Mediator

Overview of MEDIATOR test-protocols with respect to cost, efficiency, and validity

Deliverable D4.4 – WP4 – PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 814735.

Overview of MEDIATOR test-protocols with respect to cost, efficiency and validity

Work package 4, Deliverable D4.4

Please refer to this report as follows:

Karlsson, J., Thalya, P., Toffetti, A., Fiorento, A. (2023). Overview of MEDIATOR test-protocols with respect to cost, efficiency, and validity, Deliverable D4.4 of the H2020 project MEDIATOR.

Project details:	
Project start date: **Duration:** **Project name:**	01/05/2019 48 months MEDIATOR – MEdiating between Driver and Intelligent Automated Transport systems on Our Roads
Coordinator:	Prof Dr Nicole van Nes | SWOV – Institute for Road Safety Research Bezuidenhoutseweg 62, 2594 AW, The Hague, The Netherlands
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 814735.

Deliverable details:	
Version: **Dissemination level:** **Due date:** **Submission date:**	Final PU (Public) 30/04/2023 30/04/2023

Lead contractor for this deliverable:

Johan Karlsson – Autoliv Development AB

Report Author(s):	Karlsson, J. (Autoliv), Sweden Thalya, P. (Veoneer), Sweden Toffetti, A., Busiello, M. (CRF), Italy Fiorentino, A. (FCA Italy), Italy

Revision history

Date	Version	Reviewer	Description
03/04/2023	Reviewed draft	Avinoam Borowsky – Ben-Gurion University of the Negev	Internal review
25/04/2023	Reviewed draft	Linda Boyle, HFSM Lab UW	External review
30/04/2023	Final draft	Elmer van Grondelle (WP4 leader) – Delft University of Technology Ernst Verschragen (QA officer) – SWOV Institute for Road Safety Research	Final editing and checks
30/04/2023	Final deliverable	Michiel Christoph	Approved

Legal Disclaimer

Table of contents

List of Abbreviations

AD	Autonomous Driving
AI	Artificial Intelligence
ADS	Automated Driving System
CM	Continuous Mediation (automation level)
COTS	Commercial Off-The-Shelf
DDT	Dynamic Driving Task
DL	Decision Logic
HMI	Human Machine Interface
HF vehicle	Human Factors prototype vehicle
IIHS	Insurance Institute for Highway Safety
KPI	Key Performance Indicators
KSS	Karolinska Sleepiness Scale
NDRA	Non-Driving Related Activities
ODD	Operational Design Domain
OEDR	Object and Event Detection and Response
SAE	Society of Automotive Engineers
SB	Driver Standby (automation level)
TI vehicle	Technology Integration prototype vehicle
TOC	Transition of Control
TtS	Time To Sleep (automation level)
UC	Use Case
UN	United Nations
UNECE	United Nations Economic Commission for Europe
VR	Virtual Reality

List of Figures

List of Tables

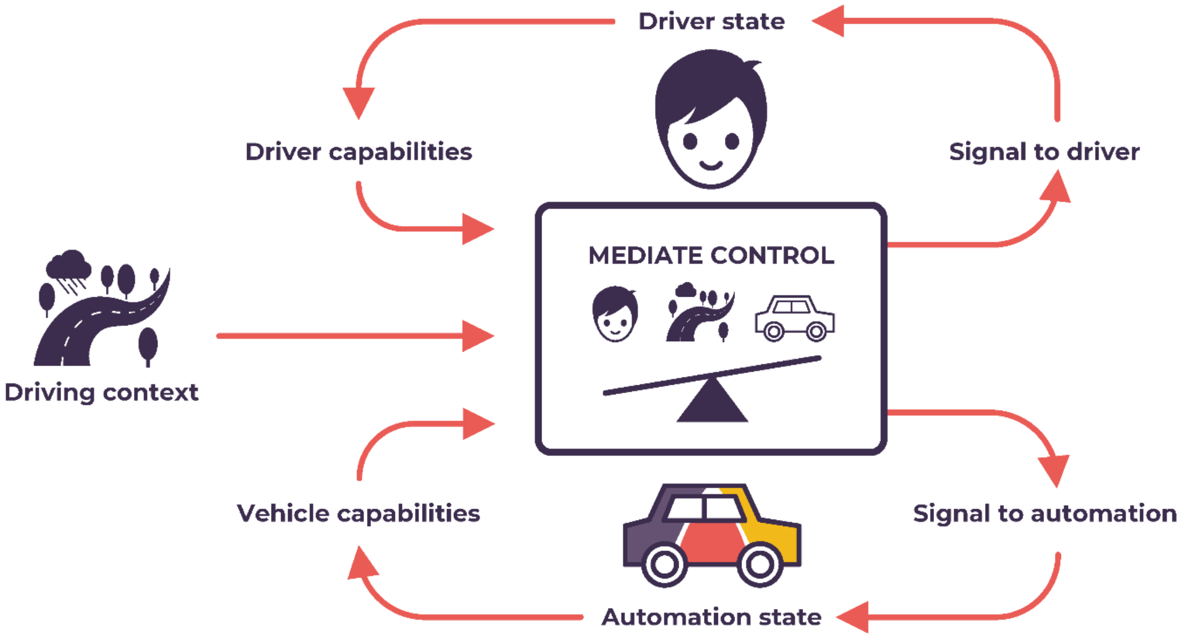
About MEDIATOR

MEDIATOR, a 4-year project coordinated by SWOV Institute for Road Safety Research, has come to an end after four years of hard work. The project has been carried out by a consortium of highly qualified research and industry experts, representing a balanced mix of top universities and research organisations as well as several OEMs and suppliers.

The consortium, supported by an international Industrial Advisory Board and a Scientific Advisory Board, represented all transport modes, maximising input from, and transferring results to aviation, maritime and rail (with mode-specific adaptations).

Vision

Automated transport technology is developing rapidly for all transport modes, with huge safety potential. The transition to full automation, however, brings new risks, such as mode confusion, overreliance, reduced situational awareness and misuse. The driving task changes to a more supervisory role, reducing the task load and potentially leading to degraded human performance. Similarly, the automated system may not (yet) function in all situations.



The MEDIATOR system will constantly weigh driving context, driver state and vehicle automation status, while personalising its technology to the drivers' general competence, characteristics, and preferences.

The MEDIATOR project aimed to develop an in-vehicle system, the Mediator system, that intelligently assesses the strengths and weaknesses of both the driver and the automation and mediates between them, while also taking into account the driving context. It assists the timely take-over between driver and automation and vice versa, based on who is fittest to drive. This Mediator system optimises the safety potential of vehicle automation during the transition to full (level 5) automation. It would reduce risks, such as those caused by driver fatigue or inattention, or on the automation side by imperfect automated driving technology. MEDIATOR has facilitated market exploitation by actively involving the automotive industry during the development process.

To accomplish the development of this support system MEDIATOR integrated and enhanced existing knowledge of human factors and HMI, taking advantage of the expertise in other transport modes (aviation, rail and maritime). It further developed and adapted available technologies for real-time data collection, storage and analysis and incorporated the latest artificial intelligence techniques. MEDIATOR has developed working prototypes, and validated the system in a number of studies, including computer simulation, virtual reality, driving simulator and on-road studies.

With MEDIATOR we further paved the way towards safe and reliable future vehicle automation that takes into account who is most fit to drive: the human or the system.

https://mediatorproject.eu/

Executive summary

This report starts out with a brief orientation of the Mediator system, what it is, its constituent modules and the functions it performs as well as the key performance indicators (KPIs) that were decided as being applicable in the beginning of the project.

The Mediator system has **four constituent modules**, The **Driver Module** for assessment of human driver driver fitness and unfitness, The **Automation Module** for assessing the Driving Automation Health, the **Decision Logic** Module which decides on appropriate actions based on Driver State, Automation Health, and current traffic context, and finally the **HMI Module** which is responsible for the communication between the driver and the Mediator system.

The **Driver Module KPIs** include sensitivity / accuracy analysis and other detection theory analysis/confusion matrix based measures using self-assessed or annotated driver state as ground truth.

The Automation Module KPIs includes to check whether time to automation (un-)fitness and availability of the automated system is correctly evaluated based on its sensing limitations, and if identification of relevant infrastructure (road markings, road signs, etc.).is correctly predicted compared to observations, and largely similar to the KPIs for the Driver Module. Further analysis can build on correlations between the automation internal states, driving context and the automation fitness. It is worth noting that the KPIs for driver and automation state assessment are not by themselves prohibiting nor supporting low-cost evaluation, since the collected values are dependent on proper data, such as driver state condition and proper sensory data for automation state.

The **Decision Logic Module KPIs** include Number of safety critical events, Total time in safety critical situation, Average duration of safety critical events, Number of times no solution was found, Total time in an uncomfortable situation, Average duration of and uncomfortable event – where a lower number indicates better performance and finally Number of events where comfort was improved where a higher number indicates higher performance. Generating data for these analyses in a computer simulation is fairly cost-effective and can be used for initial tweaking of decision logic parameters.

The **HMI Module KPIs** are Usability, Acceptance, Workload, Trust, User experience, Mode awareness, and Overreliance, all assessed by various psychometric scales and validated ranking/rating scales. These are fairly simple to collect and analyse and can be used with prototypes on different levels of fidelity and are thus well suited for low-cost testing.

A brief literature review outlines potential methods for testing of the constituent modules, with an emphasis on simulator-based testing and on-road Wizard-of-Oz testing. Following this is a brief review of three independent rating protocols for current automated driving systems, the US Consumer Reports Grading of Active Driving Assistance Systems, the US Insurance Institute for Highway (IIHS) Safety Safeguards for Assisted Driving, and the Euro NCAP protocol for Grading of Assisted Driving.

The report then provides a review of test methods used for development within the project and the merits of each method in terms of cost vs validity in relation to the KPIs, where information was gathered from test leaders responsible for the six main user studies within MEDIATOR.

Section 3 reviews in more detail one of the independent rating protocols, performs a gap analysis of what is missing from that protocol to cover the full intended functionality of the Mediator system and proposes to extend the Euro NCAP Assisted Driving Grading Protocol with a procedure to evaluate transitions-of-control (ToC) both from the driver to the automation, and from the automation to the driver, covering all combinations of transitions in terms of initiating party and direction of transition, and evaluating them for acceptance, trust, and user behaviour modification.

Some evaluation methods used in on-road tests within MEDAITOR are also proposed to support this purpose and to cover the gaps. In the final discussion a few extensions to the methods deployed within the project are proposed, such as improved wizard-of-oz testing, system-in-loop-simulations, and even extending early usability and user experience (UX) studies into the metaverse.

It is concluded that while at least initial tests of decision logic and HMI, as well as some sensor level evaluations of driver state sensing are well suited to low-cost testing, accurate validation of driver state sensing requires driver state induction in the true driving setting and thus labour intensive and thereby costly testing procedures. Furthermore, the full system validation can only be performed on a full functional prototype or (pre-) production vehicle.

1. Introduction

Road tests, involving both closed track and public roads, have been the predominant tool for the development of automotive systems but advances in modelling and Virtual Reality (VR), as well as a reduced time for developing new major vehicle systems, makes simulations a safe and cost-effective testbed for Automated Driving (AD) and the MEDIATOR project. In this deliverable the possibilities of low-cost test capabilities available today for the MEDIATOR project are summarized. We focus on the immediate challenges facing test and evaluation departments and the improvements necessary to meet the future needs of the automotive community. Furthermore, we discuss how we by shifting from the question of how to justify the costs of sustaining standard test methodologies and facilities to the question of how to transform test methodologies and facilities to reduce the cost and time required to field test systems like the Mediator.

1.1. Aim/Objectives of the deliverable

We foresee a future light vehicle market with an increased number of vehicles with some level of driver assistance and a monitoring agent matching the current situation to the vehicles Operational Design Domain (ODD), deciding whether to allow a certain level of assistance or even automation. Such vehicles will in this report be referred to as mediator-like vehicles. Assuming this development, we also foresee an increased need to assess the performance of such vehicles e.g., for consumer organizations such as Euro NCAP, Insurance Institute for Highway Safety (IIHS), Consumer Reports and similar.

The aim of this deliverable is to give a short introduction to the test methods deployed in MEDIATOR, discussing their merits in terms of validity versus cost, as well as a short review of existing independent assessment protocols used for rating assisted driving systems. Using the existing assessment protocols, we perform a gap analysis and propose what additional tests would need to be amended to the existing protocols to cover all aspects of the Mediator system.

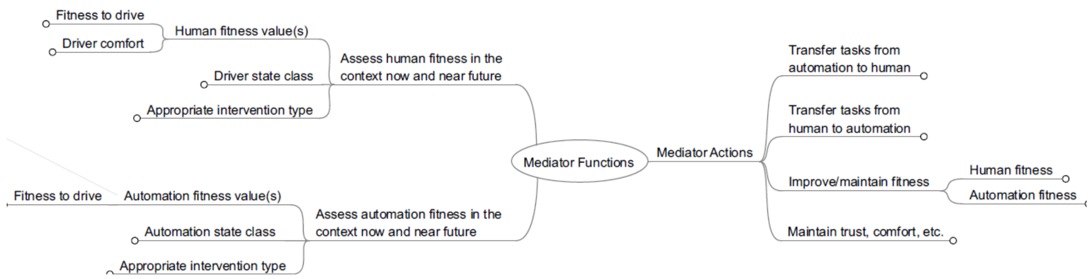


Figure 1.1 The functional requirements structure of the Mediator functions - what can be tested in a low-cost lab setting?

Since such testing will be performed on the full and final production (or pre-production) vehicle the testing protocol proposed here, will not take prototype development cost into account.

In the final discussion part of this deliverable, we provide a ranking of the relative cost of the utilized methods and compare the benefit/contribution in different stages of the project. Furthermore, in light of this we also propose a few future extensions to the methods utilized in the MEDIATOR project for future research projects, and a short commentary about what subcomponents of the Mediator system lends themselves best to low-cost testing due to the nature of the studied systems.

Delimitations

This deliverable will not deal with topics for type approval such as testing for UNR79, R157 and similar as this is beyond the scope of this deliverable.

1.2. Mediator defining features – MEDIATOR modules and KPIs

The Mediator system is intended to facilitate the transition from fully manual driving to fully automated driving by helping drivers understand the current driving mode and when it is prudent to change to a higher or lower automation level. The system consists of four main functional modules outlined below. For a more detailed description of the MEDIATOR modules see Cleij et al (2020) The modules are outlined below to help the reader understand what makes the MEDIATOR architecture unique, and why testing a mediator-like vehicle goes beyond just testing automated vehicles in general, since the value of MEDIATOR is greater than the sum of its parts.

1.2.1. Driver Module

The functionality of the driver module is to estimate the time to driver (un)fitness, and time to driver (dis)comfort to evaluate who is fit to drive. In addition, module also provides information on possible intervention strategies, in MEDIATOR labelled preventive and corrective mediation, that could improve the driver fitness or comfort and estimate the resulting fitness levels. The module also communicates with decision-logic module, information on drivers reduced fitness, driver state, context and reason for the reduced driver fitness. The work describing the driver module and algorithms are detailed in Borowsky (2020).

