eta^2 < 0.06$ , medium effects as  $\eta^2 = 0.06\text{--}0.14$  and strong effects as  $\eta^2 > 0.14$ .

As mentioned earlier (see Section 2.3.1), all parameters described in Section 2.3 were assessed within the first study and all parameters concerning the supervisory capabilities yielded strong effects. However, the mean gaze duration and the maximum gaze duration were very similar in terms of their effect strengths (see Table 4) as well as in terms of the particular results of the effect of NDRT execution. More precisely, the maximum gaze duration presented very similar findings as the mean gaze duration as presented in Figure 3: visually more distracting tasks (i.e., the browsing and the text reading task) resulted in considerably higher mean and higher maximum gaze durations than visually less distracting tasks (i.e., manual radio tuning task and SuRT). Further, in contrast to the number of gazes towards one AOI, the number of transitions is more useful as it combines the information from the number of gazes towards two AOIs and is a good indicator of drivers' compensatory behavior. Hence, the maximum gaze duration and the number of gazes were not analyzed in Study 2 and 3.

Regarding the take-over capabilities, the effects found within the first study were predominantly strong as well. However, the minimal distance to the lead vehicle at initial reaction, the maximal brake pressure, and the maximal steering angle are logically connected with the reaction time. For instance, longer reaction times logically result in a reduced minimal distance towards the lead vehicle, hence, demanding stronger initial reactions (e.g., higher brake pressure). Therefore, these variables were not assessed in Study 2 and 3. The number of crashes, which was analyzed descriptively, was a useful addition to the reaction time.

Finally, a selection of the most useful parameters was chosen for the following studies based on the first study, including: mean gaze duration, total gaze duration, number of transitions between the driving scene and NDRT, reaction time and number of crashes (RQ2). However, the number of crashes, which was analyzed descriptively, was a useful addition to the reaction time for the second study, but could not be assessed within the third study to ensure participants' safety. In order to evaluate the test protocol in light of the changes made for the test track environment, the minimal distance to the lead vehicle at initial reaction was assessed within the third study again, but yielded only weak effects.

#### 4. Discussion

This project's overarching goal was to take the initial steps towards developing a test protocol that systematically evaluates drivers' supervisory and take-over capabilities during PAD. The research questions addressed the test protocol's ability to sensitively detect differences (as expected based on the literature) in drivers' supervisory and take-over capabilities during PAD across different NDRTs (RQ1a) and display locations (RQ1b). Moreover, it was examined which parameters are sufficient to sensitively capture and evaluate drivers' take-over and supervisory capabilities (RQ2).

The three studies revealed mixed results concerning the test protocol's sensitivity to detect the effects of visual-manual NDRTs (RQ1a) and (in-vehicle) display locations (RQ1b) on drivers' supervisory and take-over capabilities during PAD. Regarding the supervisory capabilities, predominantly strong effects existed for most of the analyzed gaze parameters. This firmly indicates the test protocol's ability to sensitively detect differences in the drivers' supervisory behavior based on the executed visual-manual NDRT as well as display location (on which a NDRT is executed) (RQ1). As described in Section 3, the mean gaze duration, total gaze duration, and the number of transitions were deemed as the most useful parameters that sufficiently examine supervisory capabilities during PAD, since they yielded strong effects (RQ2). Additionally, these parameters still provide the required data to compute other parameters. For instance, the total gaze duration adds together all single gaze durations, from which the maximum gaze duration can be extracted. Moreover, with exception of the video watching task, the detected differences were congruent with the expectations as was shown exemplary for the mean gaze duration towards the NDRTs. For instance, more distractive tasks (e.g., browsing task) resulted in poorer supervision compared to the manual radio tuning task. In contrast, the video watching task appeared to be less visually distracting than expected, in terms of only slightly poorer supervision than during manual radio tuning task. However, the news video segments had low visual attraction and the content was predominantly

presented aurally rather than visually. Other videos with greater visual attraction (e.g., blockbuster videos) might be more distractive, resulting in longer gazes that might influence drivers' supervisory and take-over capabilities more negatively. Nevertheless, the results of the supervisory capabilities, based on the examined parameters, can sufficiently answer the first research question and corresponding sub questions. However, the findings concerning take-over capabilities were less clear, especially in the third study. Even though reaction time yielded the strongest effect sizes for the differentiation between the NDRTs across the simulator studies and, therefore, seemed to be the best indicator of drivers' take-over capabilities and situation criticality (RQ2), this was not replicated in the closed test track study. However, the weaker effects were likely due to the changes and adaptations made to the test protocol for applicability to the test track scenario's limited space. Especially, having the take-over situations always being executed on the same track segment greatly increased the predictability of the take-over situations compared to the simulated environment. After the first trial, participants knew where the take-over situations would occur and were then likely more attentive during these track segments in the following trials. This likely resulted in weaker effects for NDRT differentiation.

Generally, the vigilant supervision of the driving scene and system enables the drivers to notice system failures in a timely fashion and prepares them to make any necessary and timely intervention if such a case arises [2,18]. Hence, despite the partially weaker effects regarding take-over capabilities, the supervisory capabilities are strongly related to the former. Therefore, the strong effects concerning the supervisory capabilities are promising and indicate that the test protocol is useful to differentiate between the effects of different visual-manual NDRTs on drivers during PAD (RQ1). Nevertheless, it is still necessary to examine the NDRTs' effects on parameters indicative of the situation's criticality and the drivers' take-over capability, such as reaction time and the number of crashes (RQ2). Both are relevant supplements to the supervisory parameters, when drawing conclusions about NDRTs' influence on drivers during PAD.

