information

Technical Note

S.A.D.E.—A Standardized, Scenario-Based Method for the Real-Time Assessment of Driver Interaction with Partially Automated Driving Systems

Nadja Schömig [1,*], Katharina Wiedemann [1], André Wiggerich [2] and Alexandra Neukum [1]

[1] Würzburg Institute for Traffic Sciences GmbH (WIVW), Robert-Bosch-Str. 4, 97209 Veitshoechheim, Germany
[2] Federal Highway Research Institute, 51427 Bergisch-Gladbach, Germany
* Correspondence: schoemig@wivw.de

Abstract: Vehicles equipped with so-called partially automated driving functions are becoming more and more common nowadays. The special feature of this automation level is that the driver is relieved of the execution of the lateral and longitudinal driving task, although they must still monitor the driving environment and the automated system. The method presented in this paper should enable the assessment of the usability and safety of such systems in a standardized manner. It is designed to capture a driver's interaction with a system via the human–machine interface in specific scenarios in user studies. It evaluates several observable aspects of this interaction in real time and codes inadequate behavior in the categories "system operation", "driving behavior" and "monitoring behavior". A generic rating regarding the overall handling of the scenario is derived from these criteria. The method can be used with the assistance of a tablet application called the S.A.D.E. app (Standardized Application for Automated Driving Evaluation). Initial studies using driving simulators show promising results regarding its ability to detect problems related to a system or HMI, with some future challenges remaining open.

Keywords: evaluation; HMI; tablet application; partially automated driving; Level 2 automation; driver monitoring; system operation; driving performance

Citation: Schömig, N.; Wiedemann, K.; Wiggerich, A.; Neukum, A. S.A.D.E.—A Standardized, Scenario-Based Method for the Real-Time Assessment of Driver Interaction with Partially Automated Driving Systems. *Information* **2022**, *13*, 538. https://doi.org/10.3390/info13110538

Academic Editor: Joaquim Ferreira

Received: 30 September 2022
Accepted: 11 November 2022
Published: 14 November 2022

1. Background

Vehicles equipped with so-called partially automated driving functions (so-called Level 2 systems or L2 systems; see SAE definition, 2021 [1]) are becoming more and more common nowadays. Such systems are able to execute parts of the operational driving task, i.e., lateral vehicle control via steering and longitudinal vehicle control via acceleration and deceleration. These features are dedicated to being used in specific operational design domains, such as highways; however, they are also available on other road types. L2 systems require immediate driver intervention when the system reaches its limits. In this case, the system either switches off without any explicit take-over request or it remains active because its sensors have not detected the reason for necessary driver intervention (e.g., an obstacle). Therefore, the driver must continue to perform the task of object and event detection while using such systems [1]. Most of the serial systems today are hands-on systems requiring drivers to leave their hands on the steering wheel. If such a system detects that the driver takes their hands off the steering wheel for a certain period of time, it will request the driver put their hands back via a warning.

Although L2 systems are designed to increase driving comfort, some specific problems might arise during their usage. While on the one hand drivers are partly released from continuous driving tasks (steering and keeping speed/distance), they still remain responsible for intervening at system limits. In highly reliable systems, this might require the detection of and the reaction to very rare events after long periods of inactivity while the driver was simply monitoring the road and the system. Such a task is very difficult for

human operators to handle (see Bainbridge's *Ironies of Automation*, 1980 [2]). In addition, an inadequate assumption about a system's abilities and limits, i.e., an inadequate mental model, might result in overtrust and potentially safety-critical situations in the case of a non-response or the delayed or inadequate response of the operator to a system limit. A prominent example is the brand Tesla labeling its L2 feature "Autopilot" which could be associated with comparably higher automation levels in the aircraft domain (see, for example, the Tesla crash with a truck in May 2016 in Florida [3]. However, more data from real roads and from crash databases are still missing, which are needed to estimate the real impact of such potential problems.

Empirical research on L2 systems in simulators, however, has been conducted extensively. Many studies show that drivers accept and trust such systems [4,5]. When using L2 systems in traffic jams, drivers feel relieved of the driving task. However, the monitoring task which still has to be performed is experienced as rather strenuous [6,7]. It has also been observed that drivers increasingly turn their attention to non-driving-related secondary activities when driving with such systems [8]. With regard to responsiveness at system limits, some studies suggest that drivers are able to react relatively well if the system limits are announced by warnings and the driver is fully attentive [5,9,10]. However, it is typical that L2 systems are not able to predict system limits and therefore they cannot be announced in advance. Reacting to such sudden, critical situations is far more difficult and leads to greater problems with the driver's reaction [7,11,12]. The better the HMI (human–machine interface) prepares the driver for such situations and assists the driver in responding to them, the more likely it is that safety-critical consequences can be avoided [13].

In order to objectively evaluate the effects of such driver assistance systems and their HMI design in terms of usability and driving safety as efficiently as possible and in the most standardized way, suitable methods are needed that can be optimally used as a kind of "live" rating during interaction with the system. In contrast to L2 systems, some approaches were already developed for L3 systems (conditionally automated driving), e.g., by Naujoks and colleagues. They developed a method for the expert assessment of the controllability of take-over situations in L3 driving, the so-called TOC-Rating [14]. Raters are trained to use a coding sheet which includes several observation criteria to generate a global controllability rating from a video of the take-over situation. However, this method is very detailed as it is dedicated for use in subsequent data analyses and focuses exclusively on controllability but not on the usability aspects of L3 HMIs. Another method developed by [15] deals with the verification of minimum design requirements for L3 system HMIs as proposed by the National Highway Traffic Safety Administration (NHTSA, [16]). This test procedure contains a set of relevant use cases that should be tested as well as a set of evaluation criteria based on drivers' observations in the intuitive interaction with the system and on drivers' subjective reports evaluating the understandability of system mode indicators for L3 systems.

For Level 2 systems, such tools for the assessment of usability and safety do not exist so far. A new method should be able to evaluate whether a driver understands L2 system functionality and the resulting responsibilities of the driving task, whether they adequately interact with the system and are able to react to system limits effectively and safely. This requires the observation of the direct interaction of a user with a system via the HMI. Most suitable for this approach are user studies with a sample of real users who have no prior experience. In order to evaluate these interaction-related aspects of a safe human–machine interaction with partially automated driving functions, the method should fulfill the following requirements, which were estimated as meaningful by the authors' team based on an evaluation of already existing methods:

- The method should be able to identify problems related to HMI aspects as well as to system functionality aspects;
- The method should allow both a global analysis as well as a very differentiated analysis of the interaction with the system in the investigated scenarios;

- The method should be adaptable to different test environments (e.g., driving simulation, test track or real-world driving);
- The method should allow a quick and efficient data assessment and analysis;
- The method should enable a standardized procedure with regard to the testing conditions, test scenarios and evaluation criteria;
- The method should allow an objective evaluation according to clear rules;
- The method should allow the assessment of driver interaction and experiences with the system directly in the situation, i.e., in real time;
- The method should be unobtrusive (i.e., it should not disturb the user).