1.2.2. Automation Modules

The automation module is the interface between the Mediator system and the vehicle automation functions. The Module communicates to the vehicle automation and gathers additional information from the automation for them to be used in other modules. The main functionality of this module is:

- Provide automation state information
- Adjust the automation state
- Provide data for the context module

The main function of the module is not to improve the performance of the automation system in the vehicle, rather to predict when the degradation of the automation functionality. More details related to the module can be found in Mano et al., (2021).

1.2.3. Decision-logic module

The main functionality of this software module is to combine and translate information revived from driver and automation state module into Mediator action request, which is required for safe and comfortable driving. Mediator action request are either sent through HMI or automation module for them to be executed. The module has following subfunctions:

- Decide who is fittest to drive based on optimizing safety and comfort
- Select actions that enable the fittest to perform their (part of the) driving task
- Monitor action execution
- Adjust action or action parameters if needed.

Decision logic and other functional parameters are described in detail by Cleij (2020)

1.2.4. HMI module

The communication between vehicle related system, Mediator system and driver, dedicated Human-Machine Interface (HMI) was designed in MEDIATOR work package 1 and described further by Cleij (2020). The development of HMI module was based on the holistic approach (Christoph, et al., 2019). The Mediator HMI module, facilitates and manages all interaction between human and vehicle for both primary, driving related tasks as well as for secondary tasks (e.g., climate, entertainment). The main functionality of the HMI module is,

- Supporting conventional driving tasks,
- Facilitating negotiations between driver and automation,
- Guiding control transfers between driver and automation (takeovers),
- Informing on CM switch off/switch on,
- Executing corrective actions to increase driver fitness,
- Executing preventive measures to maintain driver fitness,
- Detecting and facilitating driver preferences and momentary inputs
- The HMI should also adhere to several non-functional requirements such as having high usability and transparency for its users and improving driver comfort and safety

1.3. Low-cost testing in MEDIATOR

Reducing cost at all levels without compromising safety and validity of the results, requires to consider the entire development and testing procedure. In the Human-Centred Design process, iterative testing in different kind of environment swith varying fidelity levels, make it possible to calculate the involved costs. In fact, early in the process it may make sense to use a quick test with experts and/or few participants. Using off the shelf components for applications and open-source software can also make a cost-effective development in the early-stage of testing and development. With the maturity of the concept increasing (medium and high-fidelity prototypes) the complexity and cost of testing procedures and equipment increase as well.

The use of driving simulation has multiple potential benefits for testing, both in terms of safety and cost. However, full-scale moving base simulators are expensive to build and maintain. In MEDIATOR several levels of testing methods have been used, from simple desktop simulations up to on-road studies with naive participants. The cost effectiveness the study methods employed within the MEDIATOR project is not envisioned. However, to evaluate Mediator systems, an initial set of KPIs was listed to test its functionality. Upcoming sections will give an overview of the KPIs and use cases used in testing.

1.3.1. Overview of Key Performance Indicators as ground for tests

The MEDIATOR project defined the following Key Performance Indicator (KPI) early in the project, to facilitate testing and development of the various modules. More details and reasoning for such KPIs are described by Cleij et al (2020).

KPI for Driver Module

The driver module gives an estimate for time to driver fitness and unfitness. The performance of these estimates can be evaluated based on KPIs such as,

- Confusion matrices and Sensitivity /accuracy analysis using detected driver state as ground truth.
- Subjective measures such as KSS fatigue scores and Receiver operating characteristic curves can be used to understand and visualize the results of the measurements.

- Thresholds for fitness/unfitness can be based on literature and/or assessed via driving performance measures such as time to collision and lane deviation.

KPI for Automation state module

The automation state module outputs time to automation fitness and unfitness, automation state class, current automation level and relevant context information. The KPIs for assessment could be to check whether time to automation (un)fitness is correctly predicted versus what was observed, thus leading to evaluating standard metrics such as true/false positive/negative. To do such an assessment and analysis, data collection plays an important role. The analysis of collected data can build the correlations between the automation internal states, driving context and the automation fitness. Further the assessment of automation performance was detailed in Mano et al (2020), where availability if the automated system is evaluated based on its sensing limitation (lighting conditions, weather), identification of relevant infrastructure (road markings, road signs, etc.). These KPIs were evaluated in the study carried out with the prototype developed in MEDIATOR. The details on the protypes can be found in deliverable Fiorentino A et al. (2022).

KPI for Decision-logic module

Similar to the assessment of automation state module. The KPIs for the decision-making module DL (Decision Logic) can be assessed based on the output for given use cases. Some KPIs have been proposed by Cleij et.al. (2020) and can be linked to the use cases described in the MEDIATOR project. The proposed KPIs address the performance of the module during different critical events to assess the quality of action recommended by the module. Proposed KPI’s was to evaluate the module were based the use cases defined with dummy inputs. Proposed KPI’s to evaluate the module from Cleij et.al. (2020) are listed below. They evaluation is carried out by feed the module dummy variables to evaluate the performance. The testing carried out as part of the integration effort of the decision logic module is detailed in Bakker, B et al. (2022) and Athmer, C et al. (2022)

- Number of safety critical events
  - Described by an area on the driver/automation fitness plane based on the criticality – Low number of events are desired for high performance
- Total time in safety critical situation
  - The total time during the simulation that the combination of driver and automation fitness measures were classified within one of the levels of safety criticality – low time spent
- Average duration of safety critical events
  - The average time per safety critical event spend in a state where the combination of driver and automation fitness measures were classified – low duration
- Number of times no solution was found
  - The number of times that the decision logic algorithm did not converge to a viable solution throughout the simulation – low number of times
- Number of events where comfort was improved
  - The number of times the decision logic selected an action to improve driver comfort – Higher number of events
- Total time in uncomfortable situation
  - The total time spend in a situation of which the Mediator system expects the driver to feel discomfortable – low total time
- Average duration of uncomfortable event
  - The average time per event the driver spends in an uncomfortable state – low number of actions
- Number of actions
  - Total number of actions initiated by the decision logic – low number of times
- Total time without actions

- Total time during the complete simulation where no action was being executed – long total time
- Average time between actions
  - Average time between the end of a first action and the start of a second action – long average time

KPI for HMI Module

Most of the KPIs described for the HMI module follow conventional driving task HMI functions, such as guiding takeovers, negotiations between the driver and the automation, executing corrective and preventive actions. To assess the performance the module the following key performance indicators can be used:

- Usability: e.g., System Usability Scale (Brooke, 1986)
- Acceptance: e.g., Van der Laan Acceptance Questionnaire (Van der Laan, 1997)
- Workload: e.g., NASA RTLX (Hart,1998)
- Trust: e.g., Automation Trust Scale (Jian, 2000) and/or measures based on gaze behaviour (Hergeth, 2016)
- User experience: e.g., User Experience Questionnaire (Laugwitz, 2008)
- Mode awareness: e.g., similar to the subjective measures described in Kurpiers (2020)
- Overreliance: e.g., similar to the objective measures described in Kurpiers (2020)

1.3.2. Use Cases in MEDIATOR

Ten use cases as were composed early in the project to evaluate the full potential of the Mediator system, . All testing scenarios in the project are composed out of these use-cases. The ten use cases are described further detail by Cleij et.al. (2020) and outlined here below. The main scope of defining the scenarios were limited to,

- Urban and highway driving scenario
- Detection/mitigation of specific types of driver distraction and fatigue:
  - Distraction: focus on eyes off road (and hands-off steering wheel)
  - Fatigue: focus on detection of both ‘recoverable’ fatigue (which is typically task- related, e.g., due to underload or overload) and ‘non-recoverable’ fatigue (which is typically sleep-related, e.g., due to sleep deprivation). However, mitigation actions/strategies did not cover degradation due to severe sleep-related fatigue

Considering that both safety and comfort were the focus of the MEDIATOR project, the selected ten uses cases were classified as mainly safety (red or green) or mainly comfort (blue) related. A distinction is also made between safety related use cases that describe takeovers between automation levels (red) and those that describe driving within one level of automation (green). The ten use cases are visualised in Figure 1.2.

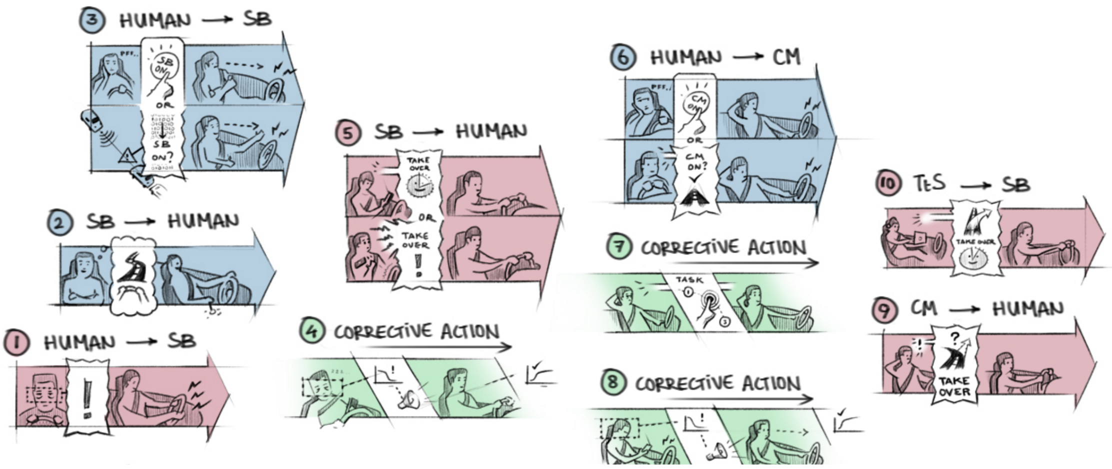


Figure 1.2 Mediator Use Cases (Extract from Cleij et.al. (2020))

A summary of the use cases is described below:

1. Mediator system initiates takeover (human to automation): Degraded human fitness, the system reacts by initiating a takeover to automation.
2. Driver takes back control: The driver uses the HMI to indicate a desire to take back.
3. Comfort takeover (human to automation): Either the driver (a) or the Mediator system (b) initiates a takeover from human to automation.
   a) The driver indicates via the HMI that he/she is not motivated to drive. The Mediator system reacts by confirming the automation fitness and guiding the takeover.
   b) The Mediator system detects an event (e.g., receiving a text message or an upcoming traffic jam). The system reacts by suggesting a takeover to automation.
4. Corrective Action (Driver Standby - SB): While driving in SB the human driver becomes drowsy. The Mediator system reacts by initiating an action to improve the driver fitness and monitors the effect.
5. Mediator initiated takeover (automation to human): A planned (a) or an unplanned (b) takeover from automation to human is initiated by the Mediator system.
   a) The automation indicates that the current route leads to automation unfitness as it will leave its operational design domain. The Mediator system reacts by preparing the driver for and guiding the driver through a non-urgent takeover.
   b) The automation indicates that its fitness is rapidly degrading and can soon no longer perform the driving task. The Mediator system reacts by informing the human driver and guiding the urgent takeover.
6. Comfort Continuous Mediation (CM) switch on: Either the driver (a) or the Mediator system (b) switches on driving in CM.
   a) The driver indicates via the HMI that he/she is not motivated to drive. The Mediator system reacts by confirming the automation fitness and switches on CM.
   b) The Mediator system detects sufficient fitness for driving in CM from which it concludes that the driver comfort could be improved and reacts by suggesting switching on CM.
7. Preventive Action (CM): While driving in CM, the driver is supported by the Mediator system in performing the monitoring task. The system does this by trying to prevent underload and keeping the driver in the loop.

8. Corrective Action (CM): While driving in CM, degraded driver fitness is detected by the Mediator system. The system reacts by initiating a corrective action to improve driver fitness.
9. CM shuts off instantly: While driving in CM, the automation fitness degrades, and automation can no longer perform its driving task. The Mediator system reacts by communicating to the driver that CM is switching off.
10. Smooth transition from Time to Sleep (TtS) to SB: while driving in TtS the driver is fully disengaged from the driving task when the automation indicates that the current route will leave the operational design domain. The Mediator system detects sufficient automation fitness for driving in SB and reacts by informing the driver that SB will be switched on and subsequently monitors the required driver fitness.

1.4. Literature review

This section is intended to introduce some previous work and efforts for testing automated vehicle functions to which the Mediator system is an extension, in context of academia, user centred development testing and independent (consumer) ratings.

1.4.1. Testing autonomous automated vehicle functions

The following section is an overview and introduction to the topic of testing of Automated vehicles in general, as well as using simulators and wizard-of-oz methodology in particular, with some emphasis on cost-savings by employing Commercial-off-the-shelf components and softwares. The intention is that the reader shall be familiarized to some potential cost-saving approaches.

The evolution of vehicle autonomous systems changes the driving experience dramatically, with drivers that may soon feel free to disengage from the primary driving task, paying attention to non-driving related tasks. This raises the needs to reproduce these novel in-car experiences, to assess usability, accessibility, and confidence provided in the interaction with the Mediator system. This evaluation process should be highly iterative, with rapid cycles through design, build and user research that allow design ideas to evolve based on user feedback. This iterative user/human-centred design process can be difficult for researchers in the context of automated driving, due to the lack of available automated vehicles, as well as some safety and economical criticises.

The road test approach surely provides an invaluable method to supporting integrated Developmental Testing and Operational Testing, as well as addressing hardware and software issues. However, recent advances in modelling and simulation methodologies could change the future effectiveness of automotive development, by dramatically reducing the overall cycle time for development of systems like the one presented in the MEDIATOR project.

Simulations surely are necessary and safe testbeds for AD. But testing in simulation alone is not enough for systems interacting directly with humans, because of various limitations such as how simulated drivers are very limited, there is a large gap between development and reality, if simulation is the only tool testing. Furthermore, testing with humans in real world traffic conditions can be unsafe and costly, while virtual driving technologies can be exploited as a tool to collect data and test autonomous systems with realistic human behaviours.

Manned driving simulator

An alternative approach to address operational and integration issues in on-road testing with physical prototypes of system components, is to translate the characteristics of the Mediator system directly into the vehicle model of a manned driving simulator, where user's actions are not simulated by and human digital model but are performed by experimental subjects who can directly interact with interfaces. However, the process to bring both prototype boards, or digital models of decision logic and automations packages, into the manned driving simulator could become resource consuming and require expensive software to get to work properly, with a significant overall impact on development cycle time. Indeed, simulator testing of human interactions with autonomous entities requires the integration of system's components in the simulator software environment and, however, the building of the physical interfaces to make subjects interaction realistic enough.

This integration process is usually performed by developing physical hardware that interfaces with the simulator systems, as well by emulating some components with codes or features already available in the simulator software. Even if this approach can provide a very controlled testing environment and great level of realism, it requires particular software for the management of a whole simulator system that is not properly low-cost, as it involve significant investments and operators with very specific and not so widespread knowledge. This issue could be addressed by running a wizard-of-oz style experiment in a driving simulator, where an experimenter imitates the autonomous systems and decision logics, by acting on specific commands on the controlling interface of the simulator, to activate or control the autonomous functionalities.