In general, the new test protocol should form the basis to assess how different NDRTs influence drivers' supervisory and take-over capabilities during PAD and, hence, to decide whether certain NDRTs are suitable for execution during PAD. Currently, conclusions can only be drawn based on the three studies conducted for this project. Nevertheless, based on comparing the tested NDRTs versus the manual radio tuning task across the three studies, some NDRTs seem less suitable than other tasks. For instance, the browsing and text reading tasks distracted drivers considerably more in terms of longer gazes towards the NDRTs and poorer take-over capability than seen with the manual radio tuning task. Following the NHTSA guidelines, stating that a task is not appropriate (for manual driving) when visual and driving performance are poorer than the manual radio tuning task, the browsing and text reading tasks would not be suitable for PAD. In contrast, the video watching task and SuRT showed similar results to the manual radio tuning task regarding drivers' supervisory behavior and take-over capabilities. Hence, the SuRT and video watching task might be rendered appropriate for PAD. However, final conclusions, especially regarding the suitability of the video watching task, should not yet be drawn. Moreover, conclusions regarding NDRT suitability during PAD should be handled cautiously since the test protocol is not yet broadly established.

#### *4.1. Future Research*

To draw conclusions concerning whether a NDRT is suitable for execution during PAD, some further steps are necessary. Firstly, further studies conducted in different environments using the developed test protocol are necessary to establish cut-off values for PAD comparable to those provided by the NHTSA for manual driving [14]. Secondly, the manual radio tuning task [34] needs to be evaluated regarding its suitability as a still reasonable reference task for PAD. Since drivers are relieved from parts of the driving task, other potentially more distractive tasks might possibly be executed during PAD without negative consequences compared to manual driving. If this is the case, the manual radio tuning task, which is perfectly congruent with the cut-off values for NDRT execution during manual driving (2 s per gaze, 12 s total gaze duration towards the NDRT), might be too conservative for PAD. Hence, if the new PAD cut-off values differ from those of manual



driving in terms of longer gazes towards the NDRTs being allowed, the manual radio tuning task might render more NDRTs unsuitable due to being too conservative. Additionally, participants in these three studies were presented with the partially automated system and secondary tasks for only short periods. The effects of prolonged PAD periods should be examined to better understand the willingness and likelihood of NDRT execution during PAD as well as the development of supervisory behavior with increasing system experience.

For these further studies, the following sections include detailed recommendations regarding test protocol usage in both driving environments.

#### 4.2. Recommendations Regarding Test Protocol Implementation

When using the developed test protocol for studies evaluating NDRT effects during PAD in a simulated or real driving environment, we, the authors, would like to provide the following recommendations. These are mainly based on the results and experiences we gathered during the three studies conducted for this project. In addition, further literature enriching these recommendations focused on standardized NDRT evaluation for manual driving e.g., [14] or higher automation levels for instance (i.e., SAE level 3) [31]. In the end, a table is provided giving an overview of the recommendations.

##### 4.2.1. Driving Scenario and Take-Over Situations

The current project employed NHTSA's [14] well-established car-following scenario [14] and extended it to PAD. Given this scenario is implementable in a simulated or real driving environment (i.e., closed test track), we recommend its usage with the necessary extensions (i.e., take-over situations) for further PAD studies. Depending on the particular driving environment, certain adaptations might be necessary.

The recommended scenario extensions include take-over situations considered necessary to examine participants' take-over and supervisory capabilities during PAD. We suggest implementing at least two types of take-over situations addressing system limitations of lateral and longitudinal vehicle control, such as the two take-over situations (*lead vehicle deceleration* and *ego vehicle drifting*) used in the current project. Other take-over situations that realistically address limitations (e.g., losing lateral control due to a curve in the road, missing lane markings, or failing to detect a road obstacle) of the partially automated system can be implemented as well. Independent of the situation type, we advise excluding any warnings or take-over requests to realistically simulate PAD (SAE level 2) as well as any external cues (e.g., trees, houses or brake lights) to reduce predictability of the take-over situations.

The driving simulator scenario can be implemented nearly identically to the NHTSA [14] specifications (see Section 2.1). We highly recommend using predominantly straight road segments for identical implementation of the two take-over situations used in the current studies. If other take-over situations are used, the test track can include curved segments as well. However, these increase the chances of simulator sickness occurring and therefore should be implemented cautiously. Corroborating NHTSA's guidelines [14], we recommend incorporating multiple lanes (i.e., two lanes in each direction) as well, especially with take-over situations addressing lateral vehicle control. Additionally, a beginning segment without take-over situations is advisable to allow participants to start the scenario, activate the partially automated system and execute the NDRTs without time pressure.

As with NHTSA [14], we used a speed of 80 km/h and a distance of 70 m to the lead vehicle in the simulator. A seven second TTC was implemented for the two take-over situations. To enhance situation criticality and scenario validity, researchers can change the speed and distance specifications or use the lead vehicle's variable speed profile [14]. However, the latter can complicate detection of system failures. For greater situation compatibility, the adaptations should result in matching TTCs.

Even though NHTSA's guidelines [14] allow for sparse (oncoming) traffic, we excluded all traffic except for the lead vehicle to reduce potential distractions (especially during the reference trial) and

to implement the take-over situations as described. For instance, when implementing the *ego-vehicle drifting* take-over situation, we recommend excluding other traffic during that interval to prohibit any traffic collisions. Other (oncoming) traffic can be included for a more realistic driving scenario or a higher situation criticality.