The following chapter describes the development and application of such a standardized scenario-based method for the real-time assessment of driver interaction with partially automated driving systems called S.A.D.E. (Standardized Application for Automated Driving Evaluation) which is intended to be employed in user studies. This method was developed within a project funded by the German Federal Highway Research Institute (German BASt: Bundesanstalt für Straßenwesen) among other instruments for the evaluation of L2 HMIs (see [17] for a detailed overview of the project contents and outputs).

2. Method Description

2.1. Evaluation Criteria—Which Behavior Should Be Observed?

The first step of the method development was to define the criteria for the evaluation of the human–machine interaction of users with partially automated systems: The human–machine interface (HMI) must be able to communicate a correct understanding of the driver's responsibilities while using the system and must be able to create an adequate awareness of the system status and the surrounding situation (mode and situation awareness). The driver must be able to efficiently use the system (which means the operation of the system with low mental and physical effort and without impairments in driving performance) and to adequately and safely handle situations in which they need to take over vehicle control. Aspects such as the efficiency (as part of usability) and safety of the interaction are therefore within the scope of evaluation. Collecting these aspects of objectively observable behavior allows the inference of underlying implicit aspects, such as comprehension of the system's logic, the level of workload in system operation, comprehension of system outputs, the level of mode and situation awareness and knowledge about the driver's responsibilities.

In order to evaluate the effort and resulting workload in system usage, the driver's system operations must be assessed. This must include the observation of driver-initiated transitions between system modes (e.g., activation or deactivation of the system or its subfunctions). Errors, imprecisions or unnecessary actions during system usage point to a high complexity of the system logic or operational logic and potential difficulties in the comprehensibility of necessary operational steps to achieve a certain goal (e.g., in the case of activation, is the system already active or still in standby mode?).

In addition, the observation of the driver's behavior in the case of system-initiated transitions at system limits or system failures helps to assess whether a driver has enough system and situation awareness to detect system limits and to adequately choose necessary interventions. This assessment can be achieved by observing whether the driver adequately deactivates the system or its subfunctions in a timely manner in cases of system limits or system failures or whether the driver adequately reacts by taking over control without impairments in driving safety. Driving errors or driver endangerment of themselves or other road users indicate that the driver has insufficient system and situation awareness and is therefore surprised and excessively challenged by system limits and failures.

While driving in certain system states (especially while driving with active L2 system) it must be determined whether the driver is aware of their responsibility to monitor the driving environment and the system. This can be observed in the degree to which the driver shows adequate monitoring behavior, both with regard to their visual attention distribution for the observation of the surrounding traffic and for checking the system

status as well as to the adequate motoric involvement in the driving task, i.e., keeping hands on the wheel and cooperating with the system on steering tasks if needed.

Taken together, it is necessary to observe a driver's system operations and driving behavior as well as driver monitoring behavior to obtain a comprehensive understanding of the assumed underlying implicit aspects named in the chapter above. In addition to the observation of the driver's behavior, the driver's subjective comprehension of the system and the HMI should also be measured via scenario-specific questions. This allows the assessment of subjective statements about the perceived level of certainty with regard to interaction with the system (e.g., knowing what to do, understanding what the system is indicating, understanding why the system behaves in a certain way). In contrast to a more generic system understanding, which can be assessed in a follow-up survey after testing, the driver is asked to evaluate specific HMI outputs and system behaviors directly in the respective scenarios.

2.1.1. Observation Categories of Driver Behavior

To achieve a standardized, objective and efficient observation, driver behavior is coded using three main categories:

- System operation;
- Driving behavior;
- Monitoring behavior.

To increase the efficiency of the observation, not all behavior is rated, but only inadequate behavior is coded and protocolled. The level or type of inadequateness is further divided into category-specific errors or problems. These error types were derived from the long-term experience of the authors' team from numerous studies about such systems (simulator studies and real driving studies with relation to the evaluation of take-over scenarios, see, for example, [14]) and resemble the most frequent problems with L2 systems. Error definitions are shown in Table 1.

Table 1. Error definitions for observation categories.

Category	Observable Error/Problem	Description/Example
System operation (especially driver-initiated operations)	Noticed nothing	Driver does not notice changes in system state
	Uncertain/delayed operation	Driver shows uncertainties in system operation, e.g., searches for a certain button; Driver takes a long time to perform an action
	Inadequate operation	Driver shows inadequate system operation, e.g., activates the system in situations where it should not be used
	Operation error	Driver initiates an incorrect operation, e.g., driver presses wrong button, driver presses correct button but not firmly enough, driver wants to activate the system when it is not possible/available
	Support by experimenter in operation required	Driver is not able to execute the expected action until a certain defined point in time so that the experimenter must give support to reach the designated system mode

Table 1. *Cont.*

Category	Observable Error/Problem	Description/Example
Driving behavior (especially relevant to driver-initiated system operations and system-initiated transitions)	No reaction	Driver does not show any reaction in a given situation which would require one
	Reaction delayed	Driver reacts to an event with a clear delay, e.g., by a braking maneuver
	Reaction too strong	Driver reacts too strongly to an event, e.g., with oversteering
	Lane exceedance	The vehicle crosses the lane marking with one wheel
	Poor lane keeping	The vehicle visibly swerves within the lane to the right and/or to the left
	Insufficient securing behavior	Driver does not execute a control glance in the mirror in the case of a lane change
	Endangerment	Safety distance below 1 s to the front/side or behind/towards other vehicles
	Collision	Vehicle collides with another traffic participant or a stationary obstacle
Monitoring behavior	Uncertainties in hands-on behavior	Driver shows clear uncertainties as to whether hands should be left on the steering wheel or not, takes them away repeatedly or rests them too weakly on the wheel
	Not attentive enough	Driver shows clear signs of inattention, e.g., no control glances to HMI for longer time intervals, direction of attention towards NDRT (non-driving-related task)
	Hands-off warning	The hands-off warning was triggered by the system
	Stage of hands-off warning	The maximum stage of the hands-off warning was reached within a scenario

2.1.2. Subjective Driver Evaluation

In addition to observational categories, drivers' subjective evaluations should be assessed in certain test scenarios (especially system- and driver-initiated transitions). With this intention, several interview questions are included in the method. The driver is asked to answer them directly after having experienced the transitions. The following questions are used to assess the subjective evaluation of the HMI by the driver:

- Comprehensibility of the required driver action: Does the driver know what to do in a certain situation, e.g., in order to activate the system, to deactivate it or to adequately react to a system limit?
- Understandability of system behavior: Does the driver understand why the system behaves in a certain way in a situation, e.g., when the lateral control is switched off?
- Comprehensibility of system outputs: Does the driver understand what the visual system indicators or acoustic signals mean?
- Perceived situation criticality: How critical does the driver perceive a certain situation as a result of the combination of the objective demands of the situation and the required reaction?