Indeed, the various test methodologies cannot be addressed and judged just by considering the cost required to perform an experiment but must be treated as an integral combination with technical expertise and methods required to achieve the desired results. However, through a confluence of proper design of experiments and emerging advances in Modelling and Simulation (M&S) techniques, several solutions are available for dramatically reducing the overall costs and time for testing of the systems presented in the MEDIATOR project. It is a widespread habitude to label those solutions with the term "low-cost" when they are implemented completely from open-source software and Commercial Off-The-Shelf (COTS) components. (Beringer, 1994; Weinberg & Harsham, 2009; Yavrucuk et al., 2011; Lu, et al., 2005; van Gent et al., 2019).

Especially for research related to driver state and Human Machine Interfaces (HMI), studies in real vehicles are rare, mainly due to the safety and liability issues inherent in testing unproven technology, and driving simulators offer distinct advantages over real vehicles, also in terms of repeatability.

However, what a "driving simulator" entails vary widely from institution to institution and study to study, moving from the low-fidelity end of the spectrum that involve a simple desktop station high-fidelity, to the high-end simulators that are the multi-million-dollar, full-motion platforms that occupy entire buildings (Bruck et al., 2020). Using COTS hardware and software it is possible to create a simulator with a degree of realism and flexibility similar to that of mid-level research simulators, but at a far lower cost (Yang et al., 2010).

Simulators using Commercial Off the Shelf components and softwares

From a software perspective, there are several open-source and commercial alternatives developed for entertainment purposes that are suitable to be the software platform for a research-grade driving simulator (Dosovitskiy et al., 2017; Wymann et al., 2000). Especially titles for virtual sport driving, defined "Sim-Racing Games" offers a convincing, realistic driving experience thanks to richly detailed graphics, accurate vehicle physics, and full support of force-feedback steering wheels. Sim-Racing software is designed to replicate the handling and feeling of an actual car as much as possible, in contrast to arcade games, in which entertainment is the key element (Narayanasamy et al., 2006).

Some of Sim-Racing software, as Assetto Corsa and rFactor, does not offer the complete flexibility of an open-source product, but allow for a deep degree of modification and customization, with a large community of enthusiasts who produce everything from custom tracks to custom vehicles. Most driving games also provide a plug-in application programming interface, whereby vehicle telemetry (including position, velocity, and acceleration), and user input (steering angle and throttle/brake positions) can be captured at sample rates. Those could be insufficient to properly study vehicle dynamics but are acceptable for human factors evaluation (Tiu et al., 2020; Koskela et al., 2011). Simulators based on Sim-Racing technologies can also benefit from the "out of the box" support for high performance motion platforms and multi-channel display systems, as highlighted by the simulator series, called Reasonably Priced Immersive Driving Simulators, that could be considered as the highest level where this gaming-derived commercial technology can be adapted for research use (Sena et al., 2016).

The choice of using Commercial Off-The-Shelf (COTS) software as the simulation engine also meant significant time savings because, rather than painstakingly modelling vehicles and roadways by scripting scenarios, is possible to build a complete scenario within several hours using editor developed for game enthusiasts, instead of using a professional modelling suite (Georgiou & Demiris, 2016). The reliance on off-the-shelf components is not without significant disadvantages, however, especially if racing simulation game are used, due to some difficulties to model complex street layouts and intersections found in urban areas. Furthermore, software developed for entertainment purposes offer very little programmatic control over the computer-controlled "AI" drivers (de Frutos & Castro, 2021).

However, in scenarios where a study's protocol calls for the subject to react to traffic cars that perform specific manoeuvres at specific times, it is possible to populate the simulation with one or more emulated AI cars, by exploiting the multiplayer design of many driving games. Those AI corridors are controlled by human "Wizard of Oz" operators who are aware of the study protocol and receive specific instructions or signals as to when and where to carry out specific manoeuvres (Weinberg & Harsham, 2009).

Simulation testbed tools composed of off-the-shelf devices, that allow humans to interact with various levels of virtuality or digitalization, provide different levels of realism and safety. We can consider three basic entities in the case of the MEDIATOR project: the driver, the autonomous vehicle and the Mediator system. Each of them can be either real, simulated or emulated.

Simulated, non-physical, vehicle and Mediator system components are used for Software-In-The-Loop (SITL) testing, where sensor data is calculated using mathematical models and command output is used to reproduce actions. Real hardware could be also physically present and can be tested alongside simulated entities using a Hardware-In-The-Loop (HITL) approach. Furthermore, humans and systems can be also emulated in several ways, by programming human comportments in a virtual human model or by reproducing the behaviour of an autonomous system by the action of real human operator, on console and interfaces, physically present in the experimental setup.

Wizard-of-Oz paradigm

This last configuration, in the context of autonomous driving, is referred as the Wizard of Oz (WoOz) paradigm. The WoOz is a research approach, born in Human Computer Interaction domain, in which the user interacts with a system that he/she believes to be real, but which in reality is controlled, completely or partially, by a human being, the driving wizard. In particular, John F. Kelley invented the WoOz paradigm in 1975 to simulate a not yet functional speech recognition system (Green & Wei-Haas, 1985).

According to a recent bibliographic collection (Bengler et al., 2020) the WoOz paradigm has allowed research on human-machine interaction in parallel with the technical development. Indeed, within the automotive research community, WoOz vehicles can be used as a method for analysing the effects of "intelligent" probabilistic systems, such as automated vehicles, even if not yet fully developed). There are different concepts around which a WoOz vehicle can be built. In some of them the participant can drive, in others not. In the not functional WoOz set-up, like the one called "the Sim-Panda" vehicle, which correspond to a real FIAT Panda car with right hand drive, equipped on left side with a full interface for driving simulator (Sena et al., 2016). This setup makes possible to both drives virtually in immersive scenarios by the inside of a real car and surely represent a low-cost approach to the WoOz paradigm, being characterized by the travelling capabilities, thanks to the right seat driving with the standard controls, and by the off-the shelfs interface for virtual driving on the left side. As described in section 2.3, a similar setup was used within the MEDIATOR project.

1.4.1.1. Driver State and Automation Health

Research on driver state sensing and assessment, including driver distraction and fatigue, has a long history. Time limitations at the writing of this deliverable prohibits even a short introduction to the topic, but Borowsky et al (2020) provides an excellent state-of-the-art report. The topic of Automation health is rather new, but crucial to the MEDIATOR project. In essence, the decision logic needs an objective assessment of the current and near future state of the Automation module in relation to the current Operational Design Domain parameters and the current traffic context. An excellent State-of-the-art report is written by Mano et.al. (2021).

1.4.1.2. Virtual testing

A current trend in safety testing is the move towards virtual testing – this means for example in restraint system development to complement the use of physical test dummies with Finite Element Models of the human body complete with physical properties for flesh and bone, to be used in crash test simulations to estimate potential damage to the human body in a crash described by certain parameters, such as a standardized crash test. For crash avoidance systems and automated driving relying on computer vision, it may mean to validate sensor performance using generated image data with known parameters from a driving simulator software.

1.4.1.3. Visual and Auditory component evaluation in virtual reality (VR)

In one of the earlier studies within MEDIATOR a Virtual Reality (VR) methodology – participants were ‘riding along’ in a Mediator-HMI-equipped cabin across a variety of use-cases and transitions between these, rating the experience on various scales on terms of usability, efficiency etc.

Enhancing trust in the system can be accomplished by clearly conveying autonomous drive events (intention of the car, driving modes, errors, etc.) while finding a balance between the amount of communication and so preventing information overload. Therefore, improving the understanding of the user’s perception is fundamental and can be achieved by continuously involving users during the design process (Singh, 2021).

A Virtual Reality (VR) environment enables an immersive environment for users without the amount of time and resources needed for a physical prototype. Therefore, the visual and auditory components are evaluated by means of a VR-environment using the Rapid Iterative Testing and Evaluation (RITE)-method (Medlock et al., 2002). By means of the Think Aloud protocol, the participants describe their experience with the HMI (following Eccles & Arsal, 2017 and Jaspers et al., 2004). After a (set of) participants the problems are identified, and a solution will be found whereafter changes will be made prior to the next (set of) participants. The evaluation focussed on the during-phases, meaning no actual driving mode transfer will happen during the evaluation.

The evaluation was split up into 6 sessions with 3 participants per session. All driving modes (figure 12, 13, 14 and 15) were shown to the participants with a 15-second break in-between. Afterwards, the participants answered several validated questionnaires (SUS, NASA-TLX, ITC-SOPI, and questions on their understanding) and a semi-structured interview was conducted, focussing on the understanding and experience with the VR HMI design. After every session, the participant’s understanding was rated by the three team members next to a discussion about overall issues that occurred and therefore what could be changed and/or emphasized in order to improve the matter.

The outcome of the evaluation was used to improve the HMI design (behaviour) and recommendations for how to employ the design in the best way.

1.4.2. Existing independent test protocols for Driver Assistance

Some independent test facilities have tested and rated/graded Automated Driving features. The ones known to the authors are the *IIHS Safeguards Assessment for Automated Driving functions*, (not yet published), the *Consumer Reports Testing of Automated driving features*, and the *Euro NCAP grading of Automated driving* (Currently at version 1.1.)

IIHS Safeguards for Assisted Driving

In 2020 the US consumer information organization Insurance Institute for Highway Safety (IIHS) first proposed that since Automated vehicle systems on the SAE Level 2 still rely on humans to be responsible for safety and humans having notoriously poor vigilance for monitoring a largely static situation for changes, the following escalation sequence of warnings should be issued :

> *The systems that are currently available either assume the driver is paying attention when his or her hands are on the wheel or use a driver-facing camera to determine if the driver's head is oriented toward the road, but neither is fool proof. The researchers recommend employing multiple monitoring methods, including using a driver-facing camera and measuring things like manual adjustments to the steering wheel and how quickly the driver responds to attention reminders.* (IIHS, 2022a)

"Partial automation systems may make long drives seem like less of a burden, but there is no evidence that they make driving safer," says IIHS President David Harkey. "In fact, the opposite may be the case if systems lack adequate safeguards."

Documentation for the full safeguards rating is not yet publicly available but some ideas for e.g. escalating attention warnings are presented in IIHS (2020a)

> *When the driver monitoring system detects that the driver's focus has wandered, that should trigger a series of escalating attention reminders. The first warning should be a brief visual reminder. If the driver doesn't quickly respond, the system should rapidly add an audible or physical alert, such as seat vibration, and a more urgent visual message* (IIHS, 2022a,)

In 2022 IIHS (2022b) outlined a proposed new rating for their Safeguards concept which includes the following parts. (paraphrased)

- Monitoring driver's gaze and hand position (with emphasis on hand availability)
- Using rapidly escalating alerts to get drivers' attention
- Fail-safe procedure to slow vehicle, notify manufacturer and turn off automation for remainder of drive [in case of failure?]
- Automated lane changes only if initiated or confirmed by human driver.
- Adaptive cruise control does not automatically resume after a lengthy stop of if the driver is not looking at the road
- Lane centring does not discourage steering by driver
- Automation features cannot be used without the belt fastened
- Automation features cannot be used with AEB/LKA disabled

The gist of the Safeguards assessment protocol is to keep the driver in control during assisted driving, as well as support the human driver in this task.

Consumer reports

The US consumer organization 'Consumer Reports' graded 12 vehicles with "ADA" – *active driving assistance (ADA) systems.* In an article (Consumer Reports 2023) in the organisations website the senior director of Auto Testing, Jake Fisher, describes:

In the simplest terms, ADA is the simultaneous use of a car's adaptive cruise control (ACC) to control speed and lane centering assistance (LCA) to control steering. ACC is an advanced form of cruise control that brakes or accelerates to keep the car a set distance from vehicles traveling ahead of you in your lane. LCA provides steering support to keep the vehicle at or near the centre of the lane.

"But they don't make a car self-driving at all," Fisher says. "Instead, they create a new way of collaboratively driving with the computers in your car. When automakers do it the right way, it can make driving safer and more convenient. When they do it the wrong way, it can be dangerous." (Consumer Reports 2023)

The Consumer Reports protocol / scoring system has five parts:

- Capabilities and Performance
- Keeping driver engaged
- Ease of use
- Clear when safe to use
- Unresponsive driver

The five parts of this assessment are rated on a scale from 1 to 10 and an overall system score is awarded – see Figure 1.3 for an excerpt of system scores for two tested vehicles.

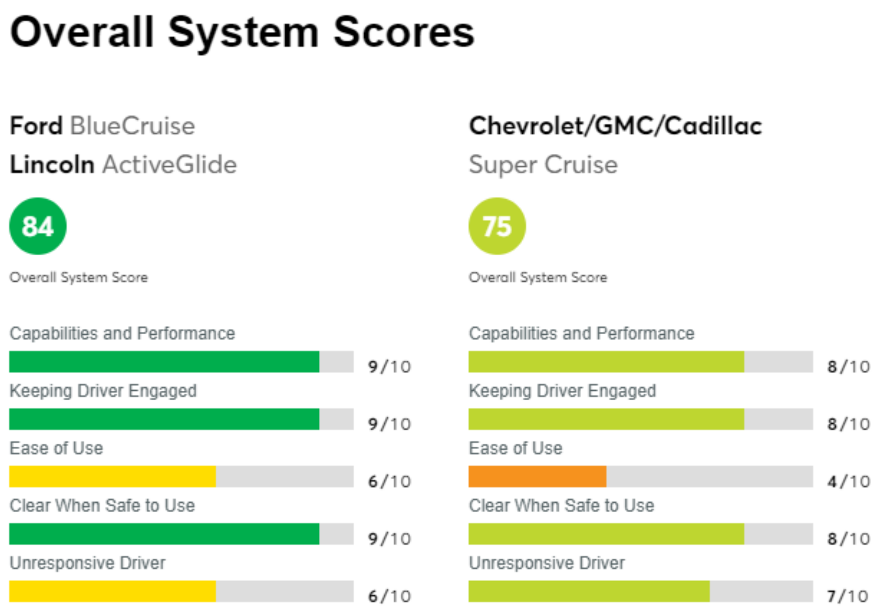


Figure 1.3 Examples of grading of assisted vehicle features from Consumer Reports.

Euro NCAP Grading of Assisted Driving

In 2020, Euro NCAP (European New Car Assessment program) released a grading scheme to assess the assisted driving. The two areas of assessment focus are a) Assistance Competence – a balance between Vehicle Assistance and Driver Engagement, and b) Safety Backup, the car's safety net in critical situations (Euro NCAP, 2020). The main objective of this grading system is to inform and educate consumers about the differences between driving assistance and automation. Assistance system always require driver supervision. Along with the previous objective, the grading scheme is also aimed at ensuring that systems offered by vehicle manufacturer provide robust assistance and do not create new crash risks.

The grading for *Assistance Competence – Driver engagement* consists of protocols for, *Consumer information, System status, Driver Monitoring* and *Driving calibration*. Similarly, *Assistance Competence – Vehicle assistance* consists of evaluating, *Speed assistance, Adaptive cruise control* performance and *steering assistance*. The *Safety backup* assessment consists of, *system failure, Unresponsive Driver intervention* and *collision avoidance*. The detailed assessment protocol is given in the Euro NCAP protocol for Assisted driving 2023. In section 3.2 of this report this existing assessment protocol is mapped to the Mediator modules and use cases to understand the how the existing assessment protocol can be used for rating mediator-like systems, and how it needs to be amended.