For real-world driving studies (e.g., closed test track) we recommend implementing the same driving scenario. Therefore, a test track allowing the application of scenario and take-over situations with similar speed or TTC specifications is highly recommended. For the current test protocol, we suggest using a straight track to implement both take-over situations as described. This also ensures that the *ego vehicle drifting* take-over situation is not mistaken for driving around a curve and that driving around a curve is not mistaken for a take-over situation itself. Provided other take-over situations are chosen, curved segments may be necessary.

The test track length depends on the number and timing of the take-over situations. Based on the simulator studies, when driving 80 km/h and implementing four take-over situations, we recommend using an 11-km test track. This allowed an analysis of a 45-s interval, equal to a 1-km route segment, prior to each take-over situation. However, combining four take-over situations on an 11-km test track results in a relatively high frequency of system failures, which might reduce external validity (see Subsection *Experimental Design* in Section 4.2.5). Hence, using an even longer test track is recommended to increase the time and distance between take-over situations to create a more realistic experience for the participants.

If such a test track is not available, adaptations become inevitable. If speed reductions are necessary, the TTC should be reduced in relation to the speed. When using a similar test track as in the current project, it is important to reduce predictability of the take-over situations in terms of time and location as much as possible, as this can strongly influence participants' supervisory and take-over behavior.

Several adaptations should always be made independent of the test track. Firstly, the take-over situations cannot result in a collision with, for instance, the lead vehicle or guardrails if the participants do not react. For this matter, fallback solutions, as described in Section 2.1.2 (i.e., programming of ego vehicle) are necessary to ensure participants' and involved researchers' safety at all times. For the same reason, additional traffic should be excluded as well or, at least, be controlled and reduced to a minimum.

#### 4.2.2. Independent Variables

The test protocol was able to discover and distinguish expected differences between different visual-manual NDRTs. This allowed evaluating guideline compliant and non-compliant tasks as well as artificial tasks and those closer to everyday life. The number of tasks that can be examined is flexible; however, it is recommended to strive for an economic study design. Moreover, we recommend comparing the effects of a partially automated drive with and without NDRT execution or comparing a partially automated drive with NDRT execution to a drive while executing a reference task (e.g., manual radio tuning task [34]). Regarding the manual radio tuning task, we recommend using the version adapted for touch displays [51] to ensure comparable task execution. As in Schömig et al. [31] and NHTSA [14], we recommend predefining the start and finish of task execution when examining distractive effects on the drivers instead of spontaneous task execution. Moreover, participants should practice the tasks to achieve comparable task understanding before each trial see [14].

Furthermore, the current project showed the test protocol's ability to distinguish between the effects of NDRT execution on different display locations. Depending on the research question, different display locations of interest can be included. In the current studies, it was not always possible to use the built-in display locations to present the NDRTs to participants. Even though we attempted to present these NDRTs in similar positions as these built-in display locations occupy and use comparable control elements for execution (e.g., touch displays), using external displays might have reduced the realism of NDRT execution during PAD. It is recommended to use available, built-

in displays as much as possible (which should be controlled in a similar manner) to strive for an economically designed study.

Moreover, it seems reasonable to validate the test protocol considering other independent variables that are meaningful for PAD (e.g., prior system experience or different HMI designs).

#### 4.2.3. Analyzed Variables

As previously discussed, several different parameters can be analyzed to evaluate drivers' supervisory capabilities and all parameters that were evaluated, provided strong effects. However, to ensure an economic study design, we suggest using mean and total gaze duration towards the NDRT (and driving scene) and the number of transitions between the driving scene and NDRT as discussed in Section 3. These three parameters can sensitively examine and reflect the supervisory capabilities and compensatory behavior during NDRT execution.

The results showed that take-over capabilities yielded weaker effects than the supervisory capabilities. Nevertheless, take-over capabilities must still be assessed and therefore different parameters can be analyzed. We recommend using reaction time to measure situation criticality, which should be enriched by the number of crashes or lane deviations for example. Other parameters can be used as well (e.g., TTC), but these parameters should be chosen based on their ability to provide additional and valuable information.

#### 4.2.4. Equipment and Materials

##### Driving Simulator and Test Vehicles

Depending on the study environment, either a driving simulator with a vehicle mock-up and corresponding simulation software or two vehicles (an ego and lead vehicle) are necessary to implement the driving scenario and take-over situations.

For both vehicle mock-up and actual ego vehicle, it is recommended that at least two (in-vehicle displays) are available, including the instrument cluster presenting (automated) system-related information and another display for NDRT execution (e.g., the head unit). The displays must be customizable for study relevant information and the participants must be able to smoothly interact with the display during NDRT execution. It is also suggested to equip the mock-up and actual vehicle with cameras facing participants, the driving scene, and the task to record study relevant behavior. Moreover, any driving input made by participants must be reflected by the simulator or ego vehicle and the corresponding partially automated system in a timely fashion to ensure a realistic system experience. This input includes braking, steering or system (de-)activation by pushing the corresponding buttons on the steering wheel for instance.

It might be useful to incorporate a self-turning steering wheel in the driving simulator to represent a more realistic PAD experience. However, this could cause participants to recognize the ego-vehicle's drifting faster than if there were no movement (especially when driving on a straight road). Moreover, in real driving environments, PAD includes hands-on warnings requiring drivers to leave one hand on the steering wheel at all times. In the current project, participants needed to remove their hands from the steering wheel to mimic an extreme situation. Both aspects must be considered based on the relevant research questions.