In addition, there might be further suitable questions depending on the specific study question that could be included in the test protocol, e.g., perceived safety when driving with the system or perceived system trust. In terms of participant response, a seven-point verbalized, bipolar rating scale is recommended with rating values from −3 to +3 which can be flexibly adapted to the specific questions. According to the literature review in [18], five to seven categories achieve the best measures with regard to reliability, validity and differentiation. In addition, this reduced number can be easily assessed in real time during the drive and easily learned. All values are verbalized to increase reliability and validity (according to [18]). The additional numerical values enhance the rapid answering of the questions during the drive. The scale should be mounted at a location in the vehicle where the drivers can assess its content with a short glance. For the question regarding the perceived criticality of the situation, the 11-point rating scale of [19] can be used.

2.1.3. Global Rating of Scenario Handling

In addition, the experimenter is requested to give a global rating per scenario, which describes how well the driver has handled the overall driving scenario (see Figure 1). The proposed rating scale was adapted from the so-called Fitness-to-Drive-Scale (FtD-Scale; [20,21], on the basis of [19]), which was originally used as a rating for driver fitness in manual driving. This scale and the categories were adapted to the context of automated driving for this test protocol.

Figure 1. Experimenter global rating of the driver's handling of a use case in a specified scenario. The cross marks an exemplary rating.

The experimenter should derive the rating from the observed errors or problems during interactions with the system in the specific test scenario. Table 2 summarizes the recommendations for global ratings dependent on the observed errors or problems. Depending on the type of scenario, different observations can lead to a classification of the rating into various categories. For example, a rating of 10 can be given if either the driver is not able to bring the system into a designated state without the support of the experimenter or if the driver caused an accident in the case of a system limit. This requires a separate interpretation of the rating per scenario type (e.g., system operation scenarios separated from take-over scenarios). The definitions of the severity of the errors are recommendations regarding which category the experimenter should assign an observed behavior. When refining the rating within a certain category, the rater should take the consequences of the erroneous behavior into account.

Table 2. Recommendations for global ratings depending on the observed errors or problems.

Error/Problem	Verbal Category	Numeric Category
• System operation: no successful operation, even after support by the experimenter • Driving behavior: occurrence of a collision with another vehicle or obstacle • Monitoring behavior: system switching off due to repeated non-reaction to hands-off warnings (last warning stage of the system)	Scenario not handled successfully	10

Table 2. *Cont.*

Error/Problem	Verbal Category	Numeric Category
• System operation: driver did not notice that a transition to a lower level of automation was required; support required, resulting in successful operation • Driving behavior: no or delayed reaction with the consequence of endangerment of self or other road users • Monitoring behavior: more than one hands-off warning or hands-off warning at stage 2	Not acceptable problems	9 8 7
• System operation: operation error or inadequate operation • Driving behavior: delayed or too strong reaction, e.g., with regard to braking, lane exceedance; insufficient securing • Monitoring behavior: clear signs of inattentive behavior, at least one hands-off warning at stage 1	Error-prone, but acceptable	6 5 4
• System operation: uncertain/delayed operation • Driving behavior: poor lane keeping • Monitoring behavior: uncertainties in hands-on behavior	good	3 2 1
• No errors/problems	perfect	0

2.1.4. Rater Training

To ensure that the method achieves high objectivity, reliability and validity, training the users on the tool beforehand in the application of the method is recommended. This should include a description of the defined categories, some examples of behaviors that should be rated as errors or problems in the different categories and some rules for the achievement of the global rating. If several raters are to be deployed within a study, a discussion round should be included before the start of the study to clarify unclear definitions and to gather a common understanding of the categories and the coded behaviors. The result should be sufficiently high interrater reliability. Training material in the form of a PowerPoint presentation is available on request from the authors' team.

2.2. Definition of Relevant Test Scenarios

The following system states and transitions between these states were selected as the most relevant states while driving with L2 systems by the authors' team. They should be used as relevant test scenarios where the driver's interaction with the system should be observed based on the above defined categories in order to assess the HMI's usability and safety:

- System activation by the driver;
- System deactivation by the driver;
- Longer driving with active L2 system;
- Driver-initiated lane change;
- Temporary standby mode of lateral control;
- System limit and/or system malfunction (system limits can be both detectable and predictable as well as not detectable and not predictable; e.g., in longitudinal control: sensors are not able to detect a stationary vehicle or any other obstacle; in lateral control: system is not able to apply the necessary steering torque to manage a situation, e.g., a sharp bend).

These more general use cases must be transferred into testable scenarios for evaluation in a driving simulator or on a test track. Driver-initiated activation and deactivation processes as well as longer phases of driving with an active L2 system and lane changes can simply be instructed by the experimenter. Depending on the specific system, a temporary standby mode of the lateral control can be achieved by the absence of lane markings. Longitudinal system limits can be realized by placing an obstacle in the ego-vehicle's lane. Realistic lateral system limits that are not manageable by the system could be construction sites or sharp bends during real driving conditions.

2.3. Implementation of the Method in a Tablet App

The assessment method described can be used as a paper–pencil tool. In order to economize the process of data entry, a tablet application called S.A.D.E. (Standardized Application for Automated Driving Evaluation; see Figure 2) was developed. The app was programmed for Windows tablets with WIVW simulation software SILAB® (v6.0). The experimenter is able to protocol the occurrence of defined errors in system operation, driving and monitoring behavior via the tablet separately for each scenario in real time during the scenario (this is performed in the left area of the app). In addition, it is possible to rate the driver's handling of a scenario globally based on the observed errors (performed in the lower area of the app) and to conduct a standardized interview during the driving course (right area of the app). The upper control buttons on the right side can be used to start, pause and stop the evaluation and to switch to previous or subsequent test scenarios. If the application is used within driving simulator studies programmed with the simulation software SILAB®, it is possible to automatically count some of the errors (e.g., endangerments, collisions, lane exceedances and hands-off warning events). The application additionally integrates specific features such as acoustic feedback at the start of a new scenario or the upcoming end of the current scenario, a visual enhancement of relevant observation categories or survey questions for a specific scenario. The data are recorded and collected in a data sheet that also includes the data of the participant and study-related information and that can be easily imported into established statistical analyzing tools.