2. Testing done within MEDIATOR

In the MEDIATOR project, multiple evaluation studies have been carried out. These studies range from computer simulation to on road studies. Within the project, three VR studies, three on-road studies, three simulator studies and one computer simulation study were fulfilled, with different aims and objectives, and with varying level of cost associated. A summary of all studies is given below, grouped by the platform used for data collection, and a short commentary on the cost driving items for each study.

2.1. Computer simulation study

A computer simulation study was carried out with the goal to simulate all aspects of driving an automated vehicle, relevant to the Decision Logic. The relevant aspects mirror the different modules of the Mediator system, including Driving context, Automation context, Driver context and Actions from the decision logic, which represent the output of the Decision Logic to be forwarded to the HMI. Furthermore, real world data was introduced in order to improve the simulator, including data from the TI-vehicle in Sweden. By parsing data from the car into the simulator parts of the simulated versions of the driving context and automation context were replaced with real data which allowed for testing the Decision Logic in a more realistic setting, or possibly differently put with a higher external validity. According to the research objective defined early in the MEDIATOR project, the computer simulation study combined with real-world data has several purposes.

First, to improve the computer simulation and make it more realistic by exploiting the real-world driving data from TI vehicle.

Second, to evaluate the decision quality of the Mediator decision logic using the offline driving data regarding the evaluation metrics related to safety, comfort, efficiency, and user acceptance.

Third, the source of the inaccuracy and the failure mode of the decision logic would be identified. Finally, the evaluation results were fed back to the design team to further optimise the Mediator Decision Logic module.

The computer simulation study should be considered a cost-effective tool for Decision Logic evaluation, due to it being able to run very many iterations at low cost of labour.

2.2. Driving simulator studies

Three main driving simulator studies were performed, each with a different study focus, ranging from driver comfort to corrective intervention for distracted drivers. All simulator studies within MEDIATOR are outlined below. All three Driving Simulator studies should be considered medium-cost evaluations, due to them utilizing existing simulator infrastructures, with fairly short lead times and prototype development times. Cost driving items are equipment rent, and protype development/modelling.

2.2.1. Simulator study 1 – Preventive Mediation of driver fatigue

This study looked at the effects of partially automated driving (i.e., SAE level 2), in a driving simulator, on fatigue progression and hazard perception performance relative to manual driving in the shorter and longer term. Further, it examined whether engagement with a secondary visual task during SAE level 2 driving conditions can mitigate fatigue progression and the decline in hazard perception performance under SAE level 2 driving conditions due to underload. Participants were asked to drive in a driving simulator two drives of 40 minutes each in one of two experimental conditions (manual, with/without a secondary task). The main results demonstrated that in the shorter-term fatigue levels under SAE level 2 driving conditions were lower when drivers engaged with a visual-auditory secondary Trivia task than when they were not engaged with a secondary task. This effect was not evident in the longer term (a week after the first experiment) In addition, hazard perception performance was similar for both conditions.

2.2.2. Simulator Study 2 – Transitions of Control and Driver Comfort

This study focused on comfort Transitions of Control (TOC) from manual to automated driving, simulated automation degradation and related TOC by the human driver as well as driver characteristics. Hence, Take Over Requests from the driver or the Mediator system due to, for instance, an upcoming traffic jam or incoming text messages (i.e., comfort-related TOC), planned or unplanned transitions from automation (i.e., driver stand-by / continuous mediation) to manual driving due to degraded automation fitness (i.e., safety critical takeovers), uncomfortable driving manoeuvres (i.e., close approach scenarios) for driving comfort estimation and the influence of driver characteristics (e.g., age, gender, driving experience) were examined. Since this study had a strong focus on drivers' evaluation of the Mediator system and its functionalities (e.g., acceptance, trust, usability, comfort, experience of TOC) and not on the technical parameters (e.g., performance of developed algorithms and technical components) it was important that all drivers experienced the same scenarios with the same reaction of the Mediator system. Thus, to ensure that drivers will experience an "optimal working" Mediator system no online detection and decision making was involved in the simulator study (e.g., no unplanned HMI reactions, no false positives or false negatives).

2.2.3. Simulator study 3 – Hazard detection and distraction in automation

This study looked at how driver's distraction affects hazard perception performance under SAE level 2 driving conditions in the shorter and longer term. Similar hazard perception scenarios were used with minor differences from those in the fatigue study. Participants were driving in two separate drives, one week apart once under SAE level 2 driving conditions without a secondary task and once under SAE level 2 driving conditions with a secondary task. The main results revealed that when drivers engaged with a secondary task drivers tended to accept, trust and appreciate the mediator's HMI at a high level (better than 80% according to the SUS scores). The scores were higher in the longer term than the shorter for most measurements, they were less likely to identify hazards and tended to deactivate the automation in response to hazardous situations.

2.3. On-road studies

Three on road studies were performed within MEDIATOR, using two different vehicle platforms, the Human Machine Interaction vehicle (HMI-vehicle) and the Techincal Integration vehicle (TI-vehicle). The HMI-vehicle was mainly used for evaluation of the HMI-concept, whereas the TI-vehicle was used for validating the technical feasibility of the Mediator concept. All on-road studies are outlined below.

2.3.1. On-road vehicle platforms

Two on-road testing platforms were equipped. The first one, a Wizard-of-Oz vehicle prototype, the HF vehicle, was created by modifying a right-hand drive Jeep Renegade. A non-functional second set of pedals and a non-functional steering wheel were integrated on the left side of the vehicle and a prototype gear shifter was ad hoc designed and integrated to allow simulating the change of the driving mode (from manual to automated and vice versa). This WoOz vehicle allowed naïve participants to have an "automated" vehicle experience, without being in a real automated vehicle. The prototype was always driven by a professional driver in the right seat, using the standard primary controls of the vehicle. The Mediator HMI, located at the other front seat, its components and solutions (visual, acoustic, vocal, …) (e.g., participant frontal display, shifter, steering wheel, seat and belt and displays) were evaluated during the on-road trials in their different states, depending on the use cases.

The TI vehicle was a conventional test vehicle equipped with state-of-the-art sensing and computing capabilities and a customizable HMI (visual components only). It was used for on-road and real-time testing of driver state sensing, automation health assessment and decision logic.

In terms of equipment costs, the comparison is a bit unfair due to the fact that the TI platform was added as new functionality to an already existing test vehicle with infrastructure for power, computation and logging, whereas the HF vehicle was developed from a a standard vehicle. For the studies as a whole, the per-participant cost was medium to high, and cost driving items are technology development for the TI vehicle whereas the HMI-vehicle draws operating cost by requiring two people working in the vehicle during all tests.

2.3.2. HF prototype vehicle - Italy on-road study

The aim of the Italy on-road study was the evaluation of the usability of the Mediator HMI and the users' acceptance and trust of the Mediator levels Continuous Mediation and Driver Stand-By in real-life conditions, through an on-road user testing, using the Human Factors (HF) prototype. vehicle Different Mediator HMI solutions (e.g., visual, acoustic, vocal, haptic) located in different parts of the vehicle (e.g., participant frontal display, shifter, steering wheel, seat and belt and centre dashboard display) were evaluated in their different states, which changed according to the tested use cases. The participants experienced, on a mixed scenario of 46 km, seven use cases in which handover and takeover were requested by the user or initiated by the system for some different reasons such as drowsiness, distraction, desire to take back the control from the user etc. In addition to handover and takeover also an improvement of driver fitness during an automated driving mode was tested. The HMI solutions were evaluated to select the most appropriate ones and refine them following the User-Centred Design process, to later be used in the on-road study in Sweden. Measures included user acceptability (before the use) and acceptance (after the use) of the automated vehicle and the acceptance and the perceived trustworthiness toward the Mediator system.

2.3.3. HF prototype vehicle – Sweden on-road study

The aim of this study was to evaluate the functionality, safety effects and user acceptance of the Mediator system under different degraded driver performance conditions, including conditions of degraded automation. Driver competences (e.g., driver experience, tendency to accept risks) also varied. The route and the scenarios were chosen to cover as many automation levels, UCs and Mediator functionalities as possible, including all ToC scenarios. The study was designed to contrast the user experience of the MEDIATOR system to a baseline HMI for automated driving. The study was conducted in the HF-vehicle platform. The limitations of the platform such as that the participant could not actually control or maneuver the vehicle made it difficult to emulate manual driving and lower levels of automation such as CM. Also, transfers of control to or from these lower levels were to some extent lacking realism. The advantage is, that it is possible to run experiments emulating higher levels of automation which are not yet available in vehicles. This was utilised in that the study included a scenario where participants were allowed to 'nap' while on the road, and then transferring to an automation level which included some level of driver responsibility.

2.3.4. TI prototype vehicle - Sweden on-road study

The aim of this study was to evaluate the functionality and user acceptance of the automation state module system under different driving contexts, including the assessment of how well the system can predict bad automation performance as defined in MEDIATOR D1.3 (Mano et al., 2021). All drivers were professional test drivers and drove the route at least three times. The study was conducted in the TI vehicle's data collection platform using a CM hands-on driving assistance system that provides longitudinal and lateral control (SAE level 2).

The research questions in this study mainly concerned the automation state module's performance in predicting bad and good automation performance on the route, and the decision logic related to that. Mainly, this on-road study will examine the various times to automation unfitness and fitness, and the reliability of the automation status detection in relation to the driving context. The study was used to examine the user perspective on the prediction of bad and good automation performance, and the timing of the Mediator system's messages. Limitations to the on-road study include that it only included static ODD changes, that is, something that is not reliant on the driving automation's sensors to be detected. This choice was made to increase the repeatability of the drives, so that a larger amount of the data was usable.

2.4. External validity

External validity is the extent to which the findings of a study can be generalized to other people, settings, and measures. The aim of scientific research is creating knowledge that is generalizable to the real world. Without high external validity, you cannot apply results obtained in a laboratory study with a certain sample of people to the real world or other people. A trade-off between external and internal validity must be done. Internal validity is the extent to which a researcher is confident that the causal relationship established in the experiment cannot be explained by other factors. Then, the trade-off between these two types of research validity: the more the study is applicable to a wider context, the fewer other (extraneous to the study) factors can be controlled in the study. This implies to put great attention in the study definition and on, for example, on aspects like the sampling, the experimental design, the setting, the timing and procedure, the choice of the experimenter...

To exemplify - the Italy study was done directly on an ecological context, the real road, using the WoOz prototype vehicle, so the findings of the study produced knowledge collected in the real world. Great attention was paid to avoid increasing external validity (ecological context) at the expense of internal validity. This was done by controlling as many confounding factors as possible, for example avoiding test drives in different weather or lighting conditions and carefully annotating any possible change in the real scenario (e.g., traffic, unforeseen events). The participants were recruited by an external company to avoid bias and the participant sample covered characteristics of target vehicle buyers. Instructions and procedure were targeted on minimizing the possible impact (situation effect) of the very first prototype of the WoOz vehicle and of the HMI solutions, used in the Italian test.

2.5. Mediator Test Leader survey

For this report, a test leader survey was performed, to collect insights from all the MEDIATOR studies in terms of cost profiles in terms both of equipment, development cost and labour made in each study, as well as a general idea of methodological trade-offs for cost and validity, and ultimately an understanding of the interplay of the factors of cost, safety estimates and external validity. In general, costs involved in advanced human machine interface and system development projects include labour time for system development, method development, recruiting and administration, performing the actual data collection, analysis, and reporting. Whatever can be done to simplify the process will have a direct bearing on the final cost of the project. In addition to this, equipment for data collection as well as prototype build will contribute to the cost. In general, method development and prototype costs are constant, while data collection and to some extend analysis costs scale with increased number of participants.

2.5.1. Method description

An email survey was sent out to each of the responsible study leaders within MEDIATOR with questions covering outlined method decisions, cost vs. validity trade-offs made and general cost and labour time estimates. Survey responses were collated in a spreadsheet and an attempt was made at creating a general cost per participant estimate. Unfortunately, responses were not sufficiently similar in resolution to make a meaningful cost estimate. Some of the descriptive responses are elaborated in the following sections. The following is not a complete representation of the survey results, but a representation of a few selected highlights using responses to the four most relevant questions.

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   The purpose of this question was to understand what, if anything, was done to increase the compatibility across studies in different countries and on different platforms across the MEDIATOR project, and thus to increase the validity while keeping costs down.

2. **What is the most defining feature of your study in terms of External Validity**

   The purpose of this question was to understand how study leaders consider their selected method to perform in terms of real-world applicability of their results – e.g., can driving simulator results really say anything about absolute levels of fatigue, or only in relative terms?

3. **What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   The purpose of this question was to learn what cost-saving trade-offs were made in each study.

4. **How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control), and : 'External validity (real world applicability)**

   The purpose of this question was to better understand if there were trade-offs for reaching better and stronger results in terms of contrasting within the study, but weaker 'evidence value' in real-world applications. The purpose of this question was to understand better the generalizability of the results, to be put in contrast to the general study cost parameters.

2.5.2. Survey response for Simulator study 2

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   Our results focus more on driver's opinion than on driver behaviour.

   Regarding driver's evaluation, the following results can be compared between different studies:

   - Benefits and drawbacks of a system like Mediator
   - Potential for improvement
   - Preferences for different use cases (e.g., Mediator recommends automated driving because of an approaching traffic jam, detected distraction, detected fatigue, etc)

- Ranking of conditions (i.e., ranking of manual driving compared to using Mediator in CM mode, SB mode or while performing secondary tasks in SM mode)
- Evaluation of HMI components and amount of presented information
- Potential of Mediator to prevent mode confusion
- Evaluation of Usability (e.g., how easy and understandable is the transfer from one driving mode to another driving mode)
- Intention to use Mediator in the future

2. **What is the most defining feature of your study in terms of External Validity**

   Some results have high external validity (e.g., evaluation of several HMI elements). For others the external validity is impaired (e.g., perceived safety, visibility of ambient lighting and LED stripes) due to the simulated environment (e.g., completely safe environment, nothing can happen even when the vehicle crashes; dark surrounding all the time, hence, no daylight conditions like direct sun can be evaluated). Drivers experience the Mediator system in exact the way as it is programmed ("optimal working" system) and not under "real world conditions" with respective unexpected events and system behaviour (or even failures).

3. **What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   Limited number of participants due to time and availability of test vehicle.

   Reduction of HMI elements (e.g., no Head-Up-Display)

   Reduction of use cases and scenarios (e.g., no inner city driving, no fatigue distraction)

   All participants experienced the different conditions in the same order

   Questions regarding Mediator and HMI were reduced to the most important / prioritized aspects in the questionnaires and interviews

   The study finished with 81 participants equally distributed to gender and age groups

   Results need to be interpreted carefully in terms of generalizability and external validity. Mediator can only be evaluated on the prioritized aspects; some questions might stay unanswered.

4. **How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control), and : 'External validity (real world applicability)**

   Internal Baseline condition can be examined under exact same environmental conditions as the experimental conditions. External Only limited results can be generalized (e.g., evaluation of HMI elements). Others need to be interpreted carefully taking the simulated environment into account.

2.5.3. Survey response for Simulator studies 1/3

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   We think that using the same questionnaires across driving simulator studies was very helpful and allowed comparing participants' subjective evaluations across studies.

