



TECHNISCHE UNIVERSITÄT MÜNCHEN
TUM SCHOOL OF ENGINEERING AND DESIGN

Usability Assessments in User Studies on Human-Machine Interfaces for Conditionally Automated Driving: Effects of the Context of Use

Deike Albers

Vollständiger Abdruck der von der TUM School of Engineering and Design
der Technischen Universität München
zur Erlangung des akademischen Grades einer
Doktorin der Ingenieurwissenschaften (Dr.-Ing.)
genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. Johannes Fottner

Prüfende der Dissertation: 1. Prof. Dr. Klaus Bengler
2. Prof. Dr. Mark Vollrath

Die Dissertation wurde am 07.12.2023 bei der Technischen Universität München eingereicht
und durch die TUM School of Engineering and Design am 19.04.2024 angenommen.

Acknowledgment

I wish to express my profound gratitude to **Prof. Klaus Bengler**, without whom this thesis would not have been possible. His support and guidance, particularly during the most critical phases of this research, are deeply appreciated.

The same applies to all my colleagues at the **Chair of Ergonomics**. This incredible team was always there to provide support and feedback, often offering their assistance without me having to ask. Their companionship, ideas and sense of humor made each day at work enjoyable. My special thanks go to **Caroline Adam, Annika Boos, Svenja Escherle, Alexander Feierle, Martin Fleischer, Nicole Fritz, Tanja Fuest, Julia Graefe, Tobias Hecht, Doris Herold, Olivia Herzog, Maximilian Hübner, Luis Kalb, Burak Karakaya, Manuel Kipp, Theresa Prinz, Michael Rettenmaier, Jonas Schulze, and Lorenz Steckhan**. Furthermore, I am grateful to **Bianca Biebl** and **Dominik Janetzko** for their valuable statistical expertise and great company. A heartfelt thanks go to **Niklas Grabbe**, who provided invaluable support during the Multilab project, helping me maintain a relaxed perspective during critical times. To my colleague and mentor, **Jonas Radlmayr**, I owe a special debt of gratitude. His guidance and motivational words played a crucial role in helping me complete this thesis.

The empirical work was made possible through the support of several external partners, particularly during the trying circumstances of the pandemic. I am grateful to **Thomas Rottmann** from the Universität der Bundeswehr in Neubiberg and Satoshi Kitazaki from Japan's National Institute of Advanced Industrial Science and Technology (AIST). **Jan-Christian Ickrath** and **Lutz Wirsig** from the Bertrandt AG demonstrated remarkable technical expertise, ensuring the experimental setup worked seamlessly and reliably. Additionally, the student theses of **Julia Graefe, Jessica Kos, Niklas Mooshofer, and Canroz Tacay** significantly contributed to my empirical work.

I am also grateful for the support of my friends from the 'Human Factors Engineering' master's program, namely **Lukas Flohr, Dominik Janetzko, and Simone Nertinger**, who provided both scientific feedback and much-needed distractions. I also want to thank my **FF-friends** from Leer, who ensured I maintained a healthy work-life balance. My special thanks go to **Amke Nimmrich** for the great 'Diss Camp' and for allowing me to call her for encouraging words literally around the clock. To **Felix-Wilhelm Siebert** I am grateful for being something like my academic mentor for almost 10 years already as well as a good friend.

My family's endless support has been the cornerstone of my life. I thank my parents, **Else** and **Dietmar Albers**, for making everything possible and encouraging me on my journey. The same gratitude extends to my brother, **Arend Albers**, and my sister, **Amke Gezeck**, who played a significant role in shaping the person I am today. Besten Dank an mien anner Verwandte un Bekannte in'd mooi Ostfreesland. Ik kiek bold weer rin för en Koppke Tee.

Finally, I want to express my deepest thanks to **Stefan Junold** for always being there for me. His unconditional support, patience, and calm demeanor have been a constant source of strength throughout these past years.

Abstract

The introduction of conditionally automated driving (CAD) entails a paradigm change in automotive mobility. For the first time, the driver is temporarily released from the responsibility of the driving task. This paradigm change challenges the development of human-machine interfaces (HMIs) facilitating the intended and safe interaction. User studies on the usability of such HMIs are commonly conducted in driving simulators and within one single culture. Identifying the potential effects of this context of use is crucial for the validity of research conducted in the HMI development. Following a review of the relevant literature, five research questions are derived that are addressed in this thesis.

A systematic literature review offers insights into common research practices of studies on the usability of HMIs for CAD. Following, a best practice advice is developed. The advice builds the basis for the experimental design for two of the three validation studies conducted in this thesis (*Exp_Testing-Environment & Exp_Culture*).

The first validation study, *Exp_Testing-Environment*, investigates the effect of the testing environment on usability assessments. An experiment conducted in a static driving simulator is compared to an otherwise identical experiment conducted in an instrumented vehicle on a test track. The findings suggest relative validity but no absolute validity. The study concludes that problems with HMI concepts identified in the driving simulator will likely be more pronounced in test track experiments. Based on the findings, driving simulators are deemed a valid tool.

The second validation study, *Exp_Culture*, investigates the effect of the users' cultural background on the usability assessment by comparing the usability ratings of U.S.-American participants to German participants. Regarding absolute validity, the database needs to be more conclusive. The findings, however, confirm relative validity. The study concludes that the results of usability assessments may be transferred across cultures of the Western industrialized world. Limitations are expected only regarding the usability facet satisfaction.

The third validation study, *Survey_Culture*, addresses the effect of the users' cultural background on the subjective importance ratings of usability factors. The comparison of U.S.-American and German ratings shows neither considerable nor systematic cultural effects. In line with *Exp_Culture*, this study concludes that usability assessments may be conducted within one culture of the Western industrialized world.

The findings of the three validation studies are consolidated in a set of preliminary recommendations. The set is discussed and refined in an expert workshop. The final 12 recommendations suggest methods for conducting user studies on the usability of HMIs in the context of CAD.

This thesis provides novel empirical findings on experimental methods in user studies on usability assessments, focusing on the validity of usability assessments in varying contexts of use. Based on prevalent literature and an expert workshop, the results are consolidated and refined. Concluding, the thesis contributes to the advancement of valid research methods for conducting usability assessments of HMIs for CAD.

Zusammenfassung

Die Einführung des Hochautomatisierten Fahrens (CAD) führt zu einem Paradigmenwechsel. Zum ersten Mal wird der Fahrer vorübergehend von der Verantwortung für die Fahraufgabe entbunden. Dieser Paradigmenwechsel stellt eine Herausforderung für die Entwicklung von Mensch-Maschine-Schnittstellen (HMIs) dar, welche die angestrebte und sichere Interaktion fördern. Nutzerstudien zur Gebrauchstauglichkeit (Usability) solcher HMIs werden üblicherweise in Fahrsimulatoren und innerhalb einer einzigen Kultur durchgeführt. Die Identifizierung möglicher Auswirkungen dieses Nutzungskontexts ist von entscheidender Bedeutung für die Validität der Forschung in der HMI-Entwicklung. Nach Sichtung relevanter Literatur werden fünf Forschungsfragen abgeleitet, die in dieser Arbeit behandelt werden.

Eine systematische Literaturrecherche bietet Einblicke in gängige Forschungspraktiken von Usability-Studien zu HMIs für das CAD. Ein Leitfaden wird entwickelt, der die Grundlage für das experimentelle Design von zwei der drei in dieser Arbeit durchgeföhrten Validierungsstudien (*Exp_Testing-Environment & Exp_Culture*) bildet.

Die erste Validierungsstudie, *Exp_Testing-Environment*, untersucht die Auswirkungen der Testumgebung auf die Usability-Bewertung. Ein in einem statischen Fahrsimulator durchgeföhrtes Experiment wird mit einem ansonsten identischen Experiment auf einem Testgelände verglichen. Die Ergebnisse deuten auf relative, aber keine absolute Validität hin. Die Studie kommt zu dem Schluss, dass Probleme mit HMI-Konzepten, die im Fahrsimulator identifiziert werden, im Testgelände stärker ausgeprägt sind. Auf Grundlage der Ergebnisse werden Fahrsimulatoren als valide Versuchsumgebung erachtet.

Die zweite Validierungsstudie, *Exp_Culture*, untersucht den Einfluss des kulturellen Hintergrunds auf die Usability-Bewertung, indem Ergebnisse von Proband*innen aus den Vereinigten Staaten Amerikas und Deutschland verglichen werden. Hinsichtlich der absoluten Validität ist die Datenbasis nicht eindeutig. Die Ergebnisse bestätigen jedoch relative Validität. Die Studie kommt zu dem Schluss, dass die Ergebnisse von Usability-Bewertungen auf andere Kulturen westlicher Industrieländer übertragbar sind. Einschränkungen sind lediglich bei der Zufriedenheit, einer Facette von Usability, zu erwarten.

Die dritte Validierungsstudie, *Survey_Culture*, befasst sich mit dem Einfluss des kulturellen Hintergrunds auf die subjektive Wichtigkeit von Usability-Faktoren. Der Vergleich der U.S.-amerikanischen und deutschen Bewertungen zeigt weder erhebliche noch systematische Effekte. Es wird abgeleitet, dass Usability-Bewertungen innerhalb einer Kultur der westlichen industrialisierten Welt durchgeföhr werden können.

Die Ergebnisse der drei Validierungsstudien führen zur Formulierung von vorläufigen Empfehlungen. Diese werden in einem Expertenworkshop diskutiert und weiterentwickelt. Die abschließenden 12 Empfehlungen schlagen Methoden für die Durchführung von Nutzerstudien zur Usability von HMIs im Kontext des CAD vor.

Diese Arbeit liefert neue empirische Erkenntnisse zu experimentellen Methoden in Nutzerstudien zur Usability-Bewertung mit Fokus auf der Validität in unterschiedlichen Nutzungskontexten. Die Ergebnisse werden auf Basis bestehender Literatur und eines Expertenworkshops konsolidiert und verfeinert. Somit leistet die Arbeit einen wertvollen Beitrag zur Weiterentwicklung valider Forschungsmethoden zur Durchführung von Bewertungen der Usability von HMIs für das CAD.

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Structure	2
2	Theoretical Foundation	3
2.1	Terms and Definitions.....	3
2.1.1	Automated Driving and Related Terms.....	3
2.1.2	Usability and Related Terms	5
2.1.3	Validity.....	7
2.1.4	Testing Environment	8
2.1.5	Culture	9
2.2	Overview of Methods for Assessing Usability.....	11
2.2.1	Field Methods.....	12
2.2.2	Inspection Methods	12
2.2.3	Usability Testing	12
2.2.4	Focus Groups, Interviews, and Surveys	13
2.3	Effects of the Testing Environment Driving Simulator	14
2.3.1	Driving Simulator as a Valid Research Tool.....	15
2.3.2	Driving Simulator Validation Studies in the Automotive Context.....	15
2.3.3	Driving Simulator Validation Studies in the Automated Driving Context... <td>16</td>	16
2.4	Effects of the Users' Cultural Background	16
2.4.1	Effects of Culture in the Data Collection Phase	17
2.4.2	Effects of Culture on Interface Designs	17
3	Research Questions	21
4	Development of a Best Practice Advice for Assessing the Usability of HMIs for L3 ADS	24
4.1	Analysis of the Status Quo of Common Research Methods and Findings.....	24
4.1.1	Definition of Usability.....	24
4.1.2	Testing Environment	24
4.1.3	Sample Characteristics	25
4.1.4	Test Cases.....	25
4.1.5	Dependent Variables	25
4.1.6	Conditions of Use.....	25
4.2	Derivation of a Best Practice Advice	25
5	Experimental Design for Validation Studies Exp_Testing-Environment and Exp_Culture	27
5.1	Definition of Usability	27
5.2	Sample Characteristics	27
5.3	Test Cases.....	28
5.4	HMI Concepts	29
5.4.1	Basic Design.....	30
5.4.2	Differences between the two HMI Concepts.....	32
5.4.3	Heuristic Expert Evaluation on Differences between the HMI Concepts	35
5.5	Study Procedure	37
5.5.1	NDRA.....	37
5.5.2	Instructions	38
5.5.3	Test Course.....	38
5.6	Data Collection.....	39
5.6.1	Sociodemographic Data	40
5.6.2	Observational Data	40

5.6.3	Self-Reported Data	42
5.6.4	Interindividual Factors	43
5.6.5	Overview and Embedding of the Dependent Variables	43
5.7	Data Analysis	45
5.7.1	Step 1: Modeling the Data Structure	45
5.7.2	Step 2: Equivalence Test for the Factor Experiment.....	46
5.7.3	Interpretation of Results	46
6	Validation Study Exp_Testing-Environment: Effect of the Testing Environment on Metrics for Assessing Usability of HMIs for L3 ADS in User Studies	48
6.1	Hypotheses	48
6.2	Sample.....	49
6.3	Results	50
6.3.1	Observational Metrics	50
6.3.2	Self-Reported Metrics	63
6.3.3	Interindividual Factors	79
6.4	Discussion	79
6.4.1	Summary of Results	79
6.4.2	Discussion of Hypotheses	81
6.4.3	Discussion of Limitations and Other Observations.....	82
6.4.4	Conclusion.....	83
7	Validation Study Exp_Culture: Effect of the Users' Cultural Background on Metrics for Assessing Usability of HMIs for L3 ADS in User Studies	85
7.1	Hypotheses	85
7.2	Sample.....	87
7.3	Results	88
7.3.1	Observational Metrics	88
7.3.2	Self-Reported Metrics	102
7.3.3	Interindividual Factors	117
7.4	Discussion	118
7.4.1	Summary of Results	118
7.4.2	Discussion of Hypotheses	119
7.4.3	Discussion of Limitations and Other Observations.....	121
7.4.4	Conclusion.....	122
8	Validation Study Survey_Culture: Effect of the Users' Cultural Background on the Subjective Importance Rating of Usability Factors in the Context of HMIs for L3 ADS	124
8.1	Hypotheses	124
8.2	Method	125
8.2.1	Sample	125
8.2.2	Data Collection.....	127
8.2.3	Study Procedure	128
8.2.4	Data Analysis	128
8.3	Results	130
8.3.1	Cultural Dimensions.....	130
8.3.2	Subjective Rating of Importance of Usability Factors regarding HMIs for L3 ADS.....	131
8.4	Discussion	134
8.4.1	Discussion of Results and Hypotheses	134
8.4.2	Discussion of Limitations.....	136
8.4.3	Conclusion.....	136

9 Expert Workshop: Recommendations for Methods for Assessing Usability of HMIs for L3 ADS in User Studies	138
9.1 Method	138
9.2 Preliminary Recommendations and Results of the Expert Workshop	139
9.3 Derivation of Final Recommendations.....	146
9.3.1 Recommendation 1.....	147
9.3.2 Recommendation 2.....	148
9.3.3 Recommendation 3.....	148
9.3.4 Recommendation 4.....	149
9.3.5 Recommendation 5.....	149
9.3.6 Recommendation 6.....	150
9.3.7 Recommendation 7.....	150
9.3.8 Recommendation 8.....	150
9.3.9 Recommendation 9.....	151
9.3.10 Recommendation 10.....	151
9.3.11 Recommendation 11.....	152
9.3.12 Recommendation 12.....	152
9.4 Discussion and Outlook	152
10 Conclusion	154
10.1 Answers to the Research Questions	154
10.1.1 RQ ₁ : Based on Common Research Methods and Findings, what is the Best Practice Advice for an Experimental Design for Assessing the Usability of HMIs for L3 ADS?.....	154
10.1.2 RQ ₂ : Which Effect has the Testing Environment on Metrics for Assessing the Usability of HMIs for L3 ADS?	155
10.1.3 RQ ₃ : Which Effect has the Users' Cultural Background on Metrics for Assessing the Usability of HMIs for L3 ADS?	155
10.1.4 RQ ₄ : Which Effect has the Users' Cultural Background on the Subjective Importance Rating of Usability Factors in the Context of HMIs for L3 ADS?156	156
10.1.5 RQ ₅ : Which Methods are recommended for Assessing the Usability of HMIs for L3 ADS?.....	156
10.2 Reflections on Limitations and Generalizability	157
10.2.1 Limitations	157
10.2.2 Generalizability	158
10.3 Future Work	159
10.4 Key Messages.....	160
11 References	161
12 Appendix I	174
13 Appendix II	181
14 Appendix III	191

List of Figures

Figure 2.1 Overview of the LoAs according to the standard J3016 (SAE International, 2021).	5
Figure 3.1 Overview of the structure of the thesis. To enhance the understandability, chapters covering research questions or empirical data are colored differently.....	21
Figure 5.1 <i>Left</i> : Structure of the basic HMI design implemented in the <i>IC</i> : (1) speed; (2) infotainment; (3) ego vehicle and its surrounding environment (L2 & L3 only); (4) scale for LoAs. <i>Right</i> : Icons for the three implemented LoAs L0 (left), L2 (center), and L3 (right).	30
Figure 5.2 <i>Left</i> : Position of the control buttons on the steering wheel. <i>Right</i> : Control buttons <i>ACT</i> and <i>MOD</i> and their respective icons.....	31
Figure 5.3 Visualization of the control logic of the HMI concepts. Transitions between LoAs are triggered via the control buttons <i>ACT</i> and <i>MOD</i> , or oversteering (OS), that is, braking or strong steering maneuvers.	31
Figure 5.4 Results of the heuristic expert evaluation for the HMI concepts <i>HC-HMI</i> and <i>LC-HMI</i>	36
Figure 5.5 Excerpts of the HMI concepts (<i>left</i> : <i>HC-HMI</i> ; <i>right</i> : <i>LC-HMI</i>) in the <i>IC</i> . <i>Top</i> : L2 has just been activated. <i>Bottom</i> : Second stage of the warning cascade during the RtI_{20s} . ..	37
Figure 5.6 <i>Top</i> : Sketch of the test course. <i>Bottom</i> : Photo of the test course and the waypoints on the test track at the Universität der Bundeswehr in Neubiberg, Germany.....	39
Figure 5.7 Photos of the experimental setup. <i>Left</i> : The participant wears the eye-tracking system Dikablis Glasses 3 by Ergoneers. <i>Right</i> : The experimenter gives instructions, triggers events, and controls the data recording.....	40
Figure 5.8 Visualization of the four AOIs: <i>Street</i> : the road environment (mainly windshield); <i>IC</i> ; <i>Controls</i> : the control buttons for the HMI on the steering wheel; <i>SuRT</i> : the tablet installed in the center console for the NDRA.	42
Figure 6.1 Boxplot diagram visualizing the results of the metric <i>Observed LoA vs. instructed LoA</i> for the study <i>Exp_Testing-Environment</i>	51
Figure 6.2 Boxplot diagram visualizing the results of the metric <i>TOT after RtI</i> for the study <i>Exp_Testing-Environment</i>	52
Figure 6.3 Boxplot diagram visualizing the results of the metric <i>Attention ratio during continuous rides in L0, L2, & L3</i> for the AOI <i>SuRT</i> for the study <i>Exp_Testing-Environment</i>	57
Figure 6.4 Bar chart visualizing the results of the metric <i>Glance allocation at start of RtI</i> for the study <i>Exp_Testing-Environment</i>	58
Figure 6.5 Boxplot diagram visualizing the results of the metric <i>Glance allocation time to IC after RtI</i> for the study <i>Exp_Testing-Environment</i>	59

Figure 6.6 Boxplot diagram visualizing the results of the metric <i>First glance duration on IC after RtI</i> for the study <i>Exp_Testing-Environment</i>	60
Figure 6.7 Boxplot diagram visualizing the results of the metric <i>Experimenter rating</i> for the study <i>Exp_Testing-Environment</i>	62
Figure 6.8 Boxplot diagram visualizing the results of the metric <i>Awareness of active LoA</i> for the study <i>Exp_Testing-Environment</i>	64
Figure 6.9 Boxplot diagram visualizing the results of the metric <i>Awareness of change of available LoAs</i> for the study <i>Exp_Testing-Environment</i>	65
Figure 6.10 Boxplot diagram visualizing the results of the metric <i>Awareness of reason for change of available LoAs</i> for the study <i>Exp_Testing-Environment</i>	65
Figure 6.11 Boxplot diagram visualizing the results of the metric <i>System understanding: allowance of NDRA</i> for the study <i>Exp_Testing-Environment</i>	66
Figure 6.12 Boxplot diagram visualizing the results of the metric <i>System understanding: allowance of H-off driving</i> for the study <i>Exp_Testing-Environment</i>	67
Figure 6.13 Boxplot diagram visualizing the results of the metric <i>Reported problems during transitions</i> for the study <i>Exp_Testing-Environment</i>	68
Figure 6.14 Boxplot diagram visualizing the results of the metric <i>SUS</i> for the study <i>Exp_Testing-Environment</i>	71
Figure 6.15 Boxplot diagram visualizing the results of the metric <i>UMUX</i> for the study <i>Exp_Testing-Environment</i>	72
Figure 6.16 Bar chart visualizing the results of the metric <i>UEQ</i> with its six dimensions for the study <i>Exp_Testing-Environment</i>	73
Figure 6.17 Boxplot diagram visualizing the results of the metric <i>Trust</i> for the study <i>Exp_Testing-Environment</i>	74
Figure 6.18 Boxplot diagram visualizing the results of the metric <i>Acceptance</i> for the study <i>Exp_Testing-Environment</i>	74
Figure 7.1 Boxplot diagram visualizing the results of the metric <i>Observed LoA vs. instructed LoA</i> for the study <i>Exp_Culture</i>	89
Figure 7.2 Boxplot diagram visualizing the results of the metric <i>TOT after RtI</i> for the study <i>Exp_Culture</i>	90
Figure 7.3 Boxplot diagram visualizing the results of the metric <i>Attention ratio during continuous rides in L0, L2, & L3</i> for the AOI SuRT for the study <i>Exp_Culture</i>	95
Figure 7.4 Bar chart visualizing the results of the metric <i>Glance allocation at start of RtI</i> for the study <i>Exp_Culture</i>	96
Figure 7.5 Boxplot diagram visualizing the results of the metric <i>Glance allocation time to IC after RtI</i> for the study <i>Exp_Culture</i>	97

Figure 7.6 Boxplot diagram visualizing the results of the metric <i>First glance duration on IC after RtI</i> for the study <i>Exp_Culture</i>	98
Figure 7.7 Boxplot diagram visualizing the results of the metric <i>Experimenter rating</i> for the study <i>Exp_Culture</i>	100
Figure 7.8 Boxplot diagram visualizing the results of the metric <i>Awareness of active LoA</i> for the study <i>Exp_Culture</i>	102
Figure 7.9 Boxplot diagram visualizing the results of the metric <i>Awareness of change of available LoAs</i> for the study <i>Exp_Culture</i>	103
Figure 7.10 Boxplot diagram visualizing the results of the metric <i>Awareness of reason for change of available LoAs</i> for the study <i>Exp_Culture</i>	104
Figure 7.11 Boxplot diagram visualizing the results of the metric <i>System understanding: allowance of NDRA</i> for the study <i>Exp_Culture</i>	104
Figure 7.12 Boxplot diagram visualizing the results of the metric <i>System understanding: allowance of H-off driving</i> for the study <i>Exp_Culture</i>	105
Figure 7.13 Boxplot diagram visualizing the results of the metric <i>Reported problems during transitions</i> for the study <i>Exp_Culture</i>	106
Figure 7.14 Boxplot diagram visualizing the results of the metric <i>SUS</i> for the study <i>Exp_Culture</i>	109
Figure 7.15 Boxplot diagram visualizing the results of the metric <i>UMUX</i> for the study <i>Exp_Culture</i>	110
Figure 7.16 Bar chart visualizing the mean results of the metric <i>UEQ</i> with its six dimensions for the study <i>Exp_Culture</i>	111
Figure 7.17 Boxplot diagram visualizing the results of the metric <i>Trust</i> for the study <i>Exp_Culture</i>	112
Figure 7.18 Boxplot diagram visualizing the results of the metric <i>Acceptance</i> for the study <i>Exp_Culture</i>	113
Figure 8.1 Visualization of the score differences for the cultural dimensions between the samples of the reference data (<i>top</i>) and between the study samples of the study <i>Survey_Culture</i> (<i>bottom</i>).	130
Figure 8.2 Bar chart visualizing the results of the importance scores for the usability factors (Hinderks et al., 2019) in the study <i>Survey_Culture</i>	132
Figure 9.1 Setting of the expert workshop on recommendations for methods for assessing the usability of HMIs for L3 ADS in user studies. Five experts participate in the workshop at the Chair of Ergonomics in February 2023.....	139
Figure 13.1 Visualization of the individual driving behavior for the metric <i>Control path of the first activation</i> for the study <i>Exp_Testing-Environment</i> (Paragraph Control Path of First Activation).	183

Figure 13.2 Visualization of the individual driving behavior for the metric <i>Take-over Path after RtI</i> for RtI _{20s} for the study <i>Exp_Testing-Environment</i> (Paragraph Take-Over Path after RtI).....	184
Figure 13.3 Visualization of the individual behavior for the metric <i>Take-over Path after RtI</i> for RtI _{6s} for the study <i>Exp_Testing-Environment</i> (Paragraph Take-Over Path after RtI).....	185
Figure 13.4 Boxplot diagram visualizing the results of the metric <i>Attention ratio during continuous rides in L0, L2, & L3</i> for all four AOIs for the study <i>Exp_Testing-Environment</i> (Paragraph Attention Ratio during Continuous Rides in L0, L2, & L3)...	186
Figure 13.5 Visualization of the individual gaze behavior for the metric <i>Gaze Behavior during RtI</i> for RtI _{20s} for the study <i>Exp_Testing-Environment</i> (Paragraph Gaze Behavior during RtI).....	187
Figure 13.6 Visualization of the individual gaze behavior for the metric <i>Gaze Behavior during RtI</i> for RtI _{6s} for the study <i>Exp_Testing-Environment</i> (Paragraph Gaze Behavior during RtI).....	188
Figure 13.7 Visualization of the mean ratings per test case for the metric <i>Experimenter rating</i> for the study <i>Exp_Testing-Environment</i> (Subsubsection Experimenter Rating).....	189
Figure 13.8 Overview of the clustered replies for praised components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Testing-Environment</i> (Subsubsection Final Interview).....	189
Figure 13.9 Overview of the clustered replies for criticized components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Testing-Environment</i> (Subsubsection Final Interview).....	190
Figure 13.10 Overview of the clustered replies for improvement suggestions for components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Testing-Environment</i> (Subsubsection Final Interview).	190
Figure 14.1 Visualization of the individual driving behavior for the metric <i>Control path of the first activation</i> for the study <i>Exp_Testing-Environment</i> (Paragraph Control Path of First Activation).	193
Figure 14.2 Visualization of the individual driving behavior for the metric <i>Take-over Path after RtI</i> for RtI _{20s} for the study <i>Exp_Culture</i> (Paragraph Take-Over Path after RtI).....	194
Figure 14.3 Visualization of the individual driving behavior for the metric <i>Take-over Path after RtI</i> for RtI _{6s} for the study <i>Exp_Culture</i> (Paragraph Take-Over Path after RtI).....	195
Figure 14.4 Boxplot diagram visualizing the results of the metric <i>Attention ratio during continuous rides in L0, L2, & L3</i> for all four AOIs for the study <i>Exp_Culture</i> (Paragraph Attention Ratio during Continuous Rides in L0, L2, & L3).....	196
Figure 14.5 Visualization of the individual gaze behavior for the metric <i>Gaze Behavior during RtI</i> for RtI _{20s} for the study <i>Exp_Culture</i> (Paragraph Gaze Behavior during RtI).....	197

Figure 14.6 Visualization of the individual gaze behavior for the metric <i>Gaze Behavior during RtI</i> for RtI _{6s} for the study <i>Exp_Culture</i> (Paragraph Gaze Behavior during RtI).....	198
Figure 14.7 Visualization of the mean ratings per test case for the metric <i>Experimenter rating</i> for the study <i>Exp_Culture</i> (Subsubsection Experimenter Rating).....	199
Figure 14.8 Overview of the clustered replies for praised components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Culture</i> (Subsubsection Final Interview). 199	
Figure 14.9 Overview of the clustered replies for criticized components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Culture</i> (Subsubsection Final Interview). 200	
Figure 14.10 Overview of the clustered replies for improvement suggestions for components of the HMI concepts in the metric <i>Final Interview</i> for the study <i>Exp_Culture</i> (Subsubsection Final Interview). 200	

List of Tables

Table 2.1 Description and empirical findings of Hofstede's model of cultural dimensions (Hofstede, 2011)	11
Table 4.1 Best practice advice for testing the usability of HMIs for L3 ADS from Albers, Radlmayr, et al. (2020)	26
Table 5.1 Description of the 12 test cases and their linkage to the NHTSA minimum requirements, adapted from Albers et al. (2021).....	29
Table 5.5.2 Overview of the eight items of Naujoks, Wiedemann, et al. (2019) that differentiate between the HMI concepts and description of their implementation in the <i>HC-HMI</i> and the <i>LC-HMI</i> concept, respectively.....	33
Table 5.3 List of heuristics used in the expert evaluation.....	35
Table 5.4 List of the dependent variables and their linkage to the items of the guidelines by Naujoks, Wiedemann, et al. (2019) violated in the <i>LC-HMI</i> , the linkage to the ISO standard 9241-11 (ISO, 2018a), and the linkage to the NHTSA minimum requirements (NHTSA, 2017).....	44
Table 5.5 Overview of potential outcomes of the inferential analysis and their simplified interpretation.....	47
Table 6.1 Descriptive analysis of the metric <i>Control path of first activation</i> for the study <i>Exp_Testing-Environment</i>	52
Table 6.2 Descriptive analysis of the metric <i>Take-over path after RtI</i> for the study <i>Exp_Testing-Environment</i>	53
Table 6.3 Summary table of the descriptive and inferential results of the quantitative metrics of the driving behavior for the study <i>Exp_Testing-Environment</i>	55
Table 6.4 Summary table of the descriptive and inferential results of the quantitative metrics of the eye-tracking for the study <i>Exp_Testing-Environment</i>	61
Table 6.5 Summary table of the descriptive and inferential results of the metric <i>Experimenter rating</i> for the study <i>Exp_Testing-Environment</i>	63
Table 6.6 Summary table of the descriptive and inferential results of the short interviews for the study <i>Exp_Testing-Environment</i>	70
Table 6.7 Summary table of the descriptive and inferential results of the questionnaires for the study <i>Exp_Testing-Environment</i>	76
Table 6.8 Distribution of responses for the interindividual factors <i>Nausea</i> and <i>Effort</i> for the experiments <i>Sim_GER</i> and <i>TT_GER</i>	79
Table 7.1 Descriptive analysis of the metric <i>Control path of first activation</i> for the study <i>Exp_Culture</i>	89
Table 7.2 Descriptive analysis of the metric <i>Take-over path after RtI</i> for the study <i>Exp_Culture</i>	91

Table 7.3 Summary table of the descriptive and inferential results of the quantitative metrics of the driving behavior for the study <i>Exp_Culture</i>	93
Table 7.4 Summary table of the descriptive and inferential results of the quantitative metrics of the eye-tracking for the study <i>Exp_Culture</i>	99
Table 7.5 Summary table of the descriptive and inferential results of the metric <i>Experimenter rating</i> for the study <i>Exp_Culture</i>	101
Table 7.6 Summary table of the descriptive and inferential results of the short interviews for the study <i>Exp_Culture</i>	108
Table 7.7 Summary table of the descriptive and inferential results of the questionnaires for the study <i>Exp_Culture</i>	114
Table 7.8 Distribution of responses for the interindividual factors <i>Nausea</i> and <i>Effort</i> for the experiments <i>TT_GER</i> and <i>TT_USA</i>	117
Table 8.1 Summary table of the descriptive analysis of the metrics on the driving background for the study <i>Survey_Culture</i>	127
Table 8.2 Description of the usability factors from Hinderks et al. (2019, pp. 1724–1726)...	129
Table 8.3 Ranking positions of the usability factors (Hinderks et al., 2019) included in the hypotheses in the study <i>Survey_Culture</i>	133
Table 8.4 Comparison of the ranking positions of the highest-ranked usability factors (Hinderks et al., 2019) in the study <i>Survey_Culture</i>	134
Table 9.1 Overview of the preliminary recommendations with supporting observations and results of the expert workshop with proposed changes.....	141
Table 9.2 Overview of the final recommendations refined after the expert workshop.	147
Table 12.1 Excerpts of the HMI concepts <i>HC-HMI</i> and <i>LC-HMI</i> for the 12 test cases and interaction errors.....	174
Table 13.1 Weather and light conditions in the study <i>Exp_Testing-Environment</i>	181
Table 13.2 Summary table of the descriptive analysis of the metrics on the sociodemographic data for the study <i>Exp_Testing-Environment</i> (Section 6.2).	181
Table 13.3 Summary table of the descriptive analysis of the metrics on the driving background for the study <i>Exp_Testing-Environment</i> (Section 6.2).....	182
Table 14.1 Weather and light conditions in the study <i>Exp_Culture</i>	191
Table 14.2 Summary table of the descriptive analysis of the metrics on the sociodemographic data for the study <i>Exp_Culture</i> (Section 7.2).....	191
Table 14.3 Summary table of the descriptive analysis of the metrics on the driving background for the study <i>Exp_Culture</i> (Section 7.2).	192

Glossary

ACC	Adaptive Cruise Control
ACT	ACTivation: a control button on the steering wheel in validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.4)
ADAS	Advanced Driver Assistance System(s)
ADS	Automated Driving System
AOI	Area Of Interest
CC	Cruise Control
CLM	Cumulative Link Model
DDT	Dynamic Driving Task
DIN	Deutsches Institut für Normung
DV	Dependent Variable: used in the formulas in the analysis of validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.7)
<i>Exp</i>	Factor EXperiment: used in the analysis of validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.7)
<i>Exp_Culture</i>	Validation study (Experiment) on the effect of the Culture on metrics in user studies (see Chapter 7)
<i>Exp_Testing-Environment</i>	Validation study (Experiment) on the effect of the Testing Environment on metrics in user studies (see Chapter 6)
GLMM	Generalized Linear Mixed Model
H _[1-3]	Hypothesis [1-3]
HC	High Compliance; refers to guidelines of Naujoks, Wiedemann, et al. (2019)
HMI	Human-Machine Interface;
HMI	Factor Human-Machine Interface: used in the analysis of validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.7)
H-off	Hands-off: referring to driving with hands off the steering wheel
IC	Instrument Cluster
ISO	International Organization for Standardization
L[0-5]	Level [0-5]; refers to SAE International (2021)
LC	Low Compliance; refers to guidelines of Naujoks, Wiedemann, et al. (2019)
LKA	Lane Keeping Assistant
LoA	Level of Automation
LRT	Likelihood-Ratio Test
<i>MOD</i>	MODE: A control button on the steering wheel in validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.4)

NDRA	Non-Driving Related Activity
NHTSA	National Highway Traffic Safety Administration
ODD	Operational Design Domain
<i>ON_GER</i>	Online survey with GERman participants in the validation study Survey_Culture (see Chapter 8)
<i>ON_USA</i>	Online survey with U.S.-American participants in the validation study Survey_Culture (see Chapter 8)
<i>Ref_GER</i>	GERman Reference data provided by Hofstede Insights (2023), used in the validation study Survey_Culture (see Chapter 8)
<i>Ref_USA</i>	U.S.-American Reference data provided by Hofstede Insights (2023), used in the validation study Survey_Culture (see Chapter 8)
RQ _[1-5]	Research Question [1-5]
RtI	Request to Intervene
<i>Sim_GER</i>	Driving Simulator experiment with GERman participants in the validation study <i>Exp_Testing-Environment</i> (see Chapter 6)
<i>SuRT</i>	Surrogate Reference Task
<i>Survey_Culture</i>	Validation study (Survey) on the effect of the Culture on subjective importance rating of usability factors (see Chapter 8)
<i>SUS</i>	System Usability Scale
TC[1-12]	Test Case [1-12]
(1 TC)	Random factor Test Case : used in the formulas in the analysis of validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.7)
TOST	Two One-Sided t-Tests
TOT	Take-Over Time
TP	Test Person : used if a specific participant is referenced
(1 TP)	Random factor Test Person : used in the formulas in the analysis of validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Section 5.7)
<i>TT_GER</i>	Test Track experiment with GERman participants in the validation studies <i>Exp_Testing-Environment & Exp_Culture</i> (see Chapter 6 & Chapter 7)
<i>TT_USA</i>	Test Track experiment with U.S.-American participants in the validation studies <i>Exp_Culture & Survey_Culture</i> (see Chapter 7 & Chapter 8)
<i>UEQ</i>	User Experience Questionnaire
<i>UMUX</i>	Usability Metric of User Experience
UX	User Experience
VSM	Values Survey Module

1 Introduction

1.1 Motivation

In 2021, a new chapter in the progress of automated driving has started: Honda launched the first vehicle that is equipped with an automated driving system (ADS) known as Level 3 (L3) ADS (SAE International, 2021; Sugiura, 2021). This ADS allows the driver to be temporarily released from the responsibility of the driving task. The term “temporarily” indicates the accompanying limitations and challenges. Repeated reallocations of the responsibility for the driving task signify the importance of well-designed human-machine interfaces (HMIs). HMIs can facilitate the intended and safe interaction between drivers and the ADS. Partly overlapping with safety-related aspects, usability comprises a more integral consideration of the interaction quality. Usability is plays a crucial role in assessing the design of HMI concepts (François et al., 2017). Advances in the research methods for assessing the usability of HMIs for L3 ADS are needed to adapt to the technological progress.

“We must recognize, however, that all of our scientific efforts fall along a continuum of fallibility. There is no investigation that can be totally lacking in its potential informativeness, nor will there ever be one that is perfect in its attainment of internal, external, and theoretical validity. Our goals, then, should be to strive toward conducting the least fallible inquiries, to cautiously interpret our experiments in accordance with their logical warrant, and to guard against the paralysis of complacency regarding the adequacy of current research methods.”

Mahoney, 1978, p. 671

According to Mahoney (1978), no perfect research method or study design exists. Instead, he advocates that researchers must be aware of pitfalls and limitations when interpreting their data. Furthermore, research methods should be selected considering the limitations of validity. Next to objectivity and reliability, validity is one of the three main quality criteria for scientific tests (Bortz & Döring, 2006, p. 195). Mahoney’s declaration motivates this thesis to learn more about these limitations and the resulting conclusions. Simultaneously, it is a constant reminder throughout this thesis’ theoretical and empirical work.

This thesis examines the potential effects of the testing environment and the users’ cultural background on usability assessments of HMIs for L3 ADS. The two factors are selected regarding their relevance for researchers and practitioners. The factor testing environment is examined by testing the validity of driving simulators. Driving simulators have many advantages, with cost-efficiency, high degrees of standardization, and low risks being only some of them (Caird & Horrey, 2011, Table 5.1). Furthermore, the availability of test vehicles equipped with ADS is limited. Therefore, most research on HMIs for ADS is conducted in driving simulators. The factor culture is examined by testing the validity of usability assessments and comparing the subjective importance ratings of usability factors across differing cultural backgrounds of potential users. To succeed in today’s globalized world, products must be available in different cultures. In examining the potential effects of

testing environment and cultural background and drawing attention to these effects, this thesis contributes to a responsible approach to usability testing on HMIs for L3 ADS.

1.2 Structure

Chapter 2 presents the theoretical foundation of this thesis. This includes the introduction of relevant terms and definitions as well as methods for assessing usability. Furthermore, existing literature on the effects of the testing environment and the culture is presented. In Chapter 3, five research questions are formulated. A short description of the approach to each research question is provided. Chapter 4 presents a systematic literature review on common practices of usability testing of HMIs for L3 ADS, thereby addressing research question RQ₁. The work is published in Albers, Radlmayr, et al. (2020), and only a summary is provided in this thesis. Chapter 5 offers insights into the experimental design applied in the subsequently presented validation studies *Exp_Testing-Environment* and *Exp_Culture*, completing research question RQ₁. The experimental design is derived from the systematic literature review presented in the previous chapter. The experimental method is published in Albers et al. (2021) and may be referred to for more details. Chapter 6 presents the findings of the validation study *Exp_Testing-Environment*, addressing research question RQ₂. Two experiments conducted in a static driving simulator and an instrumented vehicle on a test track are compared to investigate the effect of the testing environment on a selection of usability metrics. The validation study *Exp_Culture* is presented in Chapter 7, focusing on research question RQ₃. To investigate the effect of the users' cultural background on the assessment of usability, data from two experiments with German and U.S.-American samples are compared. Both experiments are conducted in an instrumented vehicle on test tracks in Germany. Chapter 8 offers insights into research question RQ₄. The effect of the users' cultural background on the subjective importance rating of usability factors in the context of HMIs for L3 ADS is examined (validation study *Survey_Culture*). Survey data of three samples comprising one German and two U.S.-American subsamples (one currently resides in the United States, and one subsample originates in the validation study *Exp_Culture*) is analyzed. Chapter 9 presents the results of an expert workshop discussing a set of preliminary recommendations for conducting user studies to assess the usability of HMIs for L3 ADS. Based on the expert workshop, a final set of recommendations is formulated, providing the answer to research question RQ₅. Chapter 10 concludes the thesis by summarizing the findings regarding the five research questions. Furthermore, the learnings are critically reflected, and an outlook on future research is presented. The thesis closes with the formulation of five key messages.

2 Theoretical Foundation

This chapter presents the thesis' theoretical foundation and identifies relevant research gaps. The chapter starts with the definition of relevant terms that are used throughout the present work. After that, an overview of methods for assessing usability is presented. The chapter closes with previous research on the effects of the testing environment driving simulator and the user's cultural backgrounds on user studies. The chapter does not cover studies in the field of usability assessments for L3 HMIs, which are presented in a literature review in Chapter 4.

2.1 Terms and Definitions

This section defines the terms relevant to the thesis. The areas of automated driving, usability, validity, testing environment, and culture are covered.

2.1.1 Automated Driving and Related Terms

Automated driving is an umbrella term referring to different degrees and application areas of automation. The framework 'Principles of Operation' by Shi et al. (2020) provides a comprehensive overview of the different types of automation, stressing the differences between continuous and discontinuous automation. The 'Principle of Operation A' comprises advanced driver assistance systems (ADAS) with informing and warning functions that indirectly influence vehicle guidance. In contrast, 'Principle of Operation B' contains ADAS that continuously and directly affect vehicle guidance through functions with varying degrees of automation. The 'Principle of Operation C' comprises ADAS that also directly influence vehicle guidance. However, these functions operate discontinuously, that is, only temporarily in accident-prone situations. This thesis focuses on continuous automation ('Principle of Operation B').

The different degrees of continuous automation are described in more detail in the standard J3016 (SAE International, 2021). Its six different levels of automation (LoAs) range between Level 0 (L0), which refers to no driving automation, and Level 5 (L5), which refers to full driving automation. The LoAs are characterized through the definition of the allocation of responsibilities between the driver and the ADS for different categories and an operational design domain (ODD), thereby determining the specific LoAs' characteristics. The categories are dynamic driving task (DDT) and DDT fallback. The DDT is subdivided into 'sustained lateral and longitudinal vehicle motion control, and object and event detection and response'. The ODD describes the operating conditions (e.g., traffic or roadway conditions) under which the ADS is designed to function. The ODD is either limited or unlimited (L5 only). Figure 2.1 provides an overview of the six LoAs presented in J3016 (SAE International, 2021).

The thesis focuses on L3, known as conditional driving automation. In 2021, Honda launched the first vehicle equipped with L3 ADS (Sugiura, 2021). According to the standard J3106, in L3 driving, the entire DDT is performed by the ADS (SAE International, 2021). The human operator is required to stay responsive in the role of the fallback-ready user. The operator reacts in cases of ADS-issued requests to intervene (RtIs) or system failures by, for example, resuming manual control of the DDT. In contrast, Level 2 (L2) is described as partial

driving automation. The driver is responsible for the ‘DDT fallback’ and one part of the DDT: object and event detection and response. The ADS is only responsible for the other part of the DDT: sustained lateral and longitudinal vehicle motion control.

The difference between L2 and L3 is substantial. Lorenz et al. (2015) describe the transition of the driver from the operator to the passenger role as a paradigm change. A simplified model for the different degrees of automation underlines this observation. The BASt introduces three different LoA (“modes”) that range between (1) assisted mode equivalent to L0, L1 (Level 1), and L2; (2) automated mode equivalent to L3; and (3) autonomous mode equivalent to L4 (Level 4) and L5 (Bundesanstalt für Straßenwesen, 2021). Following the model of the Bundesanstalt für Straßenwesen (2021), the difference between L2 and L3 may be described as fundamental compared to differences between, for example, L1 and L2. The HMI facilitates the repeated transitions between LoAs and, thus, of the DDT from the human operator to the ADS and vice versa. The design of the HMI faces new challenges with the paradigm change.

An RtI is the “alert provided by a [L3] ADS to a fallback-ready user indicating that s/he should promptly perform the DDT fallback [...]” (SAE International, 2021, p. 19). This fallback may involve resuming manual driving or pursuing a minimal risk condition (SAE International, 2021). RtIs play an essential role in L3 ADS. The term was formerly known as a take-over request and may be used as a synonym (Radlmayr, 2020).

In the automotive domain, an HMI is the location where information is transferred from the driver to the vehicle and vice versa (Bubb et al., 2015, p. 272). Bengler et al. (2020) list output channels, input channels, and dialog logic as the main elements of an HMI. The authors specify that output channels (e.g., displays or auditory signals) communicate information about the system state to the driver while input channels (e.g., buttons or pedals) transfer information from the driver to the vehicle. The dialog logic builds the relationship between input and output and the context parameters (Bengler et al., 2020).

In human factors research, HMIs play a vital role. Research has shown that well-designed HMIs reduce effects such as mode confusion (S. H. Lee & Eom, 2015) or misuse (Parasuraman & Riley, 1997) while facilitating learning effects (National Highway Traffic Safety Administration [NHTSA], 2016).

Level	Name	Narrative definition	DDT		DDT fallback	ODD
			Sustained lateral and longitudinal vehicle motion control	Object & event detection & response		
Driver performs part or all of the DDT						
Driver support	0 No driving automation	The performance by the driver of the entire DDT, even when enhanced by active safety systems.	Driver	Driver	Driver	n/a
	1 Driver assistance	The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT.	Driver and System	Driver	Driver	Limited
	2 Partial driving automation	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system.	System	Driver	Driver	Limited
ADS ("System") performs the entire DDT (while engaged)						
Automated driving	3 Conditional driving automation	The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS -issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately.	System	System	Fallback-ready user (becomes the driver during fallback)	Limited
	4 High driving automation	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Limited
	5 Full driving automation	The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene.	System	System	System	Unlimited

Figure 2.1 Overview of the LoAs according to the standard J3016 (SAE International, 2021).

2.1.2 Usability and Related Terms

The term usability has a long history originating in the software domain. According to J. R. Lewis (2012), the term usability was first used in the title of a scientific publication in 1979. Lewis reports that user friendliness and ease of use were commonly used back then. He distinguishes between two conceptions of usability, the formative and the summative conception of usability. For the formative approach, Lewis cites an early definition of usability as the ease of use in 1981 by Chapanis (p. 3, as cited by J. R. Lewis, 2012) that proposes an inversely proportional relationship between ease of use and the number and severity of difficulties people have in using software. For the summative approach, Lewis cites a definition from Bevan et al. (1991, p. 652) that considers a “class of users carrying out specific

tasks in a specific environment” while dividing usability into (1) ease of use referring to user performance and satisfaction and (2) acceptability referring to whether a product is used.

The differences between the formative and summative approaches are illustrated by Nielsen (1993) by explaining their main goals: “The main goal of formative evaluation is [...] to learn which detailed aspects of the interface are good and bad, and how the design can be improved. [...] In contrast, summative evaluation aims at assessing the overall quality of an interface” (Nielsen, 1993, p. 170). Nielsen (1993, p. 26) operationalizes usability by listing five attributes: (1) learnability, (2) efficiency, (3) memorability, (4) errors (referring to the error rate & the error’s severity), and (5) satisfaction. He argues that these attributes are “precise and measurable” and facilitate a systematic approach to usability testing (Nielsen, 1993, pp. 26–27).

This thesis applies the definition provided by the ISO standard 9241-11 (International Organization for Standardization [ISO], 2018a). In contrast to Nielsen’s operationalization (Nielsen, 1993, pp. 26–27), it provides three facets to operationalize the construct usability. The ISO definition follows the summative approach and defines usability as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO, 2018a, p. 2). Nielsen’s attributes efficiency, satisfaction, and errors correspond to the facets of the ISO definition. In addition, Nielsen’s operationalization comprises the attributes learnability and memorability, which are not directly addressed in the ISO definition. These attributes are reflected in all three facets of the ISO definition (effectiveness, efficiency, and satisfaction) with a focus on the quality of the first contact and early interaction (learnability) and the interaction quality in long-term use (memorability), respectively.

The elements of the ISO definition are further specified. The facet effectiveness is the “accuracy and completeness” (ISO, 2018a, p. 3) of the goals’ achievement. The facet efficiency relates to the resources needed to achieve the goals. The facet satisfaction refers to the match between “the user’s needs and expectations” and “the user’s physical, cognitive and emotional responses” (ISO, 2018a, p. 3). The “context of use” comprises a “combination of users, goals and tasks, resources, and environment” where the environment is further described as “the technical, physical, social, cultural and organizational environment” (ISO, 2018a, p. 4). The effect of specific features in the context of use on the usability assessment builds the focus of this thesis.

Scientific publications on usability assessments often overlap with other concepts. This paragraph provides a clear differentiation between usability and the terms user experience (UX), workload, acceptance, trust, and controllability.

UX and usability are closely related. Dumas and Salzman (2006) describe that starting in 2000, researchers expanded the meaning of the term usability by integrating “affective aspects of the user’s interaction” (p. 110). An example is proposed by Quesenberry (2004), who suggests adding “engaging” to the definition in the ISO standard 9241-11 (ISO, 2018a) to stress the importance of a pleasant, satisfying, and interesting interface. The approach of Barnum (2021) describes UX as an umbrella term that “includes usability testing, but also many other research tools” (p. 18). This thesis refers to the definition of UX as proposed by the ISO standard 9241-11: UX is “the user’s perceptions and responses that result from the use

and/or anticipated use of a system, product or service. [UX] focuses on the nature of these responses before, during and after use.” (ISO, 2018a, p. 13). In addition to the different scopes of time that these constructs refer to, the standard further explains that usability typically focuses on user groups and their goals, while UX focuses on individual users’ goals or motivations (ISO, 2018a, p. 22). Concluding, usability can be regarded as a component of UX, but the terms must not be used as synonyms.