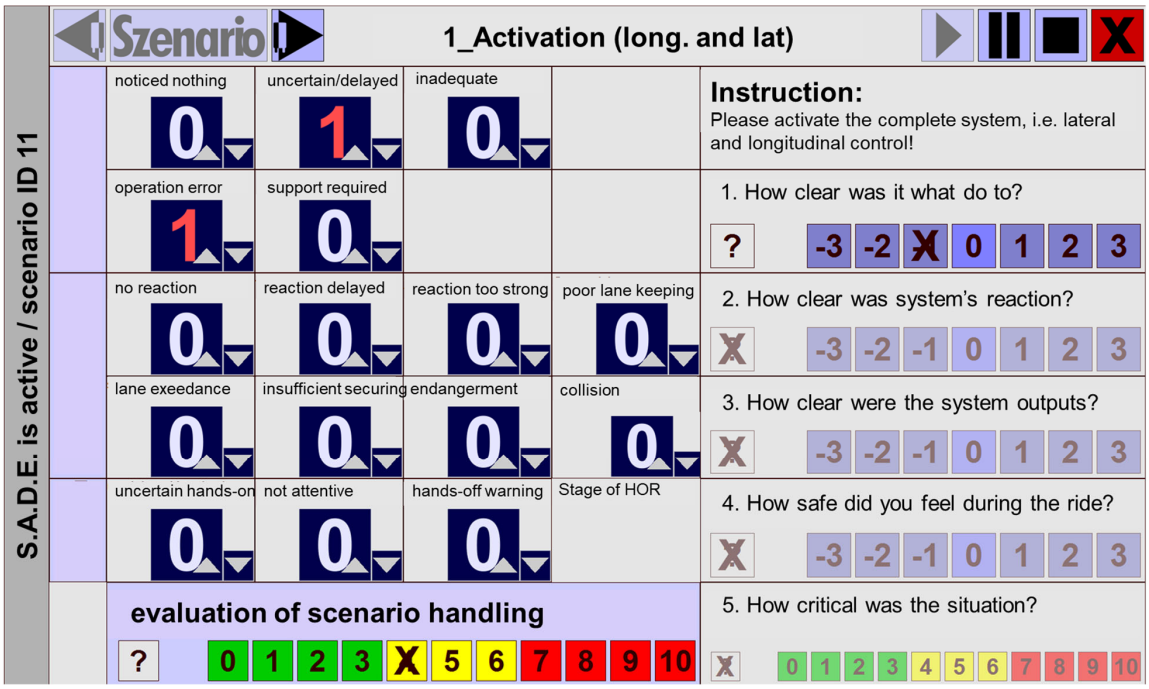

Figure 2. Tablet surface of the S.A.D.E. application.

2.4. Application of the Method

In simulator studies, the S.A.D.E. app should be used in the following way: After starting the test course, S.A.D.E. automatically starts (when using SILAB®). The experimenter can enter data that are specific to the participant (e.g., demographics) and the test trial. After that, the driver completes the test course in a predefined sequence of scenarios. During the respective scenarios, such as a system limit or a certain instruction such as to activate the L2 system, the experimenter observes driver behavior and directly protocols the observed errors or problems defined in the app. In doing so, the experimenter must have the opportunity to monitor the driver's system operations at the steering wheel as well as the driver's glance behavior. In the best instance, two cameras are fixed at proper positions in the vehicle mockup to record the driver's face and the steering wheel with the control elements as well as the driver's hand position. If possible, the driving scenery should be observable on a monitor from a bird's eye view for the experimenter to assess the driving performance and lane-keeping quality. After having recorded the observed driving errors and problems, the driver can be asked the predefined questions. After the scenario, the experimenter is expected to give the global rating of the driver handling the scenario. It is recommended to extensively train the experimenters of a user study beforehand in the usage of the application.

If the app is used in a simulator study, it is possible to make entries into the tablet only during the current scenario. After reaching the end of a scenario, the next scenario starts, all inputs are nulled and new entries can be made. Therefore, it is important that the length of a scenario is designed in a way so that the conduction of the observation and the survey are smoothly possible (i.e., there should be enough time between subsequent scenarios or simulation should be paused after each scenario). If the experimenter does not manage to enter all observations or answers from the driver, they can pause the app at the end and go through all scenarios again to make corrections before they finally stop the application.

3. Results of an Explorative Study in the WIVW Driving Simulator

An initial explorative study to test the feasibility of the developed method was performed in the motion-based driving simulator of the WIVW GmbH. The goal of the study was to evaluate whether HMI design aspects defined as possibly problematic within a previously conducted expert evaluation (using a checklist method including guidelines about the design of the human–machine interface of L2 automated vehicles, see [17]) influence real users' observable behavior and their subjective experience during interaction with the L2 system assessed by the S.A.D.E. app.

Two HMI variants A and B were investigated as part of the study. Their HMI design differed in several dimensions:

- System operation: operation logic regarding the activation of the longitudinal vehicle control (one-step vs. two-step activation);
- Control elements: labeling consistent vs. not consistent with the user manual;
- Visual indicators for active lateral vehicle control: with vs. without additional symbol of a steering wheel and text;
- Visual contrast: high vs. low contrast between foreground and background;
- Warning concept in situations with predictable system limits: presence vs. absence of a visual and acoustic warning.

These dimensions were derived from an expert checklist including several generic design guidelines derived from the literature (for more information about this checklist and its development, see [17]). HMI variant A was designed to achieve a high compliance with these guidelines (therefore named "highly compliant" HMI variant A), whereas HMI variant B was designed to achieve a low compliance with these guidelines (therefore named "low-compliant" variant B). Figure 3 shows three exemplary system outputs and how they differed between variant A and B in various scenarios.