2. **What is the most defining feature of your study in terms of External Validity**

   Regarding the HMI, we believe that the patterns of behavioural adaptation observed in the driving simulator are also valid in the external world where drivers modify their behaviour when interacting with in-vehicle systems.

3. **What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   The original plan was to use online algorithms to monitor the driver's state throughout the driver. Unfortunately, due to COVID-19, we could not implement these algorithms and thus had to use other methods. For example, in the fatigue study, we relied on participants' KSS reports provided every 10 minutes along the drive to examine drivers' fatigue levels.

   Although the original idea was to evaluate both preventive and corrective mediation concepts our results focus on preventive mediation and not corrective mediation. Preventive mediation was easier to implement as it was initiated before the driver became fatigued or distracted and thus did not rely on online algorithms to initiate an action. Corrective mediation was more difficult to implement because this concept suggests that given a driver is distracted or fatigued the HMI initiates a corrective action to help the driver to maintain his/her performance.It should be noted that the hazard notification system that we implemented in the HMI can be regarded as a type of corrective mediation in the distraction study.

4. **How would you describe advantages and drawbacks of your study in terms of 'Internal validity (contrasting w. baseline or control), and 'External validity (real world applicability)**

   Using a driving simulator with a pre-designed driving environment was an advantage in terms of internal validity. This is also a drawback because we reduced the complexity of real-world driving where drivers must handle several tasks simultaneously and in complex driving environments (external validity). Another advantage of our studies was that we tested each participant several times and could compare each one of the drivers to their own measures under various simulated driving conditions. Regarding the HMI, the patterns of behavioural adaptation observed in the driving simulator are also valid in the external world where drivers modify their behaviour when interacting with in-vehicle systems

2.5.4. Survey response for Italy on-road study with the HF vehicle

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   Psychometric instruments used during the Italian study were chosen in a set of questionnaires list decided by the project members during the first phase of MEDIATOR project a list of questionnaires and instruments to be used during users' trials were described, so different partners could use same psychometrics instruments.

   Then, for example, comparisons (at list a qualitative one) on for example,

   - Affinity for technology (Franke et al)
   - Acceptance (Van der Laan, SUaaVE questionnaire)
   - Usability (SUS questionnaire)
   - Trust (Jian et al)

   could be done among studies after the data analysis done by the single partners"

2. **What is the most defining feature of your study in terms of External Validity**

   The FCA-CRF user testing on Mediator HMI solutions was done directly in real road scenarios, using the WoOz prototype vehicle. The findings of the study produced knowledge about the real word and the external validity can be considered high. In fact, even if the WoOz was not functional and both the vehicle and the HMI solutions had some technical limits because they were in their first versions, thanks to the procedure and instructions , participants succeeded in having an "automated" vehicle experience In fact, after a while they "forgot" that the professional driver was actually driving the vehicle, and they ignored the driving wizard, and they verbalized to have loved experimenting with the experimenter as they were in an automated vehicle.

3. **What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   The first simplification was the preparation of a WoOz not functional vehicle prototype, because a functional one would have requested much more resources (time and economical) and constraints from the legal and safety point of view.

   In this WoOz prototype, there is no real vehicle automation and consequently there is a simulated Decision Logic component, which plays out pre-programmed scenarios, based on passing fixed positions in the driven evaluation route.

   Another simplification was the usage of prototype MEDIATOR components, which were not real products, because they were new concepts ideated in the project and under refinement for the following trials.

   The other simplification was due mainly to COVID delays, which requested a compression of the trial's duration and then, the sample of participants was recruited without sleep deprivation characteristics.

   The only consequence could be a possible reduced impact of the takeover criticality aspects due to the fact that users knew the vehicle was driven by a professional driver. However, difference on takeover HMIs were appreciated and coherently judged.

   The flexibility, variety and reproducibility of scenarios is limited due to the technical, safety, ethical and environmental constraints. Ethical and legal issues play a more important role than in simulated environments due to real driving risks and regulations of road authorities and the WoOz approach makes possible to overcome some of these issues.

4. **How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control), and : 'External validity (real world applicability)**

   In the study there weren't threats to internal validity such as, participants dropping out or leaving the trial (no participant left the trial), experimenter bias (the experimenters were only two and very well trained to behave the same way with all participants), or time bias (the procedure didn't annoy or tire participants).

   MEDIATOR partners decided not to develop a baseline HMI for the Italian trial, in which a first evaluation step of the HMI had to be done. Then, the comparison was done among the different MEDIATOR HMI solutions, experimented by all participants (between-subject design). This allowed to highlight advantages and disadvantages of the MEDIATOR HMI and derive proper redesign suggestions for the next project steps. The user testing was done directly in an ecological context, the real road, using the WoOz prototype vehicle, then the findings of the study produced knowledge collected in the real word. Great attention was paid to avoid increasing external validity (ecological context) at expenses of internal validity. This was done controlling as much as possible confounding factors, for example, avoiding doing trials in different weather or lighting conditions and carefully annotate any possible change in the real scenario (e.g., traffic, unforeseen events). The users' recruitment was done by an external company to avoid any bias and covering characteristics of target vehicle buyers. Instructions and procedure were targeted on minimizing the possible impact (situation effect) of the prototypal very first version of the Wizard of Oz vehicle and of the HMI solutions, used in the Italian test.

2.5.5. Survey response for Sweden on-road study with HF-vehicle

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   "The implementation of the Mediator system has been very different in all prototypes, and I am not sure the results should be compared to each other, since the participants have not evaluated the same system.

   If the same system had been implemented across studies, I believe that HMI content could be compared across studies, while HMI details such as light intensity should be evaluated outdoors in real-life light conditions.

   True driver state evaluations and driver monitoring should also be conducted in naturalistic settings with actual state inducement, at least after initial tests."

2. **What is the most defining feature of your study in terms of External Validity**

   Low external validity due to the Wizard of Oz approach (without actual dual control), due to occasional system malfunctioning (some foreseeable, such as the delay between interaction and result on HMI, and some more difficult to foresee, for example due to software licences that expired, bugs occurring in multiple systems causing errors that were difficult to locate and troubleshoot, power glitches causing random computers to reboot, computers occasionally turning off when the automatic start/stop system restarted the engine, etc.). It is also questionable to which extent a human driver is capable of driving like an automated system, with automation's strict positioning in the middle of the lane and not always smooth longitudinal control. High external validity because the tests were done on real roads in real traffic (as opposed to simulator environment), and because of the real-time interaction with (parts of) the driver monitoring system.

3. **What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   Many simplifications were done to make this study happen:

   - A simplified decision logic system was used to initiate events at pre-defined locations along the route. In an ideal setting the HMI would have been triggered by actual data.
   - Fatigue detections were triggered by a test leader instead of a driver monitoring system.
   - Distraction detections were triggered too late (due to low update frequency of the simplified decision logic system) and based on an external driver monitoring system instead of the Mediator driver monitoring system.
   - The vehicle was a right-hand steered vehicle with a mock-up steering wheel rather than an actual dual-control Wizard of Oz vehicle.

- It was not allowed to integrate the intended steering wheel, so instead a standard steering wheel was retrofitted with LEDs and biosensors.
- The implemented HMI was less cohesive than the ideal HMI design developed in the project.
- The prototype was developed to investigate isolated use cases tested one by one (for example, for the next five minutes we will test distraction detection, pause, for the next five minutes we will test a transfer of control from manual driving to Driver Standby, pause, for the next five minutes we will test fatigue detection). In the Swedish experiment, this prototype was “manipulated” to be used on a longer continuous route, which was a simplification because it was still only capable of executing one use case at the time.

The problem is that with a simplified HMI, a simplified driver monitoring system, and emulated automation, it becomes difficult to test functionality and safety effects of these systems, especially when it comes to their combined effects. As an example, the simplifications made it difficult for the participants to understand what it was, exactly, that they were supposed to evaluate and give feedback on.

It was difficult for the participants to understand how the distraction detection system worked since the warning often came after the participant had returned his/her gaze to the road again (due to the time delays in the decision logic).

It was difficult for the participants to understand why certain functionalities, such as distraction and fatigue warnings, only worked in certain places (i.e., use cases) and not everywhere.

It was difficult for the participants to understand their responsibilities since they were not able to actually control the vehicle in use cases with manual driving or continuous mediation. They also found it limiting that they could not turn the steering wheel as far as they would like. The developed prototype was not sophisticated and mature enough to represent our vision of the Mediator system and to give the illusion of actual automation.

4. **How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control), and : 'External validity (real world applicability)**

   Advantage: we have a control group (baseline HMI). A repeated measures design was used in which all participants experienced both systems (in counterbalanced order).

   Disadvantage: counterbalanced order not fully obtained as envisioned due to participant dropout. Even though we had a design with Baseline vs Mediator, the implementation of the two was not ideal and the validity of the results should be questioned.

2.5.6. Survey response for Sweden on-road study with TI in-vehicle prototype

1. **What dimensions of user performance are similar enough to make a meaningful comparison between studies?**

   Functionality, trust worthiness and user acceptance of the Mediator prototype.

   Mainly focusing on combination of user feedback in the interviews performed and vehicle data collected during the study.

2. **What is the most defining feature of your study in terms of External Validity**

   As if I take a look at the study we did. We had 7 participants (~70hrs) who experienced 4 configuration of mediator prototype. No statistical analysis can be generated in this study, despite the interview data we collected. I cannot draw any comparison to larger on-road studies. In those terms, the external validity is low. I am looking/searching for some small scale on-road studies that has been done, where I can draw inspiration to answer the external validity of the on-road study in a positive sense. I will come back to on that till then I would say external validity is low as the dataset size is externally small due to aim of the study being – “Exploratory study”

3. **What simplifications did you make compared to an ‘ideal’ study setting (ideal setting having little or no time' and financial' constraints) What consequences did that have for the final results?**

   Data collected during the study is not statistically strong.

4. **How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control), and : 'External validity (real world applicability)**

   The purpose of this question was to better understand if there were trade-offs for reaching better and stronger results in terms of contrasting within the study, but weaker ‘evidence value’ in real-world applications. The purpose of this question was to understand better the generalizability of the results, to be put in contrast to the general study cost parameters.

   Different variation of mediator prototype configurations was implemented. This would minimize the learning effect and test baseline vs changes in configuration.

3. Full system / independent assessment for consumer ratings

The main objective of this section is to compare the mediator functionality and how Mediator modules can be mapped into existing assessment protocols. In section 1.4.2 a few protocols for assessment of assisted driving systems were outlined. Since only the documentation of the Euro NCAP assessment is readily available on the internet, we chose to further analyse its content, perform a gap analysis, and propose new tests to add to the current assessment protocol to cover the full functionality of the Mediator system.

3.1. Rationale - Previous Protocol gap analysis

As explained in Section 1.4.2 this deliverable takes off from the current Euro NCAP protocol for evaluation of Assisted Driving and extends upon it to cover the full performance envelop / operational design domain of the MEDIATOR system. The Euro NCAP protocol (Euro NCAP 2022) is centred around the two central concepts of Assistance Competence and Safety backup, and in the coming sections we shall be mapping them to the Mediator system and perform a short gap analysis. The mapping also compares the gaps with respect to consumer rating and IIHS rating scheme.

Table 3.1 A summary of the independent assessment protocols and the MEDIATOR functions and use cases.

Euro NCAP	Consumer Reports	IIHS	MEDIATOR	MEDIATOR Use Case number
Assistance Competence - Driver Engagement - Consumer information	Clear When Safe to Use			UC 1, 3a, 4, 7 and 8
Assistance Competence - Driver Engagement - System Status	Ease of Use		HMI	UC 2
Assistance Competence - Driver Engagement - Driver Monitoring	Keeping Driver Engaged	Monitors both gaze and hands	Driver State	UC 1, 4, 8
Assistance Competence - Driver Engagement - Driving Collaboration	Ease of Use	Escalating alerts for driver attention	HMI	UC 2
Assistance Competence - Driver Engagement - Driver Engagement Assessment	Keeping Driver Engaged	Escalating alerts for driver attention	Preventive and Corrective Mediation	UC 7, 8, 10
Assistance Competence - Vehicle Assistance - Speed Assistance	Capabilities and Performance	-	Automation Health and decision logic	UC 9

Euro NCAP	Consumer Reports	IIHS	MEDIATOR	MEDIATOR Use Case number
Assistance Competence - Vehicle Assistance Adaptive - Cruise Control Performance	Capabilities and Performance	Fail-safe procedure slows vehicle and disables automation	Automation Health and decision logic	UC 1, 5 9
Assistance Competence - Vehicle Assistance - Steering Assistance	Capabilities and Performance		Automation Health and decision logic	UC 1, 5 9
Safety Backup - System Failure	Capabilities and Performance		Automation Health	UC 9
Safety Backup - Unresponsive Driver Intervention	Unresponsive Driver		Preventive and Corrective Mediation	UC 7,8,9
Safety Backup - Collision Avoidance	Capabilities and Performance			UC 9
Safety Backup - Safety Backup Assessment	Capabilities and Performance			UC 9
		Automated Lane changes only by driver initiative or confirmation	HMI	UC 3, 5
		ACC does not resume after lengthy stop or driver gaze off road		UC 1, 5 9
		Lane Centering Does not discourage Steering by driver		UC 1, 5 9
		Automation features cannot be used with belt unfastened		UC 7,8,9
		Automation features cannot be used with automatic AEB or LDW disabled		UC 9
			Various ToC Rituals	UC 3, 5, 6

Table 3.1 maps the use case defined in the MEDIATOR project to the existing assessment methods. Even though most of the use case can be mapped to the assessment methods, the nomenclatures that were used are not same as the ones described in the use case definition. The most prominent differentiation can be seen in terms of safety back up or type of collision avoidance systems (no mapped Mediator Module in the column corresponding to safety backup assessment) that would support the driver in case of automation failure. Similarly, when it comes to consumer information on defining “Mediator-like” system, no matching can be found in the existing protocol investigated.

3.2. Mapping gaps to MEDIATOR Modules

As detailed in section 1.21.2, Mediator system contains different module to build the functionality for the use cases details in section 1.3.21.3.2. From the Table below (ref. Table 3.2Table 3.1) and from the section 3.1 most of the Mediator module and functionality can be mapped to the Euro NCAP AD assessment protocols.

Table 3.2 A summary of the independent assessment protocols and the MEDIATOR functions

Euro NCAP	MEDIATOR
Assistance Competence - Driver Engagement - Consumer information	
Assistance Competence - Driver Engagement - System Status	Automation state
Assistance Competence - Driver Engagement - Driver Monitoring	Driver State
Assistance Competence - Driver Engagement - Driving Collaboration	HMI
Assistance Competence - Driver Engagement - Driver Engagement Assessment	Decision logic
Assistance Competence - Vehicle Assistance - Speed Assistance	Automation state
Assistance Competence - Vehicle Assistance - Adaptive Cruise Control Performance	Automation state
Assistance Competence - Vehicle Assistance - Steering Assistance	Automation state
Safety Backup - System Failure	-
Safety Backup - Unresponsive Driver Intervention	Decision logic
Safety Backup - Collision Avoidance	-
Safety Backup - Safety Backup Assessment	-
-	Various ToC Rituals

3.2.1. Assistance competence

The assessment for Assistance competence is split between how well the system informs and monitors the drivers and Assisted functionality in terms of both lateral and longitudinal control.