A similar relationship exists between the terms workload and usability. Workload is the proportion of an operator’s limited capacity required to perform a particular task (O’Donnell & Eggemeier, 1986). Workload is, therefore, directly related to efficiency—one facet of usability.

Rahman et al. (2017) define the term acceptance in the context of ADAS as “the reaction of drivers when they are exposed to an in-vehicle technology and their willingness to adopt the technology while driving” (p. 362). According to Rousseau et al. (1998), trust is generally defined as the willingness to rely on another party based on its characteristics. In the context of new technologies, trust is perceived as a critical factor for adopting these new technologies (Gefen et al., 2003), which resembles the role of construct acceptance. Both concepts are closely related to the facet satisfaction of usability building upon the users’ experiences regarding effectiveness and efficiency.

The code of practice formulated in the project Response 3 (2009) defines controllability as the “likelihood that the driver can cope with driving situations including ADAS-assisted driving, system limits and system failures” (p. 5). This safety aspect is addressed in research mainly by assessing the take-over performance (e.g., Albert et al., 2015; Naujoks et al., 2015; Naujoks et al., 2018). Controllability is related to effectiveness, which is a facet of usability. The difference between these constructs lies in their focus. While controllability studies focus on safety-relevant aspects, the assessment of effectiveness also comprises non-severe interaction errors.

The constructs presented above appear several times in the scope of this work. The definition of and differentiation between usability and related terms facilitates understanding this thesis’ research focus.

2.1.3 Validity

Newton and Shaw (2014) report that the term validity was defined in 1921 by the North American National Association of Directors of Educational Research “as the degree to which a test measures what it is supposed to measure”. According to the authors, the concept of the term validity has different meanings depending on the discipline (pp. 2–3). Furthermore, Newton and Shaw (2014) outline that over the past decades, differing concepts of validity have been developed (pp. 14–24), and numerous terms related to validity have been established (e.g., pp. 7–9). In the following, only the terms relevant to this thesis are presented.

D. T. Campbell (1957) distinguishes between internal and external validity. He describes that internal validity is given if solely the stimulus of interest is responsible for a significantly different outcome. In contrast, the author describes external validity as generalizability, that is, whether the effect of interest can be generalized over “populations, settings, and variables” (p. 297). He points out that a trade-off between concepts exists in the form of the level of control positively affecting the internal validity and negatively affecting the external.

In driving simulator research, validity may be subdivided into physical and behavioral validity (e.g., Bellem et al., 2017; Blana, 1996; Mullen et al., 2011). Bellem et al. (2017) define physical validity as “the extent to which a driving simulator is capable of reproducing physical reality” (p. 443), that is, the correspondence of physical components, such as layout, dynamic characteristics, or visual displays, to on-road driving. Behavioral validity is defined as “the behavioral correspondence between driving behavior in the simulator and that on real roads” (p. 443), which includes the behavior, performance, and experience of drivers (Bellem et al., 2017). The relationship between physical and behavioral validity is ambiguous, concerning past research on this topic providing contradicting empirical results (e.g., Bellem et al., 2017; Goodenough, 2010; Jamson, 2001). Schneider (2021) concludes by describing the relationship as non-linear and complex.

Finally, one can distinguish between absolute and relative validity. Blaauw (1982) presents a conservative approach when defining the two terms for driving simulator research. He states that absolute validity is given if numerical values are about equal in the two environments of interest. According to Blaauw (1982), relative validity requires that “differences are of the same order and direction in both systems” (p. 474). More liberal approaches, such as the one of Kaptein et al. (1996), define absolute validity as given “if the absolute size of the effect is comparable to the absolute size of the effect in reality” (p. 31), while relative validity is given if “the direction or relative size of the effect of the measure is the same as in reality” (p. 31). In this thesis, the conservative approach is applied, as Blaauw (1982) proposed.

This thesis investigates usability assessments of HMIs for L3 ADS. Thus, the focus is on behavioral validity. Furthermore, this thesis concludes on both absolute as well as relative validity. The trade-off between internal and external validity is considered during the development of the experimental design (Chapter 5) and in the discussion of the generalizability of the thesis’ findings (Chapter 10).

2.1.4 Testing Environment

In the scope of this thesis, the term testing environment describes the setting in which an experiment is conducted. These settings may be categorized differently depending on the perspective and context.

Bruder et al. (2007) distinguish between laboratory and field studies. Laboratory studies are subdivided into simple mockups (e.g., table-mounted displays) and driving simulators. Field studies are subdivided into test track studies, test drives in real traffic, and naturalistic driving studies. The different testing environments are ranked with an increasing authenticity level and decreasing experimental control: simple mockups, driving simulators, test track studies, test drives in real traffic, and naturalistic driving studies. (Bruder et al., 2007)

Other researchers use different categories (and additional dimensions) to describe the testing environment (e.g., Purucker et al., 2018; Schneider, 2021). Schmidtke and Schulze (1989) list evaluation methods that resemble the above-presented testing environments of Bruder et al. (2007), adding only mathematical models (low level of authenticity, high level of flexibility). Schneider (2021) proposes a classification approach to common settings in pedestrian research that classifies the settings on four dimensions: experimental control, scenario realism, physical fidelity, and awareness (of being observed). This thesis

distinguishes between five different testing environments, as Bruder et al. (2007) proposed. Only two of these testing environments—the driving simulator and the test track—are examined in the scope of this work.

2.1.5 Culture

The Latin origins of the term culture are associated with education or refinement (Minkov & Hofstede, 2013, p. 10). The definition of culture is complex and varies across different domains. Even within domains, researchers struggle to agree on one definition, as illustrated by Jahoda (1984), who remarks that in social sciences, the number of books covering the definition of culture is enormous. A definition often cited in social sciences is provided by Kluckhohn (1959, p. 86): “Culture consists in patterned ways of thinking, feeling and reacting, acquired and transmitted mainly by symbols, constituting the distinctive achievements of human groups, including their embodiments in artifacts; the essential core of culture consists of traditional (i.e., historically derived and selected) ideas and especially their attached values”. A more recent definition of culture is provided by the well-known social psychologist Geert Hofstede and colleagues (Hofstede et al., 2010, pp. 5–6). They describe that culture is learned in a social environment, making it a collective phenomenon. Hofstede et al. (2010) view culture as mental software and define the term as “the collective programming of the mind that distinguishes the members of one group or category of people from others” (p. 6).

Following the definition of Hofstede, the next step is operationalizing the terms group or category to conduct cultural research. According to Hofstede (2001, p. 10), groups and collectives—and thus cultures—can be formed by nations, regions, ethnicities, organizations, occupations, and even age groups or genders. Minkov and Hofstede (2013, p. 11) argue that in a pragmatic approach, culture can be defined based on the focus of the research interest. Research on cultures often uses nationality as an operationalization for culture (e.g., Barber & Badre, 1998; Hofstede & Minkov, 2013a; Minkov & Hofstede, 2013). This approach has been the subject of many controversies: Child (1981, pp. 327–328) remarks that not nationality but other phenomena like national wealth, level of industrialization, or climate may cause cultural differences. Peterson and Smith (2008) identify three main critiques for using nations in cross-cultural research: (1) the variance of individuals within nations, (2) the existence of subcultures within nations, and (3) the weaknesses of structural theories in general. Minkov and Hofstede (2013, pp. 25–26) take a stand on the points of criticism. They argue that the first and third points are irrelevant. The first point of criticism refers to the complexity of individuals within nations. Minkov and Hofstede (2013, pp. 25–26) comment that the critique shifts the focus from the level of group research to the level of individual research. Furthermore, they weaken criticism of the theoretical nature of the construct nationality by pointing out that any abstract theory could be defended without empirical evidence. The critique referring to subcultures such as regions and ethnicities is confronted by empirical data involving 299 in-country regions from 28 countries confirming the existence of national values (Minkov & Hofstede, 2013, pp. 25–26). Researchers in favor of using nations in cross-cultural research argue that nations create shared experiences regarding education, economy, and demography (Inglehart & Baker, 2000, p. 37; Parker, 1997, pp. 11–17; Minkov & Hofstede,

2013, pp. 25–27). With the complex discussion on the term culture in mind, this thesis follows the common approach of defining culture through nationality.

Cross-cultural research has a long history. One of the earliest and, up to this day, most prominent tech-related cross-cultural research is conducted by Geert Hofstede. In the 1970s, Hofstede identified cultural values and dimensions based on survey data from over 100,000 questionnaires in 50 countries provided by IBM (Hofstede, 2011). Hofstede defines values as “a broad tendency to prefer certain states of affairs over others” (Hofstede, 2001, p. 5). Over the decades, Hofstede’s model of cultural dimensions has been refined and complemented and currently holds six dimensions (Hofstede, 2011). Table 2.1 comprises the dimensions’ descriptions and empirical results for selected countries/regions.

The cultural values of Hofstede’s model can be assessed through the Values Survey Module (*VSM*), a 30-item questionnaire with six items related to sociodemographic data and 24 items related to cultural values (Hofstede & Minkov, 2013a). The thesis applies Hofstede’s model and its method to obtain data on cultural values. Numerous other cross-cultural studies focus on variations of nations, regions, and ethnicities that are not subject to this thesis. Minkov and Hofstede (2013, chapter 9) provide a comprehensive overview of major cross-cultural studies.

Table 2.1 Description and empirical findings of Hofstede's model of cultural dimensions (Hofstede, 2011).

Dimension	Description	Tendencies in empirical data
<i>Power Distance</i>	"[E]xtent to which the less powerful members of organizations and institutions (like the family) accept and expect that power is distributed unequally" (p. 9)	High scores in East European, Latin, Asian, and African countries; low scores in Germanic and English-speaking Western countries
<i>Uncertainty Avoidance</i>	"[E]xtent [to which] a culture programs its members to feel [...] uncomfortable [...] in unstructured situations" (p. 10)	High scores in East and Central European countries, Latin countries, Japan, and German-speaking countries; low scores in English-speaking, Nordic, and Chinese culture countries
<i>Individualism</i>	"[D]egree to which people in a society are integrated into groups" (p. 11)	Higher scores in developed and Western countries; neither high nor low scores in Japan; lower scores in less developed and Eastern countries
<i>Masculinity</i>	"[D]istribution of values between the genders" (p. 12)	High scores in Japan, German-speaking countries, and some Latin countries like Italy and Mexico; moderately high scores in English-speaking Western countries; moderately low scores in some Latin and Asian countries like France, Spain, Portugal, Chile, Korea and Thailand; low scores in Nordic countries and the Netherlands
<i>Long Term Orientation</i>	Connection of the past with the current and future actions/challenges	High scores in East Asian countries; moderately high scores in Eastern- and Central Europe; neither high nor low scores in South- and North-European and South Asian countries; low scores in the United States and Australia, Latin American, African, and Muslim countries
<i>Indulgence vs. Restraint</i>	Degree of freedom that societal norms give to citizens in fulfilling their human desires	High indulgence scores in South and North America, Western Europe, and parts of Sub-Saharan Africa; neither high indulgence nor high restraint scores in Mediterranean Europe; high restraint scores in Eastern Europe, Asia, and the Muslim world

2.2 Overview of Methods for Assessing Usability

This section presents an overview of the different methods for assessing usability. Emphasis is put on methods and metrics relevant to the thesis. For extensive coverage of the methods for usability assessments, please refer to Dumas and Salzman (2006), J. R. Lewis (2012), or Sarodnick and Brau (2016).

The first usability tests are reported as being “expensive, time-consuming, and rigorous” (Barnum, 2021, p. 16). Traditional usability tests were mainly conducted by experimental psychologists or cognitive scientists and typically involved 30 to 50 participants (Barnum, 2021, p. 16). In the 1990s, usability testing experienced a drastic change. Several researchers observed that sample sizes as small as $N = 5$ in a usability study discover about 80% to 85% of the usability problems that a bigger sample would have discovered (J. R. Lewis, 1994; Nielsen, 2000; Virzi, 1990). Besides the more resource-efficient way of usability testing, other methods

have been refined and developed. Dumas and Salzman (2006) assign these methods to four different categories: (1) field methods, (2) inspection methods, (3) usability testing, and (4) focus groups, interviews, and surveys.

2.2.1 Field Methods

Dumas and Salzman (2006) describe that field methods aim to study users, their needs, behaviors, and product interaction in a real-world context. The authors distinguish between explorative and evaluative field studies. Furthermore, field methods vary substantially in the degree of the users' awareness of being part of a study. Commonly used techniques are behavioral observations, interviews, or diaries.

2.2.2 Inspection Methods

Inspection methods do not involve the (potential) end user but are conducted with usability specialists or developers instead (Dumas & Salzman, 2006). The most frequently applied inspection methods are the cognitive walkthrough and the heuristic evaluation. In a cognitive walkthrough, an evaluator takes the role of a user completing a specified set of tasks while examining the cognitive demand and potential usability problems for each step (Nielsen, 1994). According to Barnum (2021, p. 46), the heuristic evaluation is the second most often selected method from a UX toolkit. In a heuristic evaluation, few evaluators assess an interface's compliance with a set of usability principles (Nielsen, 1993, p. 155). For each of the usability principles, the evaluators rate the severity of a usability problem ranging between "0: this is not a usability problem at all" to "4: usability catastrophe—imperative to fix this before product can be released" (Nielsen, 1993, p. 103). Nielsen (2005) proposes a set of 10 usability principles for the field of software usability: (1) visibility of system status; (2) match between system and the real world; (3) user control and freedom; (4) consistency and standards; (5) error prevention; (6) recognition rather than recall; (7) flexibility and efficiency of use; (8) aesthetic and minimalist design; (9) help users recognize, diagnose, and recover from errors; and (10) help and documentation. Up to this day, the set of heuristics is (with adjustments) often applied in research in the software and other domains (1,391 Google Scholar citations for the current version by Nielsen, 2005; examined 29.10.2023). The application of heuristic evaluations is recommended in the early stages of a product development process Nielsen (1993, p. 159).

2.2.3 Usability Testing

Usability testing corresponds to empirical methods with (potential) end users for identifying usability problems or for comparing or measuring the usability of specific products (Dumas & Salzman, 2006, p. 111). Further characteristics of the methods are the defined set of tasks that participants must complete, the recording and analysis of qualitative or quantitative measures, and often the involvement of the thinking-aloud technique (Dumas & Salzman, 2006, p. 111).

The thinking-aloud technique is described as one of the most essential methods of usability testing (Nielsen, 1993, p. 195). This method requires participants to think out loud while performing specific tasks (C. Lewis, 1982). The verbalized thoughts allow the experimenter insights into the users' perspectives and problems while interacting with a

product (Nielsen, 1993, p. 195). Other usability measures applied in usability testing are satisfaction ratings, error rates, or task success rates (Dumas & Salzman, 2006).

Hornbæk (2006) presents a literature review comprising 180 studies in the field of usability research on human-computer interaction. The review provides an extensive overview of usability measures applied in empirical usability studies assigned to the usability facets effectiveness, efficiency, and satisfaction. The most frequently used measures of effectiveness are accuracy, for example, error rates or precision; binary task completion, that is, the number or percentage of successfully completed tasks; and quality of outcome, that is, the “understanding or learning of information in the interface” (p. 83). Commonly used measures of efficiency are time, that is, the duration of a (part of a) task; usage patterns, for example, the use frequency or the deviation from the optimal solution; and the input rate, for example, the number of correctly entered words in a specific period. Predominant measures of satisfaction are questions regarding the satisfaction with the interface, for example, the ease of use, and questions regarding the users’ attitudes and perceptions, for example, the perception of the interaction or the perception of the relation to other persons. Usually, the response format of these questions is a scale ranging from disagreement to agreement with the respective statement. Another commonly used measure of satisfaction relates to the users’ preference, which can be inquired in a rating or ranking or observed through the users’ behavior. Only 7% of the studies in the literature review use standardized questionnaires such as the ‘Questionnaire for User Interaction Satisfaction’ (QUIS), licensed for usability measurements in human-computer interaction (Chin et al., 1988).

2.2.4 Focus Groups, Interviews, and Surveys

Dumas and Salzman (2006) summarize focus groups, interviews, and surveys into one major category. A distinction between the subcategories is made by Courage and Baxter (2005), who assign focus groups to individual users’ feedback, interviews to small samples with more in-depth data collection, and surveys to large user samples.

Focus groups comprise six to nine users discussing a product and a moderator who ensures that selected topics are covered, and every user is heard (Nielsen, 1993, pp. 214–215).

In interviews, the user and the interviewer are in direct exchange, where interviewers can respond to misunderstandings or interesting user remarks with follow-up questions (Nielsen, 1993, pp. 210–211). Nielsen (1993, pp. 210–211) points out that interviews enable in-depth data collection but are associated with a high resource demand in the data collection and analysis phase.

In contrast, questionnaires have a high resource demand in the development phase but have the advantages of allowing efficient data collection of large samples and flexible use regarding location (e.g., via mail) and time (e.g., comparisons over time) (Nielsen, 1993, pp. 212–213). Sauro and Lewis (2012, pp. 185–186) complement the above-listed advantages of standardized questionnaires with objectivity, quantification, effective communication, and scientific generalization. In addition to the *QUIS* (Questionnaire for User Interaction Satisfaction), several questionnaires with differing application areas, focuses, and lengths exist. An extensive overview is presented by Sauro and Lewis (2012, chapter 6).

- The ‘System Usability Scale’ (*SUS*) has been developed as a cost- and resource-efficient (“quick and dirty”) 10-item questionnaire with an overall score between 0 and 100, designed for a range of application contexts (Brooke, 1996). The *SUS* is frequently applied with 16,351 Google Scholar citations (examined 29.10.2023) for the original paper by Brooke (1996) introducing the questionnaire.
- The ‘Post-Study System Usability Questionnaire’ (*PSSUQ*, Sauro & Lewis, 2012) is a license-free 16-item questionnaire with one overall score and three subscales: *System Quality*, *Information Quality*, and *Interface Quality*.
- A related questionnaire is the ‘Computer System Usability Questionnaire’ (known as *CSUQ*, J. R. Lewis, 1995), which is identical to the *PSSUQ* with adjusted wordings for the adaption to research in contexts other than laboratory settings (Sauro & Lewis, 2012).
- The ‘Usability Metric of User Experience’ (*UMUX*, Finstad, 2010) is a short 4-item questionnaire directly reflecting the facets of usability in the definition of the ISO standard 9241-11 (ISO, 2018a) with an overall score between 0 and 100.
- The ‘Software Usability Measurement Inventory’ (known as *SUMI*, Kirakowski, 1996) includes 50 items that result in a global scale and the five subscales *Efficiency*, *Affect*, *Helpfulness*, *Control*, and *Learnability* (<https://sumi.uxp.ie/>, Sauro & Lewis, 2012).
- The ‘User Experience Questionnaire’ (*UEQ*, Laugwitz et al., 2008) is a 26-item questionnaire focusing on UX rather than usability with six subscales: *Attractiveness*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*.

Research by J. R. Lewis (2019) shows that the questionnaires *SUS*, *CSUQ*, and *UMUX* strongly correlate. Consequently, researchers may choose the questionnaire based on other aspects, such as comparability with previous research or length. Furthermore, the questionnaire may be selected regarding the suitability of the questionnaire items to the research subject.

In usability studies, researchers often combine several of the previously listed methods. Interviews and questionnaires are often conducted at the end of usability tests (e.g., Barnum, 2021, p. 239; Dumas & Salzman, 2006, p. 126; Hornbæk, 2006). This approach is backed by the ISO standard 9241-11 (ISO, 2018a). According to the ISO standard 9241-11 (ISO, 2018a, pp. 7, 26), no single intrinsic measure of usability exists because no measure fully represents overall usability, and usability and its facets depend on the respective user goals and context of use.

2.3 Effects of the Testing Environment Driving Simulator

This section presents literature on the validity of driving simulators as this thesis' testing environment of interest. This section's scope is not limited to usability assessments but user studies in general. After generalizing the advantages and limitations of driving simulators, selected findings of previous driving simulator validation studies are presented.

2.3.1 Driving Simulator as a Valid Research Tool

This thesis compares the testing environments, driving simulator, and test track. Both testing environments feature a lower level of authenticity and a higher level of experimental control than naturalistic driving settings (Bruder et al., 2007). Most research efforts focusing on the effects of testing environments examine driving simulators. Among others, driving simulators have the advantages of being resource- and cost-efficient, enabling a high degree of standardization and control of confounding variables and permitting risk-free testing of safety-critical situations (Caird & Horrey, 2011, Table 5.1).

Despite the advantages, several aspects require consideration when conducting research in driving simulators. Purucker et al. (2018) list several of these aspects: Regarding the modeling of the physical world, the lack of visual details and further shortcomings in the visual representation (e.g., rendering errors and luminance), as well as limitations of spatial, acoustical, physical, and cinematic cues, are listed. Furthermore, motion sickness can occur, and the participants' awareness of being in a simulator may affect their perception and behavior (e.g., perceived risk, see Ranney, 2011). Purucker et al. (2018) suggest implementing familiarization drives and training to reduce the potential effects of the abovementioned aspects.

Attempts to validate driving simulators usually involve comparisons (e.g., ANOVAs, correlations, or regressions) between driving simulator experiments and replications in instrumented cars in test track experiments or—less often—real traffic conditions. The focus of these validation studies is manifold. That is, validation studies may address specific methods (e.g., Bengler et al., 2010: lane change test in different laboratory settings), specific products (e.g., Krause et al., 2014: traffic light assistant in a static driving simulator vs. real traffic conditions), or specific settings of the testing environment (e.g., Knappe et al., 2007: lane keeping and steering performance for different field of view conditions). Review papers on driving simulator validation studies provide good overviews of the common methods and the current state of the art (Blana, 1996; Mullen et al., 2011; Wynne et al., 2019). The validation studies presented can be attributed to specific aspects of driving behavior, such as speed or drivers' perception. Relevant findings of their work are presented in the following.

2.3.2 Driving Simulator Validation Studies in the Automotive Context

One of the most common measures in validation studies is comparing drivers' speed (Mullen et al., 2011; Wynne et al., 2019). Mullen et al. (2011) conclude that most studies confirm relative, if not absolute, validity. In contrast, Wynne et al. (2019) find that more than one-third of the studies included in their review on speed validation do not demonstrate either relative or absolute validity. The differences are diverse, ranging from higher speeds (Senserrick et al., 2007; Wynne et al., 2019), lower speeds (Fors et al., 2013), or greater speed variations (Senserrick et al., 2007) in simulators compared to on-road observations. The reviews' results are similar for other aspects, such as braking behavior, lateral driving measures, overall driving performance, and physiological measures (Mullen et al., 2011; Wynne et al., 2019). The ambiguous findings reported in the reviews' studies suggest that the question of simulator validity is complex. Mullen et al. (2011) further examine validation studies covering the effects of road design and traffic control devices, complex behaviors such as divided attention tasks, and effects of specific user groups (e.g., characterized through age

or medical conditions). Mullen et al. (2011) conclude that one can assume relative validity for most of the measures but not absolute and that researchers need to be aware of the limitations and uncertainties of driving simulator validity. Wynne et al. (2019) additionally examine the relationship between findings of validity and the fidelities of driving simulators. The results indicate no clear relationship (Wynne et al., 2019). The authors conclude their review with a call for more standardization and transparent documentation in validation studies. Furthermore, they suggest including the measured speed in each validation study to enhance the comparability of different validation attempts. The two reviews by Mullen et al. (2011) and Wynne et al. (2019) stress the importance of selecting a driving simulator with appropriate features for the research question of interest. In line with Kaptein et al. (1996), Mullen et al. (2011) even urge to newly validate driving simulators for each research question of interest.

2.3.3 Driving Simulator Validation Studies in the Automated Driving Context

Validation studies for driving simulators covering the field of driver behavior in automated driving conditions are scarce. Bellem et al. (2017) investigate the validity of driving simulators for the perception of comfort in automated driving conditions. An experiment involving a test drive with lane changes and deceleration maneuvers is conducted in a moving-base simulator with two different settings and an instrument vehicle on a test track (Bellem et al., 2017). The results show relative and absolute validity for only one of the two driving simulator settings, demonstrating the importance of appropriate motion cues in research on driving comfort (Bellem et al., 2017). Another validation study for automated driving is conducted by Poisson et al. (2020), who repeat a driving simulator experiment on driver behavior for L4 driving in a Wizard of Oz vehicle on a test track. The authors observe differing take-over strategies between the two testing environments and more interruptions of non-driving related activity (NDRA) engagement while driving L4 in the Wizard of Oz experiment compared to the driving simulator experiment. No differences are found in the analysis of reaction times to RTIs.

Regarding the validity of driving simulators for usability research in ADS HMIs, the two validation studies in the context of automated driving (Bellem et al., 2017; Poisson et al., 2020) are encouraging despite the differences in single metrics and simulator settings. Furthermore, the vast body of literature on previous attempts to validate driving simulators in the general automotive context, as presented in the reviews of Mullen et al. (2011) and Wynne et al. (2019), suggests that driving simulators provide valid results. Nevertheless, several studies included in their reviews could not confirm relative or absolute validity. While most studies showed relative validity, only a minority yielded results suggesting absolute validity. Additionally, it should be noted that several studies yield results confirming some form of validity for specific metrics, while for other metrics, no validity could be found (Mullen et al., 2011; Poisson et al., 2020; Wynne et al., 2019).

2.4 Effects of the Users' Cultural Background

This section presents literature on cultural effects in user studies. The first subsection presents research on cross-cultural effects in the data collection phase. The second subsection

presents cross-cultural studies in interface design, the link between theoretical models of cultural values, and interface design. It closes with cross-cultural studies in automotive interface design.

2.4.1 Effects of Culture in the Data Collection Phase

The danger of drawing false conclusions is high if cultural effects during the data collection phase are not considered. To illustrate this, several examples of cultural effects occurring during the data collection phase are presented.

Loew et al. (2022) suggest that questionnaires may have different structures in different countries. They conduct factor analyses of the *SUS* for samples from China, the United States, and Germany and find different two-factor structures for each country. Regarding response behavior for scales such as Likert scales, Moss and Vijayendra (2020) present three response tendencies that are country-specific: (1) acquiescence response styles describe a tendency to agree: these styles are common in Latin America, the Middle East, and some African countries; (2) extreme response styles describe a tendency for using the extremes of rating scales: these styles are common in Latin America, and (3) middle response styles describe a tendency for using the mid-responses of rating scales: these styles are common in Asia. Douglas and Liu (2011, pp. 30–31) list numerous studies confirming a cultural effect on response behavior in usability tests. A common phenomenon is the reluctance to express criticism in several cultures (e.g., Chetty et al., 2007; Herman, 1996; Yeo, 2001). Herman (1996) reports an extreme example of a participant in Singapore who aborts a test and cries due to failing to complete a set of tasks. Regardless of the poor performance and the emotional stress, the overall feedback in the interview is positive. Vatrapu and Pérez-Quiñones (2006) additionally find an effect of the interviewer's cultural background on the interviewees' responses. They conduct interviews with Indian participants and either Indian or Anglo-American interviewers. Results suggest that interviewees report more usability problems and provide more detailed and forthright descriptions of these problems if the interviewer is from the same culture compared to interviews conducted with Anglo-American interviewers.

Douglas and Liu (2011, p. 33) recommend conducting tests in local contexts with local experimenters to minimize cultural effects on research methods.

2.4.2 Effects of Culture on Interface Designs

In addition to cultural effects on research methods, cultural effects on the interaction of humans and technical devices can be seen in numerous studies. To illustrate the variety of potential reasons for cultural effects and the resulting findings, selected studies are presented in this subsection. In 1991, Abed identified different scanning patterns for non-directional stimuli depending on the participants' learned reading direction. The reading direction and literacy rates influence the interaction with technical devices. Sherwani et al. (2009) find that speech interfaces are preferred over touch-tone interfaces for mobile phone applications by users with low literacy rates. In a study by Lesch et al. (2009), participants from China and the United States rate the perceived hazard of colors, words, and symbols and their combinations. The Chinese participants provide lower absolute hazard ratings than the U.S.-American participants. Furthermore, the relative levels of perceived hazard differ between the samples regarding the elements, particularly the colors. In contrast, the relative levels of perceived

hazard for combinations of the elements are similar for both samples. Honold (1999) examines cultural differences in the learning process and observes that Chinese prefer learning by imitating friends while Germans prefer individual learning by doing. Studies by Chau et al. (2002) and Frandsen-Thorlacius et al. (2009) show that the Chinese place high importance on aesthetics compared to Americans and Danish, respectively. These studies suggest the existence of cultural differences in preferences for usability aspects which are addressed in the following.

The first approaches to link culture to design aspects of technical devices appeared around the year 2000. Barber and Badre (1998) identified cultural markers in websites, such as icons or colors, and introduce ‘culturability’ to underline the strong relationship between usability and culture. Marcus and Gould (2000) developed guidelines for interface designs based on Hofstede’s model of cultural dimensions (Hofstede, 2011), illustrated with examples from website design. Twenty years later, Gong et al. (2020) took up the approach and apply five of Hofstede’s cultural dimensions to develop HMI guidelines in the automotive context. Gong et al. (2020) formulate 16 HMI guidelines for the design of automotive HMIs for the Chinese market. Sogemeier et al. (2022) map the six cultural dimensions of Hofstede’s model (Hofstede, 2011) to HMI design in the automotive context. The researchers map the cultural dimensions to a set of usability criteria and provide examples of HMI design for extreme expressions on the cultural dimensions.

In addition to cultural values such as Hofstede’s model of cultural dimensions (Hofstede, 2011), cultural differences in the context of driving may influence the drivers’ preferences and behavior. In a naturalistic driving study, Orlovska et al. (2020) observe different usage behaviors of ADAS, such as Pilot Assist and Adaptive Cruise Control (ACC) between Chinese, Swedish, and U.S.-American markets. A possible reason is provided by Large et al. (2017), explaining that road environment, local rules, and regulations (formal and informal) differ between cultures. A study by Lindgren et al. (2008) supports this argument. They compare the ratings of potentially dangerous driving behavior between Swedish and Chinese drivers. The results show that both samples mainly identify the same problems. However, the Swedish sample rates these problems more severe and stressful than the Chinese drivers. Supported by their findings of different cultural driving contexts, the authors argue that ADAS might not be accepted and might be ignored or misused if warnings occur too often in situations rated as typical or non-critical by drivers.

In addition to the driving context, the above-presented link between Hofstede’s model of cultural dimensions (Hofstede, 2011) and HMI guidelines suggests the existence of cultural differences in preferences for HMI design. Empirical studies on cultural effects regarding automotive HMIs are limited and mainly cover the design of infotainment systems. Only a few cross-cultural studies addressing automated driving exist. Selected studies from both areas are briefly presented in the following.

Roessger (2003) compares the input controls rotary push button and touch screen for samples from Germany, Japan, and the United States. The United States and Germany yield similar ratings, differing only in aspects regarding expectations and aesthetics, while the Japanese ratings differ significantly from the other two samples. Another study regarding input controls compares British and Chinese participants (Large et al., 2019). In both samples, the touch screen is preferred and rated as least demanding to use while driving. Chinese

participants, however, express more excitement for the novelty and show higher off-road glance times compared to the British participants. Young et al. (2012) compare preferences for control types and labels for Australian and Chinese drivers. The results confirm the findings of previously presented studies (Chau et al., 2002; Frandsen-Thorlacius et al., 2009) emphasizing the high importance of aesthetics for Chinese drivers. Regarding navigation systems, Heimgärtner (2007) and Large et al. (2017) find that Asian samples from China and Malaysia prefer higher information densities than Western samples from German- and English-speaking drivers. Furthermore, Heimgärtner (2007) finds that the English-speaking sample differs from the German- and the Chinese-speaking samples by preferring considerably lower display durations of maneuver advice notifications. Further differences between samples from Asian and Western cultures could be shown in studies on infotainment systems for preferences regarding the usage of quick buttons (Mehler et al., 2021: China vs. Germany), the learning behavior (Khan & Williams, 2014: India vs. UK) and the importance rating of specific HMI features (Khan et al., 2016: India vs. UK). Niehaus et al. (2020) compare Japanese to German truck drivers. They report that the Japanese sample systematically produces lower ratings while the relative ratings of the HMI variations are similar for both samples. Further analyses show no cultural effect, but the comprehensiveness of icons in the tested HMI concepts proves to be most important for the HMI design. The authors conclude that the design of understandable icons supported with descriptions is more relevant to cross-cultural HMI designs than cultural backgrounds.

Regarding the cultural effects in automated driving, Edelmann et al. (2021) compare four samples from China, Germany, Japan, and the United States in an online study examining the users' acceptance of the ADS' decisions in overtaking situations. The research shows similar results for the German and U.S.-American samples who prefer ADS decisions with only low hindrances of other traffic participants. The Japanese sample rejects all ADS decisions leading to any hindrances of other participants. The Chinese sample shows high acceptance ratings for all ADS decisions regardless of the level of hindrance. Strle et al. (2021) and Hergeth et al. (2015) conduct cross-cultural research in take-over scenarios of ADS, focusing on driving behavior and trust. Strle et al. (2021) compare U.S.-American to Slovenian drivers. The researchers observe significantly lower take-over performances and higher distractions due to engagement with voluntary NDRAs in the U.S.-American sample than in the Slovenian sample. Hergeth et al. (2015) examine the development and measurement of trust in a Chinese and a German sample concerning take-over situations. The results show similar developments of trust in both samples, while mistrust is significantly more pronounced in the Chinese sample than in the German sample. Furthermore, behavioral measures could not be related to the self-reported measures of trust.

The studies presented in this section confirm the existence of cultural differences relevant to interface designs. Most studies compare Western countries to Asian countries, specifically China. Analyses suggest that cultural differences are more pronounced between Western countries and Asian countries compared to differences between Western countries themselves. A recurring observation is the superior importance of aesthetic aspects in Chinese culture compared to Western cultures (e.g., Chau et al., 2002; Frandsen-Thorlacius et al., 2009; Young et al., 2012). Nonetheless, comparisons between Western countries suggest that differences

within these countries exist, too (e.g., expectations and aesthetics in Roessger, 2003, or take-over performances and NDRA engagement in Strle et al., 2021). The presented studies highlight the importance of the research methods when conducting cross-cultural research: The differences between cultures may only show in specific metrics. Furthermore, the existence of covariates such as comprehensiveness or language proficiency can help to explain the effects. Considering the first subsection on cross-cultural effects in the data collection phase, the greatest care must be taken in selecting methods and data interpretation.

3 Research Questions

The previous chapter provides an overview of the current state of the art. The chapter identifies research gaps concerning the effects of context on usability assessments of HMIs for L3 ADS conducted in user research. Five research questions are targeted in this thesis. The research questions and the approaches to answer them are presented in this chapter. The thesis structure aligns with the five research questions depicted in Figure 3.1.

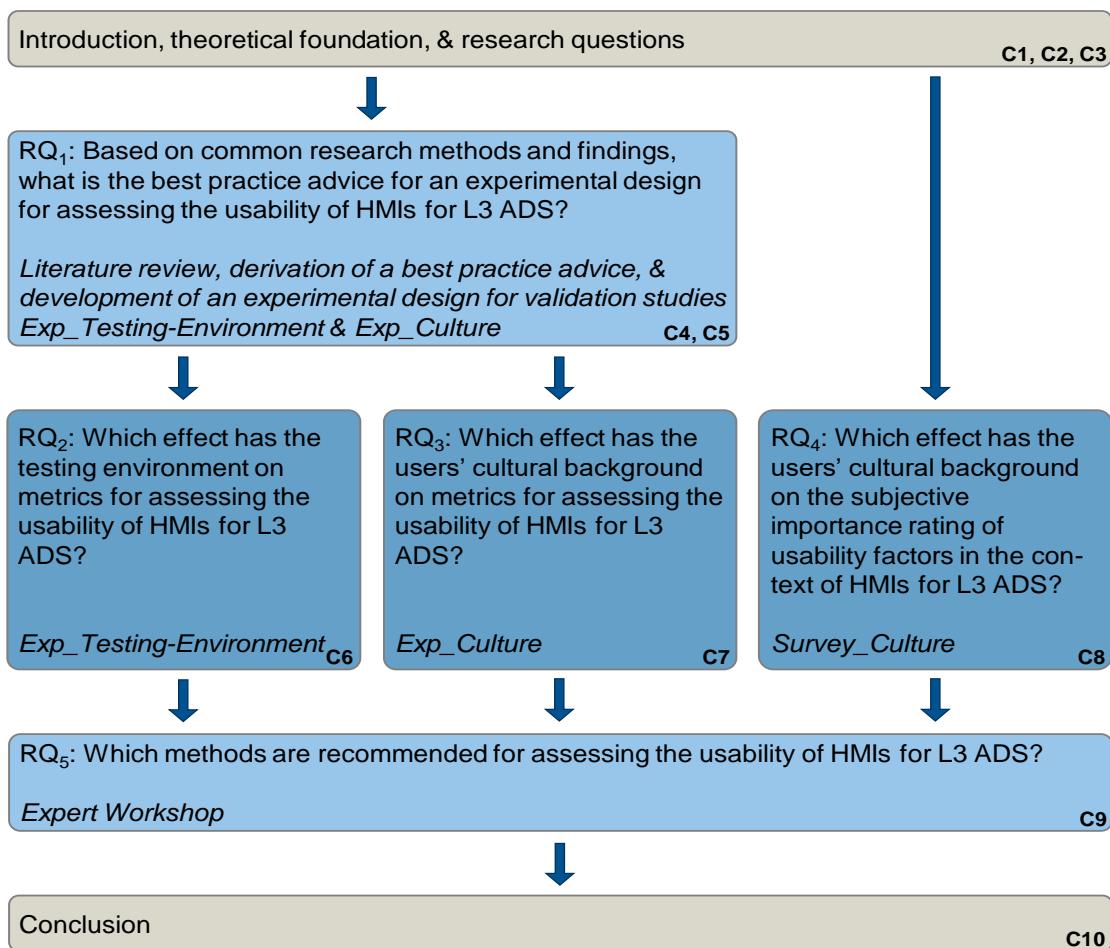


Figure 3.1 Overview of the structure of the thesis. To enhance the understandability, chapters covering research questions or empirical data are colored differently.

RQ₁ *Based on common research methods and findings, what is the best practice advice for an experimental design for assessing the usability of HMIs for L3 ADS?*

The research question RQ₁ addresses the status quo of common research methods and findings applied in usability testing of HMIs for L3 ADS. In the first step, a systematic literature review is conducted to answer this research question. Based on the literature review, a best practice advice is developed. The approach and results are presented in Chapter 4. In the second step, the best practice advice is transcribed into a study design for user tests applied in

three experiments presented in this thesis. The study design is described in Chapter 5. The three experiments provide the data basis for two validation studies presented in this thesis.

RQ₂ Which effect has the testing environment on metrics for assessing the usability of HMIs for L3 ADS?

The validation study *Exp_Testing-Environment* examines the effect of the testing environment on a selection of usability metrics. In particular, a static driving simulator is compared to an instrumented vehicle on a test track: The experiment *Sim_GER* is conducted in a static driving simulator at the Chair of Ergonomics in Garching. The experiment *TT_GER* is conducted in an instrumented vehicle on a test track at the Universität der Bundeswehr in Neubiberg. The experiments and the comparative analysis of the results are presented in Chapter 6. The chapter concludes by assessing the validity of driving simulators for assessing the usability of HMIs for L3 ADS, thereby answering research question RQ₂.

RQ₃ Which effect has the users' cultural background on metrics for assessing the usability of HMIs for L3 ADS?

The validation study *Exp_Culture* addresses the effect of the users' cultural background on a selection of usability metrics (RQ₃). The data of the experiment *TT_GER* is reused for this validation study. The participant sample consists of Germans. The experiment *TT_USA* is conducted in an instrumented vehicle on a test track at BMW Driving Academy in Maisach. The participant sample of the experiment *TT_USA* consists of U.S.-Americans¹. Chapter 7 presents the experiments and the comparative analysis of the results. This validation study aims to provide insights into the validity of usability assessments of HMIs for L3 ADS conducted in different cultural settings and, thereby, the transferability of conclusions across cultures.

RQ₄ Which effect has the users' cultural background on the subjective importance rating of usability factors in the context of HMIs for L3 ADS?

The validation study *Survey_Culture* examines the effects of the users' cultural background on the subjective importance of different usability factors in the context of HMIs for L3 ADS. Subjective data on the importance ratings and cultural values are collected. The samples are drawn from the German population (*ON_GER*), the U.S.-American population (currently residing in the USA; *ON_USA*), and from the experiment *TT_USA* conducted with U.S.-American participants in Maisach, Germany. Results from the three samples are compared and discussed in Chapter 8. This chapter aims to answer research question RQ₄, thus deepening the insights in the importance of culture in the usability testing of HMIs for L3 ADS.

¹ Initially planned experiments in the United States and Japan are canceled due to the COVID-19 pandemic. Instead, one experiment is conducted in Germany with U.S.-American participants (*TT_USA*).

RQ₅ Which methods are recommended for assessing the usability of HMIs for L3 ADS?

An expert workshop is conducted to discuss a set of preliminary recommendations for the assessment of usability in HMIs for L3 ADS. The preliminary set of recommendations is derived from the findings and experiences of the experiments presented in chapters Chapter 6, Chapter 7, and Chapter 8. The workshop's results are consolidated with literature findings and this thesis' empirical findings. Chapter 9 presents the expert workshop and the consolidation of the final methodological recommendations for usability testing of HMIs for L3 ADS, thereby answering research question RQ₅.

Chapter 10 summarizes the findings alongside the five research questions. Furthermore, the findings are critically reflected regarding their limitations and generalizability. After concluding the contribution of this thesis, potential fields for future work and the key messages are identified.

4 Development of a Best Practice Advice for Assessing the Usability of HMIs for L3 ADS

A systematic literature review is conducted to gain insights into the common research practices for assessing usability. Furthermore, a best practice advice is derived. This advice serves as the basis for the experimental design applied in the validation studies *Exp_Testing-Environment* and *Exp_Culture* presented in Chapter 6 and Chapter 7. The literature review and the experimental design based on the best practice advice address RQ₁. The literature review is published in Albers, Radlmayr, et al. (2020) and may be referred to for details. It comprises a detailed analysis of the selected sixteen articles and the derivation of the best practice advice. The approach and results are summarized in this chapter.

4.1 Analysis of the Status Quo of Common Research Methods and Findings

The review is based on the guidelines ‘Reporting Items for Systematic Reviews and Meta-Analyses’ (Moher et al., 2009). Articles are selected that feature a combination of keywords such as “usability”, “human-machine interface”, or “conditionally automated driving”. Initially, a set of 560 articles is identified during the search phase. The final selection features 16 study and theoretical articles focusing on usability for HMIs in the context of L3 automated driving. The study articles are analyzed regarding the study characteristics applied in their experiments. The theoretical articles are examined regarding the recommendations for study designs. The following six experiment characteristics serve as categories to structure the findings: definition of usability, testing environment, sample characteristics, test cases, dependent variables, and conditions of use.

4.1.1 Definition of Usability

Since the scope of the research effort is on usability assessments, the literature review analyzes the applied definitions of usability. The analysis shows that four articles do not provide a distinct definition or operationalization of usability. Five articles operationalize usability using metrics such as the SUS (Brooke, 1996). Additionally, two articles refer to the minimum requirements provided by the NHTSA as a practical guide (2017). According to the requirements, the user of an ADS HMI must be able to understand if the ADS is “(1) functioning properly; (2) currently engaged in ADS mode; (3) currently ‘unavailable’ for use; (4) experiencing a malfunction; and/or (5) requesting control transition from the ADS to the operator” (NHTSA, 2017, p. 10). Finally, four articles apply (a variation) of the definition provided by the ISO standard 9241-11 (ISO, 2018a), and two articles refer to the definition of usability provided by Nielsen (1993).

4.1.2 Testing Environment

Twelve articles provide information on the applied or recommended testing environment. Driving simulators are listed in 10 of these 12 articles. Of these 10 articles, two apply moving-base driving simulators, and four articles report using fix-base driving simulators. Two other articles describe the applied or recommended driving simulators as low-fidelity or high-fidelity

driving simulators, respectively. The use of instrumented vehicles is recommended twice in theoretical articles. One study article applies desktop methods for assessing paper and video prototypes.

4.1.3 Sample Characteristics

Regarding the sample characteristics, 14 articles provide information and either list experts or potential users as participants. Most of these articles ($n = 12$) list potential users as participants for the usability test, and only two study articles report conducting tests with experts only. Twice, both expert and user testing are recommended. Regarding the sample characteristics, the sample sizes of the expert samples vary between $N = 4$ and $N = 9$ and list ergonomics, HMIs, or ADAS as background. The sample sizes of tests with potential users range between $N = 12$ and $N = 57$. Five of the seven study articles with potential users draw samples from their own company. The age distribution mostly ranges between 20 and 62.

4.1.4 Test Cases

Thirteen articles provide information on test cases. In 10 articles, test cases cover transition scenarios. Downward transitions, for example, L3 to L0, are covered in all of these articles, while upward transitions are described in eight articles. Four articles (additionally) cover test cases on system modes and availabilities of LoAs. Six articles (additionally) cover test cases on planned maneuvers, different traffic scenarios, or the interaction with navigation systems.

4.1.5 Dependent Variables

Three articles do not provide information on dependent variables. Six articles apply or recommend observational metrics such as gaze behavior or interaction performance. Six articles (additionally) apply or recommend using standardized usability questionnaires such as the SUS (Brooke, 1996). Questionnaires for constructs affiliated with usability, such as acceptance, are (additionally) applied or recommended by seven articles. Seven articles (additionally) apply or recommend using qualitative methods such as interviews or heuristic evaluations (Nielsen & Molich, 1990).

4.1.6 Conditions of Use

The conditions of use are generally not reported in detail. In 14 articles, information indicates that first contact interaction is tested or recommended to be tested in all these cases. Five articles specifically test intuitive use without detailed instructions. Seven articles additionally report or recommend testing interactions of repeated contact.

4.2 Derivation of a Best Practice Advice

The review concludes with a best practice advice for the six study characteristics. The best practice advice is briefly described in this section and depicted in Table 4.1.

It recommends defining and operationalizing usability in the context of HMIs for L3 ADS through a combination of the definition provided by the ISO standard 9241-11 (ISO, 2018a) and the NHTSA minimum requirements (NHTSA, 2017). Regarding the testing environment,

the best practice advice recommends using high-fidelity driving simulators, which aligns with the status quo. For early prototypes, the advice acknowledges the value of desktop methods. The best practice advice further recommends conducting tests with potential end users. The sample characteristics are supposed to represent the potential user regarding the distribution of characteristics such as age, gender, prior experience, or affiliation with technical devices. The sample size is to be selected based on the planned statistical procedure. The best practice advice recommends focusing on transitions between LoAs, the availability of LoAs, and non-critical scenarios when determining the test cases. Regarding dependent variables, the best practice advice recommends the application of observational and self-reported metrics. The observational data are further specified in collecting visual behavior and interaction performance data. The advice recommends applying the *SUS*, short interviews, and supplementing standardized questionnaires for self-reported data. Finally, providing only general information on the ADS and testing the first contact interaction are recommended for the conditions of use.

Table 4.1 Best practice advice for testing the usability of HMIs for L3 ADS from Albers, Radlmayr, et al. (2020).

Study characteristic	Best practice advice
Definition of usability	General Definition: "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO, 2018a, p. 2) Practical Realization: the user understands that the ADS is "(1) functioning properly; (2) currently engaged in ADS mode; (3) currently 'unavailable' for use; (4) experiencing a malfunction; and/or (5) requesting control transition from the ADS to the operator" (NHTSA, 2017, p. 10)
Testing environment	Driving simulator
Sample characteristics	Sample group: represents the potential user population (age, gender, prior experience, affiliation with technical devices, etc.) Sample size: determined by the statistical procedure
Test cases	Scenarios: (1) transitions between different automation modes and (2) availability of different automation modes Criticality: non-critical situations
Dependent variables	General: combination of observational and subjective metrics Observational metrics: (1) visual behavior according to the ISO 15007 (ISO, 2018b) (e.g., percent on area of interest (AOI)) and (2) the interaction performance with the HMI (e.g., operating errors or reaction time for a button press) Subjective Metrics: (1) <i>SUS</i> (Brooke, 1996), (2) short interviews after test trials and questionnaires, and (3) supplementary standardized questionnaires
Conditions of use	First contact between user and ADS Instructions contain only general information on the ADS

5 Experimental Design for Validation Studies Exp_Testing-Environment and Exp_Culture

This chapter presents the experimental design that is applied in the three experiments of the validation studies *Exp_Testing-Environment* and *Exp_Culture*. The study design builds upon the best practice advice presented in the previous chapter and completes the work on RQ₁. The development of the study design is published by Albers et al. (2021). This chapter builds upon the publication and provides more detailed insights.

The validation studies are designed as between-subject studies comprising the independent variables experiment (*Exp*) and HMI concept (*HMI*). In each of the three experiments, two subsamples are formed by the independent variable of the HMI concept. Potential training and sequential effects are expected to be considerable due to the similarity of the basic structure of the HMI concepts (see Subsection 5.4.1). By choosing a between-subject design, the influence of learning effects is avoided.

The validation studies focus on assessing the effects of the testing environment and the users' cultural background. Therefore, the overall experimental design strives to achieve high internal validity, especially regarding standardization. Where possible, the experimental design reflects a realistic setting (e.g., scenarios, information availability, HMI design) to ensure the generalizability of results (see Subsection 2.1.3 for more details on the trade-off between internal and external validity).

5.1 Definition of Usability

The underlying definition of usability is provided by the ISO standard 9241-11 (ISO, 2018a) and thereby covers the three usability facets effectiveness, efficiency, and satisfaction. In addition, the NHTSA minimum requirements (NHTSA, 2017) are considered criteria for a practical approach. The operationalization of the usability assessment is realized through the selection of test cases (see Section 5.3) and metrics (see Section 5.6).