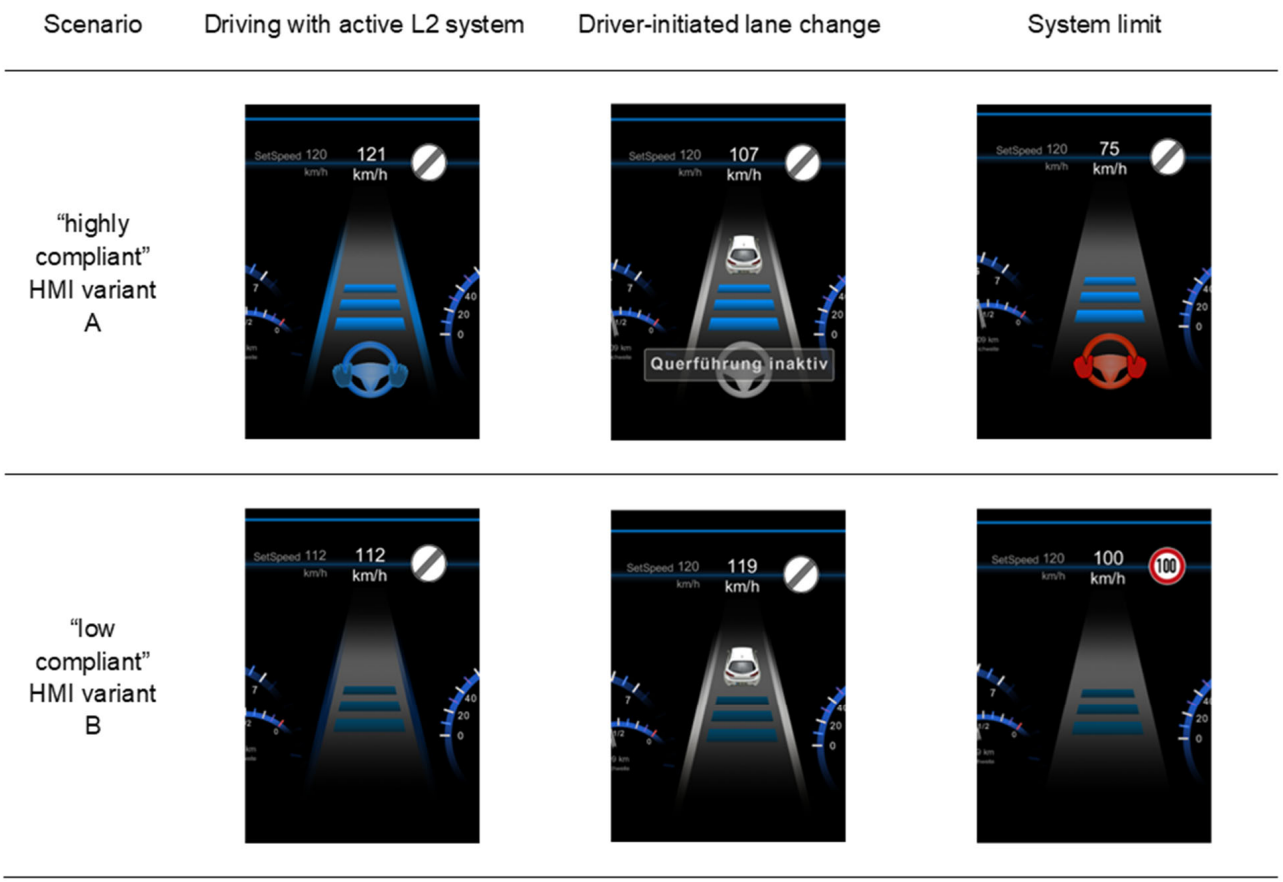

Figure 3. Exemplary scenarios in which the HMI variants differed with regard to system outputs.

The test course in the driving simulation consisted of different driving scenarios specified in the chapter "Definition of Relevant Test Scenarios". The test course was a two-lane highway with moderate traffic and a speed limit of 120 km/h. The participants were instructed to drive in the right lane. A sharp bend as well as a stationary vehicle in the ego-vehicle's lane (scenario "obstacle") served as critical use cases reaching the system limits of the L2 system. In the first scenario, "sharp bend", the necessary steering torque to manage the strong curvature could not be fully applied by the L2 system, resulting in a deactivation of lateral control. If the driver did not intervene timely enough, the vehicle would leave the lane towards the left and finally collide with the guardrail. The second scenario, "obstacle", consisted of a broken-down vehicle placed behind a hilltop on the right lane. The situation is announced to the participant by a reflective triangle in front of the hilltop (as in the case of a broken-down vehicle in real life). As the simulated sensors of the L2 system were not able to detect this stationary object, the L2 system would collide with the broken-down vehicle if the driver did not intervene by braking and/or steering. The test course took about 20 min to complete.

N = 24 drivers (11 female) recruited from the WIVW test driver panel participated in the study. The subjects were on average 41.2 years old (SD = 14.3 years). The experimental factor HMI-variant (A vs. B) varied between the subjects so that half of the drivers experienced HMI variant A, while the other half experienced variant B. Both groups had equal preconditions regarding their prior experience with assistance systems and automated driving systems from other simulator studies that the participants took part in. They did not differ in terms of driver age and driving experience.

All the test drivers experienced the following procedure: During a short manual drive, the drivers were able to get used to driving in the simulator again. After that, they read a short user manual which was adapted to the two HMI variants in terms of system operation and HMI displays. Afterwards, drivers received general instructions regarding the simulator drive and the usage of the rating scales. This was followed by the 20 min test drive. During the test drive, the participants experienced the different test scenarios and answered the survey questions directly after each scenario. The experimenter recorded the problems observed during system operation, driving behavior and monitoring behavior using the tablet app. A post-survey questionnaire (not included in the app) after the test drive contained questions on various HMI design aspects. In total, the study took approximately one hour. In addition to the observed behavioral data from the tablet, continuous driving data from the simulation software were analyzed for the two critical scenarios (minimum time-to-collision TTC to the broken-down vehicle in the scenario "obstacle", and maximum deviation from lane center during the event "sharp bend"). Only one rater evaluated all the trips, meaning that no interrater reliability could be calculated.

In brief, the study showed the following results:

- The designed differences in the two HMI variants affected drivers' behavior and subjective experiences of the system assessed via the S.A.D.E. app only to some extent. For most of the analyses, only a tendency towards the significant effects of the HMI variant was found. On the one hand, this could be due to the small sample size. On the other hand, it could be possible that some design issues did not affect driver behavior in such a significant way that real problems occurred which could be detected by the tool.
- The following results were observed on the level of single observational categories (effects with p-values < 0.15 are defined as tendentially significant, effects with p-values < 0.05 are defined as significant):
    - Effect of different warning strategies (with vs. without visual–acoustic warning) in scenario "sharp bend": In HMI variant B (without the warning), a tendency towards worse lane-keeping behavior was observed (i.e., a higher frequency of problems in the category "driving behavior" was coded; $p = 0.132$). The subjective evaluation of the drivers revealed greater problems in system understanding for HMI variant B ($p = 0.000$).
    - Effect of differences in system operation in scenario "first system activation" and scenario "deactivation": The more complex system activation and deactivation in HMI variant B resulted in a higher frequency of problems in the category "system operation" for HMI variant B (especially more frequently coded events in the category "support required by the experimenter" in scenario "first system activation": $p = 0.019$; a tendency towards this effect in scenario "system deactivation": $p = 0.140$). In addition, there was a tendency towards a higher perceived subjective difficulty for system activation in variant B ($p = 0.061$).
    - Effect of differences in visual contrast and visual indicators for the active lateral vehicle control subtask in the scenario "standby mode of lateral control": The lower distinctiveness of system states in HMI variant B did not lead to observable differences in driving performance. However, the drivers from HMI variant B subjectively reported a tendency towards greater problems in identifying the system status based on the HMI output ($p = 0.093$).