Consumer information

From the Assessment mapping it can be seen that there is a needTable 3.1 for clear information and a protocol about what the Mediator system is. Existing protocols state that there should be clear information about what the system does and what it doesn’t do, and whether oversight of the driver is needed. Maximum Points are awarded for the word “assistant”, “assistance”, “assist” or another variation of the term (NCAP). This highlights that the extended protocol is needed to give correct information about what “Mediator -like “should be called as and how it should be informed to the consumers.

System status, Driver Monitoring and Driving calibration.

The main functionality of the Mediator system is to build on understand the driving context and the usecases. The modules such as automation state and driver state estimations are working together to give a picture of the current predicted automation level and driver state, which is combined with the help of the decision logic module to inform the driver through HMI about the suitable action for a given ODD. This action could be in terms of corrective actions, automation state information (who is in control), Distraction alerts etc. The automation state also helps in controlling both longitudinal and lateral control of the vehicle. From the exiting protocol (NCAP) the different modules can be assessed as a complete system (Mediator-like system) than assessing the individual components. However, it becomes difficult to perform consumer level assessment without understand how the Mediator system is built, as this type of information can be proprietary to the vehicle manufacturer.

3.2.2. Safety backup

The Safety backup assessment consists of, system failure, unresponsive driver intervention and collision avoidance. The main functionality of the Mediator system together with its individual components is to “mediate” between the driver and the assisted systems, taking various external context into account. Components such a driver state and the decision logic module help to perform corrective actions and alert the driver during fatigue or distraction events. The use cases that were designed in the MEDIATOR project only cover the fitness and unfitness of both the driver and the automated system with contextual reasoning. Assuming that vehicles’ inbuilt safety critical systems, such a collision avoidance, or safe stop modules, will prevent any collision inducing events. Even though it is clear Table 3.2 that no Mediator functionality or modules are mapped to system failure or safety backup assessment. As the Mediator is built on top of an existing vehicle platform, assessment can be carried out for the vehicle. Thus, an extended assessment method is needed to clarify when the assessment for Mediator functionality stops, and critical system assessment begins. Thus Mediator system cannot be tested using the current NCAP assessment protocol.

3.2.3. Transfer of control (ToC)

When we map the Euro NCAP assessments of different Mediator functionalities we find ToC is missing; hence this is the primary function that needs to be added to extended assessment protocol. For a full evaluation of ToC, the following parameters needs to be considered: timing – assessment of driver in control, HMI, time budgets, direction of transition, initiating party and many more. Table 3.3 shows which MEDIATOR use cases cover transitions, which direction and initiating entity.

Table 3.3 Overview of ToC Use Cases (UC) in MEDIATOR based on initiating party and direction of transfer.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Initiating entity</td></tr>
<tr><td>Mediator</td><td>Human</td></tr>
<tr><td rowspan="2">Direction of transfer</td><td>From Mediator to Human</td><td>UC5
UC8</td><td>UC2</td></tr>
<tr><td>From Human to Mediator</td><td>UC1
UC6</td><td>UC3
UC6</td></tr>
</table>

3.3. Extended protocol

As stated in 3.2.3, in addressing the missing pieces of the AD Assessment protocol, we need to extend our current testing protocols with assessment of a transfer-of-control-procedure. Not only for starting the driving assist, but also for the handing back control to the driver.

In contemporary systems, system start and stop (transfers of control) are rather silent, mostly only changing the colour or shape of an icon on the dashboard, and in any case there is little or no standardization (Jansen et al., 2022) This is not surprising, considering that in current automated driving or assistance systems, according to the SAE J3016, automation levels do not really transfer the responsibility for the driving task. It is always the human driver who is legally in control. While type approval testing is outside the scope of this report it can be mentioned that with the arrival of higher automation levels, the responsibility for the driving task is transferred, which emphasizes the need for an appropriate transfer procedure, and hence a safety rating or even legal requirements for such transfer procedures. So far, driver assistance has been handled within UN-ECE R79 for L2 systems, but L3 automation will be regulated from within the UN-ECE R157. (ECE/TRANS/WP.29/2022/59/Rev.1.)

Figure 3.1 outlines the test settings parameters outlined within MEDIATOR WP1, with the three main branches of driver, automation and environment characteristics. As mentioned earlier, these are mostly covered by the current Euro NCAP Automated Driving protocol, but the transfer of control is the crucial missing piece of the puzzle.

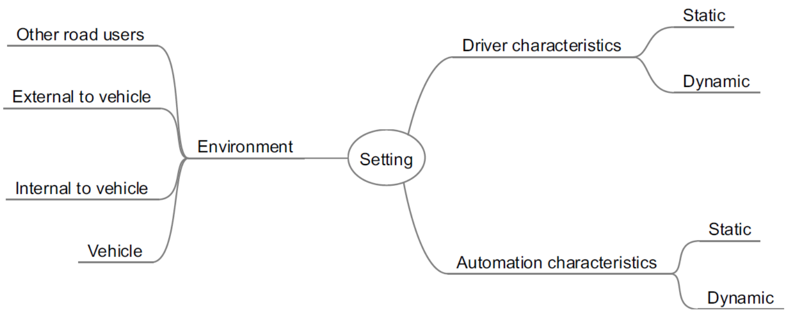


Figure 3.1 Test settings parameters from D1.1 – driver, automation and environment parameters.

3.3.1. Testing transfer of Control

Low-cost assessment of such a complex system as the MEDIATOR, does not lend itself easily to low-cost testing of the full integrated system. The project has intentionally divided the testing into smaller subsections of the full, excluding or simplifying subsystems. For our focus in a proposed protocol for future testing we chose to include all subsystems to study the particular use case of Transfer-of-control (TOC). TOC was tested in the following MEDIATOR studies: simulator study 2, and on-road studies 1, 2, and 3. While development testing of TOC is preferably done in a simulator or a wizard-of-Oz setting, as was done in several MEDIATOR studies, a rating test will obviously be performed in a road ready vehicle. A protocol for evaluating and rating a mediator-style vehicle, inspired by development and testing within the MEDIATOR project, Euro NCAP and the IIHS, safeguards rating for partially automated vehicles.

3.3.2. TOC Test Setup parameters

A brief description of a suitable setup for capturing all necessary aspects of the ToC scenarios for later analysis and application of the KPI evaluation.

Road setting

For on-road tests, performed within Mediator there were two setups, the Technical Integration (TI) vehicle road test was performed with professional drivers as participants and necessary road permits were obtained within the involved partner company . In the Human Factors (HF) test vehicle, participants were not at any point of the drive the driver in control of the vehicle.Thus, no road permits were needed. The Euro NCAP protocol for AD testing does not explicitly specify settings for testing, but it is implied in some passages that a closed test track can or should be used for testing. Meanwhile, it is stated that "*The systems being tested are those that can be broadly grouped together as Highway Assist systems as defined by Euro NCAP, or as SAE Level 2. This means that the driver retains full responsibility and shares control with the vehicle". Both vehicle and driver share OEDR, and the driver may not perform any secondary tasks over and above those permitted during normal Driving,* which implies that testing could be performed on public roads.

Vehicle setup, Data logging equipment, and Measurements

The Vehicle Under Test (VUT) should be equipped with the operational system:

To be eligible for assessment, all system status indicators must be fitted to the vehicle as part of the assistance system.

To be eligible for assessment, it must not be possible for the assistance system to be used with the primary indicator disabled by the driver. This applies to visual, audible or haptic indicator related to the system".

Safety systems should be fitted, functional and in use. The vehicle should be fitted with dual commands for accelerator and brake, see further the section on safety driver below.Testing should be documented with cameras documenting the forward roadway, a camera viewing over the shoulder of the driver including driver hands-on-wheel status. Additional data regarding speed and control inputs can be logged using suitable commercial vehicle datalogger equipment. Data collected in the vehicle may include, speed, acceleration, braking, steering, forward and lateral proximity, and lateral position on the road. Data should be time synchronized to camera recordings. Subjective evaluation can be recorded via voice recording and transcribed. Derived measures may be response time, accuracy and safety performance.

Safety Driver

A trained safety driver should be present in the passenger seat when performing transfer-of-control on a public road. On a test track setting, the driver is expected to be a trained professional driver able to take back control in the undue event of system failure.

Environment and Weather conditions

Since HMI features are crucial to TOC functions care should be taken to test in all lighting conditions ranging from darkness to full sunshine. Light conditions of particular interest are direct sunlight, in particular from low standing sun (as this can cause drivers to squint which will affect both driver monitoring and drivers' perception of visual HMI features) as well as flickering sunlight (driving in sunlight along an avenue with trees planted on the sides) or rapidly changing lighting conditions, e.g. into or out of a tunnel in sunlight.

While weather conditions normally are a critical factor in validation of safety systems, the ToC procedure as such shall be expected to function in all weathers where automated features are enabled, and thus the TOC tests can be performed any unchallenging weather.

3.3.3. MEDIATOR Use Cases with transfers of control

The full list of MEDIATOR Use Cases is found in section 1.3.2. Transfers of control in the proper sense (to and from automation above SAE L2) is experienced in the following MEDIATOR use cases: 1, 2, 3, 5, and 10.

KPI - Safety performance

We are not convinced that there is a highly accurate, yet low-cost, assessment of actual safety performance. The assessment of safety performance of the Mediator-system and mediator-like systems are discussed extensively by Chandran et al (2023). They argue that all the components of the Mediator system must be present, operational and activated and the vehicle operating within the defined limits of the Operational Design Domain (ODD) for the full system safety benefits to be reached.

3.3.4. Methodology – scenarios used to test ToC in MEDIATOR

The following procedure description is derived from the on-road test 2 in Sweden.

The tests were designed, not only to evaluate the correctness of the transfer rituals, but also to evaluate how the appropriateness of the proposed TOC was perceived by the users. A missing item is to assess the completeness or quality of the transfer – how to know that the MEDIATOR system knows if the driver is ready to take over the responsibility for the driving task. Tables below outline how to collect data for the assessment

Table 3.4 UC1 procedure.

Procedure	Description
Scenario 3 (UC1)	Switch SAE level 0 to 3 due to NDRA (TOC manual to SB, UC 1). Present a high-priority NDRA to the participant while "driving" manually after the roundabout but while still in the city. The Mediator system detects the NDRA event and reacts by suggesting a switch to SAE level 3. If the participant agrees to switch to SAE level 3, proceed to scenario 4. If the participant chooses to stay in SAE level 0, instruct him/her to make the switch anyway (after a while), to be able to proceed to scenario 4.
Expected results and expected behaviour of the system	Mediator eye tracker can detect in-cabin glances. NDRA detection algorithm can detect the NDRA tasks. Comfort, compliance, and acceptance scores are higher with the Mediator system. The participant agrees to switch to SAE level 3 more often with the Mediator system

Table 3.5 UC 5a procedure.

Procedure	Description
Scenario 4 (UC5a)	Driving in SAE level 3 (SB). The participant is informed that it is essential that he/she can take over control within 5 second notice. If the driver becomes fatigued, the system initiates a corrective action to improve driver fitness (UC 4). The system monitors the effect of the countermeasure. 1. If the countermeasure is working, do nothing. 2. If the countermeasure is not working, the Mediator system switches to SAE level 4 (TtS) (UC5a)
Expected results and expected behaviour of the system	Mediator fatigue detection algorithm detects if the driver is sleepy. Corrective actions lead to reduced KSS values and blink durations. Mediator eye tracker can detect in-cabin glances. The outcome of the corrective action is correctly captured by the Mediator Fatigue detection algorithm. Comfort, compliance, and acceptance scores for the corrective action are higher with the Mediator system compared to baseline. Comfort, compliance and acceptance scores for the switch from SB→TtS are high.

Table 3.6 UC 5b procedure

Procedure	Description
Scenario 5 (UC5b?)	Planned switch from SAE level 3 or 4 --> level 2 (SB or TtS --> CM). Initiate appropriate procedure depending on the outcome of the previous scenario 4 (UC 5a) When approaching the ODD boundary: The Mediator system reacts by preparing the participant for and guiding the driver through a non-urgent transition to SAE level 2. If the driver is too fatigued for SAE level 2 driving: Initiate corrective fatigue countermeasures. If that does not help, initiate safe stop and take a break.
Expected results and expected behaviour of the system	Are fatigued drivers detected as fatigued? Is the driver's attention directed towards traffic before the actual switch to SAE level 2? Comfort, compliance and acceptance scores for the switch from SB or TtS→manual are higher for the Mediator system compared to baseline. In case of a safe stop, the driver understands why this happens.

Table 3.7 UC 10 Procedure

Procedure	Description
Scenario 10	Smooth transition from SAE level 4 to 3 (TtS→SB, UC 10) and driving in SAE level 3. The Mediator system detects sufficient automation fitness for driving in SB and reacts by informing the driver that SB will be switched on and subsequently monitors the required driver fitness.
Expected results and expected behaviour of the system	Comfort, compliance and acceptance scores for the corrective action are higher with the Mediator system compared to baseline. Mediator fatigue detection algorithm detects if the driver is too sleepy for SB. Mediator eye tracker can detect if the participants eyes are on-road.

It may be noted that MEDIATOR use cases lack a scenario where transfer from TtS goes directly to CS or manual driving, based on the reasoning that with functioning Automation Health assessment this will never happen, but the transition will always be to the Driver Standby Automation level.

3.3.5. Evaluations and KPIs

In section 1.3, KPIs for evaluating the different modules were reviewed. These KPIs were envisioned during the development of each Mediator module.The **Driver Module KPIs** include sensitivity and accuracy analysis, and other detection theory analysis/confusion matrix-based measures using self-assessed or annotated driver state as ground truth. **The Automation Module KPIs** include to check whether time to automation (un-)fitness and availability of the automated system is correctly evaluated based on its sensing limitations, and if identification of relevant infrastructure (road markings, road signs, etc.) is correctly predicted. The **Decision Logic Module KPIs** include counting Number and duration of safety critical events, Total time in safety critical situation– where a lower number indicates better performance. Finally there are the design level **HMI Module KPIs** - Usability, Acceptance, Workload, Trust, User experience, assessed by various psychometric scales and validated ranking/rating scales. These are simple to collect and analyse and can be used with prototypes on different levels of fidelity and are thus well suited for low-cost testing but may not all be relevant for the Euro-NCAP style evaluation of Transfer of Control. The final HMI Module KPIs, Mode awareness, and Overreliance, are more difficult to assess, as there is no true consensus as to their definitions. The following tables offer more details on observations and metrics used in On-Road test 2 in Sweden.

How are TOCs from automation to human experienced?

Table 3.8 Evaluate understanding of level of urgency in TOC from automation to human

Item	Description
Sub-question	Does the driver **understand the** different **levels of urgency** (based partly on driver monitoring data)?
Scenario	2, 5, 10 (i.e., all transitions to a lower mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	Observations of behaviour or using data on whether participants are responsive to the measures that is also being used in order to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

Table 3.9 Evaluate acceptance of TOC from automation to human.

Item	Description
Sub-question	Does the driver **accept these TOCs**?
Scenario	2, 5, 10 (i.e., all transitions to a lower mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	Observations of behaviour or using data on whether participants are responsive to the measures that is also being used to trigger warnings when participants don't react to a TOR, and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

Table 3.10 Evaluate driver behaviour adjustment in TOC to lower mode.