5.2 Sample Characteristics

The samples consist of naïve participants regarding their experience with L3 ADS. This enables the assessment of intuitive usability in a first-contact interaction. The recruitment criteria strive to represent the entire population of drivers and the population of potential future users. Participants are required to hold a valid driving license. A balanced gender distribution is aimed at a minimum of 30% females. The participants' targeted age range is between 18 and 75 years. Following the NHTSA visual-manual distraction protocol, the age distribution is aimed to include a minimum of five participants in four different age groups: (1) 18-24; (2) 25-39; (3) 40-54; (4) > 54 (NHTSA, 2013). This leads to a minimum sample size of 20 participants per subsample. Participants are evenly distributed to the subsamples based on the criteria described above. Participants suffering from physical or cognitive impairments are excluded. Additionally, participants whose mobility or perception is affected by the intake of medication or drugs are excluded.

After recruitment, participants are requested to provide additional information. This includes factors potentially affecting the interactions, such as visual impairments. Information on the participants' driving experience, such as familiarity with ADAS, is expected to support the interpretation of interindividual differences or to identify subsamples relevant to future research.

5.3 Test Cases

The ADS enables L0, L2, and L3 driving. For the sake of simplicity, the system does not offer L1 driving. The L2 ADS is implemented and instructed as a L2 hands-on ADS. Only in L3 driving participants are allowed to take their hands off the steering wheel. Otherwise, a hands-off (H-off) detection warning is issued. Each experiment covers 12 test cases. Due to safety aspects, the test cases mostly comprise non-critical situations. The usability assessment focuses on situations with a high probability of occurrence, such as the interaction when using the basic functions. Instead, safety-related assessments of ADS are mainly affiliated with constructs such as controllability (see Subsection 2.1.2). These critical situations have a low probability of occurrence. The selection is based on and, therefore, linked to the NHTSA minimum requirements (NHTSA, 2017), as shown in Table 5.1. Three test cases (TC1, TC4, TC7) cover continuous rides in L0, L2, and L3 without further events. Three test cases (TC2, TC6, TC8) feature changes in the availability of LoAs. Here, twice, LoAs become available that have not been available before, and once, L3 becomes unavailable due to a malfunction. None of these availability changes affects the currently activated LoA. Three test cases (TC3, TC5, TC9) cover transitions initiated by the participant (upon request of the experimenter). Two test cases (TC10, TC12) feature RtIs. In TC10, the system reaches the end of the ODD, thus triggering an RtI with a time budget of 20 s (RtI_{20s}) before the emergency braking maneuver begins ("ODD end" in Table 5.1). In TC12, the system is degraded by a malfunction of sensors affecting the currently active L3, thus triggering an RtI requiring an immediate reaction of the driver 6 s (RtI_{6s}) before the emergency braking maneuver begins ("malfunction" in Table 5.1). One test case (TC11) features a combination of a change in the availability and a transition request initiated by the participant (upon request of the experimenter).

The HMI concepts in the experiments continuously provide information on the currently active LoA and the available LoAs. Therefore, all test cases allow to collect data on the first three NHTSA minimum requirements "functioning properly", "currently engaged in ADS mode", and "currently 'unavailable' for use" (NHTSA, 2017, p. 10). The latter two requirements, "experiencing a malfunction" and "requesting control transition from the ADS to the operator" (NHTSA, 2017, p. 10), are addressed only in two test cases each (TC6 & TC12 and TC10 & TC12, respectively).

Despite disadvantages such as potential training and sequential effects (Bortz & Döring, 2006, p. 184), the test cases have a fixed order. No full permutation could be realized with the planned study design, and most of the test cases require the precedence of specific other test cases; for example, a take-over request (TC10 or TC12) could and should not be tested before the first activation of L3 (TC3).

Table 5.1 Description of the 12 test cases and their linkage to the NHTSA minimum requirements, adapted from Albers et al. (2021).

Test case	Description	Active LoA [higher LoAs available] (LoAs according to SAE International, 2021)	NHTSA minimum requirements (NHTSA, 2017)
1	Continuous ride in L0, no events	L0 [-]	1, 2, 3
2	Change in availability	L0 [-] → L0 [L2, L3]	1, 2, 3
3	Transition: initiated by participant	L0 [L2, L3] → L3	1, 2, 3
4	Continuous ride in L3, no events	L3	1, 2, 3
5	Transition: initiated by participant	L3 → L2 [L3]	1, 2, 3
6	Change in availability (malfunction)	L2 [L3] → L2 [-]	1, 2, 3, 4
7	Continuous ride in L2, no events	L2 [-]	1, 2, 3
8	Change in availability	L2 [-] → L2 [L3]	1, 2, 3
9	Transition: initiated by participant	L2 [L3] → L3	1, 2, 3
10	Change in availability (ODD end) & transition: system-initiated	L3 → L0 [-]	1, 2, 3, 5
11	Change in availability & transition: initiated by participant	L0 [-] → L3	1, 2, 3
12	Change in availability (malfunction) & transition: system-initiated	L3 → L0 [-]	1, 2, 3, 4, 5

Note. "The NHTSA minimum requirements (NHTSA, 2017, p. 10) are: "(1) functioning properly; (2) currently engaged in ADS mode; (3) currently 'unavailable' for use; (4) experiencing a malfunction; (5) requesting control transition from the ADS to the operator."

5.4 HMI Concepts

Participants of all experiments are randomly assigned to one of two implemented HMI concepts^{2,3}. Both HMI concepts are evaluated in all three experiments. An overview of the HMI concepts is provided in Appendix I. The HMI concepts serve as the artificial research subject. Introducing two HMI concepts per experiment allows for assessing the relative validity, which refers to the agreement between the direction (and size) of effects. Furthermore, the sensitivity of metrics toward specific differences in HMI design may be assessed.

Forster and colleagues (Forster et al., 2020a, 2020b) investigate the difference between two HMI concepts that vary in their compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019). The within-subject study confirms differences in usability and acceptance measures from behavioral and self-reported data. The approach of variation between two HMI concepts is adapted from this study as presented in Forster et al. (2020b) and Forster et al. (2020a).

In this section, the design and the development of the underlying HMI concept are described. Afterward, the differences between the two HMI concepts are described, as well as the heuristic expert evaluation confirming the different degrees of compliance.

² The HMI concepts are designed and evaluated with the assistance of Canroz Tacay (2020) as part of his term paper.

³ The implementation and control of the HMI concepts in the instrumented vehicles for the experiments *TT_GER* and *TT_USA* is realized by Jessica Kos (2020) as part of her bachelor's thesis.

5.4.1 Basic Design

The underlying basic HMI concept builds upon the design of Feierle et al. (2020), originating from a design of Götze (2018). The basic concept is adapted for the LoAs L0, L2, and L3 and the selection of test cases. Due to technical constraints and simplicity, the HMI mainly consists of visual components. The instrument cluster (*IC*) structure for the two concepts is visualized in Figure 5.1 (left). Speed (Figure 5.1 (1)) is displayed on the left side of the *IC*, and infotainment features (Figure 5.1 (2)) are visualized in the right area. While the speed is displayed synchronously with the realized driving data, the infotainment area is static and not functionally implemented in the prototype. The central area of the *IC* displays the ego vehicle in its current lane (L2 & L3 only), surrounded by a ring serving as a metaphor for the vehicle's surrounding environment (introduced by van Gijssel, 2012; Figure 5.1 (3)). Following Melcher et al. (2015), the lower area in the center of the *IC* displays a scale that includes the three LoAs. The scale indicates the currently active LoA and the availability of all LoAs (Figure 5.1 (4)). Three icons are designed to represent the three LoAs (Figure 5.1, right). The icon for L0 displays a steering wheel gripped by two hands. The icon for L2 displays a steering wheel that is only touched by two hands. Above the steering wheel are two arches associated with radio waves. The icon for L3 does not show hands but three arches above the steering wheel.

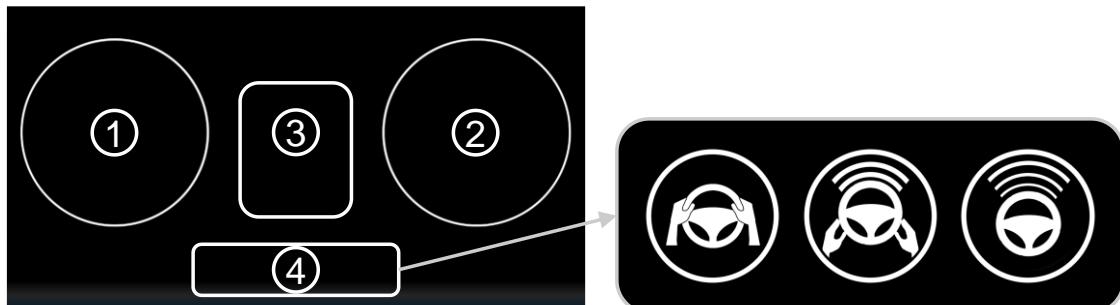


Figure 5.1 Left: Structure of the basic HMI design implemented in the *IC*: (1) speed; (2) infotainment; (3) ego vehicle and its surrounding environment (L2 & L3 only); (4) scale for LoAs. Right: Icons for the three implemented LoAs L0 (left), L2 (center), and L3 (right).

The HMI continuously provides visual information on the active LoA and the availability of the LoAs. In addition, further information, such as malfunctions and RTIs, is displayed visually. The language of the HMI is German in *Sim_GER* and *TT_GER*, and U.S.-American English in *TT_USA*.⁴

The ADS is controlled mainly via buttons on the steering wheel. A multifunction steering wheel of the BMW 3 series G21 is used (Figure 5.2, left). Only two buttons on the left spoke are relevant. These buttons are covered with stickers featuring customized labels (Figure 5.2, right).

⁴ The translation is performed through an agency.



Figure 5.2 Left: Position of the control buttons on the steering wheel. Right: Control buttons ACT and MOD and their respective icons.

Due to technical constraints, the control logic is not varied between the two HMI concepts. Two buttons allow the user to activate the different LoAs and to switch between these LoAs. The buttons and their functions are depicted in Figure 5.3. The left button, hereafter referred to as *ACT* (for activation), triggers transitions between L0 and L2 and vice versa ($L0 \rightarrow L2$; $L2 \rightarrow L0$). When pressed while L3 is active, the *ACT* button deactivates L3 driving and switches to L0 ($L3 \rightarrow L0$). The label displays an icon for a power button complemented with the letters “AUT”. The right button, hereafter referred to as *MOD* (for mode), triggers transitions between L2 and L3 and vice versa ($L2 \rightarrow L3$; $L3 \rightarrow L2$). The *MOD* button has no effect when being pressed while L0 is active. The label displays an icon with two arrows pointing up and down, complemented by the letters “AUT”. Oversteering—a braking or strong steering maneuver—also triggers a transition from L2 or L3 to L0 ($L2/L3 \rightarrow L0$).

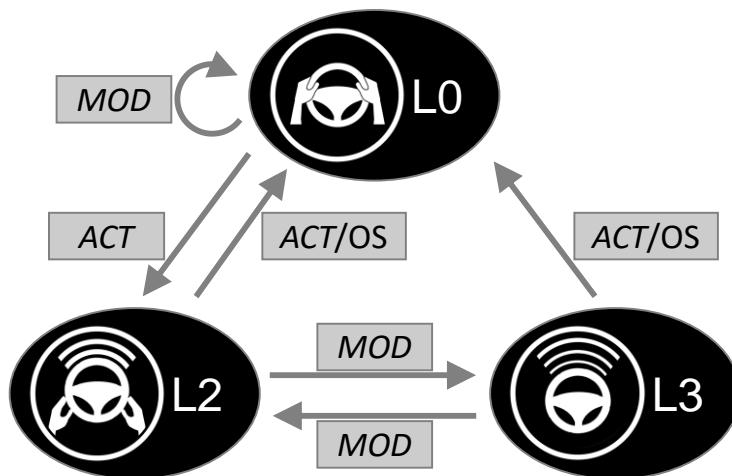


Figure 5.3 Visualization of the control logic of the HMI concepts. Transitions between LoAs are triggered via the control buttons ACT and MOD, or oversteering (OS), that is, braking or strong steering maneuvers.

5.4.2 Differences between the two HMI Concepts

Starting from the basic concept, two HMI concepts are developed. Following the framework of HMIs proposed by Bengler et al. (2020), the HMI concepts can be distinguished as follows: The input channels and dialog logic are identical. The output channel(s) providing information about the system status comprise the differences between the two concepts.

One HMI, hereafter referred to as High-Compliance-HMI (*HC-HMI*), is designed in compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019). The HMI comprises the *IC* and LED strips on the steering wheel. Furthermore, warning sounds are implemented for the multimodal communication of urgent information (Naujoks, Wiedemann, et al., 2019, item 18).

The second HMI, hereafter referred to as Low-Compliance-HMI (*LC-HMI*), features low compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019). Eight guideline items are intentionally violated, as described in Table 5.5.2. For example, only the *IC* is implemented to visually communicate with the participant. The *LC-HMI* does not use auditory or additional visual signals, such as the LED strips on the steering wheel, violating the multimodality of high-priority notifications (Naujoks, Wiedemann, et al., 2019, item 18).

Table 5.5.2 Overview of the eight items of Naujoks, Wiedemann, et al. (2019) that differentiate between the HMI concepts and description of their implementation in the *HC-HMI* and the *LC-HMI* concept, respectively.

Item of Naujoks, Wiedemann, et al. (2019, p. 129) (<i>supporting literature</i>)	Implementation in <i>HC-HMI</i>	Implementation in <i>LC-HMI</i>
Item 3: "System state changes should be effectively communicated." <i>(Kelsch et al., 2017)</i>	<p>After a transition, the now active LoA is permanently communicated via the color of the ego vehicle in the center of the <i>IC</i> and the color of the respective icon in the scale at the bottom.</p> <p>The icon of the active LoA in the scale is displayed bigger than the icons of the other LoAs.</p>	<p>After a transition, the now active LoA is permanently communicated via the color of the ego vehicle in the center of the <i>IC</i> and the color of the respective icon in the scale at the bottom.</p>
Item 5: "HMI elements should be grouped together according to their function to support the perception of mode indicators." <i>(Kelsch et al., 2017; Stevens et al., 2002)</i>	<p>The detected speed limit is displayed in the left area close to the information on the current speed.</p> <p>Notifications concerning the ADS are displayed in the central upper area of the <i>IC</i>.</p>	<p>The detected speed limit is displayed in the right area close to the infotainment area.</p> <p>Notifications concerning the ADS are displayed as an overlay in the infotainment area.</p>
Item 7: "The visual interface should have a sufficient contrast in luminance and/or color between foreground and background." <i>(ISO, 2009)</i>	<p>The colors of all LoAs fulfill the recommended ($\geq 5:1$) contrast ratio requirements when being displayed on the black display background. L0 is displayed in white (RGB 255, 255, 255) and has a contrast ratio of 21:1. L2 (green: RGB 0, 255, 0) has a contrast ratio 15.3:1. L3 (cyan: RGB 0, 255, 255) has a contrast ratio 16.7:1.</p>	<p>Not all LoAs are displayed in colors that fulfill the recommended ($\geq 5:1$) contrast ratio requirements when being displayed on the black display background. L0 (white: RGB 255, 255, 255) has a contrast ratio 21:1. L2 (dark blue: RGB 66, 51, 255) has a contrast ratio 3.1:1. L3 (yellow: RGB 255, 201, 14) has a contrast ratio 13.6:1.</p>

Item of Naujoks, Wiedemann, et al. (2019, p. 129) (supporting literature)	Implementation in HC-HMI	Implementation in LC-HMI
Item 8: "Texts (e.g., font types and size of characters) and symbols should be easily readable from the permitted seating position." (ISO, 2009; Stevens et al., 2002)	The font size (42 pt) is sufficient.	The font size (38 pt) is sufficient but smaller than the font size in HC-HMI.
Item 9: "Commonly accepted or standardized symbols should be used to communicate the automation mode. Use of non-standard symbols should be supplemented by additional text explanations or vocal phrases." (Deutsches Institut für Normung [DIN], 2003; Stevens et al., 2002)	The sans-serif font Arial is used.	The serif font Times New Roman is used, applying the style italics.
Item 14: "The colors used to communicate system states should be in accordance with common conventions and stereotypes."	After a transition, the non-standard icon of the now active LoA is temporarily displayed as an overlay in the infotainment area. The icon is supplemented with a temporary pop-up message in the central upper area of the IC announcing the currently active LoA.	Transitions are not supplemented by notifications or other information explaining the meaning of the non-standard icons representing L0, L2, and L3.
Item 15: "Design for color-blindness by redundant coding and avoidance of red/green and blue/yellow combinations." (Brandes et al., 2019, p. 760)	In accordance with the criticality, warning messages are displayed in yellow or red, while non-critical notifications are displayed in white. The LoA L3 is coded with the color cyan. In research, cyan is already commonly used (e.g., Clercq et al., 2019; Dey et al., 2021; Fuest et al., 2020; Y. M. Lee et al., 2019) and recommended (e.g., Faas & Baumann, 2019; Werner, 2018) to indicate automated driving.	Irrespective of the criticality, all notifications are displayed in white. The LoA L3 is coded with the color yellow. The color yellow is associated with warnings (e.g., J. L. Campbell et al., 2007; Green et al., 1994; Uttesch, 2014)
Item 18: "High-priority messages should be multimodal." (J. L. Campbell et al., 2007; Stevens et al., 2002)	The active LoA is redundantly communicated via the icon's color, position, and size. No red/green or blue/yellow combinations are selected for the LoAs. The non-availability of LoAs is communicated redundantly via crossing-out and grey color-coding of the icons.	The active LoA is redundantly communicated only via the icons' color and position. For the color coding of the LoAs L2 and L3, a blue/yellow combination is selected. The non-availability of LoAs is communicated only via grey color-coding of the icons.
	In accordance with the criticality, notifications are displayed in the IC and supplemented with LED lights flashing on the steering wheel and warning sounds.	Irrespective of the criticality, all notifications are displayed in the IC only. No other visual or auditory signals are used.

5.4.3 Heuristic Expert Evaluation on Differences between the HMI Concepts

A heuristic evaluation with six experts is conducted to validate the intended differences between the HMI concepts. The six experts ($n = 1$ female, $n = 5$ male) have been working in the field of HMIs for three to seven years ($M = 4.58$, $SD = 1.48$). In a permuted order, the experts experience both HMI concepts and rate each. A list of 10 heuristics is provided as a guidance (Table 5.3). The heuristics are based on the heuristics of Nielsen (2005) and the guidelines provided by (Naujoks, Wiedemann, et al., 2019).

Table 5.3 List of heuristics used in the expert evaluation.

Heuristic	Description	Source
1	“System state changes should be effectively communicated.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 3
2	“The visual interface should have a sufficient contrast in luminance and/or [color] between foreground and background.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 7
3	“Texts (e.g., font types and size of characters) and symbols should be easily readable from the permitted seating position.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 8
4	“[...] Use of non-standard symbols should be supplemented by additional text explanations or vocal phrase/s.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 9
5	“The [colors] used to communicate system states should be in accordance with common conventions and stereotypes.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 14
6	The HMI allows recognition rather than recall.	Nielsen, 2005, principle 6
7	“In case of sensor failures, their consequences and required operator steps should be displayed.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 20
8	The visual interface has an aesthetic and minimalist design.	Nielsen, 2005, principle 8
9	“HMI elements should be grouped together according to their function to support the perception of mode indicators.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 5
10	“The semantic of a message should be in accordance with its urgency.”	Naujoks, Wiedemann, et al., 2019, p. 129, item 10

The experts could indicate violations of the heuristics as well as the severity of the violation ranging between “0: this is not a usability problem at all” to “4: usability catastrophe —imperative to fix this before product can be released” (Nielsen, 1993, p. 103). After the assessment, an interview focuses on colors, icons, and the icons’ positioning. Furthermore, the experts are asked to express further feedback and comments.

The heuristic evaluation confirms the different degrees of compliance (see Figure 5.4). For the *HC-HMI*, the experts list 10 violations of the heuristics with a severity of 1 or higher ($M = 2.4$, $SD = 0.66$, $Med = 2$). For the *LC-HMI*, the experts state 35 violations with a severity of 1 or higher ($M = 3.2$, $SD = 0.95$, $Med = 4$). For the *HC-HMI*, three issues with severity scores of 3 or 4 are mentioned. The experts criticize the insufficient saliency of H-off notifications, the time-based countdown for the end of the ODD (TC10) instead of providing distance information, and the use of green color for L2, which implies no need for action and is therefore deemed more suitable for L3. Regarding the color selection the concluding interview results in mainly positive opinions (e.g., “fitting, blue and green look technical”). For the *LC-HMI*, 26 issues with severity scores of 3 or 4 are mentioned. The criticisms address all of the

implemented differences between the HMI concepts. The focus of criticism lies on the color selection (yellow used for L3 conveys caution or warning messages), too small icons, and the insufficient saliency and urgency of warnings due to the color selection and the visual implementation alone. Critique referring to the usage of serif font and the positioning of notifications is mentioned with less severity.

General feedback on the basic HMI design provokes the improvement of the icons for the control buttons (as displayed in Figure 5.2; the original icons are rated as visually cluttered) and the removal of the ego vehicle and the vehicle's surrounding environment visualized by a ring (see Figure 5.1, (3)) in L0 (no assistance systems for the vehicle surrounding are expected in L0). Two experts criticize the control logic. Due to technical limitations, the control logic is not altered. The *HC-HMI* is further improved based on the suggestions to increase the differences between the HMI concepts. Among minor changes, such as wording adjustments in single notifications, the salience of RtIs and H-off notifications is increased.

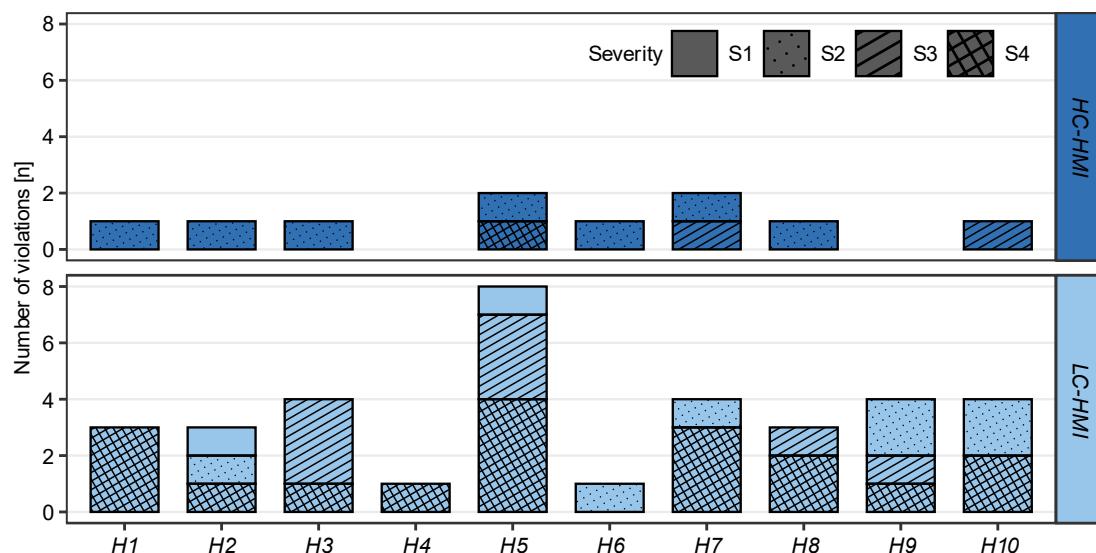


Figure 5.4 Results of the heuristic expert evaluation for the HMI concepts *HC-HMI* and *LC-HMI*.

Note. The number of violated heuristics (H1-H10) and the severity ranging between 1 and 4 is indicated.

Figure 5.5 displays excerpts of the HMI concepts in two different scenarios. Appendix I (Table 12.1) contains more excerpts of the HMI concepts.



Figure 5.5 Excerpts of the HMI concepts (*left: HC-HMI; right: LC-HMI*) in the *IC*. *Top*: L2 has just been activated. *Bottom*: Second stage of the warning cascade during the Rtl_{20s} .
Note. The second stage of the warning cascade during the Rtl_{20s} in the *HC-HMI* indicates a countdown (one box disappears per second). The warning is supplemented with yellow LED lights flashing on the steering wheel and a warning sound with low criticality.

5.5 Study Procedure

After welcoming, the participants give informed consent to participate in the experiment, to allow the data collection, and to follow the instructions and safety regulations given by the experimenter. The second part consists of a pre-questionnaire about the participants' sociodemographic and driving backgrounds. Afterward, participants familiarize themselves with the test course and the driving simulator or instrumented vehicle, respectively. After the familiarization drive, participants receive further instructions on the procedure and their driving task. The instructions include the LoAs, the HMI, and the NDRA information. After a clarification of questions, the eye-tracking system is calibrated. The following test drive, including short interviews, takes about 45 min. After the test drive, participants report their experiences through a post-questionnaire and a final interview. The total duration of the experiment varies due to interindividual and organizational differences between 1.5 hr and 2 hr in the driving simulator and 2 hr and 2.5 hr in the test track experiments, respectively.

5.5.1 NDRA

In contrast to L0 and L2, L3 driving allows to engage in NDRA. The standardized NDRA surrogate reference task (*SuRT*; ISO, 2012) is introduced as an NDRA to provide an observational measure of mode awareness or compliance with the responsibilities for the driving task. In the center console, a tablet featuring the *SuRT* is installed. Participants are instructed that engagement in the *SuRT* is only allowed in L3 driving but not in L0 or L2. While L3 is active, participants are encouraged to engage in the *SuRT*. Before the test drive, participants familiarize themselves with the *SuRT* and are encouraged to ask questions.

5.5.2 Instructions

The experiments are designed to test the intuitive usability during first contact interaction with the ADS. Therefore, no detailed information on the operation of the ADS is provided.

After welcoming, participants receive general information on the study context, procedure, and safety instructions.

After the familiarization drive, participants receive more detailed instructions. First, the three LoAs are introduced as “Manual Driving” (corresponds to L0), “Assisted Driving” (corresponds to L2), and “Automated Driving” (corresponds to L3). For each of the LoAs, information is provided explaining the abilities of the respective LoAs and the resulting responsibilities that lie with the driver. The information is based on the simplified description of the SAE LoAs for customers (Shuttleworth, 2019). Afterward, general information on the HMI is provided, informing the participant that the interaction with the ADS occurs through the HMI. The two control buttons on the steering wheel are indicated, though their function and the control logic are not explained. After that, the *SuRT* is introduced. The written instruction closes with explaining the test drive procedure and the participants’ required actions. Due to the test course features and safety aspects, the speed limit is set to 30 km/h, and each test case starts and ends at a standstill. Participants must manually accelerate at the test case’s start and reactivate the LoA with which the previous test case ended. Pre-recorded audio announcements triggered by the experimenter support the participant in this task. At the end of each test case, participants manually slow down the car to a standstill. For standardization, participants are instructed to initiate transitions only if explicitly requested by the ADS or the experimenter.

After participants finish reading the instructions, the experimenter briefly summarizes the key aspects of the instruction and encourages the participant to clarify questions.

5.5.3 Test Course

The test course in *Sim_GER* simulates the test track at the Universität der Bundeswehr in Neubiberg, Germany (used in *TT_GER*). Due to organizational constraints, *TT_USA* could not be conducted on the same test track. Instead, *TT_USA* is conducted on a test track at the BMW Driving Academy in Maisach, Germany. Discrepancies between the test courses are reduced to a minimum. Due to safety reasons, the test drives do not include surrounding traffic, obstacles, or lane change maneuvers.

The test course comprises two about 900 m long lanes with turning opportunities at each end. A sketch of the test course, including important waypoints, is depicted in Figure 5.6 (top). Test drives are conducted on the respective right lanes. For each 900 m section, one test case is performed. At the end of each test case, participants stop the vehicle before turning the vehicle around manually, starting the next test case by driving in the other direction. The first and the last 150 m sections (0 m–150 m; 750 m–900 m) are reserved for the manual acceleration and deceleration phase. When passing 150 m or 750 m, a sound marks the beginning or end of the respective test case. The data collection is limited to the central section of the straight between 200 m and 700 m. With an average speed of 30 km/h, each test case produces about 60 s of recorded data. Events occurring during the test cases are triggered at three different waypoints (375 m, 450 m, 535 m), thus reducing the predictability of events for the participant while maintaining a high degree of standardization. The test cases are only realized in the HMI

notifications. This means that cues in the test course do not accompany changes of availabilities. Thus, changes in availabilities could not be associated with a change in road type or any other feature of the test course. Consequently, participants' reactions are distinctly attributed to the respective HMI concept.

In the test track experiments, waypoints are marked for the experimenter with traffic cones or wires laid across the lanes (only waypoints at 150 m and 750 m), as visualized in Figure 5.6 (bottom). The waypoints are not marked in the simulator experiment.

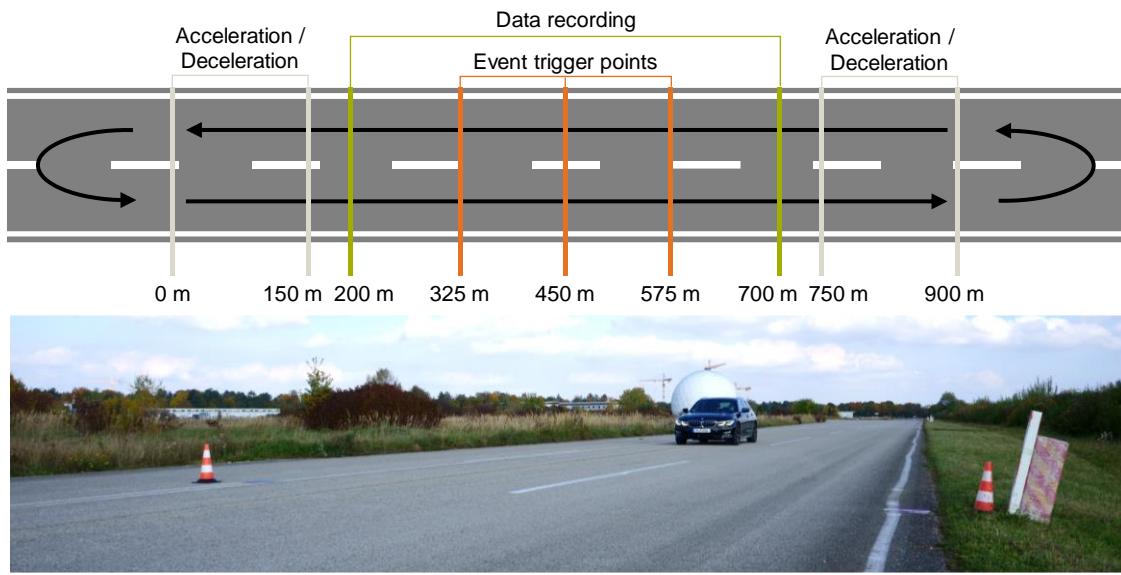


Figure 5.6 Top: Sketch of the test course. Bottom: Photo of the test course and the waypoints on the test track at the Universität der Bundeswehr in Neubiberg, Germany.

Note. After the acceleration phase (0 m-150 m), a sound marks the beginning of the test drive. Data is recorded between 200 m and 700 m. Events are triggered at three different waypoints (325 m, 450 m, 575 m). The test drive ends with another sound (750 m) and a deceleration phase (750 m-900 m). The turning points at both ends allow the test course to be driven in both directions.

5.6 Data Collection

Both observational and self-reported data are collected in the experiments. The driving simulator and the instrumented vehicles have a microphone and a camera supporting the analysis. The camera is directed at the steering wheel and catches the operation of the control buttons on the steering wheel and other movements. In the test track experiments, an additional camera is installed and directed at the driving scenery to record unexpected events potentially influencing the experiment. Figure 5.7 displays photos of the experimental setup. The simulator setup and a data gateway in the instrumented vehicle record vehicle-related data. Speed, lateral and longitudinal acceleration are recorded, as well as the operation of the gas pedal, the brake pedal, and the buttons on the steering wheel. The eye-tracking system Dikablis Glasses 3 by Ergoneers records the gaze behavior. The collection of sociodemographic and self-reported data and the documentation of the experimenter ratings is realized via LimeSurvey.

A protocol allows the experimenter to document unusual behavior, unforeseen external events, or technical issues. Furthermore, the weather and lighting conditions are documented in the test track experiments. In the driving simulator experiment, the weather and lighting conditions are set to a lightly clouded sky with bright lighting conditions.

This section briefly describes the collected data, starting with the sociodemographic data to describe the experiments' samples. Afterward, the observational and self-reported data are described, followed by the description of individual factors as potentially confounding variables. Finally, a short overview of the metrics is provided, offering a linkage to the compliance violations of the HMI concepts, the usability facets of the ISO standard 9241-11 (ISO, 2018a), and the components of the NHTSA minimum requirements (NHTSA, 2017).



Figure 5.7 Photos of the experimental setup. *Left:* The participant wears the eye-tracking system Dikablis Glasses 3 by Ergoneers. *Right:* The experimenter gives instructions, triggers events, and controls the data recording.

5.6.1 Sociodemographic Data

Sociodemographic data are collected to describe the sample and evaluate its representativity. In addition to age and gender, participants provide data on their visual deficiencies, such as the need for visual aids and color deficiencies.

Regarding the driving experience, participants report the mileage and driving frequency of the last 12 months. Afterward, participants indicate familiarity with the ADAS Cruise Control (CC), ACC, and Lane Keeping Assistant (LKA). If participants report familiarity with specific ADAS, a subsequent question inquires on the frequency of using the ADAS. Finally, participants are requested to report their prior knowledge of automated driving on a 5-point Likert scale with the anchors “0: no knowledge” and “4: expert”.

5.6.2 Observational Data

5.6.2.1 Driving Behavior

Analyzing driving behavior metrics allows for an objective assessment of the interaction quality. Data for the following metrics are collected and analyzed systematically. Other remarkable driving behavior is documented in the protocol.

The observed LoA is compared to the LoA intended by the test case schedule for all test cases.

The first contact interaction consists of an instructed transition from L0 to L3. In this test case, participants are instructed to use the input channels of the HMI for the first time (TC3). The prior test cases comprised steady rides in L0 with a change of availability in TC2. The control paths and success rate for this first contact interaction are analyzed.

The RtIs (RtI_{20s} & RtI_{6s}) require transitioning from L3 to L0. The take-over time (TOT) as the time between the start of the RtI and the transition to L0 is calculated for these test cases (TC10 & TC12).

The qualitative analysis of the take-over paths allows for identifying potential interaction problems and take-over strategies.

Throughout the test drive, the driving behavior is assessed. A focus is laid on the following two aspects: The first aspect concerns L2 driving. The number of H-off detection warnings issued by the ADS is analyzed, as well as the warning stage where participants deactivate the warning by taking their hands back to the steering wheel. The second aspect concerns unnecessary deactivations. In TC6, participants drive in L2 while receiving a notification that a sensor error has led to the non-availability of L3 driving. This information does not imply a need to act since L2 is unaffected. Deactivations following this notification may be interpreted as a misunderstanding of the notification and are therefore considered in the analysis.

5.6.2.2 Eye-Tracking

Gaze behavior comprises measures valuable to estimate the driver state, the driver's allocation of attention, and the quantification of information acquisition of stimuli such as notifications in an HMI (ISO, 2018b, p. 6). The eye-tracking data is processed and analyzed with the software D-Lab 3.60 (Ergoneers Group, 2022)

Attention ratios, that is, “the percent of time glances are within an [area of interest (AOI)]” (ISO, 2018b, p. 8), serve as a measure of trust (Körber et al., 2018) and mode awareness (Feldhütter et al., 2019) in automated driving research. Four different AOIs are defined: *Street*: the road environment (mainly windshield); *IC*; *Controls*: the control buttons for the HMI on the steering wheel; *SuRT*: the tablet installed in the center console for the NDRA. The AOIs are visualized in Figure 5.8. This experiment calculates the attention ratios for the continuous rides in all three LoAs L0, L2, and L3. The distribution of attention ratios conveys information on the mental model's correctness and the level of trust. Based on the instructions, attention ratios for the AOI *Street* are expected to be high for L0 and L2, while L3 produces high values for the *SuRT*.

In addition to the attention ratios, the gaze behavior during RtIs (TC10: RtI_{20s} & TC12: RtI_{6s}) is investigated. The glance allocation to the different AOIs at the start of the RtI is examined. The metric serves as an indicator of the development of trust and as a filter variable for the following eye-tracking metrics. RtIs are triggered (mainly⁵) by notifications displayed in the *IC*. Additionally, the gaze paths during RtIs with an emphasis on the glance allocation to the different AOIs at the end of the RtI are examined in a qualitative analysis.

⁵ In the *HC-HMI* concept, the RtI_{6s} in TC12 and the second and third stage of the warning cascade of the RtI_{20s} in TC10 are accompanied with sounds and LED lights (see Section 5.4 or Appendix I).

The glance allocation time to the *IC* reflects the visibility and saliency of these notifications. The duration of the first glance at the *IC* conveys the efficiency with which participants receive the information in the notifications.

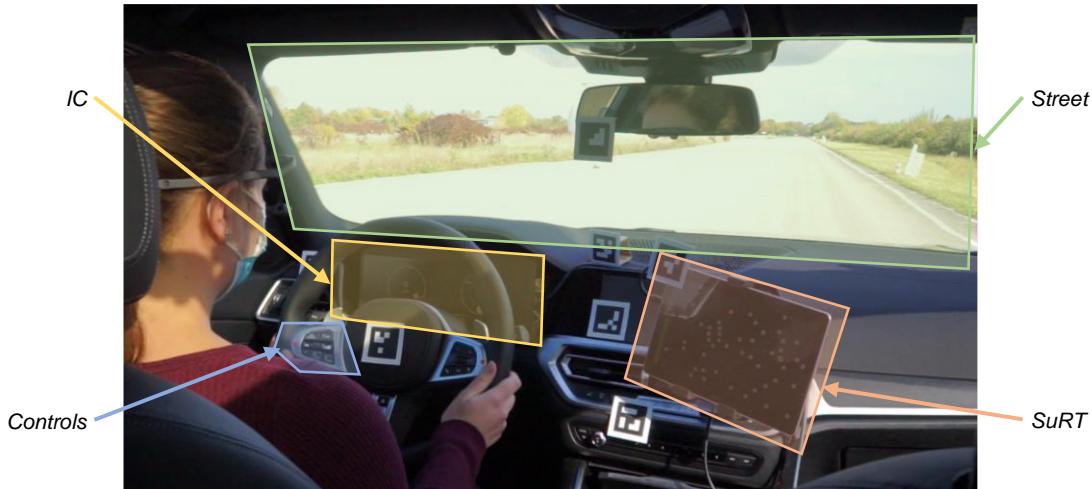


Figure 5.8 Visualization of the four AOIs: *Street*: the road environment (mainly windshield); *IC*: *Controls*: the control buttons for the HMI on the steering wheel; *SuRT*: the tablet installed in the center console for the NDRA.

5.6.2.3 Experimenter Rating

An experimenter rating is conducted based on the method reported by Forster, Hergeth, Naujoks, Beggiato, et al. (2019). After each test case, the experimenter rates the participants' interaction performance on a 5-point Likert scale ranging from (1) "no problem: quick processing" to (5) "help of experimenter: multiple problems; massive errors require restart task; help of experimenter necessary".

5.6.3 Self-Reported Data

5.6.3.1 Short Interviews

As described in Section 5.5, each test case requires a manual vehicle turning, and participants are instructed to stop the vehicle to a standstill. These short breaks are used to implore the participants' mode awareness, system understanding, and other incidents.

To assess the mode awareness, participants are requested to report the last active LoA, which is compared to the actual active LoA. Furthermore, the actual availability of LoAs is compared to the availability of LoAs reported by the participant. After RTIs or malfunctions reducing the number of available LoAs, participants are asked to report the indicated reasons for the change.

To inquire about the degree of system understanding, participants are requested to indicate whether they were allowed to take their hands off the steering wheel or answer e-mails during the last active LoA.

In test cases involving transitions, participants are requested to indicate whether they encountered problems during the interaction. If confirmed, participants are asked to describe these problems.

Furthermore, participants are encouraged to express feedback and general thoughts for a qualitative analysis supporting the interpretation of results.

5.6.3.2 Questionnaires

Once the test drive is completed, participants fill out a post-questionnaire that contains several standardized questionnaires (see Subsection 2.2.4 for more details) and is supplemented by additional questions.

Two standardized usability questionnaires are applied: the *SUS* (Brooke, 1996) and the *UMUX* (Finstad, 2010). In addition to usability questionnaires, data on affiliated constructs are collected. The *UEQ* (Laugwitz et al., 2008) is applied. Furthermore, two self-developed 1-item questions inquire about the trust and acceptance level of the participant via a 5-point Likert scale. The scale ranges between “1: strongly disagree” and “5: strongly agree”.

5.6.3.3 Final Interview

A final semi-structured interview intends to gain insights into participants’ perceptions and experiences that exceed the collected data. First, participants are reminded of the HMI components. Then, participants are requested to indicate and elaborate on aspects that—if any—they liked about the HMI. The following question refers to aspects that participants disliked. After that, participants are requested to think of improvement suggestions. Finally, participants are encouraged to express feedback and thoughts not yet covered regarding the HMI.

5.6.4 Interindividual Factors

After the experiment and the questionnaire response, two factors potentially influencing the performance and the self-reported assessments are inquired. Test courses with increased longitudinal accelerations and lateral inputs, such as urban scenarios, have shown to amplify motion sickness (Mourant et al., 2007). The test course of these experiments requires repeated maneuvers, including longitudinal accelerations and lateral inputs. In the first question, participants indicate how strenuous they rate the turning at the end of the test track. The second question refers to *Nausea* potentially triggered by the turning maneuvers. The questions are answered on a 5-point Likert scale ranging from “1: not at all strenuous/nauseous” to “5: very strenuous/nauseous”.

5.6.5 Overview and Embedding of the Dependent Variables

The experiments collect a high number of different metrics. Table 5.4 lists the collected observational and self-reported data, serving as an overview. Furthermore, the table links the differences between the HMI concepts by indicating the violated items in the *LC-HMI* expected to affect the respective metrics. In addition, the table indicates which facets of usability (ISO, 2018a) are addressed by the respective metric. The last column lists NHTSA minimum requirements if the respective metric provides information. This detailed overview allows an in-depth assessment of the usability and allows to conclude the validity and value of single metrics.

Table 5.4 List of the dependent variables and their linkage to the items of the guidelines by Naujoks, Wiedemann, et al. (2019) violated in the *LC-HMI*, the linkage to the ISO standard 9241-11 (ISO, 2018a), and the linkage to the NHTSA minimum requirements (NHTSA, 2017).

Dependent variable	Item of Naujoks, Wiedemann, et al. (2019) violated in <i>LC-HMI</i>	Facet of usability (ISO, 2018a) [†]	NHTSA minimum requirement (NHTSA, 2017) [‡]
Observational metrics			
Driving behavior			
<i>Observed LoA vs. instructed LoA</i>	all*	a	1, 2
<i>Control path of first activation</i>	3, 5, 7, 8, 9, 14, 15*	a, b	1, 2
<i>TOT after Rtl</i>	all*	b	4**, 5
<i>Take-over path after Rtl</i>	all*	(a), b	(2), 4**, 5
<i>Other observations</i>	n/a	n/a	n/a
Eye-tracking			
<i>Attention ratio during continuous rides in L0, L2, & L3</i>	all*	a, (b)	1, 2
<i>Glance allocation time to IC after Rtl</i>	5, 7, 8, 14, 18	b	4**, 5
<i>First glance duration on IC after Rtl</i>			
<i>Gaze behavior during Rtl</i>	all*	a, b	1, 2, 4**, 5
<i>Experimenter rating</i>	all*	a, b	all**, ***
Self-reported metrics			
Short interviews			
<i>Awareness of active LoA</i>	3, 5, 7, 8, 9, 14, 15*	a	2
<i>Awareness of change of available LoAs</i>	3, 5, 7, 8, 14, 18	a	3
<i>Awareness of reason for change of available LoAs</i>	5, 7, 8, 18	a	(3)
<i>System understanding: allowance of NDRA</i>	3, 7, 8, 9, 15*	a	2
<i>System understanding: allowance of H-off driving</i>			
<i>Reported problems during transitions</i>	all*	a, b, (c)	2, 4**, 5***
Questionnaires			
<i>SUS</i> (Brooke, 1996)	all*	all	all
<i>UMUX</i> (Finstad, 2010)			
<i>UEQ</i> (Laugwitz et al., 2008)			
<i>Trust</i>			
<i>Acceptance</i>			
<i>Final interview</i>	all*	all	all

* yellow-blue color-blind persons severely affected. ** TC12 only. *** TC10 and TC12 only

† a: effectiveness; b: efficiency; c: satisfaction.

‡ 1: functioning properly; 2: currently engaged in ADS mode; 3: currently “unavailable” for use;

4: experiencing a malfunction; 5: requesting control transition from the ADS to the operator.

Note. Adapted from Albers et al. (2021) with adjustments to reflect the final study design.

5.7 Data Analysis

The aggregation of driving data is conducted with the software *MATLAB* (The MathWorks Inc., 2022). All other data analysis uses *R*'s statistical software (R Core Team, 2022). The packages *tidyverse* (Wickham et al., 2019) and *reshape* (Wickham, 2007) are used for the data structuring and cleaning. Descriptive and inferential analyses are conducted with the packages *DescTools* (Signorell, 2023), *skimr* (Waring et al., 2022), *TOSTER* (Caldwell, 2022; Lakens, 2017), *car* (Fox & Weisberg, 2019), *afex* (Singmann et al., 2023), *rstatix* (Kassambra, 2023), *ordinal* (Christensen, 2022), and *compute.es* (Del Re, 2013). Data visualization is realized with the package *ggplot2* (Wickham, 2016).

The validation studies aim to conclude the relative and absolute validity of the factors testing environment and users' cultural background on different usability metrics. A two-step approach is chosen to derive these conclusions. In the first step, the data structure is modeled. It is examined whether the factors *Exp*, *HMI*, and their interaction *Exp:HMI* explain the data variance. Conclusions can be drawn on the relative validity. In the second step, an equivalence test is conducted to evaluate the data's likeness in the experimental settings. Under consideration of the descriptive data and the results of step 1, this approach allows to conclude absolute validity. In addition to the inferential statistical tests, descriptive and qualitative data analyses are conducted.

By common procedures, the significance level is set to alpha = 5% (Bortz, 2005, pp. 113–114). Multiple metrics, and consequently multiple tests, are conducted. This usually requires a correction for multiple testing, such as the Bonferroni correction, to avoid accepting hypotheses that are based on more than one test result (Bortz, 2005, pp. 271–272). Since each of the metrics is analyzed and reviewed separately before concluding on the effects of the testing environment on the single metrics, no correction is conducted in this thesis.

The two-step approach of the analysis and the subsequent analysis are described in more detail in the following.

5.7.1 Step 1: Modeling the Data Structure

The study design comprises two between-subject factors. The first factor, *Exp*, refers to the experiment determining the testing environment or the sample's cultural background. Three experiments are conducted, but only pairwise comparisons are relevant. The second factor, *HMI*, refers to the HMI concept consisting of two variants, the *LC-HMI* and the *HC-HMI*. These two factors are sufficient to model the data obtained once during the experiment, such as the questionnaires. However, it is incomplete for data obtained in several test cases, such as the experimenter ratings. A third factor, the test case TC, is introduced for these metrics. This factor is a within-subject factor.

The first step in the data analysis is building a model representing the data structure. For the two-factor case, the dependent variable is put in relation to the factors *Exp*, *HMI*, and their interaction *Exp:HMI*. The two-factor case applies to all questionnaires. Their answer format is the Likert scale. Hence, a test approach for ordinal data is chosen. Cumulative link models (CLM) are calculated for ordinal regression and tested for significance with the ANOVA method using the *R* package *ordinal* (Christensen, 2022).

For the three-factors case, a generalized linear mixed model (GLMM) is established following the theoretical instructions of Singmann and Kellen (2020). The factors *Exp*, *HMI*, and their interaction *Exp:HMI* are included as fixed effects. The test case is a random factor (1|TC). In addition, the participant is added as another random factor (1|TP) for the test case is a within-subject factor forming the following equation for the dependent variable (DV):

$$DV \sim Exp * HMI + (1|TC) + (1|TP)^6$$

Different underlying distributions and link functions are selected based on the data structure and the format; for example, a gaussian distribution is chosen for interval data, while binomial distributions are chosen for dichotomous response formats. The applied distributions are indicated in the results sections.

5.7.2 Step 2: Equivalence Test for the Factor Experiment

In a second step, equivalence tests are conducted following the procedures presented in Lakens (2017) and Lakens et al. (2018). The smallest effect size of interest is set to a medium effect size (Cohen's D = 0.5, Cohen, 1988). For the selected effect size, the sample sizes of the subsamples *HC-HMI* and *LC-HMI* in the respective experiments are too small for conducting two separate equivalence tests for the respective *HMI* subsamples. Therefore, the factor *HMI* is neglected, and only the factor *Exp* is considered in this analysis step. In cases of a significant interaction *Exp:HMI* in the first step, the neglection of the factor *HMI* in the second step impairs the meaningfulness of the equivalence test. The test cannot reflect the potential nullifying effects of the subsamples. Thus, the second step is omitted if the interaction *Exp:HMI* in the first step is significant. Regarding the data format and the fulfillment of requirements, the two one-sided t-tests (TOST) or the non-parametric alternative, the Wilcoxon TOST, are calculated with the *R* package *TOSTER* (Caldwell, 2022; Lakens, 2017).

5.7.3 Interpretation of Results

The first step considers the data structure and builds a model that checks whether the factors *Exp*, *HMI*, and their interaction *Exp:HMI* are significant predictors of the outcome. The second test, the equivalence test, considers the factor *Exp* only. The interpretation of the test results is presented in the following.