- Global rating of the experimenter per scenario: The global rating of the experimenter differed in the scenarios "first system activation" (tendentially significant; $p = 0.111$), "second system activation" (significant; $p = 0.024$), "system deactivation" (tendentially significant; $p = 0.064$) and "obstacle" (statistically significant; $p = 0.041$041). No differences were found in the scenarios "lane change", "standby mode of lateral control", "sharp bend" and "active driving with the system".
- HMI differences in the scenario "obstacle": In contrast to expectations, a tendency towards worse driving behavior was observed for drivers in the group with HMI variant A, i.e., a higher frequency of drivers produced endangerments in terms of too low a minimum distance from the obstacle ($p = 0.102$). System behavior and HMI outputs did not differ between the HMI variants in this scenario. One possible explanation for this result could be that the highly compliant HMI variant A resulted in overtrust in the system, leading to the impression of reliable system performance. As a result, drivers may have taken longer to realize that the system would not be able to handle the situation. However, this interpretation can currently only remain on a hypothetical level. Additional questions regarding driver trust would have helped to support this explanation.
- The identified trends in the observed driving behavior based on the categorial evaluation in the S.A.D.E. app corresponded with the analysis of measured continuous driving data:
    - The tendentially significant more frequent lane exceedances (observed and coded via the S.A.D.E. app; $p = 0.132$) corresponded with tendentially significant lower ratings of vehicle handling (rated by the experimenter via the S.A.D.E. app, $p = 0.145$) and with tendentially significant higher maximum measured lateral deviations in the scenario "sharp bend" (measured via the simulation software; see Figure 4 above for a comparison of the measures in the scenario "sharp bend", $p = 0.058$).
    - The significantly more frequent delayed braking reactions and endangerments (observed and coded with the S.A.D.E. app, $p = 0.012$) corresponded with a significantly lower rating of situation handling (rated by the experimenter via the S.A.D.E. app, $p = 0.024$) and with tendentially significant smaller minimum time-to-collision values in the scenario "obstacle" (indicating more critical scenarios; measured via the simulation software; see Figure 4 below for a comparison of the measures in the scenario "obstacle", $p = 0.058$).
    - These results indicate that categorial observation can partly replace the very time-intensive and resource-intensive analysis of time-based measures without losing too much information. This is an advantage of the method when used in studies with real vehicles, since it is usually time-consuming and costly to collect the necessary continuous driving data for an evaluation.

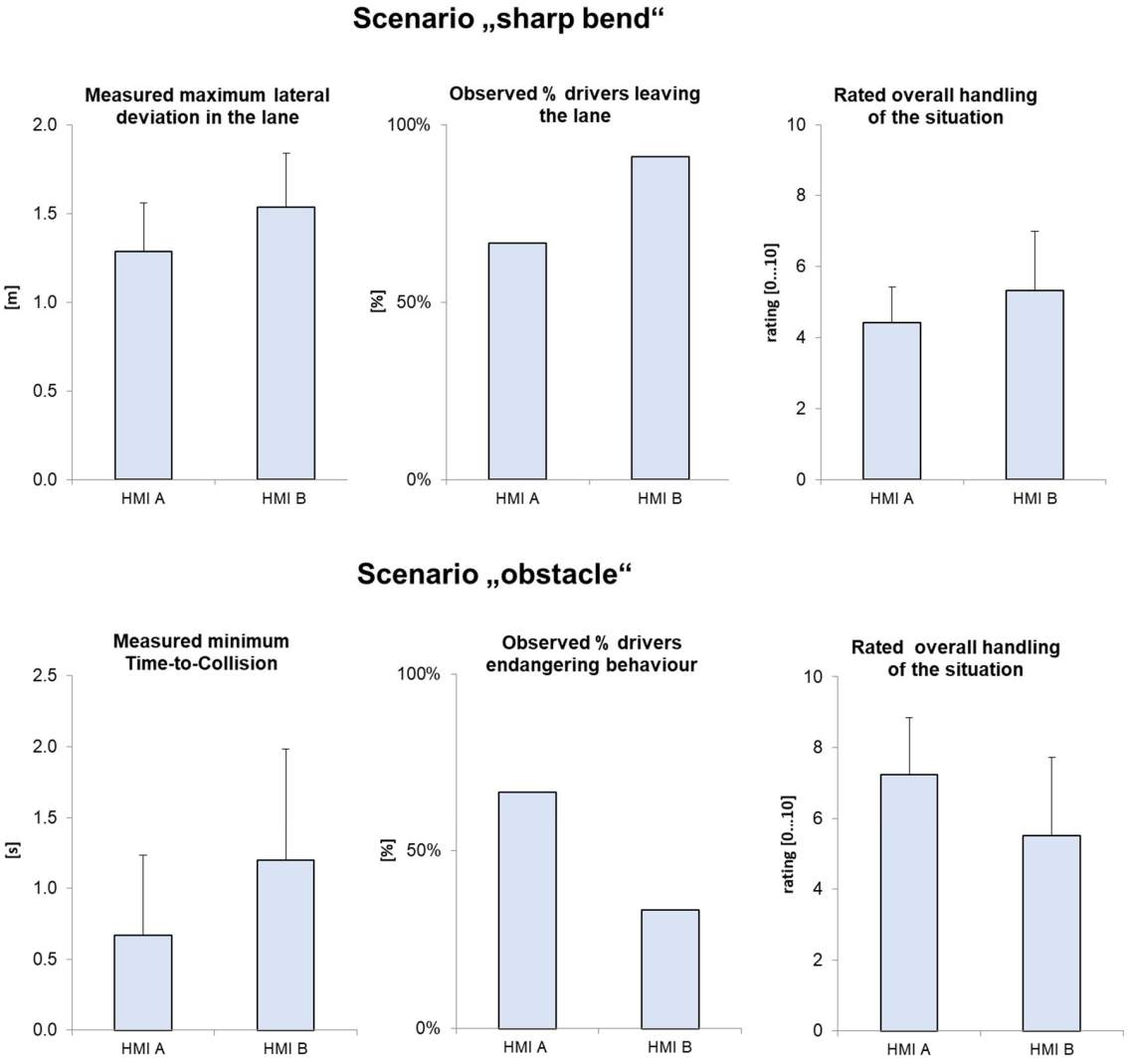

Figure 4. Comparison of measures for the two scenarios "sharp bend" (**above**: maximum lateral lane deviation; observed % of drivers leaving the lane; overall handling of the situation) and "obstacle" (**below**: minimum time-to-collision; observed % drivers showing endangering behavior; overall handling of the situation).