For the test track vehicles, we strongly recommend using high automation levels to ensure standardized and replicable driving scenario execution and take-over situations, as well as to reduce chances for human error. At a minimum, the ego vehicle should take over tasks controlled by the partially automated system and should be programmed to deliberately trigger the two take-over situations. If higher automation levels are not possible, Wizard-of-Oz approaches are reasonable alternatives; however, these reduce comparability. The ego vehicle must include sensors (e.g., LiDAR or Novatel DGPS) and devices to record driving data. The lead vehicle should at least include ADAS (i.e., cruise control), the drivers should be extensively trained on their tasks, and landmarks should exist for comparable execution of take-over situations. Additionally, it is recommended to

synchronize the vehicles. This could include using walkie-talkies; however, programmed synchronization would be preferable for standardization and replication.

### Human–Machine Interface

As mentioned by Schömig et al. [31] for SAE level 3 automation, the human–machine interface (HMI) should present participants with all relevant system states (e.g., active or inactive) and corresponding transitions between these states. The instrument cluster would be the most suitable since it presents drivers with further driving related information (e.g., speed). Additionally, the HMI must reflect participants' input (system activation and deactivation) in a timely fashion. When the goal is focusing on the effects of different visual-manual NDRTs, as with the current test protocol, we recommend using a minimal, intuitively understandable HMI that does not distract drivers from NDRT execution or cause mode confusion.

### Eye Tracking

Driver's gaze behavior must be recorded to evaluate their supervisory capabilities. Depending on the detail level (e.g., AOIs, fixations) examined, study design (i.e., study length and environment), or test sample of interest (e.g., younger vs. older participants), the researcher must decide between using a head-mounted eye tracking system or video annotations (see Section 2.4 for a more detailed discussion of the (dis)advantages of both methods).

### Questionnaires

At the least, we highly suggest collecting participants' demographic information (e.g., age, gender, and prior system experiences). In addition, further questionnaires administered before and after trials with and without NDRT execution would supplement the objective data with subjective experiences, which would help shed light on possible explanations for their past or potentially future behavior such as willingness to execute NDRTs during PAD.

### Instructions

As with NHTSA [14] and Schömig et al. [31], we recommend using written instructions regarding the following aspects to enhance standardization. Firstly, the NDRT execution should be clearly communicated, including the NDRT's goal, what constitutes successful execution, and when NDRT should be executed. When examining NDRT's distractive effects during PAD, we suggest instructing participants to continuously execute the NDRTs when the partially automated system is active and the situations allow it based on the participant's judgment, which corroborates Schömig et al. [31] and NHTSA's recommendations [14]. The instructions should also explain participants' task priorities, such as the safe execution of the driving task has the highest priority. Secondly, to ensure comparable system understanding, the partially automated system's usage and states should also be explained to participants. Only when researchers are interested in intuitive system interaction should these instructions be excluded e.g., [31]. In addition, similar to a partially automated vehicle manual, the system limits and corresponding take-over situations should be discussed with participants as well. Depending on the research questions, it might be useful to describe the most appropriate reaction to the situation, except when attempting to capture participants' spontaneous reactions.

In general, as with Schömig et al. [31] we recommend explaining system functionalities, limits and take-over situations in detail to reduce possible learning effects due to experiencing multiple take-over encounters that are recommended for the current test protocol. However, when focusing on initial contact with the system and take-over situation, reduced instructions are more suitable e.g., [31].

### 4.2.5. Experimental Design and Procedure

## Experimental Design

For both study environments, the design depends on the research question. However, we recommend including a complete, within-subjects design limiting the number of independent variables to ensure an economic study design, reduce test sample size, enhance statistic power, enable direct comparisons of participant performance across the independent variables, and exclude interindividual confounding variables. Additionally, it is highly important to randomize and counterbalance trials to reduce learning and order effects.

Regarding the number of take-over situations, it is recommended to repeat the encounters and in order to reduce learning and first contact effects it is recommended to clearly instruct the participants regarding the system's functionality and limitations [31]. Regarding the number of take-over situations, aspects such as the length of the analyzed intervals as well as the influence of the take-over situations' number on the system evaluation [31], must be considered. For a duration of 8–12 min as in the current studies, we recommend a maximum of four encounters, which should be randomized and counterbalanced across timing and situation type to reduce predictability. However, this recommendation aims at maximizing the number of take-over situations to be analyzed. This high frequency of system failures potentially lowering external validity must be considered. Depending on the research question, the number of take-over situations should be reduced and the route length should be extended (e.g., to evaluate how supervisory and take-over capabilities evolve over time and with long periods without system failures).

## Procedure

The actual procedure depends on, for instance, the study design, employed techniques, questionnaires, etc. Generally, we highly suggest including familiarization drives as mentioned in Schömig et al. [31]. In both study environments, participants should get accustomed driving manually in the simulator or actual vehicle if possible. For the former, this also allows checking for signs of simulator sickness. Depending on the research questions, participants should also be familiarized with partially automated driving and potentially with the take-over situations. We recommend familiarizing participants with partially automated driving but not with the take-over situations. This allows to achieve a comparable understanding of PAD across participants as well as to analyze the initial contact with these situations during NDRT execution. Nevertheless, the possibility of some take-over situations occurring during the trials and take-over situations itself should be described to the participants in the instructions.

Depending on the study's complexity, we suggest involving two researchers who can divide the technical tasks and participant supervision between each other to ensure a smoothly conducted study. In case of additional tasks (e.g., driving the lead vehicle), including another researcher is advisable. The researchers should receive detailed instructions and extensive trainings regarding their tasks, especially considering any driving tasks.