Item	Description
Sub-question	Does the driver **adjust his/her behaviour appropriately**?
Scenario	2, 5, 10 (i.e., all transitions to a lower mode)
Sensors	Not applicable
Data/measures/KPI	Video, observations, questionnaires
Expected results	Observations of behaviour or using data on whether participants are responsive to the measures that is also being used in order to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.

Table 3.11 Evaluate driver comprehension of TOC from higher to lower level of automation

Item	Description
Sub-question	How are these **TOCs evaluated** by the driver?
Scenario	2, 5, 10 (i.e., all transitions to a lower mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	**Observations of behaviour or using data** on whether participants are responsive to the measures that is also being used to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

How are TOCs from human to automation experienced?

Table 3.12 Evaluate driver acceptance of TOC from lower to higher level of automation

Item	Description
Sub-question	Does the driver **accept** these TOCs?
Scenario	3, 7, 9 (i.e., all transitions to a higher mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	**Observations of behaviour** or **using data** on whether participants are responsive to the measures that is also being used to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

Table 3.13 KPIs for evaluation of user behaviour in ToC rituals when transitioning to higher automation levels

Item	Description
Sub-question	Does the driver **adjust his/her behaviour appropriately**?
Scenario	3, 7, 9 (i.e., all transitions to a higher mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	**Observations of behaviour or using data** on whether participants are responsive to the measures that is also being used in order to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

Table 3.14 KPIs for evaluation of user experience of ToC to higher automation levels

Item	Description
Sub-question	How are these TOCs **evaluated by the driver?**
Scenario	3, 7, 9 (i.e., all transitions to a higher mode)
State inducement	Not applicable
Sensors	Video, observations, questionnaires
Data/measures/KPI	**Observations of behaviour or using data on whether participants are responsive** to the measures that is also being used in order to trigger warnings when participants don't react to a TOR and questionnaires after the full trip.
Expected results	The expectation is that TOCs are understood, accepted, evaluated well and that drivers adjust their behaviour appropriately as a response to the TORs. These are also understood better with Mediator compared to baseline. If this wouldn't be the case, we could explore here what could be potentially changed to improve the TOCs/response to the TORs.

In summary, much of the suggested evaluation are related to automation state module and decision logic. The expected results of the tests detailed above are on the scale of acceptance by the user/drivers. The acceptance of suggested of ToC is driven by the accuracy of the warning issued which is controlled by the combination of Mediator modules. As review in the KPIs, the evaluation of automation related tasks can only be done by data collection to understand the predicted vs observed metrics, in turn evaluating the accuracy of issues actions by the system. The issued action could vary based on use cases under testing. However, metrics such as true positive, false positive, true negative and false negative remain the same to assess the usability, safety, functionality, and acceptance of ToC. These are also evaluated in the on road-studies mentioned in section 2.3 as part of the development of Mediator systems. The analysis of collected data in the extended assessment can build the correlations between the different modules to assess effectiveness in the ToC between the driver and the system vice versa for Mediator-like systems.

4. Conclusions, Future Developments and Recommendations

This chapter takes off from the proposed protocol in section 3.3 and the MEDIATOR testing as described in section 2 as well as some elements from other protocols in section 1.41.4.2 to suggest a few potential extension of tests for mediator-like systems.

4.1. Overview of Mediator components and suitability for low-cost testing protocols

In the final of this deliverable, we rank the relative cost of the utilized methods and compare the benefit/contribution in different stages of the project, and in light of this also propose a few future extensions to the methods utilized in the MEDIATOR project for future research projects, as well as a short commentary about what subcomponents of the Mediator system lend themselves best to low-cost testing due to the nature of the studied systems.

4.1.1. Ranking methods for cost and validity

Depending on the maturity of the prototype and the development stage, following the Human-Centred Design process, the testing equipment and methods should be chosen judiciously to not tie up resources, firstly by selecting the right level of simulation fidelity, and then by determining the adequate experimental methodology. This is outlined in Table 4.1.

Very basic evaluations are referred to as desktop research, which can be a paper sketch or a storyboard, or even experiments or part-task simulations based on desktop computers with different complexity of interactions, which are used to go through sequences of events or interaction rituals with different levels of realism, immersion, and presence. As the next level of complexity driving simulator studies can take place, performed in a wide range of experimental setups, from fixed based simulator setups equipped with only basic functionalities, to high-fidelity flight simulators with all systems and instruments present and functional and full motion systems.

Obviously, by moving from low-fidelity paper charts prototypes to high-fidelity simulators, the costs increase significantly, whereas the easiness in changing, integrating, and testing a prototype decreases. Particularly for HMI concepts, passing from desktop research to simulator studies, requires an important additional effort in terms of cost and time. Between these two opposed options there is a wide-ranging spectrum of experimental setups that can be useful in human factors evaluations. For example, a head-mounted VR display (HMD) makes it possible for a user to experience an immersive digital environment, also by wearing a body or finger tracking system, providing engineers with a tool to evaluate their ideas, already in an early stage of the design process. In these kind of evaluations, human factors assessments are simplified by the architecture of the system, which is distinguishing for the easiness in integration of new cockpit concepts or single HMI components.

For ergonomic applications, it may be sufficient to have a fully static digital cockpit mock-up without any functionality, but to examine all cognitive aspects of an autonomous vehicle cockpit, an operational context is essential though. A digital cockpit mock-up approach surely can exploit the low entry barriers for developers, which is an important factor for the acceptance of a simulation environment but has some limitation when the user takes the control over the simulation and has to interact with cockpit elements.

Indeed, in the interaction in a fully virtual environment, experienced by using a VR headset, the user has no haptic feedback, which makes the usability of cockpit elements challenging. Even if the usability of the virtual cockpit elements can be enhanced by using some technological wizardry, a physical mock-up equipped with interactive hardware and connected with the simulation, continues to provide a higher level of fidelity. Adding physical hardware-controls to the simulation could be considered as a balance between a fully flexible, cost-effective virtual environment with no interactive elements, and one with costly real vehicle prototypes, which it stands out for a certainly decreased experimental flexibility.

Typically, a driving simulator consists of a physical mock-up, with driving commands (steering wheel and pedals) and a visualization system. The hardware configuration of these two components ranges from affordable desktop setups to complex installations consisting of, e.g., a 360-degree projection dome surrounding a real car and built on top of a moving system, often larger than a tennis court. The latter systems requires dedicated IT and logistics infrastructures, with multi-million investments and high running costs and this big technological investment is undertaken to provide higher fidelity, immersion, and realism to the user.

In a further attempt to rationalize the inverse relation between fidelity of the experimental environment, so its cost, and validity of results, we can consider the WoOZ approach as a specific case, where the full hardware mock-up does not weaken the flexibility of the experimental environment.

Table 4.1 An attempt at mapping cost and validity for a variety of research and testing methods/equipment. See Rauh (2023) for details.

Data collection tools	Cost	Validity	Used in MEDIATOR?
Storyboards, flowcharts and paper prototypes	Low	Low-Medium	Yes
VR ride-along	Low-medium	Medium	Yes
Low-fidelity simulator studies	Medium	Medium	Yes
Advanced full scale simulator studies	High	Medium	Yes
On road ‘non-functional’ WoOz vehicles	Medium-high	Medium	Yes
On road technical integration prototype based on computers	High	High	Yes
Actual dual-command WoOz vehicles	High	High	No

4.1.2. Cost and validity

Costs of a study can be defined by lead time and labour spent to prepare and carry out the study, as well as financial costs for developing and acquiring the needed equipment. In the context of a research project much effort is spend on method development, whereas if there is already a defined test protocol that entire cost can be omitted. Prototype or test vehicle cost remains.

4.2. Discussion about Future developments of Study/Methods and Mediator Components

This MEDIATOR Task was set up to explore and develop protocols for low-cost testing of safety aspects of mediator-like systems. Section 3 outlines an extension to the Euro NCAP protocol, based on scenarios in the MEDIATOR studies. Here in section 4, we reason about the suitability of the various methods and argue for potential future developments.

In section 2 we outlined the subcomponents of the Mediator system. In MEDIATOR we have embraced a holistic approach, meaning that all components and systems have been tested in their full, realistic, environment. Individual components testing has been limited to small, functional laboratory tests only. The extent and numbers of such a full system test could and should be possible to facilitate proper component testing, taking dependencies between modules into account.

4.2.1. Components vs methods

It can be argued that some of the different components making up a mediator-style system can be validated in sub-systems and component tests so as to take the most cost-effective approach available at each stage of a development project. Table 4.2 shows examples of different kinds of evaluations tools and their suitability for testing different MEDIATOR components. Thus, for example, some driver state sensing can only be studied and validated in the true setting, producing actual on-road human driver data, while the sensors for acquiring such data can be validated at much lower cost settings. Table 4.2

Table 4.2 Employed validation study methods for different sub-components in Mediator

Data collection tools	**Mediator sub-components**					
	Driver state assessment	**Automation health**	**Software / Decision-logic**	**Full HMI initial study**	**Full HMI validation**	**Full system validation**
VR-study	no	no	no	yes	no	no
Driving Simulators	no	partial	plausible	yes	yes	no
Computer simulations	no	no	yes	N/A	N/A	N/A
On-Road structured data collection	yes	No	no	no	N/A	N/A
On-road full Technical integration	yes	yes	plausible	no	no	no
On-road HMI WoOz study	plausible	plausible	plausible	no	yes	yes

Citing Ahlstrom et.al. (2023) “*Existing regulations typically specify that vehicle manufacturers must provide type approval authorities with documentation that describes the technical implementation of systems in sufficient detail for the authority to assess the robustness and functioning of the system.*

However, in lieu of technical regulations or guidelines, ultimate proof that a driver monitoring system meets requirements is in testing classification performance for a representative sample of human participants.

We propose that driver monitoring system testing should be conducted stepwise for (1) data quality and availability, (2) driver impairment detection performance and (3) intervention strategy, based on a combination of theory and realistic expectations. Data quality and availability are influenced by the ability of the sensors to track the driver's features reliably under as many different circumstances as possible, including variations in personal characteristics, clothing and accessories, weather related factors, etc. Using these data as input, algorithms estimate the driver's level of impairment. Ensuring good feature quality is the basis for the algorithms to be able to function properly. The impairment detection algorithms can vary in complexity"

Thus, we argue that driver state sensing systems can be validated in a layered approach where sensing systems can be validated by data collected in laboratory conditions to verify correctness of primary data such as eyelid openings or gaze directions, whereas the algorithm layer needs to be validated/verified with a representative sample of human participants. Once such data has been gathered, it can to a limited degree be used for validation of algorithms, but so far, no truly sensor-agnostic driver state sensing algorithm has been presented.

4.2.2. Automation state and decision logic

It may be argued that the final validation of a Mediator like system can only be studied in a full system technical integration vehicle. In the development stages of Mediator-like systems/modules and their parameters that affect the functionality can be tuned using hardware- and/or software-in-the-loop simulations. However, based on the extended protocol, proposed in Chapter 3.4, extended consumer rating is more suited to evaluate all components than individual modules. Despite that Vehicle-in-the-loop (VIL) simulation can be used to extract and evaluate relevant KPIs to assess the individual components on the Mediator like systems as road testing of such system are expensive and time-consuming.

VIL testing is an extension to traditional test track studies and fills the gap between software or hardware in the loop simulation and on-road testing. When applying to Mediator like vehicles, both Assistance Competence and Safety Backup can be assessed in a controlled setting. The basic principle of VIL testing involves the creation of a simulation environment that includes traffic, road signs, and road markings that would represent the use cases. The tester drives the real vehicle on the test track and the input to the sensor or algorithm comes from the simulated environment. Extending it even further would be to also involve the driver in the loop by adding a viewport into the simulated world using for example a Head-Mounted Display and the technology would in such a case be DVIL – Driver and Vehicle In the Loop. For Assisted drive testing, testers could measure the handoff is between the autonomous system and the human driver.

4.2.2.1. Automation state

The assessment Automation state during the VIL testing involves assessing the correctness of the prediction of the current automation level of the vehicle. Assessment should be done by comparing the predicted versus the designed "time budget" information, i.e., Time to Automation Unfitness or Time to Automation Fitness. These values should be computed after each automation state transition. In turn, the reliability of such metric calculation depends on sensor performance and accuracy of the data input from the simulation environment.

The data collected during VIL testing can be used to understand how much mismatch is occurring to a particular use case or ODD. The level of mismatch can then be used to grade the accuracy of performance of the automation state in isolation. Similarly, performance of the system automation state can also be assessed by number of times the tester overrides the recommended automation level by the automation state module.

4.2.2.2. Decision logic

Similar assessment methods can be followed to assess the correctness and accuracy of the recommendation, given by the decision logic module. As the main functionality of the module is to understand who is best suited to drive in the given driving condition, driver or automation. Both automation state and the decision logic module work in conjunction to give a suitable hand-over between different automation levels. Similarity can be drawn to the assessment of Assisted competence protocols. However, Decision logic also controls the corrective action recommendation and the HMI. The protocol should be extended to also cover safety backup. The overall Mediator system does not control the basic safety functionality of the vehicle. Rather, it only mediates between the automation states (manual and assisted). Thus, the recommended assessment would be to assess the corrective action, recommend by the module, versus how the system was designed for VIL testing assessment. The mismatch or accuracy can be put on a scale to grade the overall mediator like system.

4.2.3. HMI and testing the full MEDIATOR concept

The HMI has a paramount importance in current vehicles and this paramount importance will be the same in automated vehicles and, in particular, in the middle levels of automated driving, when a high synergy between the human being and the vehicle is needed to ensure a safe human-vehicle behaviour.

To guarantee the design of a usable and acceptable HMI and of a positive user experience for automated vehicles too, it is necessary to design around the human being. To reach this goal, it is advisable to follow the ISO 9241-210:2019 (Ergonomics of human-system interaction - Part 210: Humantred design for interactive systems) I.e., requirements and recommendations for human-centred design principles and activities phases throughout the life cycle of systems, form conception, to design and, implementation, up to use and maintenance phases and activities. Two of the principles of this standard are the iteration of the design process and the user-centred evaluation of the iterative design proposals. Then, throughout the design process parallel HMI solutions are designed using different kinds of prototyping modalities, so to have artifacts that users can use and evaluate. This must be done during all design phases, starting from the very beginning through the use of low-fidelity prototypes (e.g. HMI paper mock-ups) up to the involvement of higher fidelity ones, which can be HMI prototypes installed in a Wizard of Oz vehicle prototype, in case f automated vehicles are not technologically ready. passing through medium fidelity HMI prototypes in software mock-ups on computers, VR, or in driving simulator facilities. And the different levels of prototypes should be tested involving naïve users (without forgetting experts) so to reach the final HMI version, having ameliorated the human interaction with the HMI during the iterative design and evaluation cycles.

Wizard-of-Oz

The WoOz is a research approach, born in the Human Computer Interaction domain, in which the user interacts with a system that he/she believes to be real, but which is actually controlled, completely or partially, by a human being, the Wizard. The Wizard-of-Oz (WoOz) paradigm was invented in 1975 by John F. Kelley to simulate a not yet functional speech recognition system (Green & Wei-Haas, 1985). According to a recent bibliographic collection (Bengler et al., 2020) the WoOz paradigm has allowed research on human-machine interaction in parallel with the technical development. Indeed, within the automotive research community, WoOz vehicles are started to be used as a method for analysing the effects of "intelligent" probabilistic systems, as automated vehicles are, even if not yet fully developed yet.