The first step informs whether factors *Exp*, *HMI*, or their interaction *Exp:HMI* are significant predictors. Test cases and participants are included as random factors in the three-factor case and do not need further consideration. Table 5.5 displays the possible outcomes and the interpretations. For interpreting the results, the factor *HMI* is of subordinate importance and, therefore, not included in the table. If none of the factors *Exp:HMI* (GLMM), *Exp* (GLMM), and *Exp* (TOST) are significant, the database is insufficient to conclude, and only tendencies may be interpreted. If only the factor *Exp* (GLMM) is significant, relative validity can be assumed, and absolute validity is rejected. If only the factor *Exp* (TOST) is significant, both relative and absolute validity can be assumed. If the factor *Exp* is significant in both analysis steps (GLMM & TOST), relative validity can be assumed, and absolute validity is rejected. However, the effect is smaller than the smallest effect size of interest (middle) in the

⁶ The term *Exp*HMI* includes the main effects and interaction of *Exp* and *HMI* and is equivalent to the longer term *Exp + HMI + Exp:HMI*.

equivalence test and bigger than required by the model to become a significant factor. This implies that potential effects are not great or clear (high variance) enough to be proven with the provided database. However, only tendencies may be interpreted. If the interaction *Exp:HMI* is significant, absolute validity can be rejected regardless of the significance of the other factors. Furthermore, relative validity may be given if the effect is in the same direction and of similar size. However, only tendencies may be interpreted.

Table 5.5 Overview of potential outcomes of the inferential analysis and their simplified interpretation.

Step 1: GLMM		Step 2: TOST		Interpretation
<i>Exp:HMI</i>	<i>Exp</i>	<i>Exp</i>	<i>Exp</i>	
n.s.	n.s.	n.s.	n.s.	The database is insufficient to draw conclusions.
n.s.	s.	n.s.	n.s.	Relative validity can be assumed. Absolute validity can be rejected.
n.s.	n.s.	s.	s.	Relative validity can be assumed. Absolute validity can be assumed.
n.s.	s.	s.	s.	Relative validity is likely to be given. Absolute validity is likely not to be given. The potential effect is not great; only tendencies may be interpreted.
s.	n.s. or s.	n/a	n/a	Relative validity may be given if the effect is in the same direction and of similar size. Absolute validity can be rejected.

Note. The interpretation considers significant (s.) and non-significant (n.s.) results of the GLMM (factors *Exp:HMI* & *Exp*) and the TOST (factor *Exp*).

6 Validation Study Exp_Testing-Environment: Effect of the Testing Environment on Metrics for Assessing Usability of HMIs for L3 ADS in User Studies

Two experiments build the empirical basis for the validation study *Exp_Testing-Environment*. The experiment *Sim_GER*⁷ is conducted in a static driving simulator at the Chair of Ergonomics in Garching in August and September 2020. The experiment *TT_GER*⁸ is conducted in an instrumented vehicle on a test track at the Universität der Bundeswehr in Neubiberg in September and October 2020. The experimental designs are approved by the Ethical Committee of the Technical University of Munich (403/20 S-KH & 520/20 S-EB). The experiments follow the study design presented in Chapter 5. Furthermore, the results are presented and discussed regarding the validity of driving simulators for usability testing of HMIs for L3 ADS.

6.1 Hypotheses

The validation study *Exp_Testing-Environment* strives to provide answers to research question RQ₂. Most studies on driving simulator validity presented in Section 2.3 could confirm relative validity for at least some metrics. However, some studies show that even within one experiment, and thus for the same driving simulator, not all metrics yielded consistent results (Mullen et al., 2011; Poisson et al., 2020; Wynne et al., 2019). Literature reviews conducted by Mullen et al. (2011) and Wynne et al. (2019) do not provide evidence for assumptions regarding specific metrics. Only a minority of the studies indicate absolute validity of driving simulators. Regarding the validity of driving simulators for usability research in HMIs for ADS, the two validation studies in the context of automated driving (Bellem et al., 2017; Poisson et al., 2020) are encouraging even though varying differences are identified for single metrics (Poisson et al., 2020) and for specific simulator settings (Bellem et al., 2017). Therefore, the following hypotheses for the validation study *Exp_Testing-Environment* are formulated:

RQ₂ Which effect has the testing environment on metrics for assessing the usability of HMIs for L3 ADS?

- H₁ The static driving simulator does not demonstrate absolute validity compared to the test track setting regarding metrics for assessing the usability of HMIs for L3 ADS.
- H₂ The static driving simulator demonstrates relative validity compared to the test track setting regarding metrics for assessing the usability of HMIs for L3 ADS.

Additionally, an effect of the HMI concept is expected. As described in Section 5.4, the HMI concepts serve as the artificial research subject. Introducing two HMI concepts varying in

⁷ The experiment was designed and conducted with the assistance of Niklas Mooshofer (2020) as part of his master's thesis.

⁸ The experiment was designed and conducted with the assistance of Julia Graefe (2021) as part of her master's thesis.

their compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019) allows for assessing relative validity, which refers to the agreement between the direction (and size) of effects. Furthermore, the sensitivity of metrics toward specific differences in HMI design may be assessed. The approach of variation between two HMI concepts is adapted from this study as presented in Forster et al. (2020a) and Forster et al. (2020b).

H₃ The concept *HC-HMI* receives higher usability evaluations than the concept *LC-HMI*.

6.2 Sample

The final sample size of *Sim_GER* is $n = 52$ ($n_{Sim_GER-HC} = 26$; $n_{Sim_GER-LC} = 26$). Four data sets in *Sim_GER-HC* and one in *Sim_GER-LC* have missing eye-tracking data because of technical problems. Additionally, one session in *Sim_GER-LC* is aborted due to technical problems with the data recording (excluded from the final sample).

The final sample size of *TT_GER* is $n = 61$ ($n_{TT_GER-HC} = 33$; $n_{TT_GER-LC} = 28$). In *TT_GER-HC*, one session is aborted due to problems with the eye tracker (excluded from the final sample), one data set is missing eye-tracking data completely, and two data sets have incomplete eye-tracking data due to technical problems. In *TT_GER-LC*, two sessions are aborted due to problems with the eye tracker or heavy rainfall (excluded from the final sample), one data set is missing eye-tracking data completely, and three data sets have incomplete eye-tracking data due to technical problems.

The summary of the descriptive analysis of the sociodemographic data is attached in Appendix II (Table 13.2). The proportion of female participants ranges between 35.71% (*TT_GER-LC*) and 42.42% (*TT_GER-HC*) among the subsamples, thereby fulfilling the required minimum of 30% females (see Section 5.2). None of the participants indicates a diverse gender or decides not to indicate a gender. The mean age across the four subsamples ranges between 37.43 (*TT_GER-LC*, $SD = 15.12$) and 41.92 (*Sim_GER-HC*, $SD = 16.9$). The minimum age of the participants is 18, and the maximum is 73 (*Sim_GER-LC* each). One participant of *Sim_GER-HC* (*Sim_GER-HC_{TP18}*) does not provide age information. As described in Section 5.2, a minimum of five participants in four different age groups (18-24; 25-39; 40-54; > 54; NHTSA, 2013) is aimed for during the recruitment. In *Sim_GER*, this aim is not met in two cases. *Sim_GER-HC* has only four participants in the age group 18-24, and *Sim_GER-LC* has only four participants in the age group > 54.

The summary of the descriptive analysis of the driving background is attached in Appendix II (Table 13.3). The driving frequency has a slightly higher variance in *TT_GER*. Participants of *TT_GER-LC* report more often to drive “less than once a month” or “never” (21.43%) compared to the other three subsamples (between *Sim_GER-LC*: 3.85% & *TT_GER-HC*: 9.09%). The distributions for the mileage, the experience with ADAS, and the frequency of using the systems show a high variance and are similar among the study samples. Between 34.62% (*Sim_GER-LC*) and 46.43% (*TT_GER-LC*) report to have no prior knowledge in the field of automated driving. Only single participants in *TT_GER* (*TT_GER-HC*: 9.09% & *TT_GER-LC*: 3.57%) indicate expert knowledge.

6.3 Results

In the following section, the results of all metrics are described. The inferential analysis of the data follows the process described in Section 5.7.

Due to organizational reasons, the two experiments are not entirely identical but differ in several aspects. While the test track infrastructure is identical in both experiments, pylons, delineator blades, and wires are used in *TT_GER* for marking event locations. These event locations are implemented in the simulated test track in *Sim_GER* without visible markers.

The protocol documents unusual behavior, unforeseen external events, or technical issues. Such events are referred to in the analysis of outliers. In *TT_GER*, vehicles, persons, and animals in areas near the test course could not be prevented entirely. In *Sim_GER*, no such potentially distracting events are possible. In *Sim_GER*, weather conditions are constant (bright, lightly cloudy). Instead, the weather conditions in *TT_GER* vary between and within the experiment sessions. For safety reasons, experiment sessions are canceled or aborted in cases of heavy rainfall. The distributions of weather and light conditions in the experiment sessions are attached in Appendix II (Table 13.1).

6.3.1 Observational Metrics

Observational data are collected for all 12 test cases (see Section 5.3). Some of the metrics presented in this section refer to specific test cases. An emphasis is put on the two RTIs (RtI_{20s} & RtI_{6s}) triggered during the test drives.

6.3.1.1 Driving Behavior

Observed LoA vs. Instructed LoA

The test case and the resulting instructions by the experimenter or system notifications determine the active LoA at every point of the test drive. If participants fail to adhere to the instructions or to react appropriately, deviances between the instructed LoA and the observed LoA may arise. Figure 6.1 displays the match between the instructed and observed LoAs for the four subsamples.

The figure and the binomial GLMM (Table 6.3) show that the number of deviances is significantly higher in the *LC-HMI* subsamples compared to the *HC-HMI* subsamples. In *TT_GER*, the number of matching LoAs is slightly lower than in *Sim_GER*, and the variance in *TT_GER-LC* appears to be greater compared to the other three subsamples. However, neither the factor *Exp* nor the interaction *Exp:HMI* is significant. The TOST does not confirm equivalence for the factor *Exp* (Table 6.3).

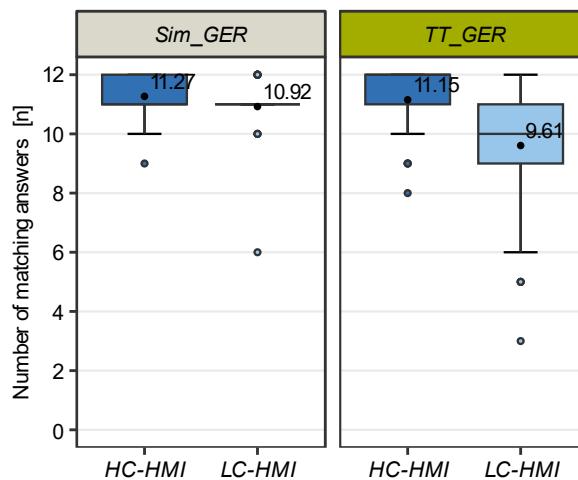


Figure 6.1 Boxplot diagram visualizing the results of the metric *Observed LoA vs. instructed LoA* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.3 for more statistics.

Control Path of First Activation

When participants are requested to activate L3 for the first time, they have no prior experience with the control logic of the HMI. Only participants are included who are in L0 at the experimenter's request to activate L3 and have not tried out the controls in prior test cases. Figure 13.1 in Appendix II displays the individual control paths of the first activation. The descriptive analysis is summarized in Table 6.1.

A higher proportion of participants of the *HC-HMI* subsamples succeed in activating L3 than participants of the *LC-HMI* subsamples. In addition, more participants of the *HC-HMI* subsamples manage to activate L3 with the minimum number of actions (*ACT → MOD*). The maximum number of actions ranges between five and six actions per subsample. Single participants repeatedly use the button *MOD* (no effect in L0) and stay in L0 or use the button *ACT* ($L0 \leftrightarrow L2$) repeatedly. One participant of *Sim_GER-HC* does not react at all.⁹ The difference between the *HC-HMI* and *LC-HMI* subsamples is more pronounced in *TT_GER* compared to *Sim_GER*.

⁹ According to the protocol, the participant (*Sim_GER-HC_{TP24}*) has overheard the experimenter's request to switch to another LoA.

Table 6.1 Descriptive analysis of the metric *Control path of first activation* for the study *Exp_Testing-Environment*.

Subsample (n)	Successful activation [% (n)]	Use of ideal path: ACT → MOD [% (n)]	Number of actions for participants not using the ideal path*	
			M (SD)	Max
Sim_GER-HC (23)	43.48% (10)	30.43% (7)	1.88 (1.36)	6
Sim_GER-LC (23)	34.78% (8)	26.09% (6)	1.76 (1.35)	6
TT_GER-HC (27)	62.96% (17)	37.04% (10)	2.35 (1.37)	5
TT_GER-LC (23)	34.78% (8)	17.39% (4)	2 (1.3)	5

Note. Only participants driving L0 at the start of the instruction are included. One participant of Sim_GER-HC does not react at all and is excluded from the analysis.

* The statistics include non-successful interactions.

TOT after Rtl

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 6.2 displays the TOT for the four subsamples and the two RtIs (Rtl_{20s} & Rtl_{6s}).

Figure 6.2 and Table 6.3 show significantly higher TOTs for the LC-HMI subsamples. Furthermore, a significant interaction shows a bigger difference between the HC-HMI and LC-HMI subsamples in TT_GER compared to Sim_GER. No TOST is conducted due to the potentially nullifying effect of the interaction *Exp:HMI* (see Section 5.7).

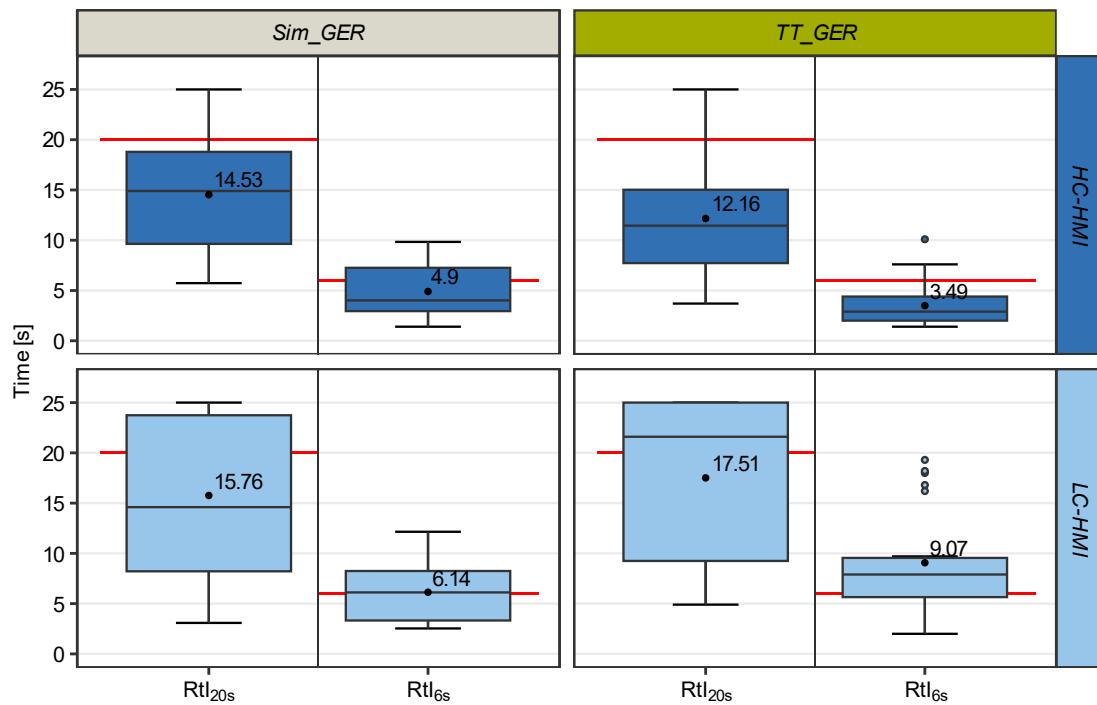


Figure 6.2 Boxplot diagram visualizing the results of the metric *TOT after Rtl* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.3 for more statistics. The red lines mark the start of the emergency braking for the respective Rtl.

Take-Over Path after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Only participants are included that drive in L3 when the respective RtI is triggered. Figure 13.2 and Figure 13.3 in Appendix II display the individual take-over paths for both RtI types (RtI_{20s} & RtI_{6s}). The descriptive analysis is summarized in Table 6.2.

Most participants conduct a successful transition to L0 with only one action. Participants of *TT_GER* tend to use the brake more often than participants of *Sim_GER*. The difference is more pronounced in the *HC-HMI* subsamples. Single participants in all four subsamples use the button *MOD* (repeatedly), which switches between L2 and L3 before switching to L0. In *Sim_GER-LC* and *TT_GER-HC*, single participants (temporarily) reactivate L2 after switching to L0. In contrast to no participants in the other three subsamples, in subsample *TT_GER-LC*, two participants in RtI_{20s} and three in RtI_{6s} do not take over at all.

Table 6.2 Descriptive analysis of the metric *Take-over path after RtI* for the study *Exp_Testing-Environment*.

RtI	Subsample (n)	Successful transition to L0 [% (n)]	1 action: Transition to L0 via ... [% (n)]:		> 1 action*: number of actions	
			Brake	ACT	M (SD)	Max
RtI _{20s}	<i>Sim_GER-HC</i> (26)	100% (26)	26.92% (7)	38.46% (10)	3.67 (1.49)	7
	<i>Sim_GER-LC</i> (26)	100% (26)	46.15% (12)	34.62% (9)	2.5 (1.26)	5
	<i>TT_GER-HC</i> (32)	100% (32)	53.13% (17)	28.13% (9)	3.5 (0.5)	4
	<i>TT_GER-LC</i> (26)	92.31% (24)	57.69% (15)	26.92% (7)	2.5 (0.5)	3
RtI _{6s}	<i>Sim_GER-HC</i> (26)	96.15% (25)	34.62% (9)	61.54% (16)	2 (n/a)**	n/a**
	<i>Sim_GER-LC</i> (26)	88.46% (23)	50% (13)	38.46% (10)	2.67 (0.47)	3
	<i>TT_GER-HC</i> (33)	96.97% (32)	72.73% (24)	24.24% (8)	3 (n/a)**	n/a**
	<i>TT_GER-LC</i> (26)	88.46% (23)	53.85% (14)	26.92% (7)	2 (0)	2

Note. Only participants driving L0 at the start of the instruction are included. None of the participants deactivates L3 through oversteering.

* Participants are excluded that do not react at all: *TT_GER-LC*: n = 2 (RtI_{20s}), n = 3 (RtI_{6s}).

** n = 1

Other Observations

If participants take their hands away from the steering wheel during L2 driving, they receive a H-off detection warning that comprises three stages depending on the H-off driving duration. The stages trigger notifications differing in their urgency (see Section 5.4).

In *Sim_GER*, only two participants of the *LC-HMI* subsample receive one H-off detection warning each (1x stage 1, 1x stage 2). In *TT_GER*, the number of H-off detection warnings is considerably higher. In *TT_GER-HC*, eight participants produce 13 H-off detection warnings (12x stage 1, 1x stage 2). Additionally, another participant of *TT_GER-HC* produces nine warnings (8x stage 1, 1x stage 2) alone. In *TT_GER-LC*, seven participants produce 13 H-off detection warnings (7x stage 1, 1x stage 2, 5x stage 3). Additionally, another participant of *TT_GER-LC* produces nine warnings (8x stage 1, 1x stage 2) alone.¹⁰

In test case TC6, participants drive in L2 and receive a notification that L3 driving is no longer available for optional activation due to a sensor error. The notification is only for

¹⁰ According to the protocol, the two participants (*TT_GER-HC_{TP8}* & *TT_GER-LC_{TP60}*) producing the nine h-off detection warnings each show no other remarkable behavior.

informational purposes and does not require an action by the participant. Nevertheless, in the *LC-HMI* subsamples, two participants each (*Sim_GER* & *TT_GER*) deactivate L2, and one of *TT_GER-HC*.

Summary

In both inferential statistical analyses, the factor *HMI* is significant. The interaction factor *Exp:HMI* is significant for the metric *TOT after RtI*. The difference between the *HC-HMI* and *LC-HMI* subsamples is more extreme in *TT_GER*. The descriptive and qualitative analysis of the driving behavior supports the findings of the inferential statistical tests: The performance scores are lower for the *LC-HMI* subsamples. There are only little differences between the experimental conditions. Where existent, performance scores are slightly lower for *TT_GER* than *Sim_GER* (e.g., number of H-off detection warnings). In the RtIs, differences in the take-over strategies could be observed, with participants of *TT_GER* tending to use the brake more often than participants of *Sim_GER*. Furthermore, differences between the *HC-HMI* and *LC-HMI* subsamples are more pronounced in *TT_GER* than in *Sim_GER*.

Table 6.3 Summary table of the descriptive and inferential results of the quantitative metrics of the driving behavior for the study *Exp_Testing-Environment*.

Metric	Subsample	Descriptive data					GLMM ^a							TOST ^b		
		n	M	SD	Min	Max	Distrib. (link)	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Obs. vs. instr. LOA	Sim_GER-HC	26	11.27	0.72	9	12	Binomial (logit)	Intercept	3.21	0.39	8.14	1 (1,356)	.051 .001** .107	1,327 .065		
	Sim_GER-LC	26	10.92	1.16	6	12		Exp	0.33	0.17	1.98					
	TT_GER-HC	33	11.15	1.12	8	12		HMI	0.54	0.17	3.24					
	TT_GER-LC	28	9.61	2.53	3	12		Exp:HMI	-0.27	0.17	-1.62					
TOT after Rtf ^{c,d}	Sim_GER-HC	26 26	14.53 4.9	5.5 2.45	5.73 1.4	25 9.83	Gaussian (identity)	Intercept	10.48	3.28	3.19	1 (216)	<.001*** .713	n/a n/a		
	Sim_GER-LC	26 26	15.76 6.14	7.88 2.92	3.08 2.53	25 12.15		Exp	-0.15	0.4	-0.37					
	TT_GER-HC	32 33	12.16 3.49	5.45 1.91	3.7 1.4	25 10.1		HMI	-1.7	0.4	-4.29					
	TT_GER-LC	24 23	17.51 9.07	7.95 4.94	4.9 2	25 19.3		Exp:HMI	1.08	0.4	2.73					.007*** ^e

^a GLMM formula: DV ~ Exp*HMI + (1 | TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c The descriptive data distinguishes between the test cases TC10 (left, Rtl_{20s}) and TC12 (right, Rtl_{6s}).

^d The maximum TOT is capped at 25 s.

^e Because of the significant interaction Exp:HMI, no TOST is calculated.

6.3.1.2 Eye-Tracking

Attention Ratio during Continuous Rides in L0, L2, & L3

Participants receive clear instructions on their responsibilities for the driving task for the different LoAs (see Subsection 5.5.2). The analysis of the attention ratio checks whether participants adhere to these instructions. Four AOIs are defined: *Street*, *IC*, *Controls*, and *SuRT* (see Subsection 5.6.2). During L0 and L2 driving, the visual attention should be focused on AOI *Street*. In L3 driving, participants are instructed to engage in the NDRA if the situation allows it. If participants adhere to the instructions, the attention ratio for the *SuRT* should be close to zero in L0 and L2 driving and considerably higher in L3 driving.

In Figure 6.3, the attention ratios for the *SuRT* are displayed for the three LoAs and the four subsamples. Attention ratios for all four AOIs are attached in Appendix II (Figure 13.4). Participants who do not drive in the instructed LoA during the specified test cases are excluded, leading to sample sizes differing within the subsamples.

In L0 and L2 driving, the mean attention ratio for the *SuRT* is below 5% for all four subsamples. The attention ratios are slightly higher in L2 driving compared to L0 driving. In L3 driving, the mean attention ratio for the *SuRT* ranges between 63.27% (*TT_GER-LC*, $SD = 28.41\%$) and 79.34% (*Sim_GER-HC*, $SD = 22.6\%$). In L3 driving, the variance in *TT_GER-HC* is considerably higher compared to the other three subsamples. In all four subsamples, single participants have an attention ratio for the *SuRT* of 30% or lower when driving in L3. The number of participants under this threshold is higher and the attention ratios lower for participants of the *LC-HMI* subsamples and participants of *Sim_GER* (*Sim_GER-HC*: $n = 2$ with $M = 15.21\%$ & $SD = 8.79\%$; *Sim_GER-LC*: $n = 5$ with $M = 1.12\%$ & $SD = 0.42\%$; *TT_GER-HC*: $n = 1$ with $AR = 26.69\%$; & *TT_GER-LC*: $n = 2$ with $M = 5.62\%$ & $SD = 5.62\%$). Table 6.4 presents the results of the GLMM and the TOST. Equivalence for the factor *Exp* is confirmed. None of the factors in the GLMM is significant.

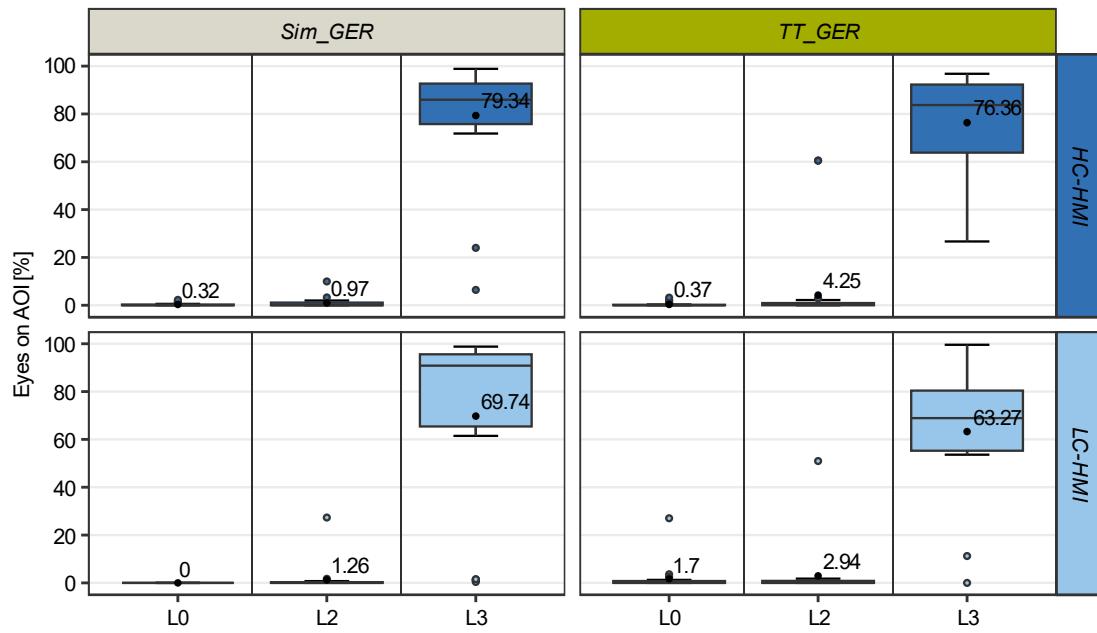


Figure 6.3 Boxplot diagram visualizing the results of the metric *Attention ratio during continuous rides in L0, L2, & L3* for the AOI *SuRT* for the study *Exp_Testing-Environment*. Note. The mean values are displayed as numbers in the figure. Refer to Table 6.4 for more statistics.

Gaze Behavior during RtI

This qualitative analysis refers to the two test cases with RtIs (see Section 5.3). Figure 6.4 displays the proportion of glances to the four AOIs at the start of the RtI for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}). Participants who do not drive in L3 at the scenario's beginning are excluded. Figure 13.5 and Figure 13.6 in Appendix II display the individual gaze paths between the start and the end of the RtIs. The end of an RtI is marked by the start of an emergency braking maneuver or the transition to L0.

At the start of the first RtI (RtI_{20s}), more than 80% of the participants in all four subsamples look at the *SuRT*. At the start of the second RtI (RtI_{6s}), participants of the *LC-HMI* subsamples tend to look at the *IC* more often and less often at the *SuRT* than at the start of the first RtI (RtI_{20s}). Participants of the *HC-HMI* subsamples do not show this effect.

Before the first glance at the *IC*, most participants look at the *SuRT* and *Street* in turns. In both RtIs, after the first glance at the *IC*, most participants look at the *IC* and other AOIs (mostly *Street*) in turns. At the end of both RtIs, no participant of the *HC-HMI* subsamples looks at the *SuRT*; most participants look at the *IC* instead. In the *LC-HMI* subsamples, two (RtI_{6s}: *Sim_GER-LC*) to six (RtI_{20s}: *Sim_GER-LC*) participants (still) look at the *SuRT*. The other participants of the *LC-HMI* subsamples mainly look at the *IC* at the end of the RtI. There are no prominent differences between the experiments.

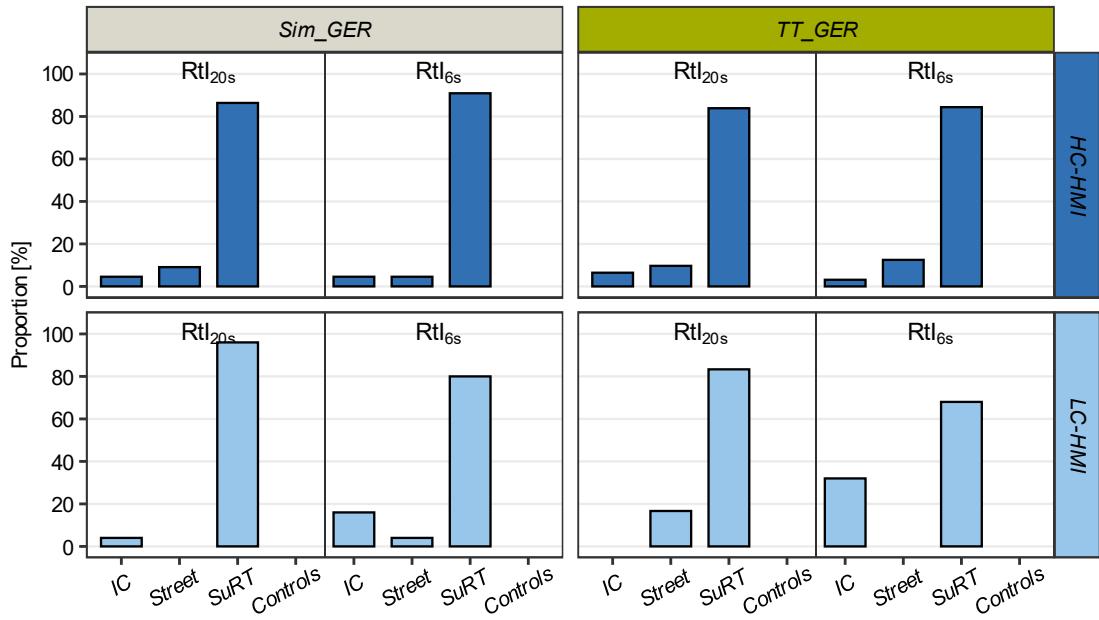


Figure 6.4 Bar chart visualizing the results of the metric *Glance allocation at start of RtI* for the study *Exp_Testing-Environment*.

Note. The sample sizes are as follows: *Sim_GER-HC*: $n = 22$ (RtI_{20s}), $n = 22$ (RtI_{6s}); *Sim_GER-LC*: $n = 25$ (RtI_{20s}), $n = 25$ (RtI_{6s}); *TT_GER-HC*: $n = 31$ (RtI_{20s}), $n = 32$ (RtI_{6s}); & *TT_GER-LC*: $n = 24$ (RtI_{20s}), $n = 25$ (RtI_{6s}).

Glance Allocation Time to IC after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 6.5 displays the glance allocation time to the *IC* for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}). This metric serves as an indicator of the salience of the RtI notifications. Participants who do not drive in L3 or that look at the *IC* already at the start of the respective RtI are excluded.

In all four subsamples, the glance allocation time to the *IC* is higher in the first RtI (RtI_{20s}) compared to the second RtI (RtI_{6s}). The mean glance allocation time is slightly higher for the *Sim_GER-HC* subsample compared to the *Sim_GER-LC* subsample in RtI_{20s}. In comparison, in the RtI_{6s} of *Sim_GER* and in both RtIs of *TT_GER*, the mean glance allocation time is slightly lower in the *HC-HMI* subsample than in the *LC-HMI* subsample. Table 6.4 presents the results of the GLMM and the TOST. Equivalence for the factor *Exp* is confirmed. None of the factors in the GLMM is significant.

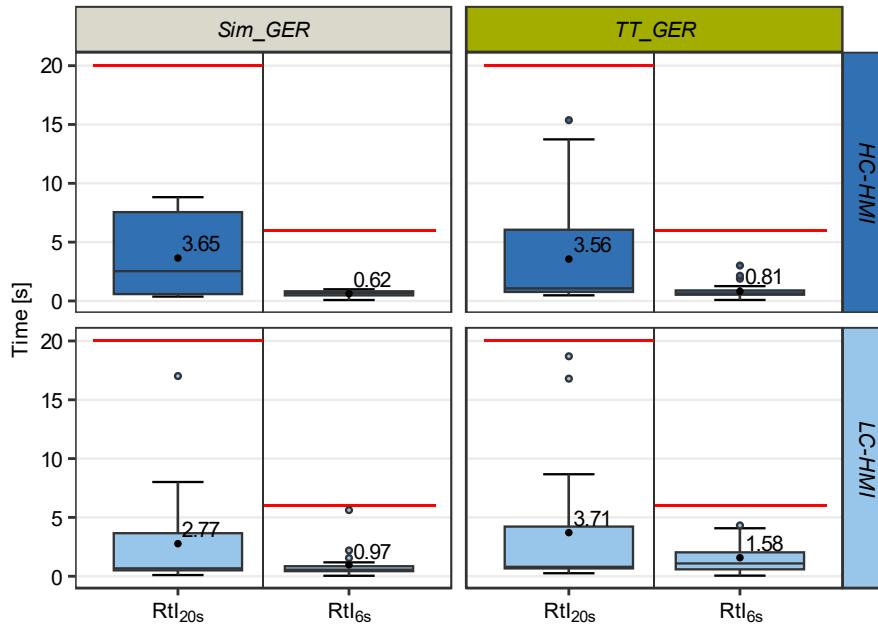


Figure 6.5 Boxplot diagram visualizing the results of the metric *Glance allocation time to IC after RtI* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.4 for more statistics. The red lines mark the start of the emergency braking for the respective RtI.

First Glance Duration on IC after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 6.6 displays the duration of the first glance to the *IC* for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}). This metric serves as an indicator for the presentation of information appropriate to the situation. Participants who do not drive in L3 or that look at the *IC* already at the start of the respective RtI are excluded.

In all four subsamples, the first glance duration to the *IC* is higher in the first RtI (RtI_{20s}) compared to the second RtI (RtI_{6s}). The first glance duration is slightly higher for the *LC-HMI* subsamples than the *HC-HMI* subsamples. There are no prominent differences between the experiments. Table 6.4 presents the results of the GLMM and the TOST. Equivalence for the factor *Exp* is confirmed. None of the factors in the GLMM is significant.

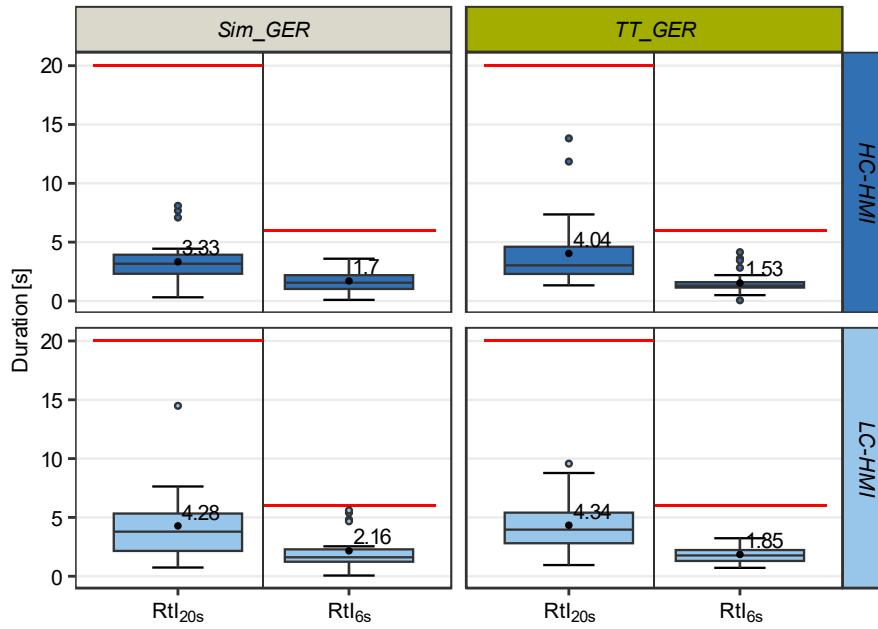


Figure 6.6 Boxplot diagram visualizing the results of the metric *First glance duration on IC* after *RtI* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.4 for more statistics. The red lines mark the start of the emergency braking for the respective *RtI*.

Summary

In all three inferential statistical analyses, equivalence for *Exp* is confirmed. In the GLMMs, none of the factors is significant. The descriptive analysis of gaze behavior shows differences between the HMI subsamples, for example, an increase of glances to the *IC* after the first *RtI* (*RtI*_{20s}) in the *LC-HMI* subsamples or the shorter glance allocation time to the *IC* in the second *RtI* (*RtI*_{6s}). In the gaze behavior analysis, several participants are excluded because they do not meet preconditions for the metric, for example, wrong LoA at the scenario start. The exclusion criteria mainly reduce the sample numbers of the *LC-HMI* subsamples. Furthermore, participants of *TT_GER* are affected more often than participants of *Sim_GER*.

Table 6.4 Summary table of the descriptive and inferential results of the quantitative metrics of the eye-tracking for the study *Exp_Testing-Environment*.

Metric	Subsample	Descriptive data					GLMM ^a (gaussian distribution with identity link function)						TOST ^b		
		n	M	SD	Min	Max	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Attention ratio during continuous rides in L0, L2, & L3: SuRT ^c	Sim_GER-HC	22	0.32	0.63	0	2.26	<i>Intercept</i>	89.98	7.11	12.64					
		22	0.97	2.19	0	9.96									
		22	79.34	22.6	6.42	98.86									
	Sim_GER-LC	25	0	0	0	0	<i>Exp</i>	1.11	0.98	1.14	1.29	1	.256	573	.015*
		25	1.26	5.45	0	27.34									
		25	69.74	36.81	0.4	98.8									
	TT_GER-HC	27	0.37	0.8	0	3.2	<i>HMI</i>	1.46	0.98	1.49	2.2	1 (283)	.238		
		32	4.25	14.77	0	60.52									
		27	76.36	18.9	26.69	96.79									
	TT_GER-LC	22	1.7	5.73	0	27.04	<i>Exp:HMI</i>	0.14	0.98	0.14	0.02	1	.888		
		20	2.94	11.32	0	50.95									
		14	63.27	28.41	0	99.57									
Glance allocation time to IC after Rtl ^c	Sim_GER-HC	21	3.65	3.37	0.37	8.82	<i>Intercept</i>	2.12	0.96	2.20					
		21	0.62	0.27	0.08	0.99									
	Sim_GER-LC	21	2.77	4.12	0.1	17.02	<i>Exp</i>	-0.16	0.24	-0.69	0.48	1 (172)	.488	386	<.001***
		19	0.97	1.23	0.04	5.62									
	TT_GER-HC	29	3.56	4.34	0.48	15.37	<i>HMI</i>	0.02	0.24	0.08	0.01	1 (172)	.934		
		31	0.81	0.6	0.09	3.01									
	TT_GER-LC	19	3.71	5.47	0.26	18.7	<i>Exp:HMI</i>	0.11	0.24	0.46	0.22	1	.643		
		11	1.58	1.46	0.05	4.32									
First glance duration on IC after Rtl ^c	Sim_GER-HC	21	3.33	2.06	0.31	8.09	<i>Intercept</i>	2.93	0.79	3.7					
		21	1.7	0.9	0.09	3.59									
	Sim_GER-LC	21	4.28	2.89	0.74	14.49	<i>Exp</i>	-0.07	0.16	-0.4	0.16	1 (172)	.690	969	.005**
		19	2.16	1.63	0.06	5.6									
	TT_GER-HC	29	4.04	2.82	1.33	13.82	<i>HMI</i>	-0.27	0.16	-1.63	2.61	1 (172)	.206		
		31	1.53	0.88	0.06	4.15									
	TT_GER-LC	19	4.34	2.23	0.95	9.57	<i>Exp:HMI</i>	-0.07	0.16	-0.43	0.19	1	.664		
		11	1.85	0.75	0.71	3.23									

^a GLMM formula: DV ~ *Exp***HMI* + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c The descriptive data distinguishes between the LoAs L0 (left), L2 (center), and L3 (right) or the test cases TC10 (left, Rtl_{20s}) and TC12 (right, Rtl_{6s}), respectively.

6.3.1.3 Experimenter Rating

After each test case, the experimenter rates the participants' interaction with the ADS. Figure 6.7 displays the participants' mean experimenter rating for the four subsamples. For simplicity reasons, the figure does not visualize the experimenter ratings in the 12 test cases but displays the distribution of the participants' mean experimenter ratings. A visualization of the mean experimenter ratings per test case is attached in Appendix II (Figure 13.7).

Figure 6.7 and Table 6.5 show significantly better mean experimenter ratings for the *HC-HMI* subsamples. Furthermore, a significant interaction shows a bigger difference between the *HC-HMI* subsamples and LC in *TT_GER* compared to *Sim_GER*. No TOST is conducted due to the potentially nullifying effect of the interaction *Exp:HMI* (see Section 5.7). The variance of the participants' mean experimenter ratings is exceptionally high for the subsample *TT_GER-LC*.

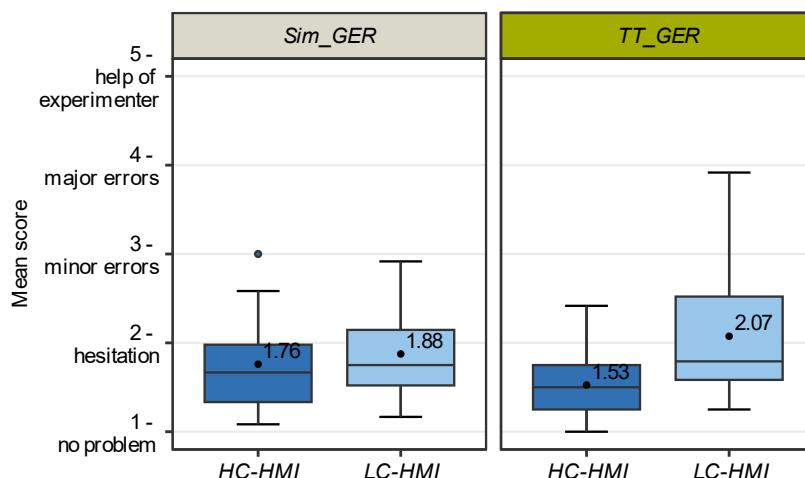


Figure 6.7 Boxplot diagram visualizing the results of the metric *Experimenter rating* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.5 for more statistics.

Table 6.5 Summary table of the descriptive and inferential results of the metric *Experimenter rating* for the study *Exp_Testing-Environment*.

Sub-sample	Descriptive data						GLMM ^a (gaussian distribution with logit link function)					
	n	M	SD	Med	Min	Max	Factor	Est.	SE	z	X ² (1, 113)	p
Sim_GER-HC	26	1.76	0.52	1.67	1.08	3	Intercept	0.51	0.09	5.69		
Sim_GER-LC	26	1.88	0.47	1.75	1.17	2.92	Exp	0.03	0.03	1.18	1.38	.240
TT_GER-HC	33	1.53	0.37	1.5	1	2.42	HMI	-0.09	0.03	-3.45	11.26	.001**
TT_GER-LC	28	2.07	0.74	1.79	1.25	3.92	Exp:HMI	0.06	0.03	2.28	5.06	.024 ^{ab}

^a GLMM formula: DV ~ *Exp***HMI* + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b Because of the significant interaction *Exp:HMI*, no TOST is calculated.

6.3.2 Self-Reported Metrics

6.3.2.1 Short Interviews

The short interviews refer to the participants' interaction with the ADS during the test drive. The set of questions refers to the mode awareness, the system understanding, and the report of interaction problems during transitions. Most metrics are calculated by assessing the ratio between correct and wrong replies. Only the metric *Reported Problems during transitions* is calculated as the number of reported problems.

Awareness of Active LoA

At the end of each test case, participants are requested to name the last active LoA. The reported LoA is compared to the observed active (not the instructed) LoA. Figure 6.8 displays the match between the reported and the observed LoAs for the two experiments and the respective HMI subsamples.

The two subsamples of *Sim_GER* and *TT_GER-HC* are very similar. The subsample *TT_GER-LC* shows considerably fewer matching answers and a higher variance. The binomial GLMM (Table 6.6) shows a significant influence of the factor *HMI*. Neither the factor *Exp* nor the interaction *Exp:HMI* is significant. The TOST confirms equivalence for the factor *Exp* (Table 6.6).

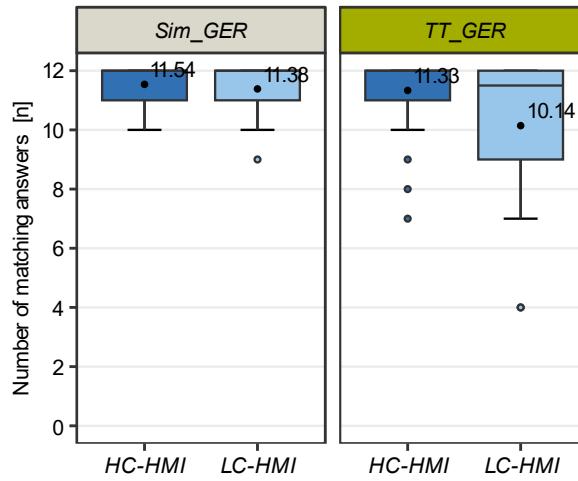


Figure 6.8 Boxplot diagram visualizing the results of the metric *Awareness of active LoA* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

Awareness of Change of Available LoAs

At the end of each test case, participants state whether a change in the available LoAs occurred. The reported change of availabilities is compared to the implemented change of availabilities. Figure 6.9 displays the match between the reported and the implemented changes of availabilities for the two experiments and the respective HMI subsamples.

The two HMI subsamples of *Sim_GER* are very similar. In *TT_GER*, the number of matching answers in both subsamples is slightly lower compared to *Sim_GER*. Furthermore, participants of *TT_GER-LC* show fewer matching answers and a higher variance than participants of *TT_GER-HC*. The binomial GLMM (Table 6.6) shows a significant influence of the factor *Exp*. Neither the factor *HMI* nor the interaction *Exp:HMI* are significant. The TOST does not confirm equivalence for the factor *Exp* (Table 6.6).

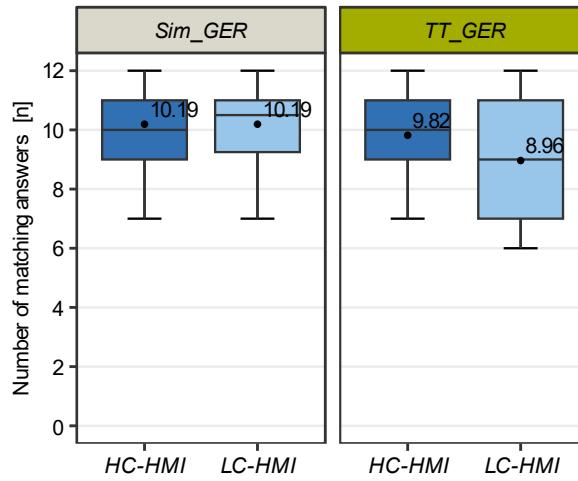


Figure 6.9 Boxplot diagram visualizing the results of the metric *Awareness of change of available LoAs* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

Awareness of Reason for Change of Available LoAs

In three test cases (see Section 5.3), a downward change of available LoAs occurs. In these test cases, the HMI provides reasons for the availability change, for example, a sensor error (see Section 5.4). Participants are requested to recall the reason for the change of availabilities. The number of matching answers between implemented and reported reasons for the change of availabilities is visualized in Figure 6.10.

In all four subsamples, the range is between zero and three matching answers. The variance is slightly lower in the subsample *Sim_GER-LC* compared to the other three subsamples. The mean value ranges between $M = 1.68$ (*TT_GER-LC*, $SD = 1.22$) and $M = 2.27$ (*Sim_GER-LC*, $SD = 0.83$). None of the factors in the binomial GLMM is significant.¹¹ The TOST does not confirm equivalence for the factor *Exp* (Table 6.6).

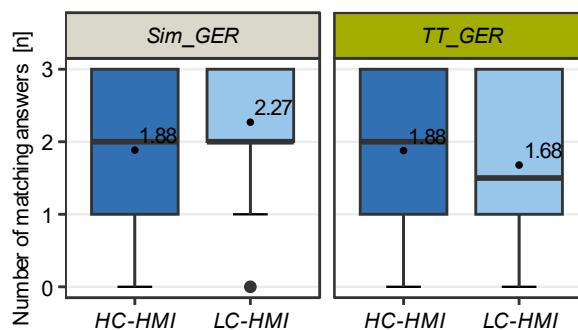


Figure 6.10 Boxplot diagram visualizing the results of the metric *Awareness of reason for change of available LoAs* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

¹¹ Due to convergence, the model does not contain the random factor (1|TC).

System Understanding: Allowance of NDRA

At the end of each test case, participants state whether it was allowed to engage in NDRA, such as writing e-mails while driving in the last active LoA. The reported allowance for this NDRA is compared to the observed LoA. Before the test drive, participants are instructed that only in L3 driving engagement in NDRA is allowed. Figure 6.11 displays the match between the reported allowance to engage in the NDRA and the observed LoA for the two experiments and the respective HMI subsamples.

The highest mean of matching answers and the lowest variance are shown by the subsample *Sim_GER-HC*. The lowest mean of matching answers and the highest variance are shown by the subsample *TT_GER-LC*. The subsamples of *Sim_GER* have slightly higher means and lower variances than the subsamples of *TT_GER*. The binomial GLMM (Table 6.6) shows a significant influence of the factor *Exp*. Additionally, the TOST confirms equivalence for the factor *Exp* (Table 6.6), implying that the potential effect of the factor *Exp* is small. Neither the factor *HMI* nor the interaction *Exp:HMI* are significant.