4. Results from the First Application of the Method in the BASt Driving Simulator

The method was also applied in a simulator study conducted at the Federal Highway Research Institute [17]. The authors explored the effects of different levels of quality of the lateral vehicle control of a partially automated system on drivers' interactions with the system. N = 56 test drivers participated in the study. The participants performed a 30 min test drive using a prototypical L2 system on a two-lane highway. The system status was indicated in a cluster display. Lateral control quality was manipulated between the subjects by two factors. The first factor was the variability of the lateral control. The systems differed in the extent that the system swerved within the lane, requiring the driver to steer cooperatively with the system without overriding it. There was a condition with "low variability", i.e., completely lane centered without any deviations, and another condition with "high variability", i.e., including deviations from the lane center to the right and to the left without leaving the lane. The second factor was the reliability of the lateral control (i.e., how often the lateral control completely fails for a longer period of time; resulting in the condition "high reliability" without any failures of lateral control vs. "low reliability" including system fails of 3–4 s every 30 s).

During the drive, the subjects experienced two scenarios including a system limit or a system malfunction in which they had to intervene to prevent a safety-critical situation (within a 12 min interval). The scenarios were the two that were already described and used in the WIVW driving simulator study (see Section 3): The first scenario was the scenario "sharp bend" in which lateral control was deactivated without any warning to the driver. For the participants, this deactivation could only be noticed from the change in the visual status indicator in the cluster HMI. The second scenario was the scenario "obstacle" where the system did not detect the stationary vehicle behind the hilltop. Again, the drivers received no warning. As the system itself was not able to detect this limit, the visual status indication did not change in the HMI.

Drivers' interactions with the system and especially drivers' reactions to the system limits were assessed using both objective driving data measured via the simulation software SILAB® as well as observed driver behavior via the S.A.D.E. app. Experimenters conducting the rating were kept blind to the experimental condition the subjects were assigned to.

The results of the study showed that a high reliability of the L2 system's lateral control (in terms of stable active system status without intermittent fails) worsened driver intervention performance at critical system limits both in the scenario "sharp bend" (measured by tendentially higher lateral deviations) and in the scenario "obstacle" (measured by tendentially lower minimum TTC). This result occurred despite the fact that drivers had been informed about the system limits beforehand. The factor variability in lateral guidance had no effects on driver intervention behavior in the study. Furthermore, the effects could also be observed in the experimenters' ratings via the S.A.D.E. app. Correlations between the measures were quite high: The correlation between the SDLP and observer rating was $r = 0.80$ ($p < 0.01$), and the correlation between minimum TTC and observer rating was $r = -0.56$ ($p < 0.01$).

The interrater reliability was calculated separately for the five studied scenarios with the Spearman rank order correlation. It reached values between $r_s = 0.602$ (scenario "second activation") and $r_s = 0.742$ (scenario "obstacle").

5. Summary

Taking the results from these initial simulator studies, the developed method is a promising tool for the evaluation of human–machine interaction in partially automated driving in the context of user studies. It showed a high although not fully perfect correlation with objective continuous driving data. The method has the advantage that it can be used in different test environments, such as a driving simulator, test track or in field studies. Instead of an elaborate analysis of continuous driving data, which requires complex measurement equipment, especially in real vehicles, the focus lies on a standardized observation of driver behavior in real time during system use. It was shown that statements can be made about driving behavior, system operation and monitoring performance, which can provide valuable information about the quality of an L2 system and its HMI in terms of usability and safety. The method can be used as a paper–pencil tool applying the proposed observational categories and the rules for the derivation of the global rating for each investigated scenario. The observation, logging and analysis of the data can be facilitated by the tablet-based S.A.D.E. application, but this is not mandatory.

By specifying a defined test procedure, test scenarios and error categories, a standardized application of the method is ensured. The global rating procedure is based on clearly defined rules, which makes the method objective. Therefore, the tool seems well suited for use in different institutions, such as universities, OEMs and testing institutions for consumer protection purposes.

By using the method in the context of user studies, it can be evaluated whether assessments of the degree of fulfillment of design guidelines based on the literature actually have practical relevance to driver interaction with a system and its HMI. The inclusion of subjective assessments by drivers can provide additional clues as to how well an HMI is able to convey an appropriate understanding of driver responsibilities and subjectively

perceived system functionalities, which is the basis of maintaining an appropriate level of mode and situation awareness.

The implementation of realistic test scenarios allows an estimation of system- and HMI-related aspects of driving safety.

6. Discussion of Limitations and Future Challenges

The current version of the tool nevertheless has some limitations and future challenges. The observation categories used with regard to system operation, driving performance and driver monitoring were defined based on the authors' expertise. The validation of the method in the two simulator studies focused predominantly on the two categories of system operation and driving behavior. It was found that the observation criteria defined in these categories were largely suitable for identifying the problems arising from inappropriate HMI design, which are serious enough that they also manifest in observable behavior. In this respect, the method seems to be valid for the detection of major problems in interaction with an existing system in a fast and efficient way. However, the method cannot accomplish the detection of relatively small differences in system design. These differences may need to be evaluated during the development process of a system. Otherwise, more detailed analyses, e.g., of gaze behavior or reaction times, are recommended.

The objectivity and reliability of the method will strongly depend on the degree to which a thorough rater training (about the general method, but also for a specific study) takes place. However, the objectivity of the method will always be lower than when a quantitative analysis of behavioral and vehicle data is performed. In addition, depending on environmental conditions, some behaviors may be difficult to observe (e.g., lane keeping if a bird's eye view is not available in a real traffic experiment). Likewise, the requirements for observing various problems simultaneously may overload the rater, causing them to overlook individual aspects.

Conversely, the method has the potential to observe and identify certain more complex cause–effect relationships that would not be apparent in the analysis of measured objective data. For example, problems with system activation are usually expressed as a prolonged reaction time until successful system activation. Whether this occurs due to a complex search for the correct operation element or due to an initially incorrect system operation can be identified far more quickly by observation.

Furthermore, there is a clear need for the optimization of the precise definition of the evaluation criteria for appropriate monitoring behavior of the drivers. In the short-term studies reported in this paper, problems in monitoring tended to be neglected or did not occur (the drivers were highly focused, were not distracted by any additional secondary activities and always had their hands on the steering wheel). The operationalization of the error category "not attentive enough" is currently still relatively vaguely defined. In the future, this will require the use of more clearly defined criteria, e.g., gaze behavior and a meaningful approach on the basis of which an attentive and situation-aware driver can be recognized (both based on data from eye tracking measurement but also purely from observation).