There are different layouts in which a WoOz vehicle can be built, and in which the user can drive or not. In the not functional WoOz set up, like the one used in the MEDIATOR HF vehicle, the vehicle is a right-hand drive one, with the user on the front left seat with siulated steering wheel and pedals, while the wizard is on the right seat driving with the true vehicle controls.

Thanks to the presence of the wizard, a lot of tasks not already implemented from the technical point of view, can be simulated, and tested by the user, who must be observed by an experimenter (the third person in a WoOz vehicle) while doing those tasks. It is crucial to also record context parameters in which each trial is conducted (e.g., speed, traffic, …), in a changing environment, different from what happened in the classical laboratory WoOz experiments.

It is fundamental to instruct the wizard how he or she needs to drive, for example with a smooth and conservative style, to consistently reproduce the automated driving style of interest, to enhance the users' impression that they are interacting with a real automated vehicle.

The advantage of the WoOz prototype vehicle was to allow the user to experience somehow the automated driving and to test the MEDIATOR HMI solutions in an ecological way on the real road.

Due to the many variables (e.g., prototype different set-ups, wizard driving style, user's impression to be on an ADS, changing context, …) in an WoOz vehicle on-road user testing simulating a real automated driving system user testing, it is advisable to work in the scientific community to define a common minimum set of method requirements.

4.2.4. VR-testing and evaluation

In the short-term, Virtual Reality as a research tool in Human Factors studies is likely to become as important as it is in vehicle design where the objective is generally user experience or brand experience, and sensory ergonomics. Former limitations e.g., the availability of appropriate environments or their quality have become marginal. In a somewhat longer term, the metaverse has great potential as a design tool. In general, the metaverse has yet to obtain its place and role, people's expectations ranging from a utopian world to explore, to merely a next social network application. For designers however, who must be systemic (not be confused with system designers) the potential of the metaverse is imminent as it allows to construct and wander around in complete systems and their interrelated levels. If and how that value also applies to Human Factors research, as with VR, depends on the extent to which HF research needs larger contexts.

A driving simulator conceived for research and development in the field of automated driving, must not only consider the increasing level of automation, but also that drivers will be able to gradually phase-out control over the driving tasks and transfer it to the automated systems. This will enable the driver to perform a variety of non-driving tasks that will become an integral part of the on-board experience. Moreover, some safety-critical issues may arise especially in transfer-of-control conditions, where the automated system and the human driver need to effectively communicate to each other their intensions and actions. Therefore, a simulator must not just enable the study of the interactions between drivers, vehicles and road users, but it is foreseeable that the future development of driving simulators will focus more on enabling onboard connectivity, driver monitoring and interaction concepts and technologies, rather than developing further complex motion systems. Indeed, the variety of motion-based driving simulators that are available nowadays, already offer a good level of realism in motion cues, with the opportunity of further development that is physically limited by the insurmountable boundaries posed by a confined workspace. Instead, driving simulators should be upgraded with multi-sensory interfaces and reconceived as living spaces where humans act out-of-the-loop

4.3. Summary and recommendations

The MEDIATOR project set out to create a new concept to bridge the gap between current driver assistance systems and future fully automated driving systems, in terms of driver state sensing, automation health and human-machine understanding, or in more academic terminology, mode awareness. From the current report it follows that not all methods used in MEDIATOR were low-cost, considereing the development cost for prototypes and method development, see section 2.62.5.

During development of a fully novel HMI-concept and even functional architecture, many hurdles must be handled along the way. Testing and validation methods used will necessarily be increasingly more costly and complex, reflecting the ever-increasing maturity of the system throughout the development process.

For final system-integration testing with target mass-produced hardware etc, the requirements for testing and validation are described in detail in a multitude of automotive industry standards. Such testing would ensure for example the SOTIF – safety of the intended function, that whatever function was designed functions as intended. It would however not cover any aspects of whether the function was designed properly for the optimal cooperation with human drivers.

This report has reviewed some of the methods throughout the development process as well as proposed future developments of such methods. In addition, we have looked at independent testing protocols for automated and/or assisted driving functions and proposed necessary extensions to cover the full utility of the Mediator system as is intended. When it is eventually time for a 'ready-for-market' system to be assessed in independent benchmarks, parameters such as hardware costs, development costs, and method development costs would be of less importance, since the vehicle is already developed, and the assessment method/protocol including pass/fail criteria would already be formed.

It should be added that, the proposed extension of the assessment protocol recommends that the "Mediator-like" system should have a clear definition on how it must be translated to consumer understanding, stating both the usability and safety aspect of such a system.

4.4. Conclusion

Independent testing of Mediator-like vehicles will need to expand their methodology to evaluate transfers-of-control, considering measurements of trust, mode-awareness and overreliance.

Furthermore, for the benefit of future research projects in the same spirit as MEDIATOR, we recommend continuing with the adoption of the Human Centred Design process in the future definition of automated vehicles HMI. This process will allow to create HMI with better usability and acceptance, thanks to:

- iterative evaluations and redesign during the design cycle (by experts and naïve users),
- using proper prototype types (from low to high fidelity)
- in proper test environments (from usability laboratory to vehicle prototypes, passing through driving simulator and WoOz vehicles),
- allowing meanwhile to maintain under control resources,
- in terms of lead time, material and labor costs keeping the effort calibrated to the each step of the design cycle

Component testing is advisable and needed, but for such a system-wide approach as the Mediator concept a full-system test is unconditionally needed. For example, for true validation of driver state sensing requires driver state induction in the true driving setting which means labour-intense and thus costly testing. However, in most other areas of the Mediator system development, major cost savings can be achieved by utilizing novel tools and technologies such as virtual reality and wizard-of-oz testing where they provide the best cost-to-validity trade-off, as well as reducing the scope and extension of sub-system tests by clever design, selection and implementation of testing, utilising divisions between sensor-, feature- and functional tests.

References

Athmer, C., Spaan, M.T.J., Li, Y., Liu, Y., Bakker, B. (2022). Simulation of Decision Making in the Mediator System, Deliverable D3.2 of the H2020 project MEDIATOR

Bakker, B., Knauss, A., Rodrigues de Campos, G., Tanov, N., Andersson, S., Mano, D., Athmer, C., Enhuber, S.M., Beggiato, M., Spaan, M. (2022). Development of integrated lab prototypes, Deliverable D2.9 of the H2020 project MEDIATOR

Bengler K., Omozik, K., Müller, A. I. (2020) The Renaissance of Wizard of Oz (WoOz) – Using the WoOz methodology to prototype automated vehicles. In Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Beringer, D. B. (1994, October). Issues in using off-the-shelf PC-based flight simulation for research and training: Historical perspective, current solutions and emerging technologies. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 38, No. 1, pp. 90-94). Sage CA: Los Angeles, CA: SAGE Publications.

Borowsky, A., Oron-Gilad, T., Chasidim, H., Ahlström, C., Karlsson J.G., Bakker, B., … Christoph, M. (2020). Behavioural Markers for Degraded Human Performance. Deliverable D1.2 of the H2020 project MEDIATOR

Bruck, L., Haycock, B., & Emadi, A. (2020). A review of driving simulation technology and applications. IEEE Open Journal of Vehicular Technology, 2, 1-16.

de Frutos, S. H., & Castro, M. (2021). Assessing sim racing software for low-cost driving simulator to road geometric research. Transportation research proceedings, 58, 575-582.

Chandran, T., Jan, H., Morando, A., Yuan, Y. et al. (2023). Assessment of safety and related societal benefits, Deliverable D4.1 of the H2020 project MEDIATOR.

Christoph, M., Cleij, D., Ahlström, C., Bakker, B., Beggiato, M., Borowsky, A., …Van Nes, C.N (2019). Mediating between human driver and automation: state-of-the art and knowledge gaps. D1.1 of the H2020 project MEDIATOR.

Cleij, D., Bakker, B., Borowsky, A., Christoph, M., Fiorentino, A., van Grondelle, E., Mano, D, van Nes, N. (2020). Mediator System and Functional Requirements, Deliverable D1.4 of the H2020 project MEDIATOR

Consumer Reports (2023) CR Rates Active Driving Assistance Systems - Consumer Reports https://www.consumerreports.org/cars/car-safety/active-driving-assistance-systems-review-a2103632203/

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017, October). CARLA: An open urban driving simulator. In Conference on robot learning (pp. 1-16). PMLR.

ECE/TRANS/WP.29/2022/59/Rev.1.https://unece.org/sites/default/files/2022-05/ECE-TRANS-WP.29-2022-59r1e.pdf

Euro NCAP (2020) Euro NCAP Vision 2030: a safer future for mobility, https://cdn.euroncap.com/media/74468/euro-ncap-roadmap-vision-2030.pdf

Euro NCAP ASSESSMENT PROTOCOL – SAFETY ASSIST SAFE DRIVING Implementation 2023 Version 10.1.1 November 2022

Fiorentino A., Bakker B., Brase D., Karlsson J., Knauss A., Santhià T., Thalya P., Toffetti A., van Grondelle E., Wang, D., (2022). Final in-vehicle prototypes, Deliverable D2.10 of the H2020 project MEDIATOR

Georgiou, T., & Demiris, Y. (2016, September). Personalised track design in car racing games. In 2016 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-8). IEEE.

Green, P., & Wei-Haas, L. (1985). The Rapid Development of User Interfaces: Experience with the Wizard of OZ Method. Proceedings of the Human Factors Society Annual Meeting, 29(5), 470–474. https://doi.org/10.1177/154193128502900515

IIHS (2020) https://www.iihs.org/news/detail/automated-systems-need-stronger-safeguards-to-keep-drivers-focused-on-the-road

IIHS (2022a) https://www.iihs.org/news/detail/iihs-creates-safeguard-ratings-for-partial-automation

IIHS (2022b) https://www.iihs.org/news/detail/consumers-want-safeguards-with-partial-automation

ISO 9241-210: 2010. Ergonomics of Human System Interaction-Part. 210: Human-Centred Design for Interactive Systems; International Standardization Organization (ISO): Geneva, Switzerland, 2019

Jansen, R. J., Tinga, A. M., de Zwart, R., & van der Kint, S. T. (2022). Devil in the details: Systematic review of TOR signals in automated driving with a generic classification framework. *Transportation Research Part F: Traffic Psychology and Behaviour*, *91*, 274–328. https://doi.org/10.1016/j.trf.2022.10.009

Koskela, K., Nurkkala, V. M., Kalermo, J., & Järvilehto, T. (2011, July). Low-cost driving simulator for driver behavior research. In proceedings of the 2011 international conference on computer graphics and virtual reality, CGVR (Vol. 11, pp. 18-21).

Lu, B., Wu, X., & Monti, A. (2005, November). Implementation of a low-cost real-time virtue test bed for hardware-in-the-loop testing. In 31st Annual Conference of IEEE Industrial Electronics Society, 2005. IECON 2005. (pp. 6-pp). IEEE.

Mano, D., Berger, N., Larsson, A., Brännström, M., Knauss, A., Toffetti, A., … Christoph, M. (2021). Quantified Markers for Degraded Automation Performance, Deliverable D1.3 of the H2020 project MEDIATOR.

Naujoks, F., Wiedemann, K., Schömig, N., Hergeth, S., & Keinath, A. (2019). Towards guidelines and verification methods for automated vehicle HMIs. Transportation Research Part F: Traffic Psychology and Behaviour, 60, 121–136. https://doi.org/10.1016/j.trf.2018.10.012

Narayanasamy, V., Wong, K. W., Fung, C. C., & Rai, S. (2006). Distinguishing games and simulation games from simulators. Computers in Entertainment (CIE), 4(2), 9-es.

Rauh, N., Springer-Teumer, S., Spaan, M., Borowsky, A., Fiorentino, A., Ahlström, C., Thalya, P. (2023). Integration of evaluation results from the MEDIATOR studies, Deliverable D3.5 of the H2020 project MEDIATOR

Sena, P., d'Amore, M., Brandimonte, M. A., Squitieri, R., & Fiorentino, A. (2016). Experimental framework for simulators to study driver cognitive distraction: brake reaction time in different levels of arousal. Transportation research proceedings, 14, 4410-4419.

Spyridakos, P. D., Merat, N., Boer, E. R., & Markkula, G. M. (2020). Behavioural validity of driving simulators for prototype HMI evaluation. IET Intelligent Transport Systems, 14(6), 601–610. https://doi.org/10.1049/iet-its.2018.5589

Tiu, J., Harmon, A. C., Stowe, J. D., Zwa, A., Kinnear, M., Dimitrov, L., ... & Carr, D. B. (2020). Feasibility and validity of a low-cost racing simulator in driving assessment after stroke. Geriatrics, 5(2), 35.

van Gent, P., Farah, H., van Nes, N., & van Arem, B. (2019). Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project. Journal of Open Research Software, 7(1).

Weinberg, G., & Harsham, B. (2009, September). Developing a low-cost driving simulator for the evaluation of in-vehicle technologies. In Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 51-54).

Weinberg, G., & Harsham, B. (2009, September). Developing a low-cost driving simulator for the evaluation of in-vehicle technologies. In Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 51-54).

Wymann, B., Espié, E., Guionneau, C., Dimitrakakis, C., Coulom, R., & Sumner, A. (2000). Torcs, the open racing car simulator. Software available at http://torcs. sourceforge.net, 4(6), 2.

Yang, L., Morland, T. B., Schmits, K., Rawson, E., Narasimhan, P., Motelow, J. E., ... & Blumenfeld, H. (2010). A prospective study of loss of consciousness in epilepsy using virtual reality driving simulation and other video games. Epilepsy & Behavior, 18(3), 238-246.

Yavrucuk, I., Kubali, E., & Tarimci, O. (2011). A low cost flight simulator using virtual reality tools. IEEE Aerospace and Electronic Systems Magazine, 26(4), 10-14.

Appendix A Study leader questionnaire

The questionnaire was intended to cover most project management related items for the individual studies, in order to get an overview of the lead time and labour time resources needed, prototype and other equipment finance/budget/cost, as well as the overall Scope, Risk and expected Deliverables for the studies.

Questions

- "What was the primary focus of your study?
- (HMI, Decision logic, Driver State sensing etc)"
- What did you do?
- Which were your main use cases/scenarios? (Note - clean up use case listing later)
- Given what you learned performing this study – what advice would you bring forward?
- What dimensions of user performance are similar enough to make a meaningful comparison between studies?
- What is the most defining feature of your study in terms of Results:
- What is the most defining feature of your study in terms of External Validity
- What is the most defining feature of your study in terms of Statistical power
- What main advantages has your selected method/equipment
- Why did you choose this over other methods?
- What simplifications did you make compared to an 'ideal' study setting (ideal setting having little or no time' and financial' constraints)
- What consequences did that have for the final results?
- How would you describe advantages and drawbacks of your study in terms of: 'Internal validity (contrasting w. baseline or control)
- How would you describe advantages and drawbacks of your study in terms of: 'External validity (real world applicability)