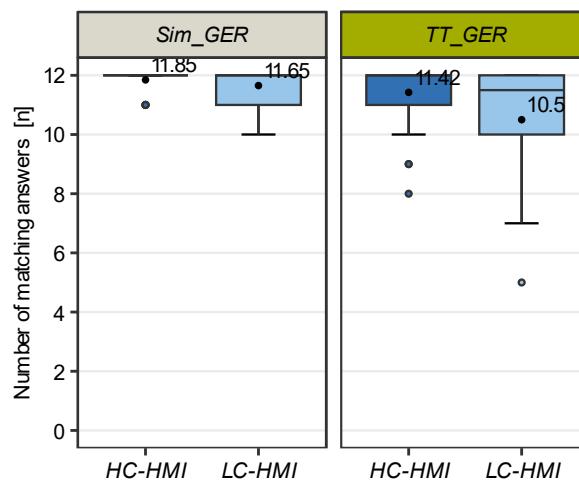


Figure 6.11 Boxplot diagram visualizing the results of the metric *System understanding: allowance of NDRA* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

System Understanding: Allowance of H-off Driving

At the end of each test case, participants state whether it was allowed to take their hands away from the steering wheel while driving in the last active LoA. The reported allowance for H-off driving is compared to the observed LoA. Before the test drive, participants are instructed that only in L3 driving it is allowed to drive H-off. Figure 6.12 displays the match between the reported allowance for H-off driving and the observed LoA for the two experiments and the respective HMI subsamples.

The highest mean of matching answers and the lowest variance are shown by the subsample *Sim_GER-HC*. The lowest mean of matching answers and the highest variance are shown by the subsample *TT_GER-LC*. The subsamples of *Sim_GER* have slightly higher means and lower variances than the subsamples of *TT_GER*. The binomial GLMM (Table 6.6)

shows a significant influence of the factor *Exp*. Additionally, the TOST confirms equivalence for the factor *Exp* (Table 6.6), suggesting that the potential effect of the factor *Exp* is small. Neither the factor *HMI* nor the interaction *Exp:HMI* are significant.

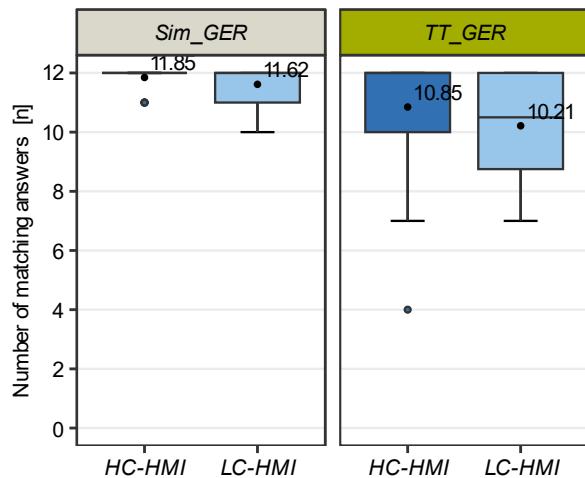


Figure 6.12 Boxplot diagram visualizing the results of the metric *System understanding: allowance of H-off driving* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

Reported Problems during Transitions

In six test cases (see Section 5.3), participants are instructed to switch between LoAs by the experimenter or the system. After each test case, participants state whether they encountered problems when switching between LoAs. The number of reported problems is visualized in Figure 6.13.

The maximum number of reported problems per participant is two (*Sim_GER-HC*) or three (*Sim_GER-LC*, *TT_GER-HC*, & *TT_GER-LC*). Between 39.39% (*TT_GER-HC*) and 65.38% (*Sim_GER-LC*) of the participants report no problem. The mean number of reported problems ranges between $M = 0.42$ (*Sim_GER-HC*, $SD = 0.58$) and $M = 0.97$ (*TT_GER-HC*, $SD = 0.98$). The variance of *TT_GER-HC* is slightly higher compared to the other three subsamples. None of the factors in the binomial GLMM is significant.¹² The TOST confirms equivalence for the factor *Exp* (Table 6.6).

¹² Due to convergence, the model does not contain the random factor (1|TC).

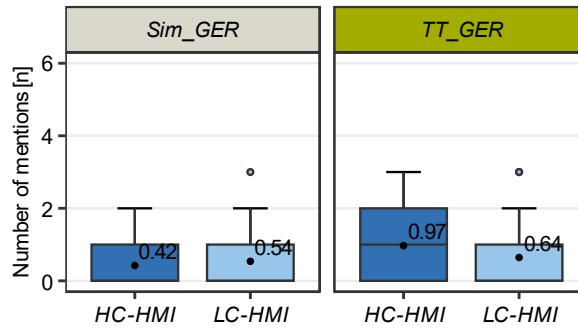


Figure 6.13 Boxplot diagram visualizing the results of the metric *Reported problems during transitions* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.6 for more statistics.

Most problems are reported after test cases with RtIs that required a time-critical action of the participants and after the test case with the first activation of the ADS: Between 27.27% (*Sim_GER-HC*) and 38.89% (*TT_GER-LC*) of the overall reported problems are reported after the first activation. After the first RtI (RtI_{20s}), between 11.11% (*TT_GER-LC*) and 63.64% (*Sim_GER-HC*) of the overall reported problems are reported.

Participants reporting a problem are requested to describe it. Between 11.11% (*TT_GER-LC*) and 18.75% (*TT_GER-HC*) of the reported problems do not refer to the LoA transition but the study procedure (e.g., *TT_GER-HC_{TP26}*: “*regarding performing no [problem], but [I] have responded incorrectly to announcement*”¹³) or more general problems (e.g., *TT_GER-LC_{TP72}*: “*have problems myself to maintain 30 km/h*”¹⁴).

Participants of all subsamples encounter similar problems with the control logic of the HMI controls. Statements referring to the control logic make up between 38.89% (*TT_GER-LC*) and 81.82% (*Sim_GER-HC*) of the reported problems. Participants make statements such as “*did not know what to do with the buttons. [...] maybe an info would be helpful what I can press, for example, by lighting up the buttons*”¹⁵ (*TT_GER-LC_{TP74}*), or “*didn't find the right button*”¹⁶ (*Sim_GER-LC_{TP51}*).

In *TT_GER-LC*, 50% of the reported problems refer to the participants’ uncertainty regarding the active LoA and its functions (e.g., *TT_GER-LC_{TP75}*: “*not sure what level I actually ended up in*”¹⁷). In contrast, only single participants of the other subsamples (*Sim_GER-HC*: 0%, *Sim_GER-LC*: 14.29%, & *TT_GER-HC*: 3.13%) report this issue.

Single participants in the subsamples *Sim_GER-LC* (14.29%) and *TT_GER-HC* (6.25%) describe that they have tried other take-over strategies during the RtIs such as putting their hands on the steering wheel which delayed the transition to L0 (e.g., *TT_GER-HC_{TP10}*: “*have*

¹³ Translated from German statement: „*vom Ausführen her nein, habe aber falsch auf Ansage reagiert*“.

¹⁴ Translated from German statement: „*habe selber Probleme, 30 km/h zu halten*“.

¹⁵ Translated from German statement: „*wusste nicht, was ich mit den Tasten machen muss. [...] hilfreich wären vielleicht Infos, was ich drücken kann, z.B. durch Aufleuchten der Tasten*“.

¹⁶ Translated from German statement: „*habe den richtigen Knopf nicht gefunden*“.

¹⁷ Translated from German statement: „*nicht sicher in welcher Stufe ich tatsächlich gelandet bin*“.

*first wanted to oversteer with hands on the steering wheel and gas, did not work, then turned off with button*¹⁸).

Summary

Equivalence is confirmed for three of the six inferential statistical tests. For the test *System understanding: allowance of NDRA*, the equivalence test and the GLMM are significant. This suggests that the effect is rather small; it is neither in the 90% confidence interval equivalence bounds nor includes zero in the 95% confidence intervals (see Section 5.7). Two additional metrics show significance for the factor *Exp* in the GLMM. These findings and the descriptive analysis show a tendency for participants of *Sim_GER* to interact better with the ADS than participants of *TT_GER*. The factor *HMI* is only significant for the metric *Awareness of active LoA*. The descriptive and qualitative analysis shows that participants of the *HC-HMI* subsamples have slightly better interaction scores than participants of the *LC-HMI* subsamples. The interaction factor *Exp:HMI* is not significant for any of the metrics.

¹⁸ Translated from German statement: „habe erst mit Händen am Lenkrad und Gas übersteuern wollen, hat nicht funktioniert, dann mit Button ausgeschaltet“.

Table 6.6 Summary table of the descriptive and inferential results of the short interviews for the study *Exp_Testing-Environment*.

Metric	Subsample	Descriptive data					GLMM ^a (binomial distribution with logit link function)						TOST ^b		
		n	M	SD	Min	Max	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Awareness of active LoA	Sim_GER-HC	26	11.54	0.71	10	12	Intercept	3.9	0.41	9.58				1,082	.001**
	Sim_GER-LC	26	11.38	0.85	9	12	Exp	0.4	0.23	1.77	3		.083		
	TT_GER-HC	33	11.33	1.27	7	12	HMI	0.48	0.23	2.09	4.42	1 (1,356)	.035*		
	TT_GER-LC	28	10.14	2.46	4	12	Exp:HMI	-0.28	0.23	-1.23	1.53		.217		
Awareness of change of available LoAs	Sim_GER-HC	26	10.19	1.3	7	12	Intercept	2.21	0.44	4.98				1,653	.652
	Sim_GER-LC	26	10.19	1.3	7	12	Exp	0.28	0.11	2.66	6.77		.009**		
	TT_GER-HC	33	9.82	1.55	7	12	HMI	0.14	0.11	1.33	1.71	1 (1,356)	.191		
	TT_GER-LC	28	8.96	2.05	6	12	Exp:HMI	-0.14	0.11	-1.33	1.7		.192		
Awareness of reason for change of available LoAs ^c	Sim_GER-HC	26	1.88	1.18	0	3	Intercept	0.96	0.25	3.84				1,328	.066
	Sim_GER-LC	26	2.27	0.83	0	3	Exp	0.33	0.23	1.42	2.01		.156		
	TT_GER-HC	33	1.88	1.17	0	3	HMI	-0.11	0.23	-0.47	0.23	1 (339)	.635		
	TT_GER-LC	28	1.68	1.22	0	3	Exp:HMI	-0.32	0.23	-1.39	1.96		.162		
System understanding: allowance of NDRA	Sim_GER-HC	26	11.85	0.37	11	12	Intercept	-1.22	0.72	-1.69				1,127	.003**
	Sim_GER-LC	26	11.62	0.57	10	12	Exp	-0.29	0.11	-2.56	6.02		.014*		
	TT_GER-HC	33	10.85	1.87	4	12	HMI	-0.05	0.11	-0.45	0.19	1 (1,356)	.666		
	TT_GER-LC	28	10.21	1.83	7	12	Exp:HMI	0.14	0.11	1.21	1.37		.242		
System understanding: allowance of H-off driving	Sim_GER-HC	26	11.85	0.37	11	12	Intercept	-0.95	0.71	-1.33				1,463	.233
	Sim_GER-LC	26	11.65	0.56	10	12	Exp	-0.48	0.11	-4.3	16.83		<.001***		
	TT_GER-HC	33	11.42	1.09	8	12	HMI	-0.06	0.11	-0.52	0.25	1 (1,356)	.617		
	TT_GER-LC	28	10.5	1.95	5	12	Exp:HMI	0.08	0.11	0.71	0.47		.493		
Reported problems during transitions ^c	Sim_GER-HC	26	0.42	0.58	0	2	Intercept	-2.38	0.2	-11.92				1,933	.020*
	Sim_GER-LC	26	0.54	0.86	0	3	Exp	-0.3	0.15	-1.91	3.7		.054		
	TT_GER-HC	33	0.97	0.98	0	3	HMI	0.07	0.15	0.44	0.19	1 (678)	.662		
	TT_GER-LC	28	0.64	0.95	0	3	Exp:HMI	-0.19	0.15	-1.25	1.57		.210		

^a GLMM formula: DV ~ Exp*HMI + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c Due to convergence, the GLMM omits the random factor (1|TC).

6.3.2.2 Questionnaires

After completing the test drive, participants fill out a post-questionnaire with standardized questionnaires and self-developed questions (see Section 5.6). The questions refer to the participants' experience with the HMI without specifying scenarios or functions.

SUS

The *SUS* score is calculated from 10 items and ranges between 0 and 100 (Brooke, 1996). Figure 6.14 displays the results for the two experiments and the respective HMI subsamples.

The distribution of the answers is similar among the four subsamples. The minimum score ranges between 40 (*Sim_GER-HC*) and 52.5 (*Sim_GER-LC*), and the maximum score ranges between 97.5 (*Sim_GER-HC & TT_GER-HC*) and 100 (*Sim_GER-LC & TT_GER-LC*). The mean score of the *SUS* ranges between $M = 77.02$ (*Sim_GER-LC*, $SD = 12.35$) and $M = 81.44$ (*Sim_GER-HC*, $SD = 16.16$). The ANOVA for the CLM results in no significant results for any factors. The TOST confirms equivalence for the factor *Exp* (Table 6.7).

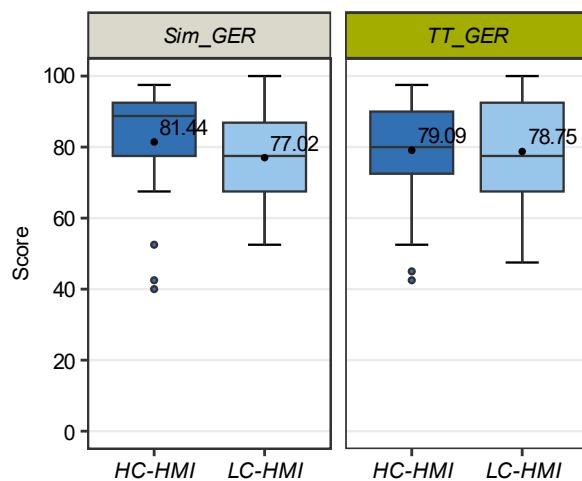


Figure 6.14 Boxplot diagram visualizing the results of the metric *SUS* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.7 for more statistics.

UMUX

The *UMUX* score is calculated from four items and ranges between 0 and 100 (Finstad, 2010). Figure 6.15 displays the results for the two experiments and the respective HMI subsamples.

The standard deviations differ distinctly between the subsamples of *Sim_GER* and range between $SD = 9.96$ (*Sim_GER-HC*) and $SD = 18.71$ (*Sim_GER-LC*). The mean *UMUX* scores are slightly higher for the *HC-HMI* subsamples than the *LC-HMI* subsamples. The *LC-HMI* subsamples differ distinctly between the experiments, with a considerably higher mean score in *TT_GER* ($M = 77.83$) than *Sim_GER* ($M = 68.11$). The ANOVA for the CLM results in

significant results for factors *Exp* and *HMI*. The TOST does not confirm equivalence for the factor *Exp* (Table 6.7).

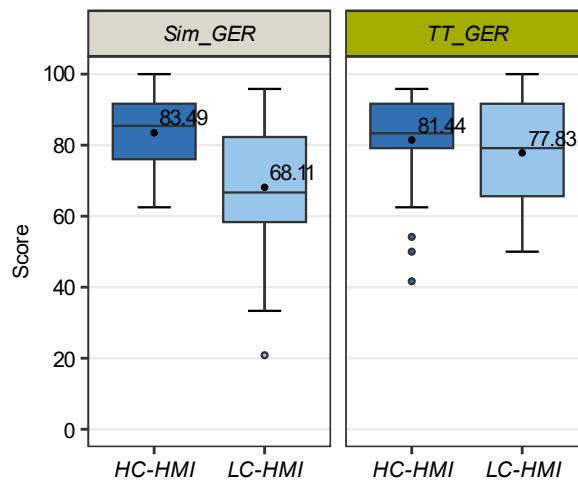


Figure 6.15 Boxplot diagram visualizing the results of the metric *UMUX* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.7 for more statistics.

UEQ

The *UEQ* comprises six subscales that result from four to six items each (Laugwitz et al., 2008). Figure 6.16 displays the results for all six subscales grouped by the four subsamples. The coloring in the figure marks the different evaluation categories (Schrepp, 2023): positive (green: > 0.8), neutral (yellow: between -0.8 and 0.8), and negative evaluation (red: < -0.8).

The descriptive data shows a tendency for higher *UEQ* mean scores (all six dimensions) and smaller standard deviations (four dimensions: *Attractiveness*, *Perspicuity*, *Efficiency*, & *Stimulation*) in the *TT_GER-LC* subsample compared to the *Sim_GER-LC* subsample. The observed trend of the standard deviations in the *TT_GER-HC* subsample is opposite to the trend of the *LC-HMI* subsamples: In *TT_GER-HC*, the standard deviations are slightly greater than the standard deviations in the *Sim_GER-HC*. For the mean scores of the *HC-HMI* subsamples, no trend is observable. The ANOVA for the CLM results in significant results only in two of the six subscales. In the subscale *Attractiveness*, the factor *Exp* is significant with higher means in *TT_GER* than *Sim_GER*. In the subscale *Dependability*, the factor *HMI* is significant with higher means in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples. The TOST confirms equivalence for the factor *Exp* for four subscales (*Perspicuity*, *Dependability*, *Stimulation* & *Novelty*; Table 6.7).

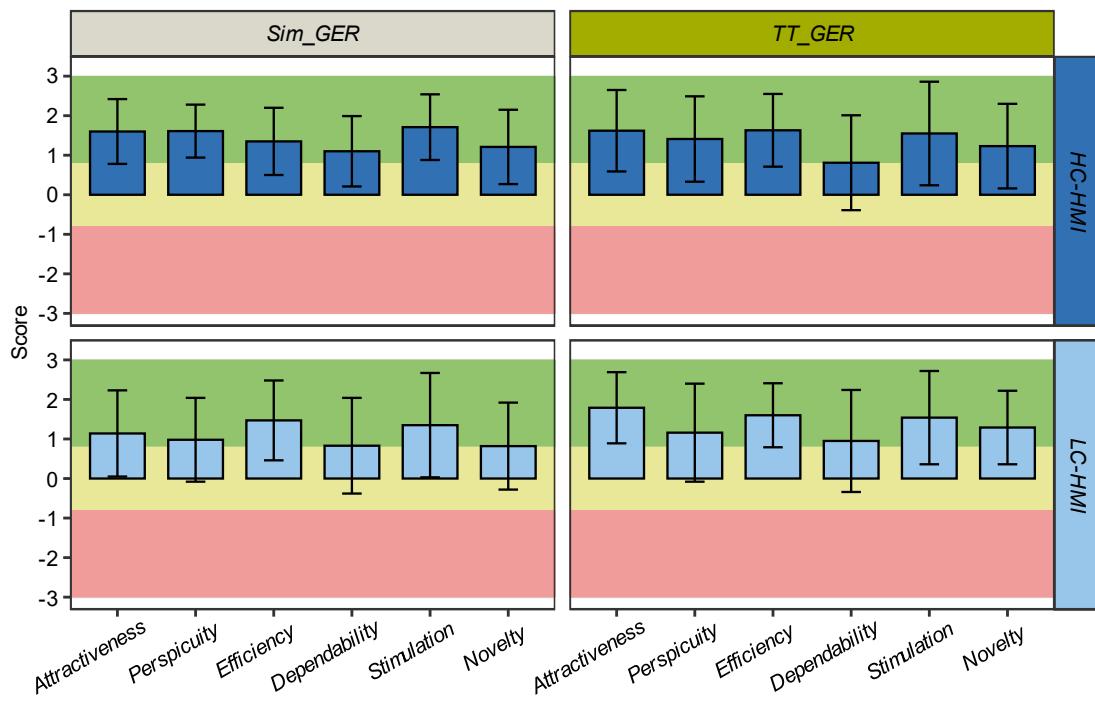


Figure 6.16 Bar chart visualizing the results of the metric *UEQ* with its six dimensions for the study *Exp_Testing-Environment*.

Note. The error bars display the *SD*. Refer to Table 6.7 for more statistics.

Trust

Trust is evaluated with a self-developed 1-item question: “I trusted the system I just used”. The response scale ranges between “1: strongly disagree” and “7: strongly agree”. Figure 6.17 displays the participants’ answers for the four subsamples.

The distribution of responses is similar in all four subsamples. Three of the four subsamples use the full range of the response scale (*Sim_GER-LC*, *TT_GER-HC* & *TT_GER-LC*). The variance among the participants is high in all four subsamples, with standard deviations between $SD = 0.83$ (*TT_GER-HC*) and $SD = 1.35$ (*TT_GER-LC*). The means of the subsamples range between $M = 3.35$ (*Sim_GER-LC*) and $M = 3.73$ (*Sim_GER-HC*). The ANOVA for the CLM results in no significant results for any factors. The TOST confirms equivalence for the factor *Exp* (Table 6.7).

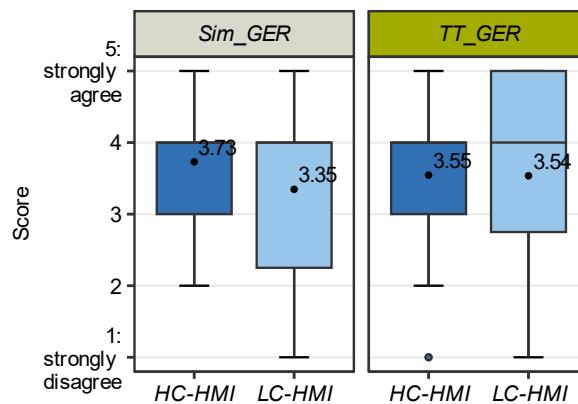


Figure 6.17 Boxplot diagram visualizing the results of the metric *Trust* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.7 for more statistics.

Acceptance

Acceptance is evaluated with a self-developed 1-item question: “If my car were fitted with a system like this, I’d use it when driving”. The response scale ranges between “1: strongly disagree” and “7: strongly agree”. Figure 6.18 displays the participants’ answers for the four subsamples.

The *LC-HMI* subsamples use the full range of the response scale (one participant each with the answer “1: strongly disagree”). The variance among the participants is high in all four subsamples, with standard deviations between $SD = 0.9$ (*TT_GER-HC*) and $SD = 1.21$ (*TT_GER-LC*). The means of the subsamples range between $M = 3.65$ (*Sim_GER-LC*) and $M = 4.14$ (*TT_GER-HC*) and show slightly higher means for *TT_GER* compared to *Sim_GER*. The ANOVA for the CLM results in significant results for the factor *Exp*. The TOST does not confirm equivalence for the factor *Exp* (Table 6.7).

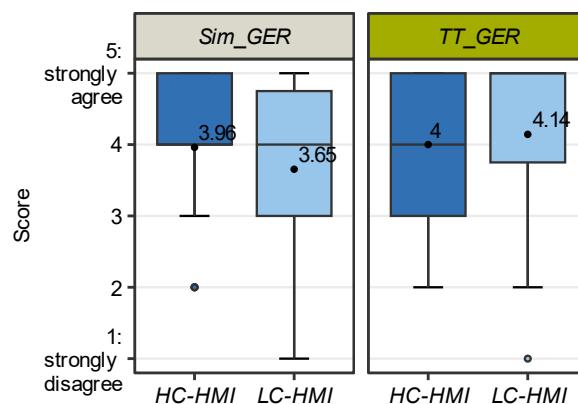


Figure 6.18 Boxplot diagram visualizing the results of the metric *Acceptance* for the study *Exp_Testing-Environment*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 6.7 for more statistics.

Summary

Overall, the results from the questionnaires are similar among all four subsamples. The equivalence of the factor *Exp* is confirmed for most of the metrics. Only the metrics *UMUX*, *Attractiveness* (subscale of the *UEQ*), and *Acceptance* produce significant results for the factor *Exp*. The factor *HMI* is not significant for most of the metrics. The interaction factor *Exp:HMI* is not significant for all of the metrics. The descriptive analysis shows a tendency for higher ratings in *TT_GER* compared to *Sim_GER* and higher ratings for the *HC-HMI* subsamples compared to the *LC-HMI* subsamples.

Table 6.7 Summary table of the descriptive and inferential results of the questionnaires for the study *Exp_Testing-Environment*.

Metric	Subsample	Descriptive data						CLM & ANOVA ^a						TOST ^b	
		n	M	SD	Med	Min	Max	Factor	Est.	SE	z	X ² (1, 113)	p	W _{max_p}	p _{max}
SUS	Sim_GER-HC	26	81.44	16.16	88.75	40	97.5	Exp:HMI	-0.03	0.23	-0.15	0.5	.481	1,205	.014*
	Sim_GER-LC	26	77.02	12.35	77.5	52.5	100		0.31	0.24	1.3	2.94	.086		
	TT_GER-HC	33	79.09	14.37	80	42.5	97.5		-0.39	0.33	-1.17	1.37	.241		
	TT_GER-LC	28	78.75	16.04	77.5	47.5	100		-0.54	0.33	-1.62	2.61	.106		
UMUX	Sim_GER-HC	26	83.49	9.96	85.42	62.5	100	Exp:HMI	0.32	0.24	1.38	4.02	.045*	1,763	.154
	Sim_GER-LC	26	68.11	18.71	66.67	20.83	95.83		0.74	0.24	3.04	10.29	.001**		
	TT_GER-HC	33	81.44	13.86	83.33	41.67	95.83		-0.54	0.33	-1.62	2.61	.106		
	TT_GER-LC	28	77.83	15.67	79.17	50	100		-0.44	0.33	-1.33	1.77	.185		
UEQ: Attractiveness	Sim_GER-HC	26	1.6	0.82	1.67	-0.17	3	Exp:HMI	0.45	0.24	1.89	4.88	.027*	1,668	.319
	Sim_GER-LC	26	1.14	1.09	1.5	-1.67	2.83		0.1	0.23	0.42	1.47	.226		
	TT_GER-HC	33	1.62	1.03	1.67	-1.5	3		-0.44	0.33	-1.33	1.77	.185		
	TT_GER-LC	28	1.79	0.9	1.92	0	3		-0.13	0.33	-0.38	0.15	.703		
UEQ: Perspicuity	Sim_GER-HC	26	1.71	0.83	2	-0.25	2.75	Exp:HMI	0.09	0.23	0.4	0.27	.601	2,054	.004**
	Sim_GER-LC	26	1.35	1.32	1.62	-1.75	3		0.17	0.23	0.71	0.59	.442		
	TT_GER-HC	33	1.55	1.31	1.75	-1.75	3		-0.13	0.33	-0.38	0.15	.703		
	TT_GER-LC	28	1.54	1.18	1.75	-0.75	3		-0.32	0.24	1.36	0.19	.665		
UEQ: Efficiency	Sim_GER-HC	26	1.35	0.85	1.25	-0.5	3	Exp:HMI	-0.06	0.23	-0.27	0.47	.495	1,735	.196
	Sim_GER-LC	26	1.47	1.01	1.5	-1.5	3		0.24	0.33	0.74	0.55	.460		
	TT_GER-HC	33	1.63	0.92	1.75	-1.75	3		-0.32	0.24	1.36	0.19	.665		
	TT_GER-LC	28	1.6	0.81	1.75	0	3		-0.32	0.24	1.36	0.19	.665		

Validation Study Exp_Testing-Environment: Effect of the Testing Environment on Metrics for Assessing Usability of HMIs for L3 ADS in User Studies

Metric	Subsample	Descriptive data					Factor	CLM & ANOVA ^a					TOST ^b		
		n	M	SD	Med	Min		Est.	SE	z	X ² (1, 113)	p	W _{max_p}	p _{max}	
UEQ: Dependability	Sim_GER-HC	26	1.61	0.67	1.5	0.25	2.75	Exp	0.09	0.23	0.39	0.64	.423	2,109	.001**
	Sim_GER-LC	26	0.98	1.06	1.25	-1.25	3		0.48	0.24	2.03	3.95	.047*		
	TT_GER-HC	33	1.41	1.08	1.5	-2	3		-0.26	0.33	-0.8	0.63	.426		
	TT_GER-LC	28	1.16	1.24	1.38	-1.25	3	Exp:HMI	-0.29	0.33	-0.87	0.76	.382		
UEQ: Stimulation	Sim_GER-HC	26	1.21	0.94	1	-1.25	3	Exp	0.34	0.24	1.44	2.53	.111	1,136	.005**
	Sim_GER-LC	26	0.82	1.1	0.75	-1.75	3		0.24	0.23	1.02	1.66	.198		
	TT_GER-HC	33	1.23	1.07	1.5	-1.75	3		-0.29	0.33	-0.87	0.76	.382		
	TT_GER-LC	28	1.29	0.93	1.12	-1	3	Exp:HMI	-0.29	0.33	-0.87	0.76	.382		
UEQ: Novelty	Sim_GER-HC	26	1.1	0.89	1	-0.25	3	Exp	-0.05	0.23	-0.2	0.06	.800	1,921	.027*
	Sim_GER-LC	26	0.83	1.21	1	-2.25	2.5		0.03	0.23	0.11	0.23	.634		
	TT_GER-HC	33	0.81	1.2	1	-1.75	2.75		-0.19	0.33	-0.57	0.33	.566		
	TT_GER-LC	28	0.95	1.29	0.88	-2.25	3	Exp:HMI	-0.38	0.35	-1.07	1.14	.285		
Trust	Sim_GER-HC	26	3.73	0.83	4	2	5	Exp	0.1	0.25	0.4	0.96	.327	1,996	.008**
	Sim_GER-LC	26	3.35	1.09	4	1	5		0.13	0.25	0.53	1.26	.262		
	TT_GER-HC	33	3.55	0.97	4	1	5		-0.38	0.35	-1.07	1.14	.285		
	TT_GER-LC	28	3.54	1.35	4	1	5	Exp:HMI	-0.52	0.35	-1.48	2.19	.139		
Acceptance	Sim_GER-HC	26	3.96	0.96	4	2	5	Exp	0.39	0.25	1.58	4.12	.042*	1,808	.098
	Sim_GER-LC	26	3.65	1.2	4	1	5		-0.08	0.25	-0.31	0.69	.408		
	TT_GER-HC	33	4	0.9	4	2	5		-0.52	0.35	-1.48	2.19	.139		
	TT_GER-LC	28	4.14	1.21	5	1	5	Exp:HMI	-0.52	0.35	-1.48	2.19	.139		

^a CLM formula: DV ~ Exp*HMI. A type 3 ANOVA is calculated with Wald chi-square tests.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5

6.3.2.3 Final Interview

At the end of the post-questionnaire participants are requested to reflect on the experienced HMI concept. Participants may praise or criticize components of the HMI concept or make improvement suggestions. An overview of the clustered replies of the participants of the four subsamples is attached in Appendix II (Figure 13.8-Figure 13.10).

Participants of all subsamples praise similar qualities of the HMI concepts. The focus is on the simple and clear design (between *TT_GER-HC*: 18.18% & *Sim_GER-HC/Sim_GER-LC*: 30.77%) as well as the easy handling (between *TT_GER-HC*: 15.15% & *TT_GER-LC*: 32.14%). In contrast to participants of the *LC-HMI* subsamples, participants of the *HC-HMI* subsamples additionally praise the usage of sounds (*Sim_GER-HC*: 26.92% & *TT_GER-HC*: 15.15%) and LED lights (*Sim_GER-HC*: 11.54% & *TT_GER-HC*: 12.12%). Single participants in the four subsamples mention the color selection and the icon design. A typical statement of the participants is: “*quick learning of operation and functionality, simple display design*”¹⁹ (*TT_GER-HC_{TP34}*).

Regarding criticism, there are differences between the experiments and between the HMI concepts. Participants of *TT_GER* appear to struggle more with the control logic (*TT_GER-HC*: 15.15% & *TT_GER-LC*: 10.71%) than participants of *Sim_GER* (no mentions). In contrast to the *HC-HMI* subsamples, participants of both *LC-HMI* subsamples criticize the missing sounds (*Sim_GER-LC*: 26.92% & *TT_GER-LC*: 21.43%) and the position of notifications (*Sim_GER-LC*: 3.85% & *TT_GER-LC*: 10.71%). In *TT_GER-HC*, 9.09% of the participants criticize the long notifications’ texts compared to none of the participants in the other three subsamples. Single participants in the four subsamples criticize the short display duration of notifications, the unclear labeling of the control buttons, or the readability of the text and icons. A typical statement of the participants is: “*no warning sound at takeover request. You had to randomly look at the display to see that you had to take over*”²⁰ (*Sim_GER-LC_{TP61}*).

When asked for improvement suggestions, there is no difference between the experiments. Participants of the *LC-HMI* subsamples express wishes for more sounds and other signals more often (*Sim_GER-LC*: 61.54% & *TT_GER-LC*: 39.29%) than participants of the *HC-HMI* subsamples (*Sim_GER-HC*: 11.54% & *TT_GER-HC*: 12.12%). Single participants of the *LC-HMI* subsamples mention the insufficient salience and urgency of the RTIs (*Sim_GER-LC*: 7.69% & *TT_GER-LC*: 10.71%). Other improvement suggestions listed by single participants are the improvement of the control logic and its labeling, the improvement of the color selection, or the notifications. A typical statement of the participants is: “*active usage of sounds when automated driving fails/driver must take over*”²¹ (*TT_GER-LC_{TP58}*).

In the interviews, differences in the HMI design are reflected in the different comments of the respective HMI subsamples. The only difference between the experiments is a more critical

¹⁹ Translated from German statement: „*Schnelle Erlernung der Bedienung und Funktionsweise, simple Display-Gestaltung*“.

²⁰ Translated from German statement: „*Kein Warnton bei Übernahmeforderung. Man musste zufällig auf die Anzeige schauen, um zu sehen, dass man übernehmen muss.*“

²¹ Translated from German statement: „*Aktiv Töne, wenn das automatisierte Fahren ausfällt, also der Fahrer übernehmen muss.*“

view toward the control logic and the length of text notifications in *TT_GER* compared to *Sim_GER*.

6.3.3 Interindividual Factors

Two 1-item questions are posed to assess whether there is a difference between the experiments regarding *Nausea* and *Effort*, respectively. The questions are answered on a 5-point Likert scale ranging from “1: not at all strenuous/nauseous” to “5: very strenuous/nauseous”. The analysis does not distinguish between participants of the *HC-HMI* and the *LC-HMI* subsamples. Table 6.8 summarizes the participants’ answers to both questions.

Table 6.8 Distribution of responses for the interindividual factors *Nausea* and *Effort* for the experiments *Sim_GER* and *TT_GER*.

Exp (n)	<i>Nausea [% (n)]</i>					<i>Effort [% (n)]</i>				
	1	2	3	4	5	1	2	3	4	5
<i>Sim_GER</i> (52)	48.08 (25)	19.23 (10)	15.38 (8)	13.46 (7)	3.85 (2)	57.69 (30)	13.46 (7)	5.77 (3)	13.46 (7)	9.62 (5)
<i>TT_GER</i> (61)	98.36 (60)	1.64 (1)	0 (0)	0 (0)	0 (0)	88.52 (54)	9.84 (6)	0 (0)	1.64 (1)	0 (0)

Note. The scale ranges from “1: not at all nauseous/strenuous” to “5: very nauseous/strenuous”.

At the end of the final questionnaire, participants rate whether the turning after every test case has made them nauseous.

The mean score in *Sim_GER* ($M = 2.06$, $SD = 1.24$) is considerably higher than in *TT_GER* ($M = 1.02$, $SD = 0.13$). A Wilcoxon rank sum test for difference shows a significant difference between *Sim_GER* ($Med = 2$) and *TT_GER* ($Med = 1$) with $p < .001$ ($W = 2,392$). The Wilcoxon rank sum test for equivalence is not significant, with $p = .627$ ($W = 1,637$).

At the end of the final questionnaire, participants rate whether the turning after every test case has been strenuous for them.

The mean score in *Sim_GER* ($M = 2.04$, $SD = 1.44$) is higher than in *TT_GER* ($M = 1.15$, $SD = 0.48$). A Wilcoxon rank sum test for difference shows a significant difference between *Sim_GER* ($Med = 1$) and *TT_GER* ($Med = 1$) with $p < .001$ ($W = 2,117.5$). The Wilcoxon rank sum test for equivalence is also significant, with $p = .031$ ($W = 1,283$) suggesting a rather small effect (see Section 5.7).

6.4 Discussion

This section starts with a summary of the results leading to answering the hypotheses. Afterward, limitations and other observations on the experimental design are reflected. The section closes with the conclusion of the study’s results.

6.4.1 Summary of Results

The observational data comprise driving behavior, eye-tracking data, and experimenter ratings. Only data points that fulfilled the requirements could be included, for example, the

correct LoA at the scenario start or not looking at the *IC* during the start of an RtI. Thus, the database for driving behavior and eye-tracking is reduced, leading to a lower statistical power. More data points in the *LC-HMI* subsamples are excluded than in the *HC-HMI* subsamples.

The general driving behavior and the driving behavior in specific situations (the first activation & during the RtIs) are assessed. The inferential and descriptive analyses show that the driving performance is considerably better in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples. Overall, differences between the experimental conditions are only little and non-significant in the inferential tests. Where differences exist, the driving performance is lower in *TT_GER* than in *Sim_GER*. Furthermore, the differences between the HMI subsamples are more pronounced in *TT_GER* than *Sim_GER*, resulting in a significant interaction *Exp:HMI* in the metric *TOT after RtI*. In the RtIs, differences in the take-over strategies could be observed, with participants of *TT_GER* tending to use the brake more often than participants of *Sim_GER*.

The eye-tracking data are assessed in test cases with continuous rides in L0, L2, and L3 and during RtIs. None of the factors in the GLMMs is significant, implying that neither the factors *Exp*, *HMI*, nor their interaction *Exp:HMI* show significant differences. Furthermore, equivalence between the experiments is confirmed for all three TOSTs. The descriptive analysis of the metrics shows that there are differences between the HMI subsamples, for example, an increase of glances to the *IC* after the first RtI (*RtI_{20s}*) in the *LC-HMI* subsamples or the shorter glance allocation time to the *IC* in the second RtI (*RtI_{6s}*). The difference between the HMI subsamples is more pronounced in *TT_GER* compared to *Sim_GER*.

The *Experimenter rating* shows a significant difference between the HMI subsamples, which is more pronounced in *TT_GER* than *Sim_GER*, resulting in a significant interaction *Exp:HMI*. The variance in the subsample *TT_GER-LC* is considerably higher than the variances of the other three subsamples.

The descriptive results of the short interviews imply a tendency for better performance scores in *Sim_GER* and the *HC-HMI* subsamples, respectively. The six inferential statistical tests support this observation: three of the metrics confirm equivalence, and half the metrics show a significant influence of the factor *Exp* (the metric *System understanding: allowance of NDRA* is significant in the equivalence test & the binomial GLMM). The factor *HMI* is significant only for the metric *Awareness of active LoA*. There are no interactions.

The questionnaires show a rather high variance within the four subsamples. The results themselves appear to be similar among all subsamples. The equivalence test is significant for six of the ten metrics, and the factor *Exp* is significant for only three of the metrics in the ANOVAs of the CLM. The factor *HMI* is significant only twice. Descriptive analysis suggests a tendency for higher ratings in *TT_GER* compared to *Sim_GER*.

In the *Final interview*, the differences in the HMI design are reflected in the participants' comments differing between the respective subsamples, for example, criticizing the missing sounds and overall salience of notifications in the *LC-HMI* concept. Participants of *TT_GER* report having problems with the control logic and the length of the text notifications considerably more often than participants of *Sim_GER*. Otherwise, the participants' answers

are very informative and valuable for future development efforts but do not differ much between the four subsamples.

6.4.2 Discussion of Hypotheses

Two hypotheses regarding the relative and absolute validity of driving simulators are formulated.

- H₁ The static driving simulator does not demonstrate absolute validity compared to the test track setting regarding metrics for assessing the usability of HMIs for L3 ADS.
- H₂ The static driving simulator demonstrates relative validity compared to the test track setting regarding metrics for assessing the usability of HMIs for L3 ADS.

Equivalence is confirmed for most metrics, and no differences are found. However, few performance metrics show lower performance scores for *TT_GER* compared to *Sim_GER*, while some questionnaires show higher ratings for *TT_GER* compared to *Sim_GER*. Since several usability metrics do not indicate absolute validity, hypothesis H₁ is confirmed.

Regarding relative validity, the HMI's effect and direction are essential. Based on the inferential and descriptive analysis, the HMI predominantly affects observational metrics. Self-reported metrics are less affected. In both experiments, the same effects are found differing in their magnitude only in a few of the metrics. Hypothesis H₂ is confirmed.

It should be noted that single metrics, such as take-over strategies in RTIs, yield different results in the experiments, illustrating the limits of a general conclusion on the validity of driving simulators. This study's results align with previous driving simulator validation studies (see reviews of Mullen et al., 2011; Wynne et al., 2019), indicating that driving simulators mostly demonstrate relative but no absolute validity. It is a remarkable tendency that in *TT_GER*, participants show lower performance ratings and report more problems in the final interviews while giving higher ratings for the HMI compared to participants of *Sim_GER*. The findings suggest that the more complex setting of a test track experiment (possibly due to animals, persons, and objects in the surroundings; differing light & weather conditions; or a more complex handling of the instrumented vehicle compared to the simulator mock-up) increases the overall workload (see Purucker et al., 2018). The more complex setting leads to lower performance scores in the experiment *TT_GER*. The higher self-reported ratings in *TT_GER* can be attributed to the general enthusiasm for experiencing an innovative technology "in the real world".

A third hypothesis is formulated on differences in the usability assessment between the HMI concepts:

- H₃ The concept *HC-HMI* receives higher usability evaluations than the concept *LC-HMI*.

In several metrics, differences between the HMI concepts demonstrate a higher usability of the *HC-HMI* than the *LC-HMI*. More of the observational metrics than the self-reported metrics show an apparent effect. Overall, the effect is smaller than expected after the unambiguous heuristic expert evaluation conducted in the HMI development process (see

Subsection 5.4.3). Following the framework of Bengler et al. (2020), the input channel and the dialog logic are identical in both concepts. The output channel differs in the number of included modalities and the design of the visual output channel. Furthermore, the information content is identical in both concepts. The two-factor theory of Herzberg et al. (1967) may be transferred to this observation to illustrate the limited variance of the two concepts. The covering of the information demand takes the role of hygiene factors. According to Herzberg et al. (1967, pp. 113–114), the absence of hygiene factors causes dissatisfaction, but the presence of hygiene factors does not increase satisfaction. The presentation of the information and the HMI's overall design take the motivators' role. These factors increase the satisfaction (Herzberg et al., 1967, pp. 113–114). The two HMI concepts do not differ in the hygiene factors, and the variance in the motivators is limited. Thus, the observed results of less clear differences between the HMI concepts may be explained. Considering this limitation, hypothesis H₃ is confirmed.

6.4.3 Discussion of Limitations and Other Observations

The study compares one static driving simulator to one test track study. Therefore, the transferability of the results to other driving simulators and, more importantly, naturalistic driving situations in the field is limited. Past research has shown that changing specific settings of the driving simulator, such as the motion cues, may severely affect both relative and absolute validity (Bellem et al., 2017). Furthermore, the study setting is simple with low speeds, straight roads without obstacles or interactions with other road users, and breaks between test cases, limiting the generalizability of the study's findings.

The smaller or non-present effects of differences between the HMI concepts aggravate the derivation of statements regarding relative validity, which is concerned with the direction and magnitude of effects. The statistical power is further lowered due to limitations in the data availability: due to crashes, technical problems, data quality issues, and exclusions (e.g., wrong LoA at scenario start), the number of data points is considerably reduced in the observational metrics, primarily affecting the *LC-HMI* subsamples. This issue is partly compensated by the descriptive data analysis stressing the importance of different data types and sources.

The database needs to be more sufficient to include confounding factors present in *TT_GER* (e.g., extreme weather conditions such as glare or heavy rain) in the statistical models. However, these factors are considered in inspecting outliers and are expected to increase the participants' overall workload.

Equivalence tests could not be conducted with subsamples but with total samples (*Sim_GER* vs. *TT_GER*), only ignoring the variance through different HMI concepts. Tests with subsamples, for example, *Sim_GER-LC* versus *TT_GER-LC*, could not be conducted due to the small sample sizes limiting the statistical power.

The effort of turning the vehicle around at the end of a test drive is rated slightly higher in *Sim_GER* than *TT_GER*. Since the difference and equivalence tests are statistically significant, the effect is rather small and may be neglected. *Nausea*, however, is more prevalent among participants of *Sim_GER*. Only one participant of *TT_GER* reports to feel a little nauseous. The effect is highly significant and implies that *Nausea* might be a potential influencing factor.

The metrics *Experimenter rating* and the *Final interview* could be affected by the experimenter who is aware of the *HMI* and *Exp* condition of the respective participants. To a

lesser magnitude, the participants' behavior and, therefore, other metrics could also be unconsciously affected by the experimenter's knowledge (e.g., Rosenthal effect, see Bortz & Döring, 2006, pp. 82–83) of the experimental conditions.

The experiments are conducted in different areas near Munich (*Sim_GER*: TUM campus in Garching vs. *TT_GER*: campus of the Universität der Bundeswehr in Neubiberg). The samples might originate from slightly differing populations. Due to similar sample characteristics regarding sociodemographic features and driving behavior of the subsamples, a potential effect is considered negligible.

Regarding the study procedure, participants report feeling exhausted due to the duration of the experiment and the extensive instructions. At the same time, participants note that the study setting is too simple because of low speeds and missing obstacles, interactions with other road users, or curves. Several participants expressed their excitement for automated driving before, during, and after the experiment (and their privilege to be part of the development process by partaking in this study). Participants even answer questions referring to the HMI design with general statements attributed to automated driving. Participants are not able to separate these feelings of excitement from the interaction with an HMI and provide generally high/positive ratings. Participants appear to tolerate flaws and problems (to some degree) when the research subject is obviously in a prototype state. The overall scores in the standardized questionnaires are high for both HMI concepts. In the *Final interview*, participants mention different aspects in their critique, but their overall rating is similar and positive for both HMI concepts. Another aspect is that several times, participants activate the wrong LoA without noticing their mistake. Because of missing feedback from the system or the experimenter, they could not consider these problems in the ratings but refer to their self-estimated performance when evaluating the HMI. Mistakes may be left unconsidered if no feedback is provided (Drew et al., 2018). No clear decision could be made for overall satisfaction in the between-subject experiments. If participants had compared the two HMI concepts, the satisfaction ratings may have shown apparent differences. In planning the experimental design, a within-subject design was excluded because of learning effects ruling out testing naïve participants. Furthermore, the study duration of a between-subject design is considerably shorter.

The observations described in this subsection will be used to derive recommendations for the experimental method in Chapter 10.

6.4.4 Conclusion

The validation study *Exp_Testing-Environment* confirms relative validity for driving simulators for usability assessments of HMIs for L3 ADS (H_2 confirmed). While several metrics show absolute validity, overall absolute validity is rejected (H_1 confirmed). Furthermore, the study provides valuable insights into the study design for usability assessments of HMIs for L3 ADS. The study's results confirm differences between the HMI concepts (H_3 confirmed). However, limitations of specific metrics detecting differences in HMI concepts are identified.

This study may draw a practical conclusion: Driving simulators are deemed a valid tool to assess the usability of HMIs for L3 ADS. Problems with HMI concepts that arise in driving simulator experiments will likely be more pronounced in more complex environments such as

test track or field studies. Therefore, it is recommended to test and refine HMI concepts in risk-free and resource-efficient driving simulator experiments before testing them in the field.

7 Validation Study Exp_Culture: Effect of the Users' Cultural Background on Metrics for Assessing Usability of HMIs for L3 ADS in User Studies

Two experiments build the empirical basis for the validation study *Exp_Culture*. The data of the experiment *TT_GER* (see Chapter 6) is reused for this validation study. This experiment²² is conducted in an instrumented vehicle on a test track at the Universität der Bundeswehr in Neubiberg in September and October 2020. The participant sample consists of Germans. The experiment *TT_USA*²³ is conducted in an instrumented vehicle on a test track at the BMW Driving Academy in Maisach in July and August 2021. The participant sample of the experiment *TT_USA* consists of U.S.-Americans currently residing in Germany. The experimental designs are approved by the Ethical Committee of the Technical University of Munich (520/20 S-EB & 394/21 S-KH). The experiments follow the study design presented in Chapter 5. Furthermore, the results are presented and discussed regarding the validity of usability assessments of HMIs for L3 ADS conducted in different cultural settings and, thereby, the transferability of conclusions across cultures.

7.1 Hypotheses

The validation study *Exp_Culture* seeks answers to research question RQ₃. The literature presented in Section 2.4 confirms the existence of cultural differences. Cross-cultural studies suggest that differences are more pronounced between Western countries and Asian countries compared to differences between Western countries. Nevertheless, previous studies identify differences within Western countries, such as differences regarding expectations and aesthetics in interface design (Roessger, 2003), differences in take-over performances and NDRA engagement (Strle et al., 2021), and differences in the preferred display durations of maneuver advice notifications (Heimgärtner, 2007). Furthermore, several studies show that cultural differences may show in specific metrics only (e.g., Heimgärtner, 2007; Hergeth et al., 2015).

Following the proposed mapping of cultural dimensions to usability by Sogemeier et al. (2022), the following tendencies in HMI design can be expected in this experiment: *Long Term Orientation* is significantly more pronounced in Germany compared to the United States (Hofstede Insights, 2023). Consequently, participants from the United States might show stronger preferences for processes that can be influenced independently (abortions, alternative options) (Sogemeier et al., 2022). The dimensions *Individualism*, *Uncertainty Avoidance*, and *Indulgence vs. Restraint* have slightly differing scores in the United States compared to Germany, that is, more pronounced *Individualism*, more pronounced *Indulgence*, and less pronounced *Uncertainty Avoidance* in the United States compared to Germany (Hofstede Insights, 2023). Thus, participants from Germany might show stronger preferences for simple and organized interfaces that meet the participants' expectations, while participants from the United States might have stronger wishes for options for customization and individualization

²² The experiment was designed and conducted with the assistance of Julia Graefe (2021) as part of her master's thesis.

²³ Initially planned experiments in the United States and Japan had to be canceled due to the COVID-19 pandemic. Instead, one experiment was conducted in Germany with U.S.-American participants.