### 4.2.6. Data Preparation and Analyses

Regarding the supervisory behavior, the camera recordings or the eye tracking data must be annotated or mapped concerning the relevant AOIs: the NDRT, driving scene, instrument cluster, and vehicle interior. Other AOIs can be included if needed. The take-over capability data must be extracted from the simulator or ego vehicle and prepared for further analyses.

When examining supervisory capabilities during PAD, we recommend using an interval prior to the take-over situation. In that interval, the partially automated system must be active and should exclude any parts of earlier take-over situations. Therefore, the interval length depends on the time between the take-over situations. For instance, the current project included a 45-s segment in the simulator studies and a 10-s segment in the test track study, whereas Dogan et al. [25] chose a 15-s segment before a take-over situation occurred. In general, the interval length should be long enough to include at least one complete NDRT execution trial. In the current case, the NDRT trials were designed to take no longer than 20 s. Since participants are unlikely to complete a trial within 20

consecutive seconds, we recommend using a generous interval of 45 s for instance. Moreover, NHTSA [14] specifies that a NDRT trial should be completed within a total gaze duration of 12 s. With a 45 s interval, it should be possible to find these cumulative 12 s of total gaze duration as well. Moreover, if new PAD cut-off values are less conservative and result in longer total gaze durations towards the NDRTs, the 45-s intervals might also provide enough buffer for this. In contrast, to examine take-over capabilities, we suggest using an interval from the moment the take-over situation is triggered until participants' initial reaction. If participants do not react, the interval should last until the collision occurs or the researcher terminates the situation.

The current project analyzed the first and third situation (see Section 2.6). Depending on the research questions, other analyses can be done as well, such as comparing the first and last take-over situation or all take-over situations. However, the latter is only possible if the predictability of the take-over situations is low. Comparisons between the trials with and without NDRTs as well as between the trial with the reference task (e.g., manual radio tuning task) and the trials with other NDRTs are recommended to evaluate drivers' supervisory and take-over capabilities. The concrete analyses depend on the chosen research design.

#### 4.2.7. Participants

Concerning the participants, the following aspects must be considered. Firstly, the *sample size*. NHTSA [14] recommends including 24 participants to examine the distractive effects of visual-manual NDRTs. For studies involving conditional automated driving (SAE level 3), a sample size of at least  $n = 20$  is recommended when assessing the suitability of in-vehicle systems or at least  $n = 12$  participants per experimental test condition [31]. In general, desired sample size depends on the research question and intended statistical power. As with Schömig et al. [31], it is recommended to include at least  $n = 12$  participants per experimental test condition or  $n = 20$  participants depending on the research design. Secondly, the age distribution must be considered. The current studies aimed to follow NHTSA's guidelines of distributing the participants evenly across four recommended age groups: 18–24 years, 25–39 years, 40–54 years, and older than 55 [14]. The age distribution had no effect in either of the three studies. Nevertheless, we recommend involving all relevant age groups in the sample to control for age effects and reflect on different levels of driving experience. As with Schömig et al. [31], it is advisable to use the four age groups NHTSA highlights to achieve a heterogeneous age group. However, it must be taken into account that evenly distributing participants across these four age groups does not realistically reflect the populations' age distribution. Thirdly, the gender distribution must be considered. NHTSA [14] recommends having an even gender distribution. Similar to the age distribution, gender did not affect the results in either of the current three studies. When examining subjective PAD or NDRT execution experiences, it might still be useful to obtain an even gender distribution as Schömig et al. [31] recommend. We also recommend including an even gender distribution to control for gender effects. In addition to these three aspects, it might be reasonable to examine other sample characteristics as well, such as prior system experience, depending on the research questions.

## 5. Conclusions

In conclusion, the current project's overarching goal was to fill the methodological gap and take initial steps towards developing a test protocol for the systematic evaluation of the effects of NDRT execution on the drivers' supervisory and take-over capabilities during PAD. We believe that the systematic evaluation of the NDRTs' effects during PAD using the new test protocol developed within this project enhances comparability between different studies and generalizability of the studies' results, as well as provides a basis for developing cut-off values for deciding whether certain NDRTs are applicable for PAD. For the matter of using the test protocol, we provide a summarizing overview of the most important recommendations in Table 6.