Kircher and Ahlström [22] list expert judgments as one possible evaluation method for the assessment of the minimum required attention to the driving task aside from other evaluation methods, such as the usage of eye tracking sensors to directly measure drivers' glance directions and eye blinks. They cite a publication [23] which includes recommendations for eye scanning rules. The named examples are: "keep the eyes moving", "scan the entire traffic scene", "center the gaze on the travel path" and "look at mirrors and instruments" [22]. However, they found that the recommendations of driving instructors in a laboratory condition regarding how attention should be distributed on a highway were not reliable for the assessment of driver attentional distribution in traffic if the situational context was unknown.

What the method cannot provide is a detailed explanatory model of why certain problems occur during system interaction, since only behavior itself is observed. If conclusions

are to be drawn about the underlying psychological mechanisms, it is advisable to supplement the method with other methods that can go deeper into the analysis of psychological processes, such as precise analyses of gaze and reaction times or the addition of specific surveys of drivers, e.g., about their mental model, system trust and acceptance, etc.

Lastly, the use of the method has so far been limited to simulators. Initial tests in real traffic had mainly focused on the applicability of the tool. Currently, the method is also used and validated in real vehicles for the evaluation of driver–vehicle interaction in L2 systems by BASt. The results will be made available to the public and used to further develop the method.

Author Contributions: N.S. and K.W.: project administration, study conceptualization, methodology and validation; N.S.: writing—original draft preparation; A.N.: supervision and funding acquisition; A.W.: conceptualization and validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Federal Highway Research Institute within a project with the overall goal of developing evaluation methods for the assessment of human–machine interaction for partially automated driving functions (Grant Number: FE82.0708).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *J3016_202104*; SAE On-Road Automated Vehicle Standards Committee. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. SAE International: Warrandale, PA, USA, 2021.
2. Bainbridge, L. Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems*; Elsevier: Amsterdam, The Netherlands, 1983; pp. 129–135.
3. National Transportation Safety Board. *Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck near Williston, Florida, May 7, 2016*; Highway Accident Report NTSB/HAR-17/02; National Transportation Safety Board: Washington, DC, USA, 2017.
4. Schaller, T.; Schiehlen, J.; Gradenegger, B. Stauassistenz–Unterstützung des Fahrers in der Quer- und Längsführung: Systementwicklung und Kundenakzeptanz. In *3 Tagung Aktive Sicherheit durch Fahrerassistenz*; Lehrstuhl für Fahrzeugtechnik, Technische Universität: München, Germany, 2008.
5. Kircher, K.; Larsson, A.; Hultgren, J.A. Tactical driving behavior with different levels of automation. *IEEE Trans. Intell. Transp. Syst.* **2014**, *1*, 158–167. [CrossRef]
6. Begiatto, M.; Hartwich, F.; Schleinitz, K.; Krems, J.; Othersen, I.; Petermann-Stock, I. What would drivers like to know during automated driving? Information needs at different levels of automation. In *7. Tagung Fahrerassistenzsysteme*; Lehrstuhl für Fahrzzeugtechnik, Technische Universtität: München, Germany, 2015.
7. Large, D.R.; Banks, V.A.; Burnett, G.; 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; pp. 20–22.
8. Llaneras, R.E.; Salinger, J.; Green, C.A. 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, Bolton, NY, USA, 17–20 June 2013.
9. Naujoks, F.; Purucker, C.; Neukum, A. Secondary task engagement and vehicle automation—Comparing the effects of different automation levels in an on-road experiment. *Transp. Res. Part F* **2016**, *38*, 67–82. [CrossRef]
10. Kleen, A.T. Beherrschbarkeit von teilautomatisierten Eingriffen in die Fahrzeugführung. Ph.D. Thesis, Technische Universität Braunschweig, Braunschweig, Germany, 2014.
11. 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. Part F Traffic Psychol. Behav.* **2019**, *62*, 870–882. [CrossRef]
12. Stanton, N.A.; Young, M.S.; Walker, G.H.; Turner, H.; Randle, S. Automating the driver's control tasks. *Int. J. Cogn. Ergon.* **2001**, *5*, 221–236. [CrossRef]
13. Wulf, F.; Zeeb, K.; Rimini-Döring, M.; Arnon, M.; Gauterin, F. Effects of human-machine interaction mechanisms on situation awareness in partly automated driving. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013; pp. 2012–2019.
14. Naujoks, F.; Wiedemann, K.; Schömig, N.; Jarosch, O.; Gold, C. Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* **2018**, *5*, 579–592. [CrossRef] [PubMed]

15. Naujoks, F.; Hergeth, S.; Wiedemann, K.; Schömig, N.; Forster, Y.; Keinath, A. Test procedure for evaluating the human–machine interface of vehicles with automated driving systems. *Traffic Inj. Prev.* **2019**, *20*, 146–151. [CrossRef] [PubMed]
16. National Highway Traffic Safety Administration. *Federal Automated Vehicles Policy 2.0. Washington, DC: National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT)*; National Highway Traffic Safety Administration: Washington, DC, USA, 2017.
17. Wiggerich, A.; Hoffmann, H.; Schömig, N.; Wiedemann, K.; Segler, K. Bewertung der Sicherheit der Mensch-Maschine-Interaktion teilautomatisierter Fahrfunktionen (Level 2). In Proceedings of the Beitrag auf dem 13, Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren, digital conference, 16–17 July 2020.
18. Menold, N.; Bogner, K. Gestaltung von Ratingskalen in Fragebögen. *Mannh. GESIS—Leibniz-Inst. Für Soz. (SDM Surv. Guidel.)* **2015**. [CrossRef]
19. Neukum, A.; Lübbeke, T.; Krüger, H.-P.; Mayser, C.; Steinle, J. ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In Proceedings of the Workshop Fahrerassistenzsysteme—FAS2008, Walting, Germany, 2–4 April 2008; Maurer, M., Stiller, C., Eds.; fmrt Karlsruhe: Walting, Germany, 2008; pp. 141–150.
20. Kaussner, Y. Assessment of driver fitness: An alcohol calibration study in a high-fidelity simulation. In Proceedings of the Fit to Drive 7th International Traffic Expert Congress, Berlin, Germany, 25–26 April 2013.
21. Kenntner-Mabiala, R.; Kaussner, Y.; Jagiellowicz-Kaufmann, M.; Hoffmann, S.; Krüger, H.-P. Driving performance under alcohol in simulated representative driving tasks: An alcohol calibration study for impairments related to medicinal drugs. *J. Clin. Psychopharmacol.* **2015**, *35*, 134–142. [CrossRef] [PubMed]
22. Kircher, K.; Ahlstrom, C. Evaluation of methods for the assessment of attention while driving. *Accid. Anal. Prev.* **2018**, *114*, 40–47. [CrossRef] [PubMed]
23. Zwahlen, H.T. Eye scanning rules for drivers: How do they compare with actual observed eye scanning behaviour. *Transp. Res. Rec. J. Transp. Res. Board* **1991**, *1403*, 14–22.