(Sogemeier et al., 2022). The dimensions *Power Distance* and *Masculinity* have similar scores in both cultures (Hofstede Insights, 2023). Therefore, no different preferences for HMI design are derived from the mapping between cultural dimensions and usability criteria (Sogemeier et al., 2022).

The differences between the two HMI concepts are limited and mainly concern the multimodality (only in *HC-HMI*) and the visual appearance (see Section 5.4). German participants might rate the simple designs of both concepts better than participants from the United States. Regarding the low saliency of RTIs in the *LC-HMI* concept, German participants might be more critical than participants from the United States due to a mismatch in their expectations. Furthermore, participants from the United States might be more critical of the missing customization and individualization options than those from Germany. Concluding, slightly better overall ratings for both HMI concepts with a more pronounced effect between the concepts can be expected for participants from Germany compared to participants from the United States. The listed aspects mainly concern the usability facet satisfaction, while no conclusions can be drawn for the facets effectiveness and efficiency. Satisfaction is mainly assessed through self-reported metrics, while the other two facets are predominantly assessed through observational metrics. Therefore, the following hypotheses for the validation study *Exp_Culture* are formulated:

RQ₃ Which effect has the users' cultural background on metrics for assessing the usability of HMIs for L3 ADS?

H₁ Relative and absolute validity are demonstrated for cross-cultural research between the United States and Germany compared to the test track setting regarding observational metrics for assessing the usability of HMIs for L3 ADS.

H₂ Relative and absolute validity are not demonstrated for cross-cultural research between the United States and Germany compared to the test track setting regarding self-reported metrics for assessing the usability of HMIs for L3 ADS.

Additionally, an effect of the HMI concept is expected. As described in Section 5.4, the HMI concepts serve as the artificial research subject. Introducing HMI concepts varying in their compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019) allows for assessing relative validity, which refers to the agreement between the direction (and size) of effects. Furthermore, the sensitivity of metrics toward specific differences in HMI design may be assessed. The approach of variation between two HMI concepts is adapted from this study as presented in Forster et al. (2020a) and Forster et al. (2020b).

H₃ The concept *HC-HMI* receives higher usability evaluations than the concept *LC-HMI*.

7.2 Sample

The final sample size of *TT_GER* is $n = 61$ ($n_{TT_GER-HC} = 33$; $n_{TT_GER-LC} = 28$). In *TT_GER-HC*, one session is aborted due to problems with the eye tracker (excluded from the final sample), one data set is missing eye-tracking data completely, and two data sets have incomplete eye-tracking data due to technical problems. In *TT_GER-LC*, two sessions are aborted due to problems with the eye tracker or heavy rainfall (excluded from the final sample), one data set is missing eye-tracking data completely, and three data sets have incomplete eye-tracking data due to technical problems.

The final sample size of *TT_USA* is $n = 42$ ($n_{TT_USA-HC} = 21$; $n_{TT_USA-LC} = 21$). In *TT_USA-HC*, one data set lacks eye-tracking data because of technical problems. Additionally, one session in *TT_USA-LC* is aborted due to heavy rain (excluded from the final sample).

The summary of the descriptive analysis of the sociodemographic data is attached in Appendix III (Table 14.2). The proportion of female participants ranges between 35.71% (*TT_GER-LC*) and 47.62% (*TT_USA-LC*) among the subsamples, thereby fulfilling the required minimum of 30% females (see Section 5.2). None of the participants indicates to be diverse or decides not to indicate a gender. The mean age across the four subsamples ranges between 37.43 (*TT_GER-LC*, $SD = 15.12$) and 38.43 (*TT_USA-HC*, $SD = 9.70$). The minimum age of the participants is 20 (*TT_GER-LC* & *TT_USA-LC*), and the maximum is 69 (*TT_GER-HC*). As described in Section 5.2, a minimum of five participants in four different age groups (18-24; 25-39; 40-54; > 54; NHTSA, 2013) is aimed for. In *TT_USA*, this aim is not met for two age groups in either subsample. Only two participants in each subsample represent age group 18-24. The age group > 54 is represented by two participants (*TT_USA-HC*) and one participant (*TT_USA-LC*), respectively.

The summary of the descriptive analysis of the driving background is attached in Appendix III (Table 14.3). The driving frequency and mileage are lower in *TT_USA* compared to *TT_GER*. The reported experience with ADAS is similar among the study samples. Participants of *TT_GER* report a slightly higher frequency of using the ADAS than *TT_USA*. Considerably fewer participants of *TT_GER* (*TT_GER-HC*: 42.42% & *TT_GER-LC*: 46.43%) report having no prior knowledge in the field of automated driving compared to *TT_USA* (*TT_USA-HC*: 23.81% & *TT_USA-LC*: 9.52%). Only single participants in *TT_GER* (*TT_GER-HC*: 9.09% & *TT_GER-LC*: 3.57%) and in *TT_USA-LC* (9.52%) indicate expert knowledge.

The recruitment criteria for *TT_USA* required participants to hold U.S.-American citizenship and not live in Germany permanently. Participants are allowed to live in Germany for eight years or less. The duration of living in Germany ranges between one week and eight years (one participant each), with an average duration of 3.98 years ($SD = 2.27$).

The questionnaire VSM for cultural values is applied (Hofstede & Minkov, 2013b). The results are compared to a sample of participants who hold U.S.-American citizenship and are based in the United States. The results are presented and discussed in detail in Chapter 8 (see Subsection 8.3.1 & Section 8.4). The analysis concludes that it may be assumed that the *TT_USA* represents the United States regarding its cultural values.

7.3 Results

In the following section, the results of all metrics are described. The inferential analysis of the data follows the process described in Section 5.7.

Due to organizational reasons, the two experiments are not entirely identical but differ in several aspects. The test courses in both experiments comprise two lanes marked with pylons, delineator blades, and wires. In *TT_GER*, only one adjacent lane is present. In *TT_USA*, about five adjacent lanes (with traffic cones but partially without lane markings) are present.

The protocol documents unusual behavior, unforeseen external events, or technical issues. Such events are referred to in the analysis of outliers. In *TT_GER* and *TT_USA*, vehicles, persons, and animals in areas near the test course could not be prevented entirely. Weather conditions vary between and within the experiment sessions in both experiments. For safety reasons, experiment sessions are canceled or aborted in cases of heavy rainfall. The distributions of weather and light conditions in the experiment sessions are attached in Appendix III (Table 14.1).

7.3.1 Observational Metrics

Observational data are collected for all 12 test cases (see Section 5.3). Some of the metrics presented in this section refer to specific test cases. An emphasis is put on the two RTIs (RtI_{20s} & RtI_{6s}) triggered during the test drives.

7.3.1.1 Driving Behavior

Observed LoA vs. Instructed LoA

The test case and the resulting instructions by the experimenter or system notifications determine the active LoA at every point of the test drive. If participants fail to adhere to the instructions or to react appropriately, deviances between the instructed LoA and the observed LoA may arise. Figure 7.1 displays the match between the instructed and observed LoAs for the two experiments and the respective HMI subsamples.

The figure and the binomial GLMM (Table 7.3) show that the number of deviances is significantly higher in the *LC-HMI* subsamples compared to the *HC-HMI* subsamples. In *TT_USA*, the variance appears to be greater compared to *TT_GER*. Additionally, the variance is greater in the *LC-HMI* subsamples compared to the *HC-HMI* subsamples. Neither the factor *Exp* nor the interaction *Exp:HMI* are significant. The TOST confirms equivalence for the factor *Exp* (Table 7.3).

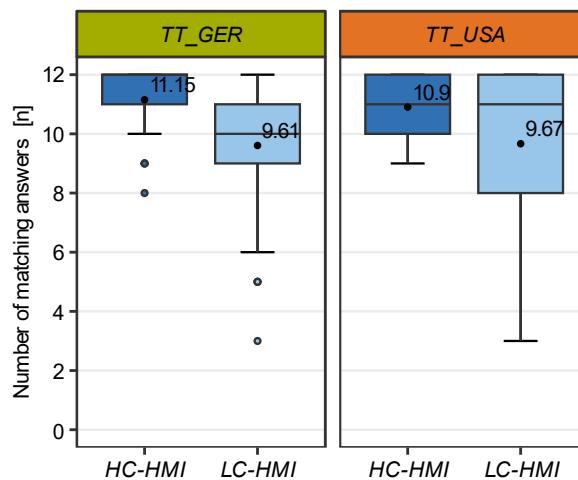


Figure 7.1 Boxplot diagram visualizing the results of the metric *Observed LoA* vs. *instructed LoA* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.3 for more statistics.

Control Path of First Activation

When participants are requested to activate L3 for the first time, they have no prior experience with the control logic of the HMI. Only participants are included who are in L0 at the experimenter's request to activate L3 and who have not tried out the controls in prior test cases. Figure 14.1 in Appendix III displays the individual control paths of the first activation. The descriptive analysis is summarized in Table 7.1.

A higher proportion of participants of the *HC-HMI* subsamples succeed in activating L3 than participants of the *LC-HMI* subsamples. More participants of *HC-HMI* subsamples manage to activate L3 with the minimum number of actions (*ACT* → *MOD*). The maximum number of actions ranges between five and six actions per subsample. Single participants repeatedly use the button *MOD* (no effect in L0) and stay in L0, or use the button *ACT* (L0 ↔ L2). The difference between the *HC-HMI* and *LC-HMI* subsamples is slightly more pronounced in *TT_GER* than in *TT_USA*.

Table 7.1 Descriptive analysis of the metric *Control path of first activation* for the study *Exp_Culture*.

Subsample (n)	Successful activation [% (n)]	Use of ideal path: ACT → MOD [% (n)]	Number of actions for participants not using the ideal path*	
			M (SD)	Max
TT_GER-HC (27)	62.96% (17)	37.04% (10)	2.35 (1.37)	5
TT_GER-LC (23)	34.78% (8)	17.39% (4)	2 (1.3)	5
TT_USA-HC (14)	71.43% (10)	42.86% (6)	3.5 (2.06)	8
TT_USA-LC (14)	50% (7)	28.57% (4)	2.4 (1.62)	6

Note. Only participants driving L0 at the start of the instruction are included.

* The statistics include non-successful interactions.

TOT after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 7.2 displays the TOT for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}).

Figure 7.2 and Table 7.3 show significantly higher TOTs for the *LC-HMI* subsamples. Participants of *TT_USA* have slightly higher TOTs than participants of *TT_GER* (except for RtI_{6s} : *TT_USA-LC*). Neither the factor *Exp* nor the interaction *Exp:HMI* of the GLMM are significant. The TOST does not confirm equivalence. The results indicate that the data are inconclusive for both tests.

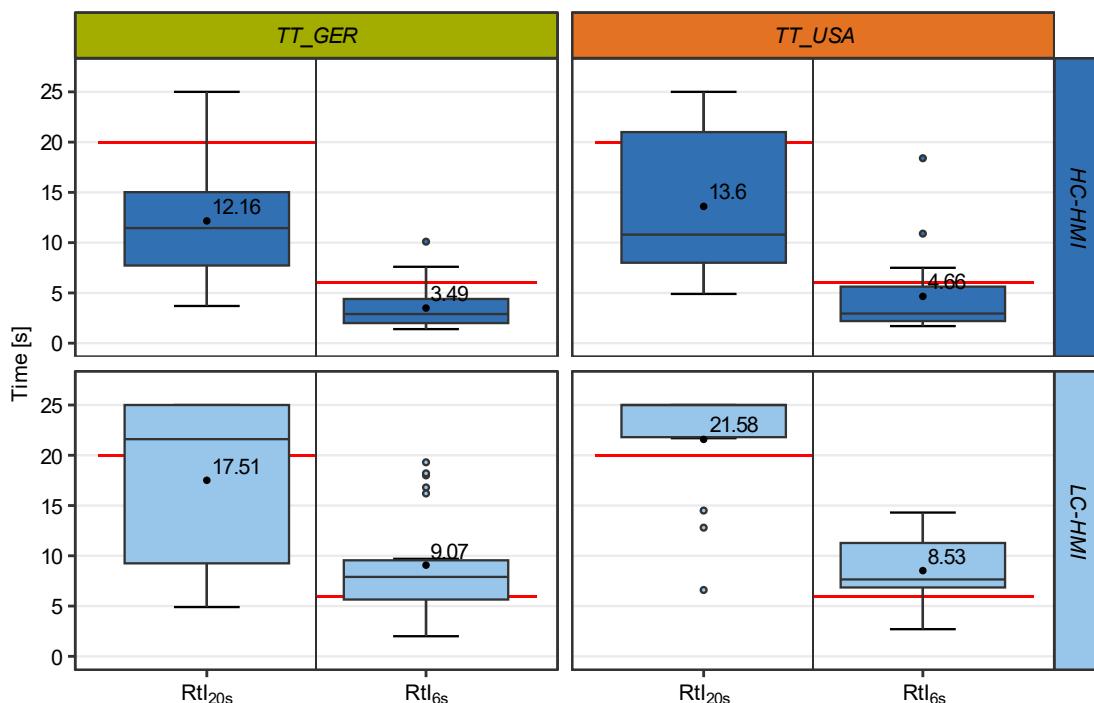


Figure 7.2 Boxplot diagram visualizing the results of the metric *TOT after RtI* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.3 for more statistics. The red lines mark the start of the emergency braking for the respective RtI.

Take-Over Path after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Only participants are included that drive in L3 when the respective RtI is triggered. Figure 14.2 and Figure 14.3 in Appendix III display the individual take-over paths for both RtI types (RtI_{20s} & RtI_{6s}). The descriptive analysis is summarized in Table 7.2.

Most participants conduct a successful transition to L0 with only one action. Participants of *TT_USA* tend to use the brake more often than participants of *TT_GER*. Single participants in all four subsamples use the button *MOD* (repeatedly), which switches between L2 and L3 before switching to L0. Primarily in the *LC-HMI* subsamples, single participants do not take

over at all (RtI_{20s}: TT_GER-LC: $n = 2$ & TT_USA-LC: $n = 2$; RtI_{6s}: TT_GER-LC: $n = 3$ & TT_USA-HC: $n = 2$).

Table 7.2 Descriptive analysis of the metric *Take-over path after RtI* for the study *Exp_Culture*.

RtI	Subsample (n)	Successful transition to L0 [% (n)]	1 action: Transition to L0 via ... [% (n)]:		> 1 action*: number of actions	
			Brake	ACT	M (SD)	Max
RtI _{20s}	TT_GER-HC (32)	100% (32)	53.13% (17)	28.13% (9)	3.5 (0.5)	4
	TT_GER-LC (26)	92.31% (24)	57.69% (15)	26.92% (7)	2.5 (0.5)	3
	TT_USA-HC (21)	95.24% (20)	76.19% (16)	14.29% (3)	2.5 (0.5)	3
	TT_USA-LC (16)	93.75% (15)	68.75% (11)	12.5% (2)	2 (0)	2
RtI _{6s}	TT_GER-HC (33)	96.97% (32)	72.73% (24)	24.24% (8)	3 (n/a)**	n/a**
	TT_GER-LC (26)	88.46% (23)	53.85% (14)	26.92% (7)	2 (0)	2
	TT_USA-HC (21)	95.24% (20)	85.71% (18)	9.52% (2)	n/a (n/a)**	n/a**
	TT_USA-LC (16)	100% (16)	75% (12)	25% (4)	n/a (n/a)**	n/a**

Note. Only participants driving L0 at the start of the instruction are included. None of the participants deactivates L3 through oversteering.

* Participants are excluded that do not react at all: TT_GER-LC: $n = 2$ (RtI_{20s}), $n = 3$ (RtI_{6s}); TT_USA-HC: $n = 1$ (RtI_{6s}); & TT_USA-LC: $n = 1$ (RtI_{20s}).

** $n = 1$

Other Observations

If participants take their hands away from the steering wheel during L2 driving, they receive a H-off detection warning that comprises three stages depending on the H-off driving duration. The stages trigger notifications differing in their urgency (see Section 5.4).

In TT_GER-HC, eight participants produce 13 H-off detection warnings (12x stage 1, 1x stage 2). Additionally, another participant of TT_GER-HC produces nine warnings (8x stage 1, 1x stage 2) alone. In TT_GER-LC, eight participants produce 13 H-off detection warnings (7x stage 1, 1x stage 2, 5x stage 3). Additionally, another participant of TT_GER-LC produces nine warnings (8x stage 1, 1x stage 2) alone. In TT_USA-HC, six participants produce eight H-off detection warnings (7x stage 1, 1x stage 2). In TT_USA-LC, five participants produce 12 H-off detection warnings (9x stage 1, 2x stage 2, 1x stage 3). Additionally, another participant of TT_USA-LC produces six warnings (2x stage 1, 2x stage 2, 2x stage 3) alone. Considering the smaller sample size in TT_USA and the outliers in three of the four subsamples (TT_GER-HC, TT_GER-LC, TT_USA-LC), both experiments' distribution of H-off detection warnings is similar. Participants of TT_USA-LC produce repeated H-off detection warnings slightly more often than participants of the other three subsamples.²⁴

In test case TC6, participants drive in L2 and receive a notification that L3 driving is no longer available for optional activation due to a sensor error. The notification is only for informational purposes and does not require an action by the participant. Nevertheless, in TT_GER-LC, two participants deactivate L2. In the other three subsamples, one participant each deactivates L2.

²⁴ According to the protocol, the two participants (TT_GER-HC_{TP8} & TT_GER-LC_{TP60}) of TT_GER producing the nine h-off detection warnings each show no other remarkable behavior. The participant of (TT_USA-LC_{TP60}) producing six h-off detection warnings reports that he registered the warnings but did not feel the need to react. He further elaborates that he cannot tell any difference in the behavior of L2 and L3 driving and consequently took his hands away from the steering wheel.

Summary

In both inferential statistical analyses, the factor *HMI* is significant. The factor *Exp* and the interaction factor *Exp:HMI* are not significant in either of the tests. The equivalence test is significant for the metric *Observed LoA vs. Instructed LoA*. The descriptive and qualitative analysis of the driving behavior supports the findings of the inferential statistical tests: The performance scores are lower for the *LC-HMI* subsamples. In the descriptive analysis, no tendency for systematic differences between the experimental conditions is observable. In the RtIs, differences in the take-over strategies could be observed, with participants of *TT_USA* tending to use the brake more often than participants of *TT_GER*.

Table 7.3 Summary table of the descriptive and inferential results of the quantitative metrics of the driving behavior for the study *Exp_Culture*.

Metric	Subsample	Descriptive data					GLMM ^a							TOST ^b		
		n	M	SD	Min	Max	Distrib. (link)	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Observed LoA vs. instructed	TT_GER-HC	33	11.15	1.12	8	12	Binomial (logit)	Intercept	2.96	0.4	7.35				667 < .001***	
	TT_GER-LC	28	9.61	2.53	3	12		Exp	0.06	0.2	0.31	0.09	1	.760		
	TT_USA-HC	21	10.9	1.14	9	12		HMI	0.63	0.21	3.05	8.96	(1,236)	.003**		
	TT_USA-LC	21	9.67	2.97	3	12		Exp:HMI	0.2	0.21	0.97	0.96		.328		
TOT after R _{t^{c,d}}	TT_GER-HC	32 33	12.16 3.49	5.45 1.91	3.7 1.4	25 10.1	Gaussian (identity)	Intercept	11.3	3.43	3.3				1,044 .204	
	TT_GER-LC	24 23	17.51 9.07	7.95 4.94	4.9 2	25 19.3		Exp	-0.68	0.47	-1.44	2.04		.154		
	TT_USA-HC	21 20	13.6 4.66	7.14 3.97	4.9 1.7	25 18.4		HMI	-2.86	0.47	-6.07	31.06	1 (184)	< .001***		
	TT_USA-LC	15 16	21.58 8.53	5.49 3.28	6.6 2.7	25 14.3		Exp:HMI	0.09	0.47	0.18	0.03		.854		

^a GLMM formula: DV ~ Exp*HMI + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c The descriptive data distinguishes between the test cases TC10 (left, R_{t120s}) and TC12 (right, R_{t6s}).

^d The maximum TOT is capped at 25 s.

7.3.1.2 Eye-Tracking

Attention Ratio during Continuous Rides in L0, L2, & L3

Participants receive clear instructions on their responsibilities for the driving task for the different LoAs (see Subsection 5.5.2). The analysis of the attention ratio checks whether participants adhere to these instructions. Four areas of interest (AOIs) are defined: *Street*, *IC*, *Controls*, and *SuRT* (see Subsection 5.6.2). During L0 and L2 driving, the visual attention should be focused on AOI *Street*. In L3 driving, participants are instructed to engage in the NDRA if the situation allows it. If participants adhere to the instructions, the attention ratio for the *SuRT* should be close to zero in L0 and L2 driving and considerably higher in L3 driving.

In Figure 7.3, the attention ratios for the *SuRT* are displayed for the three LoAs and the four subsamples. Attention ratios for all four AOIs are attached in Appendix III (Figure 14.4). Participants who do not drive in the instructed LoA during the specified test cases are excluded, leading to sample sizes differing within the subsamples.

In L0 and L2 driving, the mean attention ratio for the *SuRT* is below 5% for three of the four subsamples (slightly higher in L2 driving compared to L0 driving). Only in *TT_USA-LC* and L2 driving the mean attention ratio for the *SuRT* is considerably higher, with $M = 15.68\%$ ($SD = 28.27\%$). In L3 driving, the mean attention ratio for the *SuRT* ranges between 39.56% (*TT_USA-LC*, $SD = 33.52\%$) and 76.36% (*TT_GER-HC*, $SD = 18.9\%$). The mean attention ratios for the *SuRT* are higher in the subsamples of *TT_GER* compared to the respective subsamples in *TT_USA*. Furthermore, the variance in the subsamples of *TT_USA* is considerably higher than that of *TT_GER*. In all four subsamples, single participants have an attention ratio for the *SuRT* of 30% or lower when driving in L3. The number of participants under this threshold is higher and the attention ratios lower for participants of the *LC-HMI* subsamples and participants of *TT_USA* (*TT_GER-HC*: $n = 1$ with $AR = 26.69\%$; *TT_GER-LC*: $n = 2$ with $M = 5.62\%$ & $SD = 5.62\%$; *TT_USA-HC*: $n = 4$ with $M = 9.62\%$ & $SD = 8.36\%$; & *TT_USA-LC*: $n = 5$ with $M = 5.31\%$ & $SD = 8.18\%$).

Table 7.4 presents the results of the GLMM and the TOST. In the GLMM, the factors *Exp* and *HMI* are significant. The interaction *Exp:HMI* is not significant. Equivalence for the factor *Exp* is not confirmed.

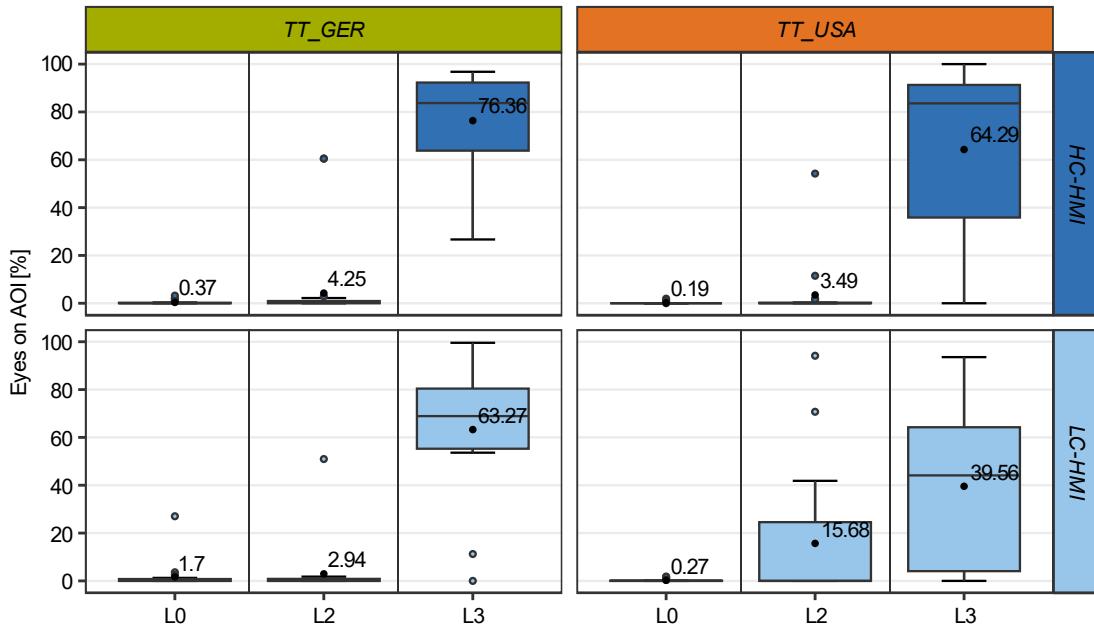


Figure 7.3 Boxplot diagram visualizing the results of the metric *Attention ratio during continuous rides in L0, L2, & L3* for the AOI *SuRT* for the study *Exp_Culture*.
Note. The mean values are displayed as numbers in the figure. Refer to Table 7.4 for more statistics.

Gaze Behavior during RtI

This qualitative analysis refers to the two test cases with RtIs (see Section 5.3). Figure 6.4 displays the proportion of glances to the four AOIs at the start of the RtI for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}). Participants who do not drive in L3 at the scenario's beginning are excluded. Figure 14.5 and Figure 14.6 in Appendix III display the individual gaze paths between the start and the end of the RtIs. The end of an RtI is marked by the start of an emergency braking maneuver or the transition to L0.

At the start of the first RtI (RtI_{20s}), more than 70% of the participants in all four subsamples look at the *SuRT*. At the start of the second RtI (RtI_{6s}), fewer participants of the *LC-HMI* subsamples look at the *SuRT* than at the start of the first RtI (RtI_{20s}). Instead, more participants of *TT_USA-LC* look at the *Street* than the *IC*, while none of the participants of *TT_GER-LC* look at the *Street*.

Before the first glance at the *IC*, most participants look at the *SuRT* and *Street* in turns. After the first glance at the *IC*, most participants look at the *IC* and other AOIs (mostly *Street*) in turns. At the end of the RtIs, no participant of the *HC-HMI* subsamples looks at the *SuRT* (except for two participants of *TT_GER* in RtI_{20s}). In all subsamples, most participants look at the *IC* at the end of the RtI. In the *LC-HMI*, five (RtI_{20s} & RtI_{6s}; *TT_GER-LC*) to seven (RtI_{20s}; *TT_USA-LC*) participants (still) look at the *SuRT*. No participants of the *HC-HMI* subsamples look at the *SuRT*. In the *TT_GER* subsamples, participants show more glances compared to the *TT_USA* subsamples. Apart from that, there are no prominent differences between the experiments.

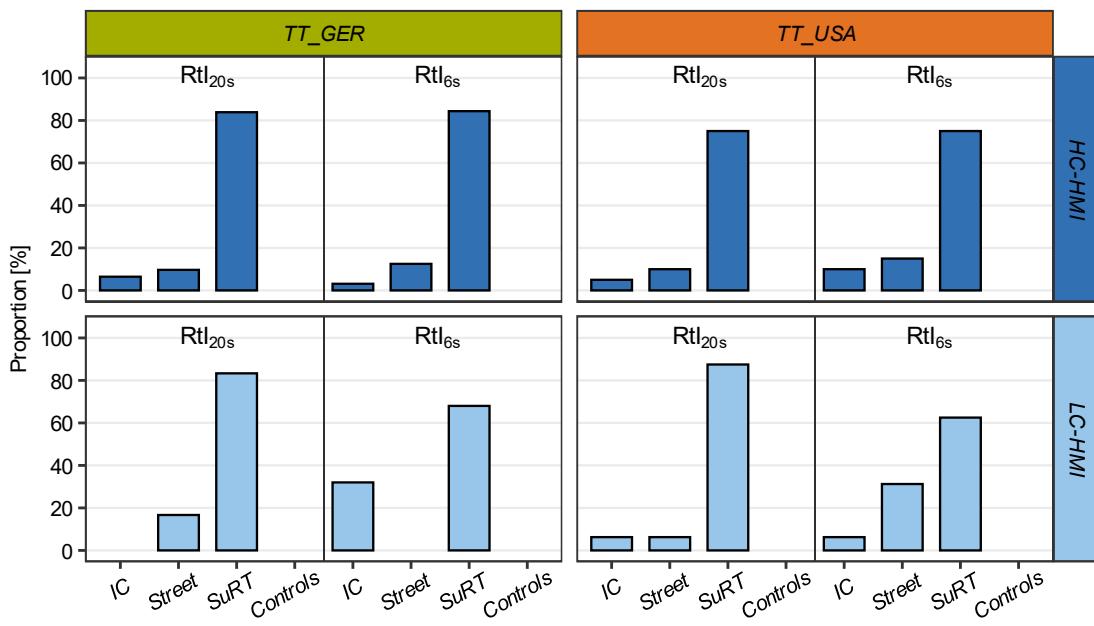


Figure 7.4 Bar chart visualizing the results of the metric *Glance allocation at start of RtI* for the study *Exp_Culture*.

Note. The sample sizes are as follows: *TT_GER-HC*: $n = 31$ (RtI_{20s}), $n = 32$ (RtI_{6s}); *TT_GER-LC*: $n = 24$ (RtI_{20s}), $n = 25$ (RtI_{6s}); *TT_USA-HC*: $n = 20$ (RtI_{20s}), $n = 20$ (RtI_{6s}); & *TT_USA-LC*: $n = 16$ (RtI_{20s}), $n = 16$ (RtI_{6s}).

Glance Allocation Time to IC after RtI

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 7.5 displays the glance allocation time to the *IC* for the four subsamples and the two RtIs (RtI_{20s} & RtI_{6s}). This metric serves as an indicator of the salience of the RtI notification. Participants who do not drive in L3 or look at the *IC* already at the start of the respective RtI are excluded.

In all four subsamples, the glance allocation time to the *IC* is higher in the first RtI (RtI_{20s}) compared to the second RtI (RtI_{6s}). In both experiments and RtIs, the mean glance allocation time is higher in the *LC-HMI* subsamples than in the *HC-HMI* subsamples. The variance in the first RtI (RtI_{20s}) is slightly higher in the *HC-HMI* subsamples. In the second RtI (RtI_{6s}), the variance is higher in the *LC-HMI* subsamples. For both HMI concepts and RtIs, the glance allocation times are higher for the *TT_USA* subsamples than for the *TT_GER* subsamples. Table 7.4 presents the results of the GLMM and the TOST. The factor *Exp* is significant. The factors *HMI* and *Exp:HMI* are not significant. Equivalence for the factor *Exp* is not confirmed.

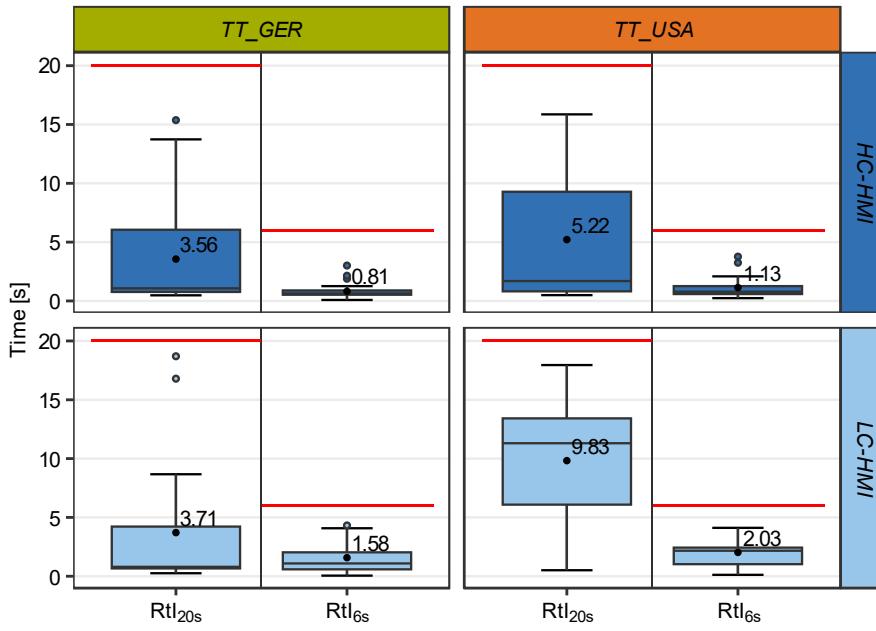


Figure 7.5 Boxplot diagram visualizing the results of the metric *Glance allocation time to IC after Rtl* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.4 for more statistics. The red lines mark the start of the emergency braking for the respective Rtl.

First Glance Duration on IC after Rtl

This metric refers to the two test cases with RtIs (see Section 5.3). Figure 7.6 displays the duration of the first glance to the *IC* for the four subsamples and the two RtIs (Rtl_{20s} & Rtl_{6s}). This metric serves as an indicator for the presentation of information appropriate to the situation. Participants who do not drive in L3 or look at the *IC* already at the start of the respective Rtl are excluded.

In all four subsamples, the first glance duration to the *IC* is higher in the first Rtl (Rtl_{20s}) compared to the second Rtl (Rtl_{6s}). The first glance duration is slightly higher for the *LC-HMI* subsamples than the *HC-HMI* subsamples. There are no prominent differences between the experiments. Table 7.4 presents the results of the GLMM and the TOST. Equivalence for the factor *Exp* is confirmed. None of the factors in the GLMM is significant.

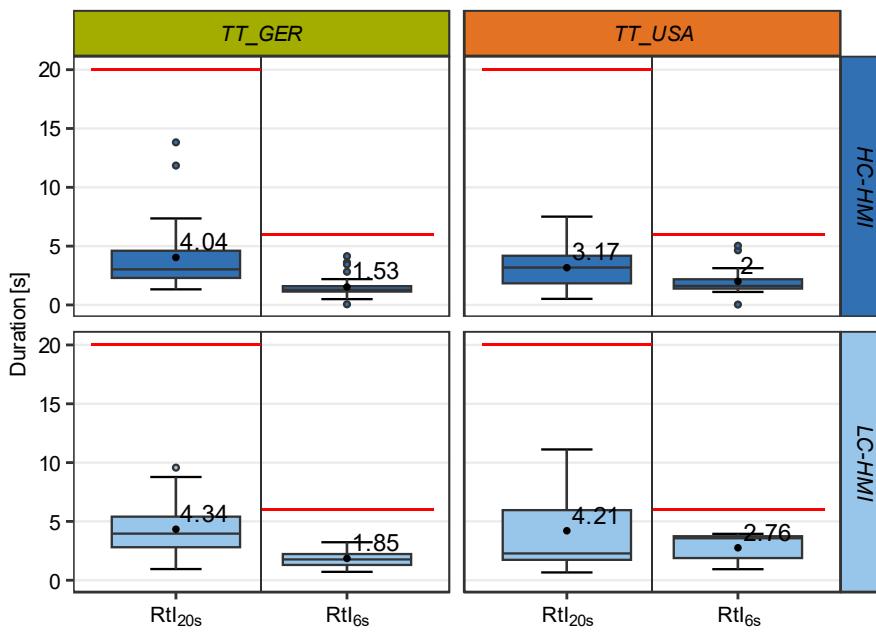


Figure 7.6 Boxplot diagram visualizing the results of the metric *First glance duration on IC after RtI* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.4 for more statistics. The red lines mark the start of the emergency braking for the respective RtI.

Summary

In the GLMMs, the factor *HMI* is significant only for the metric *Attention Ratio during continuous rides in L0, L2, & L3: SuRT*. The descriptive analysis of gaze behavior shows differences between the HMI subsamples, for example, an increase of glances to the *IC* after the first RtI (*RtI_{20s}*) in the *LC-HMI* subsamples or the shorter glance allocation times to the *IC* in RtIs in the *HC-HMI* subsamples. The factor *Exp* is significant for two of the three metrics included in inferential statistical tests (*Attention ratio during continuous rides in L0, L2, & L3: SuRT*, & *Glance allocation time to IC after RtI*), reflecting better performance scores for *TT_GER* compared to *TT_USA*. Equivalence for the factor *Exp* is confirmed only for the metric *First glance duration on IC after RtI*. The interaction factor *Exp:HMI* is not significant for any metric. Descriptive analyses of the attention ratios and the gaze paths during RtIs show differences in the gaze behavior for the factor *Exp*. In the gaze behavior analysis, several participants are excluded because they do not meet preconditions for the metric, for example, wrong LoA at the scenario start. The exclusion criteria mainly reduce the sample numbers of the *LC-HMI* subsamples.

Table 7.4 Summary table of the descriptive and inferential results of the quantitative metrics of the eye-tracking for the study *Exp_Culture*.

Metric	Subsample	Descriptive data					GLMM ^a (gaussian distribution with identity link function)						TOST ^b		
		n	M	SD	Min	Max	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Attention ratio during continuous rides in L0, L2, & L3: SuRT ^c	TT_GER-HC	27	0.37	0.8	0	3.2	Intercept	85.25	9.26	9.21					
		32	4.25	14.77	0	60.52									
		27	76.36	18.9	26.69	96.79									
	TT_GER-LC	22	1.7	5.73	0	27.04	Exp	3.05	1.17	2.61	6.59	1	.010*	433	.356
		20	2.94	11.32	0	50.95									
		14	63.27	28.41	0	99.57									
	TT_USA-HC	19	0.19	0.5	0	1.92	HMI	3.39	1.17	2.9	8.07	(246)	.005**	433	.356
		20	3.49	12.22	0	54.23									
		15	64.29	36.65	0	100									
	TT_USA-LC	20	0.27	0.56	0	1.82	Exp:HMI	-1.55	1.17	-1.33	1.75	1	.186	433	.356
		18	15.68	28.27	0	94.14									
		12	39.56	33.52	0	93.58									
Glance allocation time to IC after Rtl ^c	TT_GER-HC	29	3.56	4.34	0.48	15.37	Intercept	3.24	1.38	2.35					
		31	0.81	0.6	0.09	3.01									
	TT_GER-LC	19	3.71	5.47	0.26	18.7	Exp	-1.06	0.33	-3.24	9.92	1	.002**	419	.264
		11	1.58	1.46	0.05	4.32									
	TT_USA-HC	17	5.22	5.58	0.49	15.86	HMI	-0.6	0.33	-1.83	3.28	(142)	.070	419	.264
		18	1.13	0.97	0.24	3.77									
	TT_USA-LC	7	9.83	6.03	0.51	17.95	Exp:HMI	0.64	0.33	1.97	3.79	1	.051	419	.264
		10	2.03	1.28	0.12	4.11									
First glance duration on IC after Rtl ^c	TT_GER-HC	29	4.04	2.82	1.33	13.82	Intercept	3.08	0.74	4.15					
		31	1.53	0.88	0.06	4.15									
	TT_GER-LC	19	4.34	2.23	0.95	9.57	Exp	-0.07	0.18	-0.39	0.14	1	.696	257	.040*
		11	1.85	0.75	0.71	3.23									
	TT_USA-HC	17	3.17	1.77	0.52	7.51	HMI	-0.34	0.18	-1.88	3.44	(141)	.064	257	.040*
		18	2	1.19	0.03	5.03									
	TT_USA-LC	7	4.21	3.74	0.66	11.12	Exp:HMI	0.13	0.18	0.07	0.49	1	.483	257	.040*
		9	2.76	1.16	0.94	3.94									

^a GLMM formula: DV ~ Exp*HMI + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c The descriptive data distinguishes between the LoAs L0 (left), L2 (center), and L3 (right) or the test cases TC10 (left, Rtl_{20s}) and TC12 (right, Rtl_{6s}), respectively.

7.3.1.3 Experimenter Rating

After each test case, the experimenter rates the participants' interaction with the ADS. Figure 7.7 displays the participants' mean experimenter rating for the four subsamples. For simplicity reasons, the figure does not visualize the experimenter ratings in the 12 test cases but displays the distribution of the participants' mean experimenter ratings. A visualization of the mean experimenter ratings per test case is attached in Appendix III (Figure 14.7).

Figure 7.7 and Table 7.5 show significantly better mean experimenter ratings for the *HC-HMI* subsamples. The variance of the participants' mean experimenter ratings is higher in the subsample *TT_GER-LC* compared to the other three subsamples. Otherwise, no differences between the experiments are observed. The factors *Exp* and *Exp:HMI* in the GLMM are not significant. The TOST confirms equivalence.

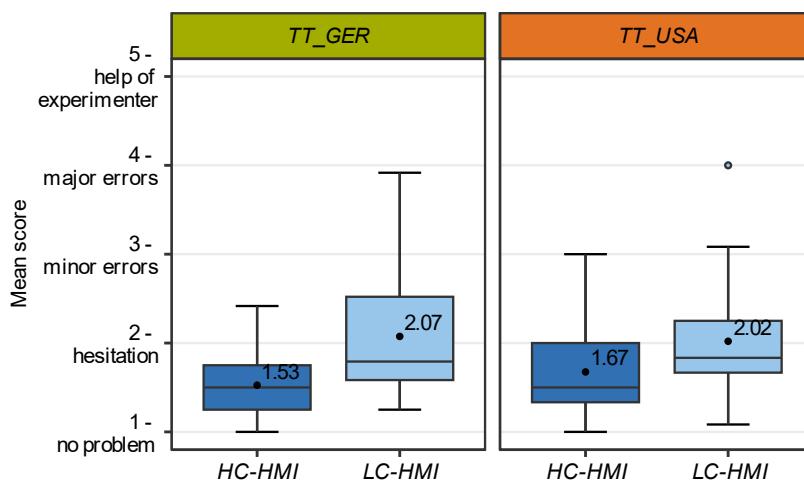


Figure 7.7 Boxplot diagram visualizing the results of the metric *Experimenter rating* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.5 for more statistics.

Table 7.5 Summary table of the descriptive and inferential results of the metric *Experimenter rating* for the study *Exp_Culture*.

Subsample	Descriptive data						GLMM ^a (gaussian distribution with logit link function)					TOST ^b		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Med</i>	<i>Min</i>	<i>Max</i>	Factor	<i>Est.</i>	<i>SE</i>	<i>z</i>	<i>X</i> ² (1, 113)	<i>p</i>	<i>W</i> _{max, <i>p</i>}	<i>p</i> _{max}
TT_GER-HC	33	1.53	0.37	1.5	1	2.42	<i>Intercept</i>	0.53	0.08	6.32				
TT_GER-LC	28	2.07	0.74	1.79	1.25	3.92	<i>Exp</i>	-0.02	0.03	-0.7	0.5	.480	1,603	.015*
TT_USA-HC	21	1.67	0.52	1.5	1	3	<i>HMI</i>	-0.13	0.03	-4.2	16.22	< .001***		
TT_USA-LC	21	2.02	0.68	1.83	1.08	4	<i>Exp:HMI</i>	-0.03	0.03	-0.87	0.75	.390		

^a GLMM formula: DV ~ *Exp***HMI* + (1 | TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

7.3.2 Self-Reported Metrics

7.3.2.1 Short Interviews

The short interviews refer to the participants' interaction with the ADS during the test drive. The set of questions refers to the mode awareness, the system understanding, and the report of interaction problems during transitions. Most metrics are calculated by assessing the ratio between correct and wrong replies. Only the metric *Reported Problems during transitions* is calculated as the number of reported problems.

Awareness of Active LoA

At the end of each test case, participants are requested to name the last active LoA. The reported LoA is compared to the observed active (not the instructed) LoA. Figure 7.8 displays the match between the reported and the observed LoAs for the two experiments and the respective HMI subsamples.

The *LC-HMI* subsamples show considerably fewer matching answers and a higher variance than the *HC-HMI* subsamples. In the *LC-HMI* subsamples, single participants have less than 50% correct matching answers. There are no prominent differences between the experiments. The binomial GLMM (Table 7.6) shows a significant influence of the factor *HMI*. Neither the factor *Exp* nor the interaction *Exp:HMI* are significant. The TOST confirms equivalence for the factor *Exp* (Table 7.6).

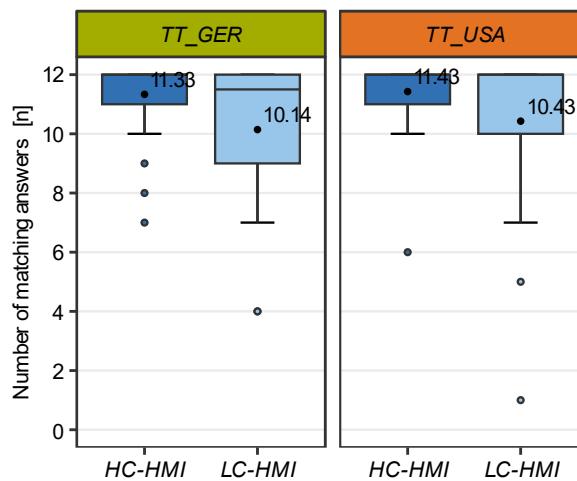


Figure 7.8 Boxplot diagram visualizing the results of the metric *Awareness of active LoA* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

Awareness of Change of Available LoAs

At the end of each test case, participants state whether a change in the available LoAs occurred. The reported change of availabilities is compared to the implemented change of availabilities. Figure 7.9 displays the match between the reported and the implemented changes of availabilities for the two experiments and the respective HMI subsamples.

The number of matching answers is lower in the *LC-HMI* subsamples than in the *HC-HMI* subsamples. Furthermore, the *LC-HMI* subsamples have a higher variance than the *HC-HMI* subsamples. The difference between the subsamples is slightly more pronounced in *TT_USA* than in *TT_GER*. The binomial GLMM (Table 7.6) shows a significant influence of the factor *HMI*. Neither the factor *Exp* nor the interaction *Exp:HMI* are significant. The TOST does not confirm equivalence for the factor *Exp* (Table 7.6).

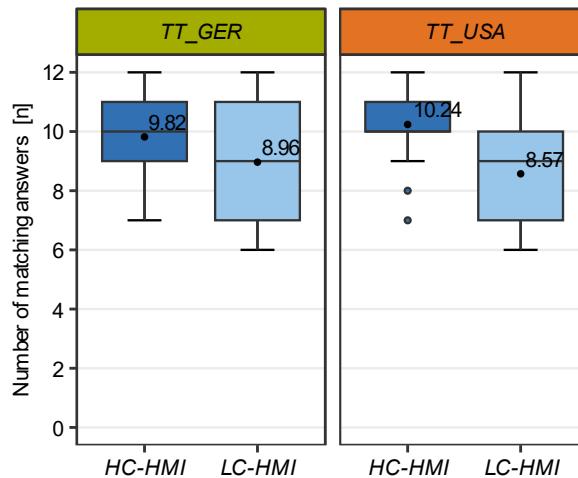


Figure 7.9 Boxplot diagram visualizing the results of the metric *Awareness of change of available LoAs* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

Awareness of Reason for Change of Available LoAs

In three test cases (see Section 5.3), a downward change of available LoAs occurs. In these test cases, the HMI provides reasons for the availability change, for example, a sensor error (see Section 5.4). Participants are requested to recall the reason for the change of availabilities. The number of matching answers between implemented and reported reasons for the change of availabilities is visualized in Figure 7.10.

In three subsamples, the range is between zero and three matching answers. In *TT_USA*, none of the participants of the *HC-HMI* subsample and only one participant of the *LC-HMI* subsample reports the correct reason in all three scenarios. In comparison, 14 participants of *TT_GER-HC* and 11 of *TT_GER-LC* report the correct reason in all three scenarios. Hence, the subsamples of *TT_USA* have lower means of matching answers than the subsamples of *TT_GER*. Furthermore, the *LC-HMI* subsamples have slightly lower means of matching answers than the *HC-HMI* subsamples. The factor *Exp* in the binomial GLMM is significant, but none of the other factors.²⁵ The TOST does not confirm equivalence for the factor *Exp* (Table 7.6).

²⁵ Due to convergence, the model does not contain the random factor (1|TC).

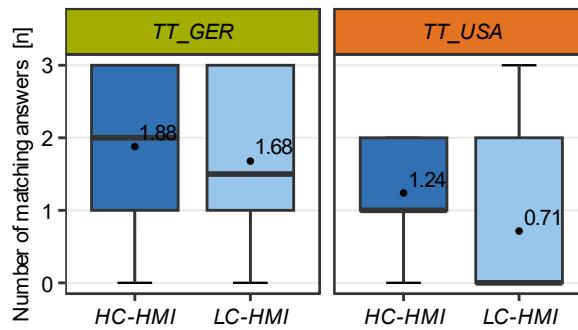


Figure 7.10 Boxplot diagram visualizing the results of the metric *Awareness of reason for change of available LoAs* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

System Understanding: Allowance of NDRA

At the end of each test case, participants state whether it was allowed to engage in NDRA, such as writing e-mails while driving in the last active LoA. The reported allowance for this NDRA is compared to the observed LoA. Before the test drive, participants are instructed that only in L3 driving it is allowed to engage in NDRA. Figure 7.11 displays the match between the reported allowance to engage in the NDRA and the observed LoA for the two experiments and the respective HMI subsamples.

The *HC-HMI* subsamples have slightly higher means and lower variances than the *LC-HMI* subsamples. There are no prominent differences between the experiments. None of the factors in the binomial GLMM (Table 7.6) is significant. The TOST confirms equivalence for the factor *Exp* (Table 7.6).

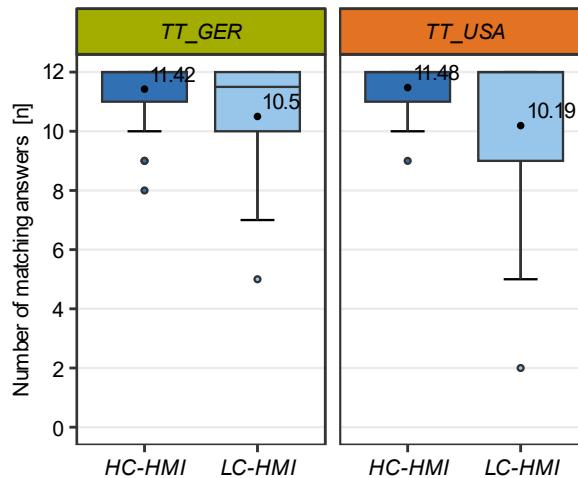


Figure 7.11 Boxplot diagram visualizing the results of the metric *System understanding: allowance of NDRA* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

System Understanding: Allowance of H-off Driving

At the end of each test case, participants state whether it was allowed to take their hands away from the steering wheel while driving in the last active LoA. The reported allowance for H-off driving is compared to the observed LoA. Before the test drive, participants are instructed that only in L3 driving it is allowed to drive H-off. Figure 7.12 displays the match between the reported allowance for H-off driving and the observed LoA for the two experiments and the respective HMI subsamples.