**Table 6.** Short summary of the recommendations for test protocol usage.

	Recommendations
Driving scenario	<ul style="list-style-type: none"> <li>- Simulator: Driving scenario (car-following task) as described in Section 2.1, without other traffic and curved road segments</li> <li>- Test track: Same scenario as in simulator, requires appropriate test track (e.g., straight segment with minimum 11 km length)</li> </ul>
Take-over situations	<ul style="list-style-type: none"> <li>- Types: Responding to both lateral and longitudinal vehicle control (e.g., <i>deceleration of lead vehicle and ego vehicle drifting</i>)</li> <li>- Specifications: Exclusion of warnings and take-over requests, matching time to collisions (e.g., 7 s), multiple, counterbalanced, and randomized encounters</li> </ul>
Non-driving related tasks (NDRTs)	<ul style="list-style-type: none"> <li>- Types: Visual-manual NDRTs in comparison with reference task (e.g., manual radio tuning task [34]) and reference trial without NDRT execution</li> <li>- Specifications: Predefined start and finish, continuous execution while system is active</li> </ul>
Analyzed variables	<ul style="list-style-type: none"> <li>- Supervisory capabilities: mean gaze duration, total gaze duration, number of transitions</li> <li>- Take-over capabilities: reaction time and parameters for situation criticality (e.g., number of crashes)</li> </ul>
Equipment	<ul style="list-style-type: none"> <li>- Vehicle: Simulated mock-up or (automated) vehicles (see Section 2.4)</li> <li>- Human-machine interface: simple, intuitive design with relevant system states, timely reactions to input</li> <li>- Eye-tracking: head-mounted or video-based depending on level of detail required</li> <li>- Questionnaires: at a minimum demographic information</li> <li>- Instructions: written form, regarding partially automated system, take-over situations, NDRTs, task priorities</li> </ul>
Design and procedure	<ul style="list-style-type: none"> <li>- Study design: economic (e.g., limitation of number of NDRTs), for instance within-subjects design</li> <li>- Number of take-over situations: max. 4 per trial of a length of 8–12 min</li> </ul>
Procedure	<ul style="list-style-type: none"> <li>- Familiarization: Manual and partially automated drives (without take-over situations), NDRT execution</li> </ul>
Data preparation and analysis	<ul style="list-style-type: none"> <li>- Analyzed intervals: 45 s prior to take-over situations for supervisory capabilities, start of take-over situations until participant reaction, collision or fallback solutions</li> <li>- Analyzed take-over situations: in case of 4 take-over situations, first and third take-over situation</li> </ul>
Participants	<ul style="list-style-type: none"> <li>- Number: at least <math>n = 12</math> per experimental group or at least <math>n = 20</math> in total</li> <li>- Demographics: age and gender controlled (e.g., following NHTSA [14])</li> </ul>

**Author Contributions:** Conceptualization, C.H., N.R., F.N., S.H., J.F.K., and A.K.; methodology, C.H., N.R., F.N., and S.H.; investigation, C.H.; formal analysis, C.H. and N.R.; resources, F.N., S.H., J.F.K., and A.K.; writing—original draft preparation, C.H.; writing—review and editing, N.R., F.N., S.H., J.F.K., and A.K.; supervision, J.F.K. and A.K.; project administration, C.H., N.R., F.N., and S.H.; funding acquisition, J.F.K., C.H., and N.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded and supported by the BMW Group, Germany.

**Acknowledgments:** All opinions expressed in this paper are those of the authors and not necessarily those of the BMW Group. We thank Sebastian Scholz for his technical support during the simulator studies and Benjamin Jähn for his technical support and his support during the execution of the test track study.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funder (BMW Group) was left unmentioned during the acquisition of the participants and the study execution. Moreover, tasks such as the data analyses were done by the authors belonging to the Technical University of Chemnitz as an independent research institute with the authors belonging to the BMW Group providing advice and support from a solely scientific point of view.

## References

1. Flemisch, F.; Kelsch, J.; Löper, C.; Schieben, A.; Schindler, J. Automation spectrum, inner/outer compatibility and other potentially useful human factors concepts for assistance and automation. In *Human Factors for Assistance and Automation*; de Waard, D., Oberheid, F.O., Flemisch, B., Lorenz, H., Brookhuis, K.A., Eds.; Shaker Publishing: Maastricht, The Netherlands, 2008; pp. 1–16.
2. Merat, N.; Jamson, A.H.; Lai, F.C.; Carsten, O. Highly automated driving, secondary task performance, and driver state. *Hum. Factors* **2012**, *54*, 762–771, doi:10.1177/0018720812442087.



3. SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles* (No. J3016); SAE International: Warrendale, PA, USA, 2016.
4. Stanton, N.A.; Young, M.S. Vehicle automation and driving performance. *Ergonomics* **1998**, *41*, 1014–1028, doi:10.1080/001401398186568.
5. Van den Beukel, A.P.; van der Voort, M.C. Design considerations on user-interaction for semi-automated driving. In Proceedings of the 35th FISITA World Automotive Congress, Maastricht, The Netherlands, 2–6 June 2014; pp. 1–8.
6. Körber, M.; Weißgerber, T.; Kalb, L.; Blaschke, C.; Farid, M. Prediction of take-over time in highly automated driving by two psychometric tests. *DYNA* **2015**, *82*, 195–201, doi:10.15446/dyna.v82n193.53496.
7. Saxby, D.J.; Matthews, G.; Warm, J.S.; Hitchcock, E.M.; Neubauer, C. Active and passive fatigue in simulated driving: Discriminating styles of workload regulation and their safety impacts. *J. Exp. Psychol.: Appl.* **2013**, *19*, 287–300, doi:10.1037/a0034386.
8. Coughlin, J.F.; Reimer, B.; Mehler, B. Monitoring, managing, and motivating driver safety and well-being. *IEEE Pervasive Computing* **2011**, *10*, 14–21, doi:10.1109/MPRV.2011.54.
9. Regan, M.A.; Hallett, C.; Gordon, C.P. Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accid. Anal. Prev.* **2011**, *43*, 1771–1781, doi:10.1016/j.aap.2011.04.008.
10. Spence, C.; Ho, C. Crossmodal information processing in driving. In *Human Factors of Visual and Cognitive Performance in Driving*; Castro, C., Ed.; CRC Press: Boca Raton, FL, USA, 2009; pp. 187–200.
11. Renge, K. The effects of driving experience on a driver's visual attention. An analysis of objects looked at: Using the 'verbal report' method. *Intern. Asso. Traffic Safety Sci. Res.* **1980**, *4*, 95–106.
12. Hughes, P.K.; Cole, B.L. What attracts attention when driving? *Ergonomics* **1986**, *29*, 377–391, doi:10.1080/00140138608968272.
13. Ahlstrom, C.; Kircher, K. Changes in glance behavior when using a visual eco-driving system – A field study. *Appl. Ergon.* **2017**, *58*, 414–423, doi:10.1016/j.apergo.2016.08.001.
14. National Highway Traffic Safety Administration. *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*; Department of Transportation: Washington, DC, USA, 2016.
15. Birrel, S.A.; Fowkes, M. Glance behaviours when using an in-vehicle smart driving aid: A real-world, on-road driving study. *Transp. Res. F Traffic Psychol. Behav.* **2014**, *22*, 113–125, doi:10.1016/j.trf.2013.11.003.
16. Ayama, M.; Hasegawa, H.; Kawaguchi, M.; Ihata, N.; Ikegami, M.; Kasuga, M. A study to measure spare capacity of driver's attention payable to cognitive subtask. In Proceedings of the 5th IEEE International Conference on Intelligent Transportation Systems, Singapore, 6 September 2002; pp. 279–283, doi:10.1109/ITSC.2002.1041229.
17. Uno, H.; Hiramatsu, K. Effects of auditory distractions on driving behavior during lane change course negotiation: Estimation of spare mental capacity as a index of attention distraction. *JSAE Rev.* **2000**, *21*, 219–224, doi:10.1016/S0389-4304(99)00093-4.
18. Jamson, A.H.; Merat, N.; Carsten, O.M.; Lai, F.C.H. Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transp. Res. C Emerg. Technol.* **2013**, *30*, 116–125, doi:10.1016/j.trc.2013.02.008.
19. Llaneras, R.; Salinger, J.; Green, C. Human factors issues associated with limited ability autonomous driving systems: drivers' allocation of visual attention to the forward roadway. In Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, New York, NY, USA, 17–20 June 2013; pp. 92–98.
20. Carsten, O.; Lai, F.C.H.; Barnard, Y.; Jamson, A.H.; Merat, N. Control Task Substitution in Semiautomated Driving: Does It Matter What Aspects Are Automated? *Hum. Factors* **2012**, *54*, 747–761, doi:10.1177/0018720812460246.
21. Hergeth, S.; Lorenz, L.; Vilimek, R.; Krems, J.F. Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving. *Hum. Factors* **2016**, *58*, 509–519, doi:10.1177/0018720815625744.
22. Cunningham, M.L.; Regan, M.A. Driver distraction and inattention in the realm of automated driving. *IET Intell. Transp. Syst.* **2017**, *12*, 1–7, doi:10.1049/iet-its.2017.0232.
23. Martens, M.; van den Beukel, A. The road to automated driving: Dual mode and human factors considerations. In Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems, The Hague, The Netherlands, 6–9 October 2013; pp. 2262–2267.

24. Eriksson, A.; Stanton, N.A. Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Hum. Factors* **2017**, *59*, 689–705, doi:10.1177/0018720816685832.
25. Dogan, E.; Rahal, M.C.; Deborne, R.; Delhomme, P.; Kemeny, A.; Perrin, J. Transition of control in a partially automated vehicle: Effects of anticipation and non-driving-related task involvement. *Transp. Res. F Traffic Psychol. Behav.* **2017**, *46*, 205–215, doi:10.1016/j.trf.2017.01.012.
26. Lin, R.; Liu, N.; Ma, L.; Zhang, T.; Zhang, W. (2019). Exploring the self-regulation of secondary task engagement in the context of partially automated driving: A pilot study. *Transp. Res. F Traffic Psychol. Behav.* **2019**, *64*, 147–160, doi:10.1016/j.trf.2019.05.005.
27. Large, D.R.; Banks, V.A.; Burnett, G.E.; Baverstock, S.; Skrypchuk, L. Exploring the behaviour of distracted drivers during different levels of automation in driving. In Proceedings of the 5th international conference on driver distraction and inattention (DDI2017), Paris, France, 20–22 March 2017.
28. Lee, J.; Hirano, T.; Hano, T.; Itoh, M. Conversation during partially automated driving: How attention arousal is effective on maintaining situation awareness. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3718–3723, doi:10.1109/SMC.2019.8914632.
29. Louw, T.; Kuo, J.; Romano, R.; Radhakrishnan, V.; Lenné, M.G.; Merat, N. Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures. *Transp. Res. F Traffic Psychol. Behav.* **2019**, *62*, 870–882, doi:10.1016/j.trf.2019.03.020.
30. Naujoks, F.; Höfling, S.; Purucker, C.; Zeeb, K. From partial and high automation to manual driving: Relationship between non-driving related tasks, drowsiness and take-over performance. *Accid. Anal. Prev.* **2018**, *121*, 28–42, doi:10.1016/j.aap.2018.08.018.
31. Schömig, N.; Befelein, D.; Wiedemann, K.; Neukum, A. *Methodische Aspekte und aktuelle inhaltliche Schwerpunkte bei der Konzeption experimenteller Studien zum hochautomatisierten Fahren* [Methodological Aspects and Current Focus Areas Concerning the Conception of Studies for Highly Automated Driving]; FAT Forschungsvereinigung Automobiltechnik: Berlin, Germany, 2020.
32. Naujoks, F.; Wiedemann, K.; Schömig, N.; Hergeth, S.; Keinath, A. Towards guidelines and verification methods for automated vehicle HMIs. *Transp. Res. F Traffic Psychol. Behav.* **2019**, *60*, 121–136, doi:10.1016/j.trf.2018.10.012.
33. Wiggerich, A. Development of a modular tool for safety assessments of human-machine-interaction for assisted driving functions (SAE level 2). In Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV), Eindhoven, The Netherlands, 10–13 June 2019.
34. Alliance of Automobile Manufacturers. *Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems Including*; Alliance of Automobile Manufacturers: Washington, DC, USA, 2006.
35. Lorenz, L.; Hergeth, S. Einfluss der Nebenaufgabe auf die Überwachungsleistung beim teilautomatisierten Fahren [Influence of secondary tasks on the supervisory performance during partially automated driving]. In *Der Fahrer im 21. Jahrhundert*; VDI, Ed.; VDI-Verlag: Düsseldorf, Germany, 2015; Volume 8, pp. 159–172.
36. Green, D.M.; Swets, J.A. *Signal Detection Theory and Psychophysics*; John Wiley and Sons: New York, NY, USA, 1966.
37. Wickens, C.D. Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* **2002**, *3*, 159–177, doi:10.1080/14639220210123806.
38. Horrey, W.J.; Wickens, C.D. Driving and side task performance: The effects of display clutter, separation, and modality. *Hum. Factors* **2004**, *46*, 611–624, doi:10.1518/hfes.46.4.611.56805.
39. Jamson, A.H.; Merat, N. Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transp. Res. F Traffic Psychol. Behav.* **2005**, *8*, 79–96, doi:10.1016/j.trf.2005.04.002.
40. Metz, B.; Schömig, N.; Krüger, H.P. Attention during visual secondary tasks in driving: Adaptation to the demands of the driving task. *Transp. Res. F Traffic Psychol. Behav.* **2011**, *14*, 369–380, doi:10.1016/j.trf.2011.04.004.
41. Vollrath, M.; Krems, J. Fahren [Driving]. In *Verkehrspsychologie. Ein Lehrbuch für Psychologen, Ingenieure und Informatiker*; Vollrath, M., Krems, J., Eds.; Kohlhammer: Stuttgart, Germany, 2011; pp. 22–40.
42. Mattes, S. The lane change task as a tool for driver distraction evaluation. In *Quality of Work and Products in Enterprises of the Future*; Strasser, H., Kluth, K., Rausch, H., Bubb, H., Eds.; Ergonomia: Stuttgart, Germany, 2003; pp. 57–60.