The *HC-HMI* subsamples have slightly higher means and slightly lower variances than the *LC-HMI* subsamples. There are no prominent differences between the experiments. None of the factors in the binomial GLMM (Table 7.6) is significant. The TOST confirms equivalence for the factor *Exp* (Table 7.6).

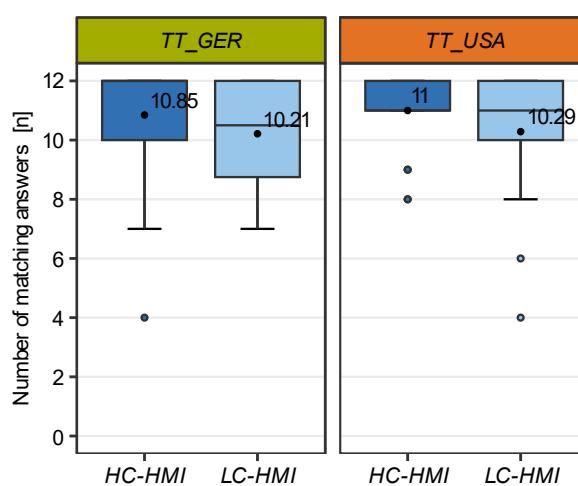


Figure 7.12 Boxplot diagram visualizing the results of the metric *System understanding: allowance of H-off driving* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

Reported Problems during Transitions

In six test cases (see Section 5.3), participants are instructed to switch between LoAs by the experimenter or the system. After each test case, participants state whether they encountered problems when switching between LoAs. The number of reported problems is visualized in Figure 7.13.

The maximum number of reported problems per participant is two (*TT_USA-HC* & *TT_USA-LC*) or three (*TT_GER-HC* & *TT_GER-LC*). Between 39.39% (*TT_GER-HC*) and 76.19% (*TT_USA-LC*) of the participants report no problem. The mean number of reported problems is lower in the *HC-HMI* subsamples than in the *LC-HMI* subsamples. Additionally, the mean number of reported problems is lower in the subsamples of *TT_USA* than in the subsamples of *TT_GER*. The two factors *Exp* and *HMI* in the binomial GLMM are significant.²⁶ The interaction factor *Exp:HMI* is not significant. The TOST confirms

²⁶ Due to convergence, the model does not contain the random factor (1|TC).

equivalence for the factor *Exp* (Table 7.6). The factor *Exp* is significant in both models, suggesting a small effect.

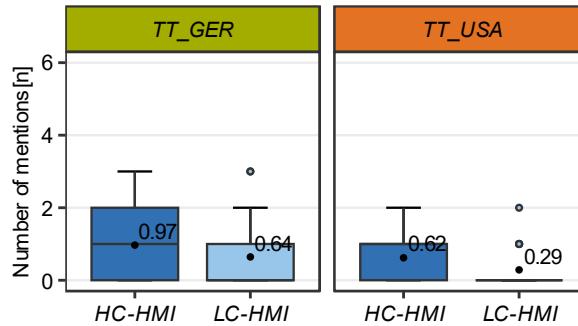


Figure 7.13 Boxplot diagram visualizing the results of the metric *Reported problems during transitions* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.6 for more statistics.

Most problems are reported after test cases with RTIs that required a time-critical action of the participants and after the test case with the first activation of the ADS: Between 33.33% (*TT_USA-LC*) and 38.89% (*TT_GER-LC*) of the overall reported problems are reported after the first activation. After the first RTI (RtI_{20s}), between 11.11% (*TT_GER-LC*) and 25% (*TT_GER-HC*) of the overall reported problems are reported.

Participants reporting a problem are requested to describe it. Between 0% (*TT_USA-LC*) and 18.75% (*TT_GER-HC*) of the reported problems do not refer to the LoA transition but the study procedure (e.g., *TT_GER-HC_{TP26}*: “*regarding performing no [problem], but [I] have responded incorrectly to announcement*”²⁷) or more general problems (e.g., *TT_GER-LC_{TP72}*: “*have problems myself to maintain 30 km/h*”²⁸).

Participants of all subsamples encounter similar problems with the control logic of the HMI controls. Statements referring to the control logic make up between 38.89% (*TT_GER-LC*) and 100% (*TT_USA-LC*) of the reported problems. Participants make statements such as “*did not know what to do with the buttons. [...] maybe an info would be helpful what I can press, for example, by lighting up the buttons*”²⁹ (*TT_GER-LC_{TP74}*), or “*had to guess which button to use*” (*TT_USA-HC_{TP2}*).

In *TT_GER-LC*, 50% of the reported problems refer to the participants’ uncertainty regarding the active LoA and its functions (e.g., *TT_GER-LC_{TP75}*: “*not sure what level I actually ended up in*”³⁰). In contrast, only once a participant of *TT_GER-HC* and no participants of the *TT_USA* subsamples report this kind of issue.

Single participants in the *HC-HMI* subsamples (*TT_GER-HC*: 6.25% & *TT_USA-HC*: 7.69%) describe that they have tried other take-over strategies during the RTIs such as putting their hands on the steering wheel which delayed the transition to L0 (e.g., *TT_GER-HC_{TP10}*:

²⁷ Translated from German statement: „*vom Ausführen her nein, habe aber falsch auf Ansage reagiert*“.

²⁸ Translated from German statement: „*habe selber Probleme, 30 km/h zu halten*“.

²⁹ Translated from German statement: „*wusste nicht, was ich mit den Tasten machen muss. [...] hilfreich wären vielleicht Infos, was ich drücken kann, z.B. durch Aufleuchten der Tasten*“.

³⁰ Translated from German statement: „*nicht sicher in welcher Stufe ich tatsächlich gelandet bin*“.

"have first wanted to oversteer with hands on the steering wheel and gas, did not work, then turned off with button"³¹).

Summary

Equivalence is confirmed for four of the six inferential statistical tests. The equivalence test and the GLMM are significant for the metric *Reported problems during transitions*. This suggests that the effect is rather small and that it is neither in the 90% CI equivalence bounds nor includes zero in the 95% CI (see Section 5.7). One additional metric, *Awareness of change of available LoAs*, shows significance for the factor *Exp* in the GLMM. The descriptive analysis shows a tendency for participants of *TT_GER* to have a better interaction with the ADS than participants of *TT_USA*. The only exception is the metric *Reported problems during transitions*, where the interaction scores are not calculated through answers that could be right or wrong but where participants can freely report problems when switching between LoAs. In this metric, participants of *TT_USA* report fewer problems than participants of *TT_GER*. Additionally, participants of *TT_USA* almost entirely report problems referring to the transitions, while participants of *TT_GER* also report more general problems. The factor *HMI* is significant for three of the six metrics. The descriptive analysis shows that participants of the *HC-HMI* subsamples have (slightly) better interaction scores than participants of the *LC-HMI* subsamples. The interaction factor *Exp:HMI* is not significant for any of the metrics.

³¹ Translated from German statement: „habe erst mit Händen am Lenkrad und Gas übersteuern wollen, hat nicht funktioniert, dann mit Button ausgeschaltet“.

Table 7.6 Summary table of the descriptive and inferential results of the short interviews for the study *Exp_Culture*.

Metric	Subsample	Descriptive data					GLMM ^a (binomial distribution with logit link function)						TOST ^b		
		n	M	SD	Min	Max	Factor	Est.	SE	z	X ²	df (N)	p	W _{max_p}	p _{max}
Awareness of active LoA	TT_GER-HC	33	11.33	1.27	7	12	Intercept	3.76	0.41	9.25					
	TT_GER-LC	28	10.14	2.46	4	12	Exp	-0.11	0.27	-0.41	0.17		.680	1,968 < .001***	
	TT_USA-HC	21	11.43	1.36	6	12	HMI	0.71	0.28	2.55	6.49	1 (1,236)	.011*		
	TT_USA-LC	21	10.43	2.82	1	12	Exp:HMI	0.07	0.27	0.27	0.07		.787		
Awareness of change of available LoAs	TT_GER-HC	33	9.82	1.55	7	12	Intercept	1.81	0.38	4.71					
	TT_GER-LC	28	8.96	2.05	6	12	Exp	-0.02	0.11	-0.19	0.04		.848	1,075 .083	
	TT_USA-HC	21	10.24	1.34	7	12	HMI	0.39	0.11	3.6	12.14	1 (1,236)	< .001***		
	TT_USA-LC	21	8.57	2.04	6	12	Exp:HMI	-0.13	0.11	-1.22	1.45		.229		
Awareness of reason for change of available LoAs ^c	TT_GER-HC	33	1.88	1.17	0	3	Intercept	-0.25	0.21	-1.23					
	TT_GER-LC	28	1.68	1.22	0	3	Exp	0.8	0.22	3.65	15.29		< .001***	1,525 .952	
	TT_USA-HC	21	1.24	0.62	0	2	HMI	0.39	0.21	1.88	3.71	1 (309)	.054		
	TT_USA-LC	21	0.71	1.01	0	3	Exp:HMI	-0.2	0.21	-0.97	0.97		.325		
System understanding: allowance of NDRA	TT_GER-HC	33	10.85	1.87	4	12	Intercept	-1	0.65	-1.53					
	TT_GER-LC	28	10.21	1.83	7	12	Exp	0	0.14	0.04	0		.972	867 .002**	
	TT_USA-HC	21	11	1.34	8	12	HMI	-0.17	0.14	-1.25	1.47	1 (1,236)	.225		
	TT_USA-LC	21	10.29	2.17	4	12	Exp:HMI	0.02	0.14	0.17	0.03		.866		
System understanding: allowance of H-off driving	TT_GER-HC	33	11.42	1.09	8	12	Intercept	-0.66	0.61	-1.07					
	TT_GER-LC	28	10.5	1.95	5	12	Exp	0.07	0.13	0.56	0.3		.582	970 .017*	
	TT_USA-HC	21	11.48	0.81	9	12	HMI	-0.06	0.13	-0.47	0.21	1 (1,236)	.643		
	TT_USA-LC	21	10.19	2.71	2	12	Exp:HMI	-0.05	0.13	-0.4	0.15		.695		
Reported problems during transitions ^c	TT_GER-HC	33	0.97	0.98	0	3	Intercept	-2.38	0.21	-11.54					
	TT_GER-LC	28	0.64	0.95	0	3	Exp	0.35	0.17	2.15	4.77		.029*	1,003 .028*	
	TT_USA-HC	21	0.62	0.67	0	2	HMI	0.34	0.17	2.05	4.41	1 (618)	.036*		
	TT_USA-LC	21	0.29	0.56	0	2	Exp:HMI	-0.09	0.17	-0.54	0.29		.588		

^a GLMM formula: DV ~ Exp*HMI + (1|TC) + (1|TP). The GLMM is fitted using the Laplace approximation. A type 3 ANOVA is calculated applying the LRT method.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

^c Due to convergence, the GLMM omits the random factor (1|TC).

7.3.2.2 Questionnaires

After completing the test drive, participants fill out a post-questionnaire with standardized questionnaires and self-developed questions (see Section 5.6). The questions refer to the participants' experience with the HMI without specifying scenarios or functions.

SUS

The *SUS* score is calculated from 10 items and ranges between 0 and 100 (Brooke, 1996). Figure 7.14 displays the results for the two experiments and the respective HMI subsamples.

The mean scores range between $M = 78.75$ (*TT_GER-LC*, $SD = 16.04$) and $M = 86.19$ (*TT_GER-HC*, $SD = 11.2$). One outlier in *TT_USA-LC* has a score of 17.5. The other minimum scores range between 42.5 (*TT_GER-HC*) and 47.5 (*TT_GER-LC*, *TT_USA-HC*). The maximum score ranges between 97.5 (*TT_GER-HC*) and 100 (*TT_GER-LC*, *TT_USA-HC*, & *TT_USA-LC*). Except for the outlier, the distribution of answers appears to be similar among the four subsamples. The ANOVA for the CLM results in no significant results for any factors. The TOST does not confirm equivalence for the factor *Exp* (Table 7.7).

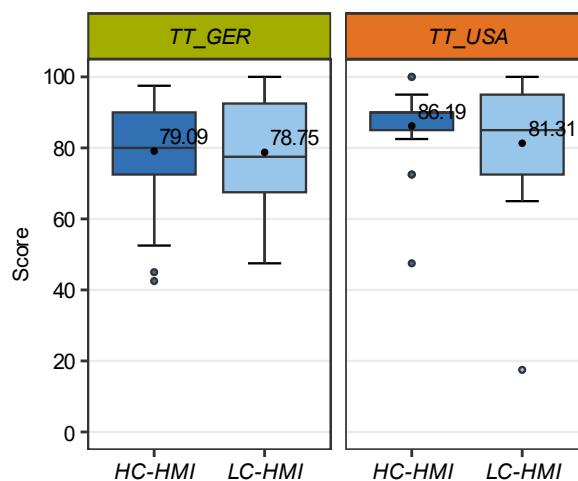


Figure 7.14 Boxplot diagram visualizing the results of the metric *SUS* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.7 for more statistics.

UMUX

The *UMUX* score is calculated from four items and ranges between 0 and 100 (Finstad, 2010). Figure 7.15 displays the results for the two experiments and the respective HMI subsamples.

The distribution of answers appears to be similar among the four subsamples. The mean *UMUX* scores are slightly higher for the *HC-HMI* subsamples than the *LC-HMI* subsamples. Additionally, the mean *UMUX* scores are slightly higher in the *TT_USA* subsamples than in the *TT_GER* subsamples. The mean scores range between $M = 77.83$ (*TT_GER-LC*, $SD = 15.67\%$) and $M = 86.31$ (*TT_USA-HC*, $SD = 10.8$). The minimum scores range between 41.67

(*TT_GER-HC*) and 54.17 (*TT_USA-HC*). The maximum score ranges between 95.83 (*TT_GER-HC*) and 100 (*TT_GER-LC*, *TT_USA-HC*, & *TT_USA-LC*). The ANOVA for the CLM results in no significant results for any factors. The TOST does not confirm equivalence for the factor *Exp* (Table 7.7).

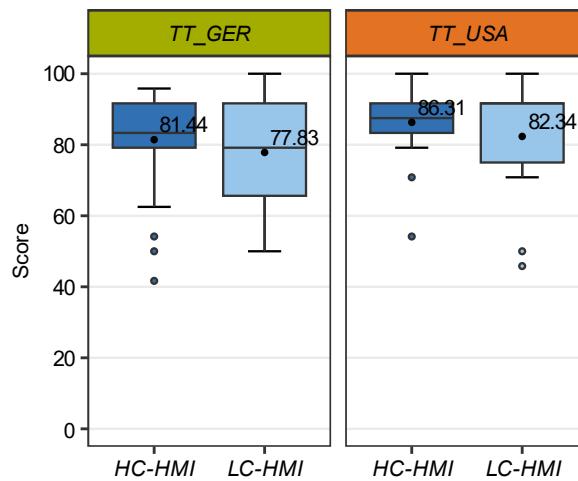


Figure 7.15 Boxplot diagram visualizing the results of the metric *UMUX* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.7 for more statistics.

UEQ

The *UEQ* comprises six subscales that result from four to six items each (Laugwitz et al., 2008). Figure 7.16 displays the results for all six subscales grouped by the four subsamples. The coloring in the figure marks the different evaluation categories (Schrepp, 2023): positive (green: > 0.8), neutral (yellow: between -0.8 and 0.8), and negative evaluation (red: < -0.8).

The descriptive data shows a tendency for higher *UEQ* mean scores (all dimensions) and smaller standard deviations (all dimensions) in the *TT_USA-HC* subsample compared to the *TT_GER-HC* subsample. In the *LC-HMI* subsamples, the tendency is not as pronounced. In four dimensions (*Perspicuity*, *Efficiency*, *Dependability*, & *Novelty*), the mean scores of *TT_USA-LC* are higher than the respective mean scores of *TT_GER-LC*. The dimensions *Attractiveness* and *Stimulation* result in similar mean ratings in both *LC-HMI* subsamples. For the standard deviations of the *LC-HMI* subsamples, no trend is observable. The ANOVA for the CLM results in non-significant results for all six subscales. Neither does the TOST confirm equivalence for the factor *Exp* for any of the six subscales (Table 7.7).

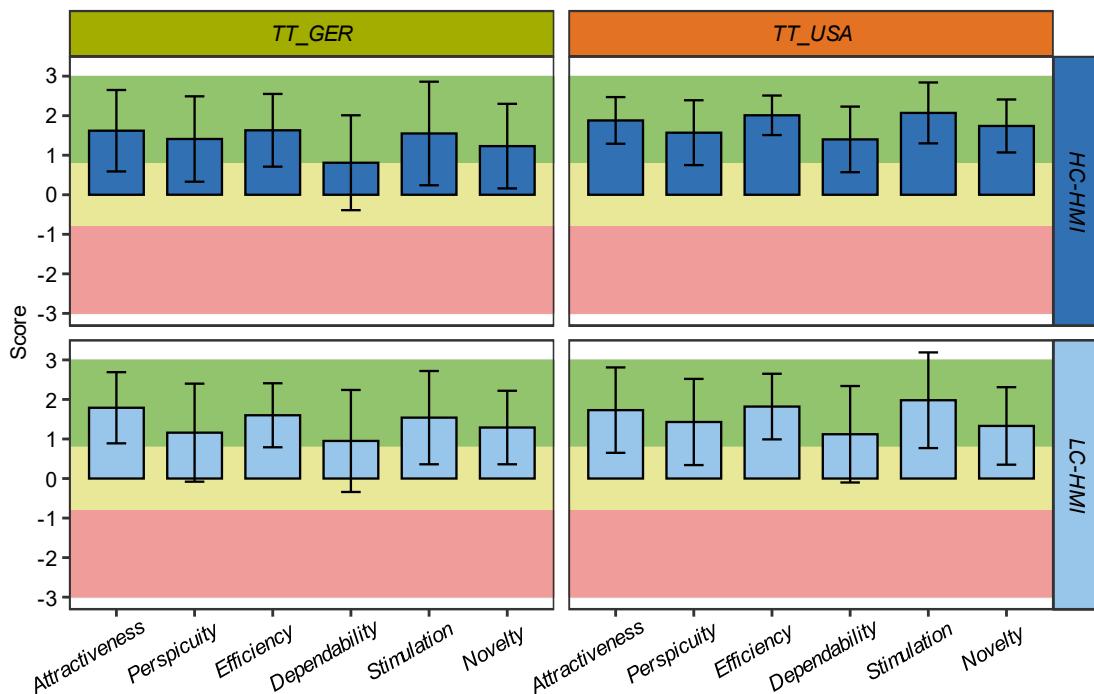


Figure 7.16 Bar chart visualizing the mean results of the metric *UEQ* with its six dimensions for the study *Exp_Culture*.

Note. The error bars display the *SD*. Refer to Table 7.7 for more statistics.

Trust

Trust is evaluated with a self-developed 1-item question: “I trusted the system I just used”. The response scale ranges between “1: strongly disagree” and “7: strongly agree”. Figure 7.17 displays the participants’ answers for the four subsamples.

Three of the four subsamples (*TT_GER-HC*, *TT_GER-LC*, & *TT_USA-LC*) use the full range of the response scale. The variance among the participants is high in all four subsamples with standard deviations between $SD = 0.91$ (*TT_USA-HC*) and $SD = 1.35$ (*TT_GER-LC*). The means of the subsamples range between $M = 3.54$ (*TT_GER-LC*) and $M = 3.86$ (*TT_USA-HC*). There are no prominent differences between the experiments. The ANOVA for the CLM results in no significant results for any factors. The TOST does not confirm equivalence for the factor *Exp* (Table 7.7).

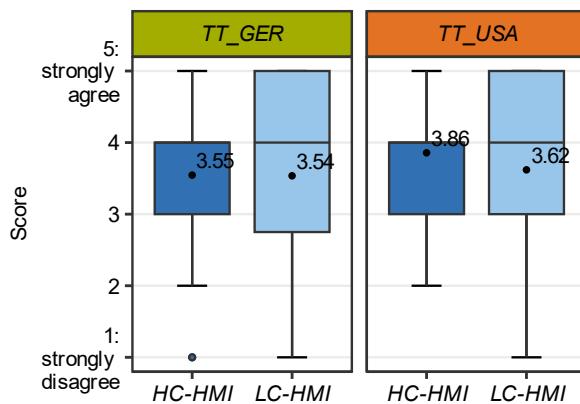


Figure 7.17 Boxplot diagram visualizing the results of the metric *Trust* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.7 for more statistics.

Acceptance

Acceptance is evaluated with a self-developed 1-item question: “If my car were fitted with a system like this, I'd use it when driving”. The response scale ranges between “1: strongly disagree” and “7: strongly agree”. Figure 7.18 displays the participants' answers for the four subsamples.

The *LC-HMI* subsamples use the full range of the response scale (one participant each with the answer “1: strongly disagree”). The variance among the participants is high in all four subsamples with standard deviations between $SD = 0.85$ (*TT_USA-HC*) and $SD = 1.24$ (*TT_USA-LC*). The means of the subsamples range between $M = 3.95$ (*TT_USA-LC*) and $M = 4.29$ (*TT_USA-HC*). In *TT_GER*, the mean score of the *HC-HMI* subsample ($M = 4$) is slightly lower compared to the *LC-HMI* subsample ($M = 4.14$). In *TT_USA*, the mean score of the *HC-HMI* subsample ($M = 4.29$) is slightly higher than the *LC-HMI* subsample ($M = 3.95$). The distribution of answers appears to be similar in all four subsamples. The ANOVA for the CLM results in no significant results for any factors. The TOST confirms equivalence for the factor *Exp* (Table 7.7).

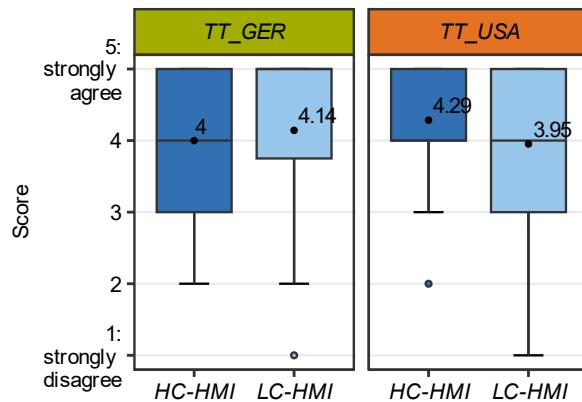


Figure 7.18 Boxplot diagram visualizing the results of the metric *Acceptance* for the study *Exp_Culture*.

Note. The mean values are displayed as numbers in the figure. Refer to Table 7.7 for more statistics.

Summary

Overall, the questionnaire results are similar among all four subsamples. In almost all metrics, neither the ANOVA for the CLM nor the equivalence test yields significant results for any factor. Only the metric *Acceptance* shows equivalence for the factor *Exp*. The interaction factor *Exp:HMI* is not significant for any metric. Regarding the effect of the HMI, a slight tendency for higher ratings in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples is observed.

Table 7.7 Summary table of the descriptive and inferential results of the questionnaires for the study *Exp_Culture*.

Metric	Subsample	Descriptive data						Factor	CLM & ANOVA ^a					TOST ^b	
		n	M	SD	Med	Min	Max		Est.	SE	z	X ² (1, 113)	p	W _{max_p}	p _{max}
SUS	TT_GER-HC	33	79.09	14.37	80	42.5	97.5	Exp	0.44	0.25	1.77	0.74	.39	1,478	.093
	TT_GER-LC	28	78.75	16.04	77.5	47.5	100		HMI	0.15	0.25	0.6	0.01	.933	
	TT_USA-HC	21	86.19	11.2	90	47.5	100		Exp:HMI	0.17	0.35	0.5	0.25	.62	
	TT_USA-LC	21	81.31	18.12	85	17.5	100								
UMUX	TT_GER-HC	33	81.44	13.86	83.33	41.67	95.83	Exp	0.41	0.25	1.62	1.43	.232	1,413	.188
	TT_GER-LC	28	77.83	15.67	79.17	50	100		HMI	0.29	0.25	1.15	1.08	.299	
	TT_USA-HC	21	86.31	10.8	87.5	54.17	100		Exp:HMI	-0.06	0.35	-0.17	0.03	.867	
	TT_USA-LC	21	82.34	15.36	91.67	45.83	100								
UEQ: Attractiveness	TT_GER-HC	33	1.62	1.03	1.67	-1.5	3	Exp	0.12	0.25	0.48	0	1.000	1,555	.595
	TT_GER-LC	28	1.79	0.9	1.92	0	3		HMI	-0.08	0.25	-0.32	0.37	.542	
	TT_USA-HC	21	1.88	0.59	1.83	0.83	3		Exp:HMI	0.17	0.35	0.48	0.23	.629	
	TT_USA-LC	21	1.73	1.08	1.83	-1.33	3								
UEQ: Perspicuity	TT_GER-HC	33	1.55	1.31	1.75	-1.75	3	Exp	0.52	0.25	2.05	2.77	.096	521	.143
	TT_GER-LC	28	1.54	1.18	1.75	-0.75	3		HMI	-0.05	0.25	-0.21	0.02	.878	
	TT_USA-HC	21	2.07	0.77	2.25	-0.25	3		Exp:HMI	-0.14	0.35	-0.42	0.17	.677	
	TT_USA-LC	21	1.98	1.21	2.5	-1.25	3								
UEQ: Efficiency	TT_GER-HC	33	1.63	0.92	1.75	-1.75	3	Exp	0.43	0.25	1.69	1.08	.298	691	.221
	TT_GER-LC	28	1.6	0.81	1.75	0	3		HMI	0.15	0.25	0.59	0.13	.716	
	TT_USA-HC	21	2.01	0.5	2	1	3		Exp:HMI	0.05	0.35	0.13	0.02	.897	
	TT_USA-LC	21	1.82	0.83	2	0.5	3								

Metric	Subsample	Descriptive data					Factor	CLM & ANOVA ^a					TOST ^b		
		n	M	SD	Med	Min		Est.	SE	z	X ² (1, 113)	p	W _{max_p}	p _{max}	
UEQ: Dependability	TT_GER-HC	33	1.41	1.08	1.5	-2	3	Exp	0.22	0.25	0.89	0.64	.425	1,628	.417
	TT_GER-LC	28	1.16	1.24	1.38	-1.25	3		0.18	0.25	0.74	0.64	.425		
	TT_USA-HC	21	1.57	0.82	1.75	0	3		0.10	0.35	-0.29	0.08	.773		
	TT_USA-LC	21	1.43	1.09	1.75	-1	3	Exp:HMI	-0.10	0.35	-0.29	0.08	.773		
UEQ: Stimulation	TT_GER-HC	33	1.23	1.07	1.5	-1.75	3	Exp	0.28	0.25	1.14	0	.963	667	.187
	TT_GER-LC	28	1.29	0.93	1.12	-1	3		0.29	0.25	1.16	0.01	.933		
	TT_USA-HC	21	1.74	0.67	2	0.25	3		0.38	0.35	1.06	1.13	.287		
	TT_USA-LC	21	1.33	0.98	1	-0.25	3	Exp:HMI	-0.10	0.35	-0.29	0.08	.773		
UEQ: Novelty	TT_GER-HC	33	0.81	1.2	1	-1.75	2.75	Exp	0.37	0.25	1.5	0.3	.586	1,424	.199
	TT_GER-LC	28	0.95	1.29	0.88	-2.25	3		0.06	0.25	0.24	0.13	.715		
	TT_USA-HC	21	1.4	0.83	1.25	-0.25	2.75		0.25	0.35	0.71	0.51	.475		
	TT_USA-LC	21	1.12	1.22	1	-1.5	3	Exp:HMI	-0.10	0.35	-0.29	0.08	.773		
Trust	TT_GER-HC	33	3.55	0.97	4	1	5	Exp	0.18	0.26	0.68	0	.961	1,505	.064
	TT_GER-LC	28	3.54	1.35	4	1	5		0.05	0.26	0.19	0.11	.744		
	TT_USA-HC	21	3.86	0.91	4	2	5		0.22	0.37	0.61	0.37	.544		
	TT_USA-LC	21	3.62	1.24	4	1	5	Exp:HMI	-0.10	0.35	-0.29	0.08	.773		
Acceptance	TT_GER-HC	33	4	0.9	4	2	5	Exp	0.04	0.27	0.14	0.56	.453	1,651	.006**
	TT_GER-LC	28	4.14	1.21	5	1	5		-0.07	0.27	-0.26	1.39	.238		
	TT_USA-HC	21	4.29	0.85	4	2	5		0.48	0.38	1.26	1.59	.207		
	TT_USA-LC	21	3.95	1.24	4	1	5	Exp:HMI	-0.10	0.35	-0.29	0.08	.773		

^a CLM formula: DV ~ Exp*HMI. A type 3 ANOVA is calculated with Wald chi-square tests.

^b The TOST applies the Wilcoxon rank sum test with continuity correction. The smallest effect size of interest is set to d = 0.5.

7.3.2.3 Final Interview

At the end of the post-questionnaire participants, are requested to reflect on the experienced HMI. Participants may praise or criticize components of the HMI or make improvement suggestions. An overview of the clustered replies of the participants of the four subsamples is attached in Appendix III (Figure 14.8–Figure 14.10).

Participants of *TT_GER* mainly praise the easy handling (*TT_GER-HC*: 15.15% & *TT_GER-LC*: 32.14%) and the simple and clear design (*TT_GER-HC*: 18.18% & *TT_GER-LC*: 28.57%). Only single participants of *TT_USA-HC* (9.52%) and no participants of *TT_USA-LC* praise the simple and clear design. Instead, considerably more participants of *TT_USA* praise the easy handling (*TT_USA-HC*: 52.38% & *TT_USA-LC*: 76.19%). In contrast to participants of the *LC-HMI* subsamples, participants of the *HC-HMI* subsamples additionally praise the usage of sounds (*TT_GER-HC*: 15.15% & *TT_USA-HC*: 19.05%) and LED lights (*TT_GER-HC*: 12.12% & *TT_USA-HC*: 19.05%). Single participants in the four subsamples mention the color selection and the icon design. A typical statement of the participants is: “*symbols were easy to understand after a short familiarization period, and the system was very tidy and not cluttered*”³² (*TT_GER-LC_{TP59}*).

Several participants in all four subsamples criticize the control logic (between *TT_GER-LC*: 10.71% & *TT_USA-HC/TT_USA-LC*: 19.05%) and the insufficient salience of notifications (more prevalent among *LC-HMI* subsamples; *TT_GER-HC*: 6.06%, *TT_GER-LC*: 10.71%, *TT_USA-HC*: 9.52%, & *TT_USA-LC*: 23.81%). In contrast to the *HC-HMI* subsamples, participants of both *LC-HMI* subsamples criticize the missing sounds (*TT_GER-LC*: 21.43% & *TT_USA-LC*: 28.57%). Additionally, single participants in the four subsamples criticize that the display duration of notifications is too short. Single participants of *TT_GER* criticize that the labels of the control buttons are unclear. The position of notifications in the right area of the display is criticized by 10.71% of the participants in *TT_GER-LC*. In contrast, none of the participants in the other three subsamples mention this issue. A typical statement of the participants is, for example, “*The instructions to changes should be backed up with a sound, since instructions can be missed when answering emails, for example*”³³ (*TT_GER-LC_{TP61}*).

When asked for improvement suggestions, the wish for more sounds, light, and haptic signals is expressed most often. Considerably more participants in *TT_USA* compared to participants in *TT_GER*, and more participants in *LC-HMI* subsamples compared to participants in *HC-HMI* subsamples express this wish (*TT_GER-HC*: 12.12%, *TT_GER-LC*: 39.29%, *TT_USA-HC*: 52.38%, & *TT_USA-LC*: 71.43%). Participants in *TT_USA* wish to improve the control logic more often than participants in *TT_GER* (*TT_GER-HC*: 6.06%, *TT_GER-LC*: 7.12%, *TT_USA-HC*: 33.33%, & *TT_USA-LC*: 14.29%). The same observation is true for improvement suggestions regarding the overall design (*TT_GER-HC*: 6.06%, *TT_GER-LC*: 0%, *TT_USA-HC*: 19.05%, & *TT_USA-LC*: 23.81%). In *TT_GER-LC*, 14.29% of the participants wish for a central positioning of notifications and an improvement in the color

³² Translated from German statement: „*Symbole waren nach kurzer Eingewöhnungszeit gut verständlich, das System sehr aufgeräumt und nicht überladen.*“

³³ Translated from German statement: „*Die zu ändernden Anweisungen sollten mit einem Ton hinterlegt werden, da Hinweise z.B. beim E-Mails beantworten untergehen können.*“

selection. Furthermore, 10.71% of the participants of *TT_GER-LC* wish to increase the salience and urgency of RTIs. None of the participants of the other three subsamples mention one of these suggestions. A typical statement of the participants is, for example, “*I would have liked more sound alerts to draw my attention back to the steering wheel if a problem with the system occurs*” (*TT_USA-LC_{TP61}*).

Differences between the *HC-HMI* subsamples and the *LC-HMI* subsamples are evident in the interviews. Furthermore, participants of *TT_USA* criticize the overall design more often and express wishes for more sound, light, and haptic signals more often than participants of *TT_GER*.

7.3.3 Interindividual Factors

Two 1-item questions are posed to assess whether there is a difference between the experiments regarding *Nausea* and *Effort*, respectively. The questions are answered on a 5-point Likert scale ranging from “1: not at all strenuous/nauseous” to “5: very strenuous/nauseous”. The analysis does not distinguish between participants of the *HC-HMI* and the *LC-HMI* subsamples. Table 7.8 summarizes the participants’ answers to both questions.

Table 7.8 Distribution of responses for the interindividual factors *Nausea* and *Effort* for the experiments *TT_GER* and *TT_USA*.

Exp (n)	<i>Nausea [% (n)]</i>					<i>Effort [% (n)]</i>				
	1	2	3	4	5	1	2	3	4	5
<i>TT_GER</i> (61)	98.36 (60)	1.64 (1)	0 (0)	0 (0)	0 (0)	88.52 (54)	9.84 (6)	0 (0)	1.64 (1)	0 (0)
<i>TT_USA</i> (42)	95.24 (40)	2.38 (1)	2.38 (1)	0 (0)	0 (0)	78.57 (33)	11.90 (5)	7.14 (3)	2.38 (1)	0 (0)

Note. The scale ranges from “1: not at all nauseous/strenuous” to “5: very nauseous/strenuous”.

At the end of the final questionnaire, participants rate whether the turning after every test case has made them nauseous.

The mean scores are low and similar in both experiments (*TT_GER*: $M = 1.02$, $SD = 0.13$ & *TT_USA*: $M = 1.07$, $SD = 0.34$). Only single participants in both experiments report having a *Nausea* score higher than 1 (“not at all nauseous”). A Wilcoxon rank sum test for difference shows no significant difference between *TT_GER* ($Med = 1$) and *TT_USA* ($Med = 1$), with $p = .357$ ($W = 1,240.5$). The Wilcoxon rank sum test for equivalence is significant, with $p < .001$ ($W = 2,441$).

At the end of the final questionnaire, participants rate whether the turning after every test case has been strenuous for them.

The mean score in *TT_USA* ($M = 1.33$, $SD = 0.72$) is slightly higher than in *TT_GER* ($M = 1.15$, $SD = 0.48$). A Wilcoxon rank sum test for difference shows no significant difference between *TT_GER* ($Med = 1$) and *TT_USA* ($Med = 1$), with $p = .150$ ($W = 1,145.5$). The Wilcoxon rank sum test for equivalence is significant, with $p < .001$ ($W = 1,283$).

7.4 Discussion

This section starts with a summary of the results leading to answering the hypotheses. Afterward, limitations and other observations on the experimental design are reflected. The section closes with the conclusion of the study's results.

7.4.1 Summary of Results

The observational data comprise driving behavior data, eye-tracking data, and experimenter ratings. Only data points that fulfill the requirements could be included, for example, the correct LoA at the scenario start or not looking at the *IC* during the start of an RtI. Thus, the database for driving behavior and eye-tracking is reduced, leading to a lower statistical power. More data points in the *LC-HMI* subsamples are excluded than in the *HC-HMI* subsamples.

The general driving behavior and the driving behavior in specific situations (the first activation and RtIs) are assessed. The inferential and descriptive analyses show that the driving performance is considerably better in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples. Overall, differences between the experimental conditions are only little and non-significant in the inferential tests. The equivalence test is only significant for the metric *Observed LoA vs. instructed LoA*. There is a tendency that in *TT_USA*, the TOTs after RtIs are slightly higher than *TT_GER*. The database is inconclusive since the GLMM and the equivalence test for the factor *Exp* are non-significant. Overall, there are no interactions in the GLMMs or observable in the descriptive analysis. In the RtIs, differences in the take-over strategies could be observed, with participants of *TT_USA* tending to use the brake more often than participants of *TT_GER*.

The eye-tracking data are assessed in test cases with continuous rides in specific LoA and during RtIs. The factor *HMI* is significant only for the metric *Attention ratio during continuous rides in L0, L2, & L3: SuRT*. The descriptive analysis of gaze behavior shows differences between the HMI subsamples, for example, an increase of glances to the *IC* after the first RtI (*RtI_{20s}*) in the *LC-HMI* subsamples or the shorter glance allocation times to the *IC* in RtIs in the *HC-HMI* subsamples. Additionally, the descriptive analysis of the attention ratios and the gaze paths during RtIs show differences in the gaze behavior for the factor *Exp*. In the inferential tests, the factor *Exp* is significant for two metrics (*Attention ratio during continuous rides in L0, L2, & L3: SuRT & Glance allocation time to IC after RtI*), reflecting better performance scores for *TT_GER* than *TT_USA*. Equivalence for the factor *Exp* is confirmed only for the metric *First glance duration on IC after RtI*. The interaction factor *Exp:HMI* is not significant for any metric.

The *Experimenter rating* shows a significant difference between the HMI subsamples. The variance in the subsample *TT_GER-LC* is considerably higher than the variances of the other three subsamples. Neither the factor *Exp* nor the interaction factor *Exp:HMI* is significant. Instead, the equivalence test is significant.

The descriptive results of the short interviews imply a tendency for better interaction scores in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples, where variances tend

to be greater and mean interaction scores lower. In the inferential statistical analyses, the factor *HMI* is significant in three out of six tests. Regarding the influence of the experiment, equivalence is confirmed in four of the six inferential statistical tests, and two of the metrics show a significant influence of the factor *Exp* (metric *Reported problems during transitions* is significant in the equivalence test & the binomial GLMM). Participants of *TT_GER* report more problems in total and more problems referring to other issues than the control logic of the HMI concepts. There are no interactions.

The questionnaires show a high variance within the four subsamples, decreasing the power of the inferential statistical tests. The distributions appear to be similar among all subsamples. In almost all metrics, neither the ANOVA for the CLM nor the equivalence test yields significant results for any factor. Only the metric *Acceptance* shows equivalence for the factor *Exp*. The interaction factor *Exp:HMI* is not significant for any of the metrics. The descriptive analysis shows a slight tendency for higher ratings in the *HC-HMI* subsamples compared to the *LC-HMI* subsamples. The results suggest that the database is insufficient to conclude whether effects for the factor *Exp* are present. However, the results imply that the potential effects are minor.

In the *Final interview*, the differences in the HMI design are reflected in the participants' comments differing between the HMI subsamples, for example, criticizing the missing sounds and overall salience of notifications in the *LC-HMI* concept. Participants of *TT_USA* criticize the overall design more often and express wishes for more sound, light, and haptic signals considerably more often than participants of *TT_GER*. Otherwise, the participants' answers are very informative and valuable for future development efforts but do not differ much between the four subsamples.

7.4.2 Discussion of Hypotheses

Two hypotheses on relative and absolute validity regarding observational and self-reported metrics are formulated.

- H₁ Relative and absolute validity are demonstrated for cross-cultural research between the United States and Germany compared to the test track setting regarding observational metrics for assessing the usability of HMIs for L3 ADS.
- H₂ Relative and absolute validity are not demonstrated for cross-cultural research between the United States and Germany compared to the test track setting regarding self-reported metrics for assessing the usability of HMIs for L3 ADS.

Equivalence for the factor *Exp* is confirmed for about half of the observational metrics. While the *Experimenter rating* yields similar results for both experimental conditions, differences are observed in the driving behavior and are even more pronounced in the gaze behavior. Descriptive analyses indicate differences in strategies reflected in the gaze behavior or the take-over strategies. However, the analyses do not paint a clear picture, with single metrics indicating less trust or system understanding (e.g., *Attention ratio during continuous rides in L0, L2, & L3: SuRT*) or lower performance scores (e.g., *Glance allocation time to IC after RTI*) in *TT_USA* compared to *TT_GER*. Nonetheless, differences between the HMI concepts could be identified in most of the metrics in both inferential and descriptive analyses.

The database is insufficient to conclude the matter of validity. The rather small power of the inferential statistical tests—increased by exclusions of single data points due to wrong prerequisites—limits the confidence with which a statement on relative validity may be drawn. However, both the inferential statistical tests and the descriptive analyses do not show systematic differences between the experimental conditions. Furthermore, there are no interactions between factors *Exp* and *HMI*. Regarding the first hypothesis, relative validity for observational metrics may be assumed, while no conclusion on absolute validity shall be drawn. Hypothesis H₁ is only partially confirmed.

The results of the self-reported metrics appear to be inconclusive. The questionnaires produce almost no significant results (except for the equivalence test of the metric *Acceptance*), and only tendencies are observable for differences between the HMI concepts. Regarding differences between the experiments, only the *UEQ* shows a tendency (descriptive only) for higher ratings in *TT_USA* compared to *TT_GER*. The short interviews and the *Final interview* demonstrate more distinctly the differences between the HMI concepts. Here, small differences between the experiments are observable, for example, differing performance scores and ratings for *TT_GER* and *TT_USA*, though without a clear direction. In the *Final interview*, participants focus on different aspects of their evaluation, which aligns with previous research (e.g., Roessger, 2003). As for the observational metrics, there are no interactions between factors *Exp* and *HMI*. Again, the rather small power of the inferential statistical tests—due to high variances—limits the confidence with which a statement on the relative validity may be drawn. Concluding, both the inferential statistical tests and the descriptive analyses do not show systematic differences between the experimental conditions. Therefore, the conclusion for the hypothesis H₂ is the same as for the observational metrics: Relative validity for self-reported metrics may be assumed. In contrast, no conclusion on absolute validity shall be drawn. Consequently, hypothesis H₂ is partially rejected.

The results do not contradict previous research in Section 2.4, indicating that cultural differences between Western countries are less pronounced and may even show in specific metrics only (e.g., Heimgärtner, 2007).

A third hypothesis is formulated on differences in the usability assessment between the HMI concepts:

H₃ The concept *HC-HMI* receives higher usability evaluations than the concept *LC-HMI*.

In several metrics, differences between the HMI concepts demonstrate a higher usability of the *HC-HMI* than the *LC-HMI*. Only single metrics of the self-reported data reflect differences in the HMI concepts. All descriptive and most inferential analyses of the observational data indicate differences in the HMI concepts. Overall, the effect is—just as in the validation study *Exp_Testing-Environment*—smaller than expected after the unambiguous heuristic expert evaluation conducted in the HMI development process (see Subsection 5.4.3). Following the framework of Bengler et al. (2020), the input channel and the dialog logic are identical in both concepts. The output channel differs in the number of included modalities and the design of the visual output channel. Furthermore, the information content is identical in both concepts. The two-factor theory of Herzberg et al. (1967) may be transferred to this observation to illustrate the limited variance of the two concepts. The covering of the

information demand takes the role of hygiene factors. According to Herzberg et al. (1967, pp. 113–114), the absence of hygiene factors causes dissatisfaction, but the presence of hygiene factors does not increase satisfaction. The presentation of the information and the HMI's overall design take the motivators' role. These factors increase the satisfaction (Herzberg et al., 1967, pp. 113–114). The two HMI concepts do not differ in the hygiene factors, and the variance in the motivators is limited. Thus, the observed results of less clear differences between the HMI concepts may be explained. Considering this limitation, the hypothesis H₃ is confirmed.

7.4.3 Discussion of Limitations and Other Observations

The study compares usability assessments of two cultures. Furthermore, these cultures belong to Western industrialized countries and resemble each other more than other countries regarding their cultural values (Hofstede, 2011; Hofstede Insights, 2023). Due to the COVID-19 pandemic, the experiment *TT_USA* is conducted in Germany and includes U.S.-American citizens currently based in Germany. The assignment of the participants to the U.S.-American culture is checked via the *VSM* by Hofstede and Minkov (2013b). The analysis concludes that it may be assumed that the *TT_USA* represents the United States regarding its cultural values (see Subsection 8.3.2 & Section 8.4).

The smaller or non-present effects of differences between the HMI concepts aggravate the derivation of statements regarding relative validity, which is concerned with the direction and magnitude of effects. The statistical power is further lowered due to limitations in the data availability: due to crashes, technical problems, data quality issues, and exclusions (e.g., wrong LoA at scenario start), the number of data points is considerably reduced in the observational metrics, mainly affecting the *LC-HMI* subsamples. This issue is partly compensated by the descriptive data analysis stressing the importance of different data types and sources.

The database is not sufficient to include confounding factors present in both experiments (e.g., extreme weather conditions such as glare or heavy rain) in the statistical models. However, these factors are considered in the inspection of outliers.

Equivalence tests could not be conducted with subsamples but with total samples (*TT_GER* vs. *TT_USA*), only ignoring the variance through different HMI concepts. Tests with subsamples, for example, *TT_GER-LC* versus *TT_USA-LC*, could not be conducted due to the small sample sizes limiting the statistical power.

The *Experimenter rating* and the *Final interview* could be affected by the experimenter's awareness of the *HMI* and *Exp* condition of the respective participants. To a lesser magnitude, the participants' behavior and, therefore, other metrics could also be unconsciously affected by the experimenter's knowledge (e.g., Rosenthal effect, see Bortz & Döring, 2006, pp. 82–83) of the experimental conditions.

In contrast to the experiments' language in *TT_GER* (German), the experimenter in *TT_USA* is not a native speaker (U.S.-American English). As shown in a study by Vatrapu and Pérez-Quiñones (2006), this may have affected the extent and content of self-reported data. An agency translated the materials. Nonetheless, misunderstandings of single terms (e.g., “emergency brake”) could not be entirely prevented. An effect on specific metrics, such as *TOT after RtI* or self-reported metrics covering understandability or satisfaction, is possible.

An agency recruited U.S.-Americans living in Germany for a maximum of eight years before the experiment. Due to the COVID-19 pandemic, travel and work visits were limited; therefore, no stricter recruitment criteria could be applied. The samples are possibly drawn from differing populations in experiments *TT_GER* and *TT_USA*. Due to similar sample characteristics regarding sociodemographic features and driving behavior of the subsamples, a potential effect is considered negligible. However, unprompted comments of participants reveal potentially important sample characteristics that are not considered in the current experiment: Single participants of *TT_USA* describe that they transferred prior knowledge of experiences with Tesla vehicles to the interaction with the tested HMI concepts, thus posing a confounding variable that is not systematically collected. Another potential confounding variable is the familiarity with vehicles featuring automatic transmission, which is more common in the United States than Germany. Thus, the different degrees of familiarity with automatic transmission potentially reduce the cognitive demand and thereby increase the subjective well-being of participants in *TT_USA* compared to *TT_GER*.

Regarding the study procedure, participants report feeling exhausted due to the duration of the experiment and the extensive instructions. At the same time, participants note that the study setting is too simple because of low speeds and missing obstacles, interactions with other road users, or curves.

Several participants expressed their excitement for automated driving before, during, and after the experiment (and their privilege to be part of the development process by partaking in this study). Participants even answer questions about the HMI design with general statements about automated driving. Participants seem to not separate these feelings of excitement from the interaction with an HMI and provide generally high/positive ratings. Participants seem to tolerate flaws and problems (to some degree) when the research subject is obviously in a prototype state. The overall scores in the standardized questionnaires are high for both HMI concepts. In the *Final interview*, participants mention different aspects in their critique, but their overall rating is similar and positive for both HMI concepts.

Another aspect is that several times, participants activate the wrong LoAs without noticing their mistake. Because of missing feedback from the system or the experimenter, they could not consider these problems in the ratings but refer to their self-estimated performance when evaluating the HMI. Mistakes may be left unconsidered if no feedback is provided (Drew et al., 2018). No clear decision could be made for overall satisfaction in the between-subject experiments. If participants had compared the two HMI concepts, the satisfaction ratings may have shown clear differences. During the planning of the experimental design, a within-subject design was excluded because of learning effects ruling out testing naïve participants. Furthermore, the study duration of a between-subject design is considerably shorter.

The observations described in this subsection will be used to derive recommendations for the experimental method in Chapter 10.

7.4.4 Conclusion

The validation study *Exp_Culture* confirms relative validity for observational data (H_1 partially confirmed) and self-reported data (H_2 partially rejected). The database is inconclusive

to conclude absolute validity for observational or self-reported data. Based on the empirical data available, it is believed that absolute validity is likely not to be achieved in potential future studies, especially for studies conducted with more diverse cultures. Therefore, more cultures should be examined that show greater differences in their cultural values than Germany and the United States. However, the study provides valuable insights into the study design for usability assessments of HMIs for L3 ADS. The study's results confirm differences between the HMI concepts (H_3 confirmed). However, limitations of specific metrics detecting differences in HMI concepts are identified.

This study may draw a practical conclusion: Usability assessments for HMIs for L3 ADS may be conducted within one culture of the Western industrialized world. If the focus is on the refinement of design elements mainly contributing to the facet satisfaction, the transferability of results is expected to be limited.