43. Fry, E.B. *Teaching Faster Reading: A Manual*; Cambridge University Press: Cambridge, UK, 1963.
44. Victor, T.W.; Harbluk, J.L.; Engström, J.A. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transp. Res. F Traffic Psychol. Behav.* **2005**, *8*, 167–190, doi:10.1016/j.trf.2005.04.014.
45. Ecker, R. Der verteilte Fahrerinteraktionsraum [The Distributed Driver Interaction Space]. Ph.D. Thesis, Ludwig-Maximilians-Universität (LMU) München, München, Germany, 14 June 2013.
46. Vollrath, M.; Huemer, A.K.; Teller, C.; Likhacheva, A.; Fricke, J. Do German drivers use their smartphones safely?—Not really! *Accid. Anal. Prev.* **2016**, *96*, 29–38, doi:10.1016/j.aap.2016.06.003.
47. National Highway Traffic Safety Administration. *Traffic Safety Facts Research Note: Distracted Driving 2014*; Report No. DOT HS 812 260; National Highway Traffic Safety Administration: Washington, DC, USA, 2016; Available online: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812260> (accessed on 5 May 2020).
48. Wittmann, M.; Kiss, M.; Gugg, P.; Steffen, A.; Fink, M.; Pöppel, E.; Kamiya, H. (2006). Effects of display position of a visual in-vehicle task on simulated driving. *Appl. Ergon.* **2006**, *37*, 187–199, doi:10.1016/j.apergo.2005.06.002.
49. Lambale, D.; Laakso, M.; Summala, H. Detection thresholds in car following situations and peripheral vision: Implications for positioning of visually demanding in-car displays. *Ergonomics* **1999**, *42*, 807–815, doi:10.1080/001401399185306.
50. Weinberg, G.; Harsham, B.; Medenica, Z. Evaluating the usability of a head-up display for selection from choice lists in cars. In Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Salzburg, Austria, 30 November–2 December 2011; pp. 39–46, doi:10.1145/2381416.2381423.
51. Krause, M.; Angerer, C.; Bengler, K. Evaluation of a Radio Tuning Task on Android while Driving. *Procedia Manuf.* **2015**, *3*, 2642–2649, doi:10.1016/j.promfg.2015.07.334.
52. Brillenstudie 2014 [Glasses Study 2014]. Available online: <https://www.zva.de/brillenstudie> (accessed on 20 March 2020).
53. *Tobii Pro Glasses 2 [Apparatus and software]*; Tobii AB: Stockholm, Sweden, 2016.
54. ISO 15007-1. *Road Vehicles—Measurement of Driver Visual Behaviour with Respect to Transport Information and Control Systems Part 1: Definitions and Parameters*; ISO: Geneva, Switzerland, 2014.
55. Sedlmeier, P.; Renkewitz, F. *Forschungsmethoden und Statistik: Ein Lehrbuch für Psychologen und Sozialwissenschaftler [Research Methods and Statistics: A Textbook for Psychologists and Social Scientists]*, 2nd ed.; Pearson: München, Germany, 2013.
56. Hensch, A.-C.; Rauh, N.; Schmidt, C.; Hergeth, S.; Naujoks, F.; Krems, J. F.; Keinath, A. Effects of secondary tasks and display position on glance behavior during partially automated driving. *Transp. Res. F Traffic Psychol. Behav.*, **2020**, *68*, 23–32, doi:10.1016/j.trf.2019.11.014.
57. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Laurence Erlbaum Associates: Hillsdale, NJ, USA, 1988.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).