8 Validation Study Survey_Culture: Effect of the Users' Cultural Background on the Subjective Importance Rating of Usability Factors in the Context of HMIs for L3 ADS

The validation study *Survey_Culture* examines the effects of the users' cultural background on the self-reported importance of different usability factors in the context of HMIs for L3 ADS. Self-reported data on cultural values and importance ratings are collected via online surveys. The samples are drawn from the German population, the U.S.-American population (currently residing in the USA), and from the experiment *TT_USA* conducted with U.S.-American participants in Maisach, Germany. The data of the U.S.-American participants in Maisach is collected in July and August 2021. The data from the other two samples is collected between December 2021 and February 2022. This chapter aims to answer research question RQ₄ and thus provides more insights into the importance of culture in usability testing. Furthermore, the data are used to check whether the U.S.-American sample in *TT_USA* represents the United States regarding its cultural values (see Section 7.2).

8.1 Hypotheses

The validation study *Survey_Culture* seeks answers to research question RQ₄. The literature presented in Section 2.4 confirms the existence of cultural differences. Cross-cultural studies suggest that differences are more pronounced between Western countries and Asian countries compared to differences between Western countries. Nonetheless, a previous study by Roessger (2003) identifies differences between samples from the United States and Germany regarding expectations and aesthetics in interface design—usability-related aspects.

Differences in the preferences of usability aspects are examined in this study. A list of usability factors provided by Hinderks et al. (2019) is applied in this study (see Subsection 8.2.2). Following the proposed mapping of cultural dimensions to usability criteria by Sogemeier et al. (2022), the following tendencies for subjective importance ratings of usability factors can be expected: *Long Term Orientation* is significantly more pronounced in Germany compared to the United States (Hofstede Insights, 2023). Consequently, participants from the United States are expected to show stronger preferences for factors relating to processes that can be influenced independently (Sogemeier et al., 2022), such as the usability factor *Flexibility* (Hinderks et al., 2019, see Table 8.2). The dimensions *Individualism*, *Uncertainty Avoidance*, and *Indulgence vs. Restraint* have slightly differing scores in the United States compared to Germany; that is, the United States show more pronounced *Individualism*, more pronounced *Indulgence*, and less pronounced *Uncertainty Avoidance* compared to Germany (Hofstede Insights, 2023). Thus, participants from Germany might show stronger preferences for simple and organized interfaces that meet the participants' expectations, while participants from the United States might have stronger wishes for options for customization and individualization (Sogemeier et al., 2022). These differences may be expressed in higher subjective importance ratings of the factors *Appearance/Attractiveness*, *Emotion/Affect*, *Fun*, *Identity*, and *Stimulation* (Hinderks et al., 2019, see Table 8.2) for participants from the United States compared to participants from Germany. Furthermore, participants from Germany might emphasize the factors *Controllability/Dependability* and

Trust/Credibility (Hinderks et al., 2019, see Table 8.2) compared to participants from the United States. As the differences in the cultural values of the respective dimensions are only minor, differences in the importance ratings of the factors are expected to be minor, too. The dimensions *Power Distance* and *Masculinity* have similar scores in both cultures (Hofstede Insights, 2023). Therefore, no expectations toward differing preferences are derived from the mapping between cultural dimensions and usability criteria (Sogemeier et al., 2022).

Overall, the cultural values between the United States and Germany are similar. Therefore, no substantial differences in the importance ratings of the usability factors are expected. Nevertheless, small differences in the factors mentioned above are predicted based on the literature findings and the model for mapping cultural values to usability criteria by Sogemeier et al. (2022). Therefore, the following hypotheses for the validation study *Survey_Culture* are formulated:

RQ₄ Which effect has the users' cultural background on the subjective importance rating of usability factors in the context of HMIs for L3 ADS?

- H₁ The factor *Flexibility* receives considerably higher importance ratings from participants from the United States than from Germany.
- H₂ The factors *Appearance/Attractiveness*, *Emotion/Affect*, *Fun*, *Identity*, and *Stimulation* receive slightly higher importance ratings from participants from the United States than from Germany.
- H₃ The factors *Controllability/Dependability* and *Trust/Credibility* receive slightly higher importance ratings from participants from Germany than from the United States.

8.2 Method

The survey method is presented in the following section, covering the sample, the metrics, the study procedure, and the data analysis. The goal is to gain insights into the relationship between cultural values and individual ratings of the importance of usability factors in the context of HMIs for L3 ADS. Unless stated otherwise, the translation of survey materials is conducted by a team at the Chair of Ergonomics following the procedure of Jones et al. (2001).

8.2.1 Sample

The study comprises three samples: *TT_USA*, *ON_USA* and *ON_GER*. The sample *TT_USA* partakes in the validation study *Exp_Culture* presented in the previous chapter (see Chapter 7). In addition to the metrics presented in Section 5.6, the data presented in the following subsection is collected. The other two samples only participate in this study via online surveys. The sample *ON_USA* is recruited via Amazon Mechanical Turk (Amazon Mechanical Turk, Inc., 2023), a crowdsourcing marketplace widely used in social sciences to recruit participants (Cheung et al., 2017). The participation requirements ensure that residency in the United States and citizenship as a U.S.-American are fulfilled. The participants receive an incentive of 1 USD as compensation. The sample *ON_GER* is recruited via social media platforms and the Chair of Ergonomics' participant database.

Sociodemographic data are collected to describe the sample and evaluate its representativity. In addition to age and gender, participants provide data on their citizenship.

Regarding the driving background, participants report the mileage and driving frequency of the last 12 months. Afterward, participants indicate their familiarity with the ADAS CC, ACC, and LKA. If participants report familiarity with specific ADAS, a subsequent question inquires on the frequency of using the ADAS. Finally, participants are requested to report their prior knowledge of automated driving on a 5-point Likert scale with the anchors “0: no knowledge” and “4: expert”.

The total sample of the validation study consists of $N = 110$ participants. In *TT_USA*, $n = 42$ participants are included in the sample (male: $n = 23$; female: $n = 19$; diverse: $n = 0$). In *ON_USA*, $n = 30$ participants are included in the sample (male: $n = 20$; female: $n = 9$; diverse: $n = 1$). In *ON_GER*, $n = 38$ participants are included in the sample (male: $n = 28$; female: $n = 9$; diverse: $n = 1$). The mean age ranges between 38.14 (*TT_USA*, $SD = 9.8$) and 51.08 (*ON_GER*, $SD = 21.99$). The minimum age of the participants is 20 (*TT_USA*), and the maximum is 80 (*ON_GER*). While the age distribution of *TT_USA* and *ON_USA* are similar, the mean age and the standard deviation of *ON_GER* are considerably higher.

The descriptive analysis of the driving background is summarized in Table 8.1. The driving frequency is slightly higher in the sample *ON_USA* (“every day” or “several times a week”: 86.67%) compared to *ON_GER* (“every day” or “several times a week”: 71.05%) and considerably higher compared to *TT_USA* (“every day” or “several times a week”: 52.38%). The mileage of *TT_USA* is lower (“< 5,000 km”: 50%) compared to the other two samples (“< 5,000 km”: *ON_USA* = 50%; *ON_GER* = 28.95%). The experience with ADAS and the frequency of using the systems is similar among the study samples. The reported prior knowledge in the field of automated driving is slightly lower in *TT_USA* compared to the other two samples. Only single participants in all three study samples indicate to have expert knowledge.

Table 8.1 Summary table of the descriptive analysis of the metrics on the driving background for the study *Survey_Culture*.

Metric	Response	Proportion [% (n)]		
		TT_USA (42)	ON_USA (30)	ON_GER (38)
<i>Frequency of driving (12 months)</i>	Every day	16.67 (7)	46.67 (14)	18.42 (7)
	Several times a week	35.71 (15)	40 (12)	52.63 (20)
	Several times a month	14.29 (6)	0 (0)	10.53 (4)
	Less than once a month	30.95 (13)	3.33 (1)	18.42 (7)
<i>Mileage (12 months)</i>	Never	2.38 (1)	10 (3)	0 (0)
	> 20,000 km	4.76 (2)	6.67 (2)	2.63 (1)
	10,001 km-20,000 km	14.29 (6)	26.67 (8)	26.32 (10)
	5,001 km-10,000 km	30.95 (13)	40 (12)	42.11 (16)
<i>No ADAS experience</i>	< 5,000 km	50 (21)	26.67 (8)	28.95 (11)
	Yes	4.76 (2)	16.67 (5)	10.53 (4)
	No	95.24 (40)	83.33 (25)	89.47 (34)
	Several times a day	0 (0)	6.67 (2)	7.89 (3)
<i>Usage frequency (12 months): CC</i>	Every day	7.14 (3)	0 (0)	5.26 (2)
	Every week	11.9 (5)	20 (6)	13.16 (5)
	Every month	23.81 (10)	10 (3)	10.53 (4)
	Seldom	50 (21)	33.33 (10)	34.21 (13)
<i>Usage frequency (12 months): ACC</i>	Never	2.38 (1)	6.67 (2)	13.16 (5)
	No prior experience	4.76 (2)	23.33 (7)	15.79 (6)
	Several times a day	2.38 (1)	3.33 (1)	2.63 (1)
	Every day	0 (0)	3.33 (1)	5.26 (2)
<i>Usage frequency (12 months): LKA</i>	Every week	2.38 (1)	10 (3)	7.89 (3)
	Every month	9.52 (4)	6.67 (2)	7.89 (3)
	Seldom	19.05 (8)	10 (3)	13.16 (5)
	Never	2.38 (1)	0 (0)	10.53 (4)
<i>Prior knowledge in the field of automated driving</i>	No prior experience	64.29 (27)	66.67 (20)	52.63 (20)
	Several times a day	2.38 (1)	3.33 (1)	2.63 (1)
	Every day	4.76 (2)	6.67 (2)	15.79 (6)
	Every week	2.38 (1)	3.33 (1)	15.79 (6)
<i>Usage frequency (12 months): LKA</i>	Every month	7.14 (3)	6.67 (2)	2.63 (1)
	Seldom	28.57 (12)	10 (3)	18.42 (7)
	Never	2.38 (1)	6.67 (2)	10.53 (4)
	No prior experience	52.38 (22)	63.33 (19)	34.21 (13)
<i>Prior knowledge in the field of automated driving</i>	4: expert	4.76 (2)	6.67 (2)	5.26 (2)
	3	9.52 (4)	30 (9)	26.32 (10)
	2	28.57 (12)	16.67 (5)	18.42 (7)
	1	40.48 (17)	30 (9)	23.68 (9)
<i>Prior knowledge in the field of automated driving</i>	0: no prior knowledge	16.67 (7)	16.67 (5)	26.32 (10)

Note. The mode values of each metric are indicated in bold.

8.2.2 Data Collection

In addition to sociodemographic data and the data on the driving background to describe the sample, data on the cultural values and the usability factors are collected.

As presented in Subsection 2.1.5, different models aim to make cultural values and categorization measurable. In this study, the model of cultural values by Hofstede is applied (Hofstede, 2011; see Subsection 2.1.5 for more details). The *VSM* (Hofstede & Minkov, 2013b) is applied to gain information on the cultural values of the samples. It consists of 24 items with differing response scales in Likert format. The questionnaire is available in both German and English.

In a study by Hinderks et al. (2019), the authors identify 25 usability aspects listed in Table 8.2. In this study, the authors provide a context and request that participants rate the importance of each of the 25 factors on a 7-point Likert scale from “-3: not important at all” to “+3: very important”. This procedure is repeated here with an adaptation of the context. The context information requests the participant to imagine a situation where he or she drives a car equipped with an ADS comprising three LoAs: L0, L2, and L3. Further information on the ADS explains that an HMI indicates the active LoA and the availability of the LoAs L2 and L3. The participant is informed that transitions between the LoAs may be conducted via the HMI voluntarily or upon the ADS' request. Explanations and examples complement the context information. Finally, the participant is requested to indicate for each factor how important the factor or product quality is for him/her for the just described HMI.

8.2.3 Study Procedure

The survey starts with the informed consent of the participant. After that, the participant provides sociodemographic data, followed by the *VSM* (Hofstede & Minkov, 2013b) and the questionnaire on the importance rating of the usability factors (Hinderks et al., 2019). The survey closes with a free text field for comments. In *TT_USA*, the survey additionally serves as the prequestionnaire for the validation study *Exp_Culture*. Thus, the participant information for the informed consent differs slightly from that for *ON_GER* and *ON_USA*. The duration of the survey is about 15 min.

8.2.4 Data Analysis

The analysis of the data is performed on a descriptive basis. The responses for the *VSM* (Hofstede & Minkov, 2013b) are used to calculate scores for the six dimensions *Power Distance*, *Individualism*, *Masculinity*, *Uncertainty Avoidance*, *Long Term Orientation*, and *Indulgence vs. Restraint*. The formulas for each dimension include arbitrary constants that prevent the interpretation of scores without comparing with other samples using the same constants (Hofstede & Minkov, 2013a, 2013b). Thus, the scores are interpreted as relative values related to the values obtained in the other samples in this study. These differences are hereafter referred to as delta. The differences between the samples are visually compared to a set of reference data provided by Hofstede Insights (2023). This reference data compares the scores in the cultural dimensions from a German sample to a U.S.-American sample. The constants applied to calculate scores in the reference data are unknown. Therefore, the comparison between the reference data and the data obtained in this study is limited to reviewing similarities and differences between the deltas in the six dimensions.

The three samples' mean scores and standard deviations for each usability factor are calculated. For each sample, the usability factors are ranked based on the mean scores. These

rankings are compared to each other. Furthermore, noticeable differences between the samples or standard deviations are examined.

The data analysis and visualization are conducted with the statistical software *R* (R Core Team, 2022) using the packages *tidyverse* (Wickham et al., 2019), *reshape* (Wickham, 2007), *skimr* (Waring et al., 2022), and *ggplot2* (Wickham, 2016).

Table 8.2 Description of the usability factors from Hinderks et al. (2019, pp. 1724–1726).

Usability Factor	Description
<i>Appearance/Attractiveness</i>	“The product is attractive, beautiful and/or designed in an appealing way.”
<i>Completeness</i>	“The user considers the information and/or functionality provided and/or offered to the user by the product to be complete.”
<i>Controllability/Dependability</i>	“The product always responds to user interaction in a predictable and consistent way.”
<i>Convenience</i>	“The product makes life easier and/or makes performing a task easier.”
<i>Craftsmanship</i>	“The product appears to be of high quality and robust.”
<i>Ease of use</i>	“The product is easy to operate.”
<i>Efficiency</i>	“The user can reach their goals with minimum time required and minimum physical effort.”
<i>Emotion/Affect</i>	“The product causes positive or negative emotions in the user.”
<i>Flexibility</i>	“The user can adapt the product to their personal needs and requirements and/or their working style.”
<i>Fun</i>	“Interacting with the product brings the user fun.”
<i>Helpfulness</i>	“The product helps the user.”
<i>Identity</i>	“The user can relate to the product and adopt properties of the product for himself.”
<i>Immersion</i>	“When interacting with the product, the user loses track of time.”
<i>Intuitive Operation</i>	“The user is able to operate the product with their present skills immediately and without any training or instruction by others.”
<i>Learnability/Perspicuity</i>	“It is easy for the user to perform their tasks with the product.”
<i>Loyalty</i>	“The user is so convinced of the product that they tell others about it in a positive way and use the product again and again themselves.”
<i>Novelty</i>	“The product is new or innovative.”
<i>Originality</i>	“The product is designed in an interesting and unusual way.”
<i>Overall</i>	“Describes the overall impression of the product in general. The product is good or poor in summary. This is typically a valence factor.”
<i>Relevancy</i>	“The information provided to the user by the product is relevant and/or significant to the user.”
<i>Pragmatic Quality</i>	“The product is practical and functional overall.”
<i>Simplicity</i>	“The product is simple in operation.”
<i>Social Influences</i>	“Interacting with the product helps the user to socialize or present themselves in a favorable way.”
<i>Stimulation</i>	“Working with the product encourages the user to work with it again and again.”
<i>Trust/Credibility</i>	“The product appears trustworthy to the user.”

Note. The German version is partially based on Winter et al. (2017), supplemented by translations conducted at the Chair of Ergonomics following Jones et al. (2001).

8.3 Results

This section describes the results of the cultural dimensions and the subjective ratings of the importance of usability factors.

8.3.1 Cultural Dimensions

The results are described for each of the dimensions. First, the expected differences between the U.S.-American samples *TT_USA* and *ON_USA* and the German sample *ON_GER* are described based on the reference data (Hofstede Insights, 2023). After that, the differences between the samples in the study are examined. The results are visualized in Figure 8.1. The deltas are calculated as the difference between the German sample (*Ref_GER*) and the U.S.-American sample (*Ref_USA*) in the reference data (Δ_{Ref_GER}), and the difference of *ON_USA* and *ON_GER* to *TT_USA* in the study data (Δ_{ON_USA} & Δ_{ON_GER}), respectively.

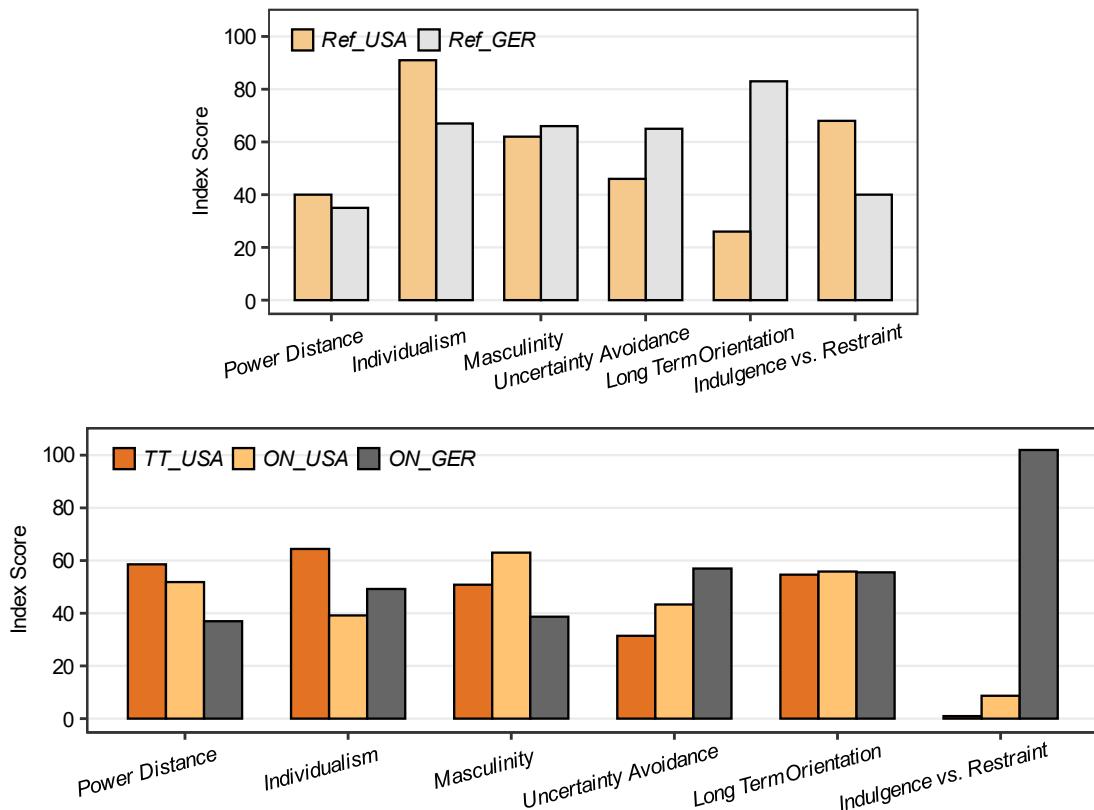


Figure 8.1 Visualization of the score differences for the cultural dimensions between the samples of the reference data (top) and between the study samples of the study Survey_Culture (bottom).

The reference data (Hofstede Insights, 2023) shows that the *Power Distance* score is slightly smaller in Germany than in the United States with $\Delta_{Ref_GER} = -5$. The trend in the study is the same with $\Delta_{ON_GER} = -21.6$, though the difference is more pronounced. The difference between the U.S.-American samples in the study is small, with $\Delta_{ON_USA} = -6.7$.

The reference data (Hofstede Insights, 2023) shows that the *Individualism* score is smaller in Germany than in the United States, with $\Delta_{Ref_GER} = -24$. The trend in the study is the same with $\Delta_{ON_GER} = -15.2$, though the difference is less pronounced. The difference between the U.S.-American samples in the study is considerable, with $\Delta_{ON_USA} = -25.2$. This results in a lower *Individualism* score in *ON_USA* compared to *ON_GER*, showing a different trend than the reference data and Δ_{ON_GER} .

The reference data (Hofstede Insights, 2023) shows that the *Masculinity* score is slightly higher in Germany than in the United States with $\Delta_{Ref_GER} = 4$. The trend in the study is the opposite, with $\Delta_{ON_GER} = -12.2$. The overall distance is rather small. The difference between the U.S.-American samples in the study is small, with $\Delta_{ON_USA} = 12.2$. This results in an even more pronounced difference between the samples *ON_USA* and *ON_GER* compared to the reference data than Δ_{ON_GER} .

The reference data (Hofstede Insights, 2023) shows that the *Uncertainty Avoidance* score is higher in Germany than in the United States with $\Delta_{Ref_GER} = 19$. The trend in the study is the same with $\Delta_{ON_GER} = 25.5$, though the difference is slightly more pronounced. The difference between the U.S.-American samples in the study is small, with $\Delta_{ON_USA} = 11.9$. This results in a less pronounced difference between *ON_USA* and *ON_GER* compared to the reference data or Δ_{ON_GER} .

The reference data (Hofstede Insights, 2023) shows that the *Long Term Orientation* score is considerably higher in Germany than in the United States, with $\Delta_{Ref_GER} = 57$. The trend in the study is different: All three samples yield about identical scores, with $\Delta_{ON_GER} = 1.2$ and $\Delta_{ON_USA} = 0.9$.

The reference data (Hofstede Insights, 2023) shows that the *Indulgence vs. Restraint* score is considerably smaller in Germany than in the United States, with $\Delta_{Ref_GER} = -28$. The trend in the study is different: The difference between the U.S.-American samples is small, with $\Delta_{ON_USA} = 7.7$. The difference between the U.S.-American and German samples is considerable and in the opposite direction with $\Delta_{ON_GER} = 101$. In the study, the *Indulgence vs. Restraint* score is considerably higher in *ON_GER* than in the U.S.-American samples.

8.3.2 Subjective Rating of Importance of Usability Factors regarding HMIs for L3 ADS

The mean scores and the standard deviations for each of the usability factors and the three samples are displayed in Figure 8.2. The sample of *TT_GER* has the highest overall mean and the lowest standard deviation ($M_{TT_USA} = 1.44$, $SD_{TT_USA} = 1.14$), and the sample of *ON_USA* has the lowest overall mean and the highest standard deviation ($M_{ON_USA} = 1.11$, $SD_{ON_USA} = 1.68$). The overall mean and the standard deviation of *ON_GER* is between the two U.S.-American samples ($M_{ON_GER} = 1.28$, $SD_{ON_GER} = 1.21$). The variance of the single factor ratings increases as the mean rating decreases. In general, the ratings of the single factors are very similar among the samples. Only the factor *Fun* receives considerably differs between the samples with $M_{TT_USA} = 0.64$ ($SD_{TT_USA} = 1.45$), $M_{ON_USA} = -0.3$ ($SD_{ON_USA} = 1.88$), and $M_{ON_GER} = 1.18$ ($SD_{ON_GER} = 1.33$). Due to a technical problem, data for the factor *Relevancy* are missing, and the factor is therefore not included in the analysis.

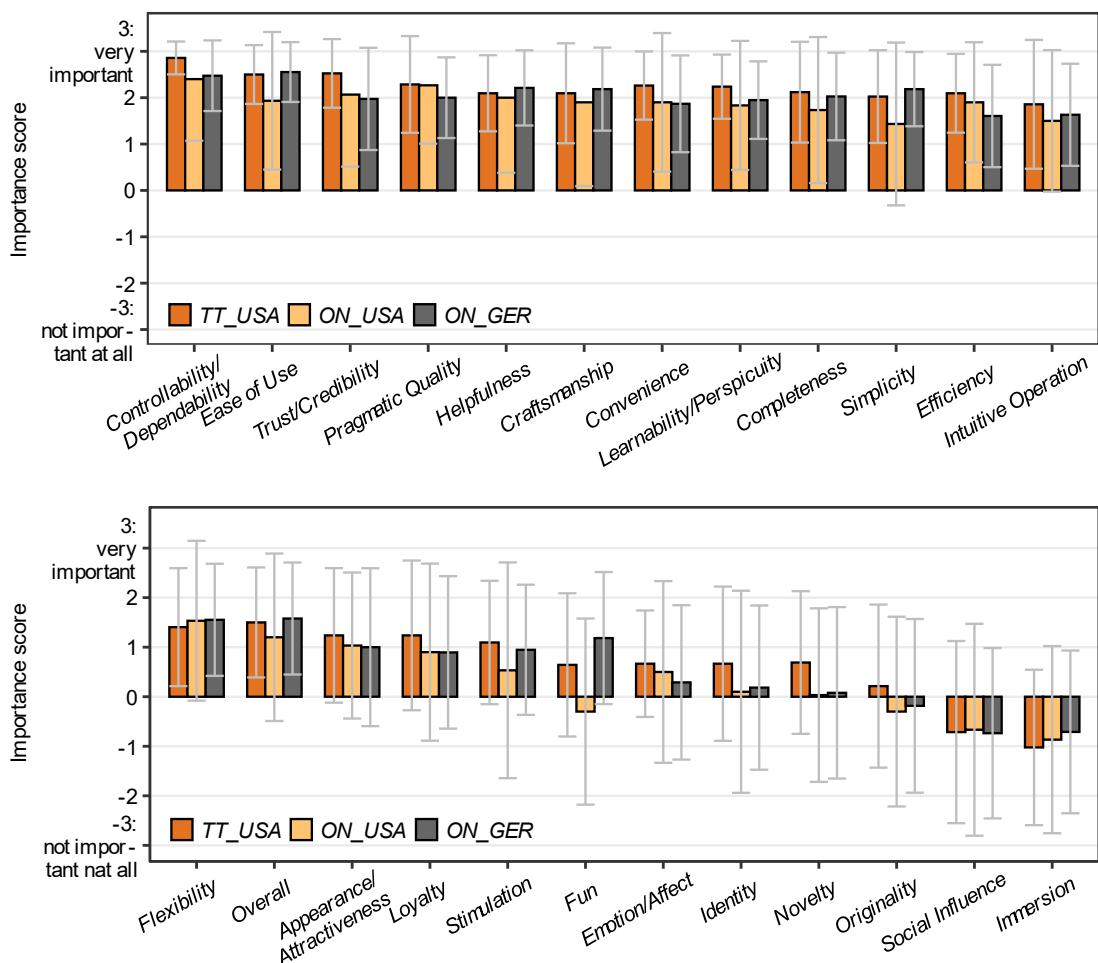


Figure 8.2 Bar chart visualizing the results of the importance scores for the usability factors (Hinderks et al., 2019) in the study *Survey_Culture*.

Note. The scores are ordered based on the overall mean scores. The error bars display the SD. Due to a technical problem, data for the factor *Relevancy* are missing.

8.3.2.1 Factors Included in the Hypotheses

The hypotheses concern differences in the importance ratings of eight usability factors of Hinderks et al. (2019). Table 8.3 presents these eight factors' mean ratings and respective ranking positions. Considering the small differences in the overall means between the samples, the ranking positions are included to provide information on the importance rating of the respective factors in relation to the other factors—the relative importance ranking.

Hypothesis H₁ predicts considerably higher importance ratings for the factor *Flexibility* from participants from the United States compared to participants from Germany. The results show similar mean ratings and similar relative importance rankings. The ranking position of the *ON_GER* sample is 14. Thus, the factor *Flexibility* receives a relative importance ranking lower or the same as in the U.S.-American samples (*TT_USA*: 14 & *ON_USA*: 11).

Hypothesis H₂ predicts slightly higher importance ratings for the factors *Appearance/Attractiveness*, *Emotion/Affect*, *Fun*, *Identity*, and *Stimulation* from participants from the United States compared to participants from Germany. As described before, the factor

Fun shows considerable differences between the samples. In the relative importance rating, the factor is ranked at position 15 in *ON_GER* and at position 21 in both U.S.-American samples, while the mean of *ON_GER* lies between the means of the U.S.-American samples. The other factors do not paint a clear picture. The ranking positions among the three samples are the same or deviate only about one position. The means of the factors *Identity* and *Stimulation* reflect, like the factor *Fun*—the general tendency in the response behavior with the *ON_GER* mean lying between the means of *TT_USA* and *ON_USA*. For the factors *Appearance/Attractiveness* and *Emotion/Affect*, the means of *ON_GER* are slightly lower than those of the U.S.-American samples.

Hypothesis H₃ predicts slightly higher importance ratings for the factors *Controllability/Dependability* and *Trust/Credibility* from participants from Germany compared to participants from the United States. The factor *Controllability/Dependability* is ranked in the second position in *ON_GER* and the first in the U.S.-American samples. The mean of *ON_GER* ($M = 2.47$, $SD = 0.76$) is lower than the mean of *TT_USA* ($M = 2.86$, $SD = 0.35$) and very similar to the mean of *ON_USA* ($M = 2.4$, $SD = 1.33$). The factor *Trust/Credibility* is ranked at position 8 in *ON_GER* and positions 2 and 3 in *TT_USA* and *ON_USA*, respectively. The mean of *ON_GER* is lower than the means of the U.S.-American samples.

Table 8.3 Ranking positions of the usability factors (Hinderks et al., 2019) included in the hypotheses in the study *Survey_Culture*.

Hypothesis	Usability factor (Hinderks et al., 2019)	Ranking position (<i>M</i> ; <i>SD</i>)		
		<i>TT_USA</i>	<i>ON_USA</i>	<i>ON_GER</i>
H ₁ : considerably higher importance ratings in the United States than in Germany	<i>Flexibility</i>	14 (1.4; 1.19)	11 (1.53; 1.62)	14 (1.55; 1.13)
	<i>Appearance/Attractiveness</i>	15 (1.24; 1.36)	15 (1.03; 1.47)	16 (1; 1.59)
	<i>Emotion/Affect</i>	19 (0.67; 1.08)	18 (0.5; 1.83)	19 (0.29; 1.56)
H ₂ : slightly higher importance ratings in the United States than in Germany	<i>Fun</i>	21 (1.5; 1.45)	21 (-0.3; 1.88)	15 (1.18; 1.33)
	<i>Identity</i>	20 (0.67; 1.56)	19 (0.1; 2.04)	20 (0.18; 1.66)
	<i>Stimulation</i>	17 (1.1; 1.25)	17 (0.53; 2.18)	17 (0.95; 1.31)
H ₃ : slightly higher importance ratings in Germany than in the United States	<i>Controllability/Dependability</i>	1 (2.86; 0.35)	1 (2.4; 1.33)	2 (2.47; 0.76)
	<i>Trust/Credibility</i>	2 (2.52; 0.74)	3 (2.07; 1.55)	8 (1.97; 1.1)

8.3.2.2 Highest-Ranked Factors

For each sample, the usability factors are ranked based on the mean scores. The five highest-ranked usability factors per sample are compared to each other in Table 8.4. The factors *Controllability/Dependability* and *Ease of Use* are rated among the most important

factors in all samples. Additionally, the factors *Trust/Credibility* and *Pragmatic Quality* are ranked highly in the U.S.-American samples *TT_USA* and *ON_USA*. In *ON_GER*, these factors rank 8 and 7, respectively. In *ON_GER*, the factors *Helpfulness* and *Craftsmanship* rank 3 and 4, respectively. *ON_USA* ranks *Helpfulness* at position 4 (*TT_USA*: position 10). *Craftsmanship* receives ranking positions 8 in *TT_USA* and 7 in *ON_USA*.

Table 8.4 Comparison of the ranking positions of the highest-ranked usability factors (Hinderks et al., 2019) in the study *Survey_Culture*.

Usability factor	Ranking position (M; SD) in		
	<i>TT_USA</i>	<i>ON_USA</i>	<i>ON_GER</i>
<i>Controllability/Dependability</i>	1 (2.86; 0.35)	1 (2.4; 1.33)	2 (2.47; 0.76)
<i>Trust/Credibility</i>	2 (2.52; 0.84)	3 (2.07; 1.55)	8 (1.97; 1.1)
<i>Ease of Use</i>	3 (2.5; 0.63)	5 (1.93; 1.48)	1 (2.55; 0.65)
<i>Pragmatic Quality</i>	4 (2.29; 1.04)	2 (2.27; 1.26)	7 (2; 0.87)
<i>Convenience</i>	5 (2.26; 0.73)	6 (1.9; 1.49)	10 (1.87; 1.04)
<i>Craftsmanship</i>	8 (2.1; 1.08)	7 (1.9; 1.81)	4 (2.18; 0.9)
<i>Helpfulness</i>	10 (2.1; 0.82)	4 (2; 1.62)	3 (2.21; 0.81)
<i>Simplicity</i>	11 (2.92; 1)	13 (1.43; 1.76)	5 (2.18; 0.8)

Note. The five highest-ranked factors per sample are indicated in bold.

8.4 Discussion

This section discusses the results under consideration of the hypotheses and subsequently reflects the limitations of this validation study. The section closes with the conclusion of the study's results.

8.4.1 Discussion of Results and Hypotheses

The descriptive analysis of the cultural values shows that, overall, the differences between the cultural values of *TT_USA* and *ON_USA* are minor. Therefore, it may be assumed that the U.S.-American sample recruited in Germany for *TT_USA* represents the United States regarding its cultural values. Furthermore, most dimensions show the same or similar trends for the differences between German and U.S.-American samples compared to the reference data (Hofstede Insights, 2023). There are two exceptions: The dimension *Long Term Orientation* shows no differences between the samples in the study. In contrast, the reference data shows a considerably higher score in Germany compared to the United States. For the dimension *Indulgence vs. Restraint*, the reference data shows considerably higher scores for the United States than Germany. This contradicts the study results where *ON_GER* yields considerably higher scores than the U.S.-American samples *TT_USA* and *ON_USA*. The reasons for these differences are unclear. Previous studies have encountered similar problems naming attitudes, selection processes, and small sample sizes as possible reasons (e.g.,

Hofstede, 2013 on a study of Fischer & Al-Issa, 2012; Khan et al., 2016; Young et al., 2012). Possible influences could be attributed to differences in the sample, such as age. While the age distribution is similar for *TT_USA* and *ON_USA*, *ON_GER* features a considerably higher mean age and standard deviation. Differences in the demographic features of the samples might severely affect the scores on the cultural dimensions or, as Hofstede (2013, p. 5) puts it: “*Valid cross-cultural studies compare apples with apples across countries; basing a country comparison upon apples in one country versus oranges in another (or even in the same) country or countries produces fruit salad.*”

Regarding the subjective ratings of usability factors, the ratings are very similar among the samples. The variance is high in all samples and increases as the mean ratings decrease. The high means and small standard deviations for some of the usability factors (e.g., *Controllability/Dependability*) may be an indicator of a ceiling effect; that is, the rating scale is not sufficient to reflect variances in extreme response areas (Bortz & Döring, 2006, p. 558).

The usability factors *Controllability/Dependability* and *Ease of Use* are rated among the most important factors in all samples. Additionally, *Trust/Credibility* and *Pragmatic Quality* are ranked highly in the U.S.-American samples, while the German sample attributes high importance ratings to the factors *Helpfulness* and *Craftsmanship*. Only one factor—the factor *Fun*—receives importance ratings considerably differing among the three samples. The survey results indicate that the subjective importance of different usability factors in the context of HMIs for L3 ADS does not differ substantially between different cultures.

This leads to the answer to the three hypotheses formulated based on the literature findings in Section 2.4.

- H₁ The factor *Flexibility* receives considerably higher importance ratings from participants from the United States than from Germany.
- H₂ The factors *Appearance/Attractiveness*, *Emotion/Affect*, *Fun*, *Identity*, and *Stimulation* receive slightly higher importance ratings from participants from the United States than from Germany.
- H₃ The factors *Controllability/Dependability* and *Trust/Credibility* receive slightly higher importance ratings from participants from Germany than from the United States.

The descriptive analysis of the data comprises examining the means and standard deviations, and ranking the single usability factors relative to the other usability factors. As described, the results do not indicate clear differences between the samples. Instead, seven of the eight factors receive quite similar means. Only the factor *Fun* shows different means, with the mean rating in *ON_USA* differing considerably from the mean ratings in *TT_USA* and *ON_GER*. Moreover, the relative importance ratings do not support the hypotheses. Regarding hypothesis H₁, the relative importance rating compared to other usability factors does not indicate the expected higher importance ratings compared to *TT_USA* and *ON_USA* (positions 14 & 11, respectively). In hypothesis H₂, slightly higher importance ratings for the five factors *Appearance/Attractiveness*, *Emotion/Affect*, *Fun*, *Identity*, and *Stimulation* are expected for the U.S.-American samples compared to the German sample. Instead, the ranking positions are the same or deviate only about one position in all three samples. The only exception is the factor

Fun—the rankings (*TT_USA* & *ON_USA*: 21 & *ON_GER*: 15) contradict the hypothesis. The same is evident for H_3 , where slightly higher importance ratings for the two factors *Controllability/Dependability* and *Trust/Credibility* are expected for the German sample compared to the U.S.-American samples. Again, the rankings (*Controllability/Dependability*: *TT_USA* & *ON_USA*: 1, *ON_GER*: 2; & *Trust/Credibility*: *TT_USA*: 2, *ON_USA*: 3, & *ON_GER*: 8) are contrary to the hypothesis.

8.4.2 Discussion of Limitations

The results need to be considered in light of the study's limitations. Only two Western industrialized cultures are compared. Differences between more diverse cultures likely exist for the subjective importance rating of usability factors in the provided context (e.g., Chau et al., 2002; Frandsen-Thorlacius et al., 2009; Young et al., 2012). The samples are recruited via different methods, which might have affected specific sample characteristics. Furthermore, the age distribution of *ON_GER* differs considerably from the age distributions of *TT_USA* and *ON_USA*. The proportion of females ranges between 24% in *ON_USA* and 45% in *TT_USA*. This might have affected the importance ratings since gender and age are known to influence usability and related constructs for ADAS and ADS (Liu et al., 2021; Piao et al., 2005; Rödel et al., 2014). The small size of the samples limits the data analysis to descriptive examinations of trends with rather high variances. The existence of user groups could explain the high variances. Due to the sample sizes, no analysis targeted to identify user groups could be conducted. No other information, for example, via interviews, could have supported the interpretation of the results, such as the trends for the dimensions *Long Term Orientation* and *Indulgence vs. Restraint* deviating from the reference data. Differences in the overall means and standard deviations between the samples could be identified, with *ON_GER* having the highest means and lowest standard deviations and *ON_USA* having the lowest means and highest standard deviations. This observation could indicate a difference in the response behavior described by previous researchers (Douglas & Liu, 2011; Moss & Vijayendra, 2020). Potential biases in the response behavior limit the value of analyzing the means alone. Thus, the contemplation of the relative importance ratings is recommended as the superior analysis tool. Due to a technical problem, data are missing for one of the twenty-five usability factors (*Relevancy*). Thus, the analysis of the usability factors, as presented by Hinderks et al. (2019), could not be fully repeated.

8.4.3 Conclusion

The results of the validation study *Survey_Culture* allow the conclusion that the U.S.-American sample recruited in Germany for *TT_USA* represents U.S.-American citizens regarding its cultural values. Furthermore, comparing between the reference data for Germany and the United States provided by Hofstede Insights (2023) and the study samples yields the same or similar trends for four dimensions. The study results for the two dimensions *Long Term Orientation* and *Indulgence vs. Restraint* differ considerably from the reference data. Possible influences could be attributed to differences in the sample, such as age.

Regarding the research question and the posed hypotheses, no considerable and systematic influence of the users' cultural background on the subjective importance rating of usability factors (Hinderks et al., 2019) in the context of HMIs for L3 ADS could be identified.

Thus, all three hypotheses are rejected. The usability factors *Controllability/Dependability* and *Ease of Use* are rated among the most important factors in all samples.

This study may draw a practical conclusion: Usability assessments for HMIs for L3 ADS may be conducted within one culture of the Western industrialized world. However, the transferability of results to more diverse cultures is expected to be limited.

9 Expert Workshop: Recommendations for Methods for Assessing Usability of HMIs for L3 ADS in User Studies

This chapter presents the results of an expert workshop. In this workshop, a set of preliminary recommendations for assessing usability in HMIs for L3 ADS is discussed. The preliminary recommendations are derived from the findings and experiences of the studies *Exp_Testing-Environment*, *Exp_Culture*, and *Survey_Culture*. They serve as a starting point for the expert discussion and do not claim to be final. The expert workshop is conducted in February 2023 after a preparation phase requiring input from the experts between December 2022 and January 2023. The results are incorporated into deriving the final recommendations presented at the end of this chapter.

9.1 Method

The expert workshop is conducted with five employees of the Chair of Ergonomics. The experts are selected due to their prior experiences with user tests in the field of HMIs for ADS.

The experts are required to provide input during a preparation phase of four weeks. Information is provided in written form and video format, covering the workshop procedure, the instructions, and background information. The background information includes an overview of the studies *Exp_Testing-Environment*, *Exp_Culture*, and *Survey_Culture*. After familiarizing the studies, the experts are requested to provide their inputs to the recommendations. The experts fill out a table that lists 15 recommendations, complemented with comments and supporting observations from the empirical data. Furthermore, a preface clarifies the goal and application context of the recommendations.

The experts rate their level of approval for each of the 15 recommendations on a 7-point Likert scale (-3: "I completely disagree"; -2: "I disagree"; -1: "I rather disagree"; 0: "I am uncertain"; +1: "I rather agree"; +2: "I agree"; +3: "I completely agree") and comment their thoughts for each of the recommendations. The instructions encourage the experts to critically reflect on the recommendations based on their experiences and suggest additional recommendations.

The experts' input is prepared for the second phase of the workshop. This phase is conducted in a 4 hr session at the Chair of Ergonomics in February 2023. The recommendations, their supporting observations, the experts' comments, and their ratings are printed out and pinned on whiteboards. These whiteboards are installed for the workshop, leaving space for editing and further comments (see Figure 9.1). The session is moderated by the author of this thesis. The experts discuss the recommendations based on their professional experience and the provided information. The discussion results are consolidated and documented by the moderator. The final set of recommendations is derived following the results of the expert workshop.



Figure 9.1 Setting of the expert workshop on recommendations for methods for assessing the usability of HMIs for L3 ADS in user studies. Five experts participate in the workshop at the Chair of Ergonomics in February 2023.

9.2 Preliminary Recommendations and Results of the Expert Workshop

This section presents the preliminary set of recommendations and the workshop results. The recommendations are assigned to different categories intended to provide a structure for the discussion. The four categories are Experimental Design, Testing Environment, Procedure, and Sample. Each recommendation is complemented by one or more observations obtained in the studies *Exp_Testing-Environment*, *Exp_Culture*, and *Survey_Culture*. Single recommendations include comments reminding of the limitations of the recommendations as stated in the preface. The experts' ratings in the workshop's preparation phase and the results of the workshop session are included. The summary is presented in Table 9.1.

Besides the recommendations, the preface is discussed, and suggestions for rephrasing it are made. The preface informs the experts about the goal and application context of the recommendations. It is formulated as follows:

The recommendations refer to experiments aiming to assess the usability of HMIs for L3 ADS that are still in a prototype state. Experiments following these recommendations may help to identify flaws in the design of specific HMI components and compare different design approaches. The recommendations do not cover validations of HMIs integrated into series vehicles for the final approval for market launch.

During the discussion of the preface and the recommendations, the experts identify several aspects that should be added to the preface or that need rephrasing. First, it should be specified that the goal and application context are the assessment of intuitive and relative

usability. This means that the assessments focus on usability during first-contact interactions. Furthermore, the recommendations are formulated for comparisons of different HMI concepts and do not claim to provide absolute usability ratings sufficient for assessing one HMI concept alone. Second, the experts want to stress that safety and ethical considerations in planning the experimental design are always prioritized over the recommendations. Additionally, every recommendation should be weighed against whether it is compatible with the research questions of the planned experiment: An experimental design compatible with the research question must be prioritized over the recommendations. Third, the experts suggest including the definition of usability applied in the recommendations (i.e., ISO standard 9241-11, ISO, 2018a) in the preface instead of stating it in a recommendation (preliminary recommendation 1). Moreover, the experts suggest structuring the recommendations by assigning them to the definition's elements. Thereby, it becomes clear which recommendations are usability-specific and which could be transferred to experimental designs with other research focuses.

Table 9.1 Overview of the preliminary recommendations with supporting observations and results of the expert workshop with proposed changes.

# Category	Preliminary recommendation [optional comment]	Supporting observation	Rating [M (SD); Med]	Proposed changes by experts
1 Experimental Design	Apply the definition of ISO standard 9241-11 (ISO, 2018a, p. 2). Investigate usability by covering all three facets effectiveness, efficiency, and satisfaction.	- The facet satisfaction is covered in standardized questionnaires and the <i>Final interview</i> . The overall ratings for both HMI concepts are high, while the <i>Final interviews'</i> results reflect differences between the concepts.	1.8 (1.17); 2	<p>Discard:</p> <ul style="list-style-type: none"> - The applied definition of usability and the related facets of usability should be included in the preface, not in a recommendation.
2 Experimental Design	If the test focuses on satisfaction, apply a within-subject design. If the test focuses on effectiveness and efficiency, apply a between-subject design.	<ul style="list-style-type: none"> - Several participants express their excitement for automated driving. They seem to superimpose these feelings, leading to high/positive ratings while tolerating identified flaws and problems. - Naïve participants have no background knowledge and, therefore, no internal reference system that they may apply in ratings of ADS HMIs. - In between-subject designs, HMI concepts can be tested with (fully) naïve participants, and no learning effects need to be considered. 	2 (1.55); 3	<p>Reformulate for mitigation:</p> <ul style="list-style-type: none"> - Instead of recommending a specific research design, the advantages and disadvantages of each design regarding the different facets of usability should be focused on.
3 Experimental Design	Define a set of test cases that covers the most important transitions (such as switching between L0 & L3) and the NHTSA minimum requirements (NHTSA, 2017).	- Only test cases addressing implemented differences in the HMI concepts reflect these differences (e.g., RtlIs).	1.4 (1.02); 1	<p>Reformulate and shorten:</p> <ul style="list-style-type: none"> - The testing of all HMI elements, as presented in Bengler et al. (2020), should be covered, ensuring that transitions, etc. are covered.

# Category	Preliminary recommendation [optional comment]	Supporting observation	Rating [<i>M (SD); Med</i>]	Proposed changes by experts
4 Experimental Design	Define a set of observational and self-reported metrics that covers all three facets of usability as defined in the ISO standard 9241-11 (ISO, 2018a), AND the most important transitions (such as switching between L0 & L3), AND the NHTSA minimum requirements (NHTSA, 2017).	- Different metrics uncover different usability problems in both HMI concepts.	1.2 (0.98); 2	Shorten: - A set of metrics should be defined, covering all three facets of usability, as defined in the ISO standard 9241-11 (ISO, 2018a). - Due to repetition with other recommendations, the section on transitions and the NHTSA minimum requirements (NHTSA, 2017) should be omitted. - The specification of metrics into observational and self-reported metrics should be included in Recommendation 5.
5 Experimental Design	Do not rely on inferential statistics alone, but also consider descriptive and qualitative analyses, especially qualitative interviews, open questions, or other possibilities for feedback.	- The qualitative analysis of control paths provides both success rates and type of errors while the TOT confirms differences between the HMI concepts. - Participants notice HMI flaws and problems. They express these in the interviews. However, their critique does not affect the generally high ratings in questionnaires.	1.8 (1.94); 3	Reformulate for clarification: - The collection of qualitative vs. quantitative data and self-reported vs. observational data should be recommended. - The analysis through descriptive as well as inferential statistics should be advised. - Future research should be initiated to obtain absolute usability assessments with thresholds such as "acceptable" or "good" usability.
6 Experimental Design	If the study design allows, provide feedback on participants' errors and interaction problems.	- Some participants do not notice interaction mistakes (e.g., wrong LoA). Consequently, their objective performance measurements may deviate from their subjective performance. Mistakes may be left unconsidered if no feedback is provided (Drew et al., 2018).	1.8 (1.47); 2	Reformulate for mitigation: - Attention should be drawn to the potential discrepancy between objective and subjective performance.

# Category	Preliminary recommendation [optional comment]	Supporting observation	Rating [M (SD); Med]	Proposed changes by experts
7 Testing Environment	Conduct tests in driving simulators. The environment should simulate a realistic setting, including other road users, curves, etc. [HMIs should be validated in naturalistic settings before their final approval.]	- The study <i>Exp_Testing-Environment</i> confirms relative validity between the environments driving simulator and test track. - A driving simulator experiment is risk-free, allows higher standardization and adaptability (e.g., weather conditions, data quality, & test route), and is more efficient (e.g., time & costs) than a test track experiment.	1.6 (0.49); 2	Split into two recommendations: - A first recommendation should recommend conducting tests in driving simulators instead of field studies providing reasons for the choice. - A second recommendation should focus on the design of the simulated environment. A realistic setting (e.g., implementation of curves, other traffic participants, & higher speeds) is suggested.
8 Procedure	Test the first contact interaction and give only general instructions on the ADS functions and responsibilities beforehand without detailed descriptions of the HMI or the ADS operation. [Resource-intensive experiments on long-term usage should be conducted only if good usability of the HMI is established for intuitive use.]	- Participants receive no detailed instruction or training on handling the HMI concepts. This allows testing whether the handling of a system is intuitively understandable without consulting the manual ("rental car scenario", see Albers, Radlmayr, et al., 2020 or Forster, Hergeth, Naujoks, Krems, & Keinath, 2019), which is highly important from a safety point of view.	2.6 (0.8); 3	Shorten: - It should be clarified that the first contact interaction refers to the research subject, i.e., the HMI in total or a specific HMI element. Prior interactions (e.g., familiarization drive) with the ADS in general or elements of the HMI that are not focused on are accepted. - The specification instructions should be included in Recommendation 9.
9 Procedure	Provide instructions that are as short as possible and briefly explain the ADS functions and resulting responsibilities in plain language.	- Several participants comment that they were overwhelmed and overloaded with information in an unknown domain (amplified due to tension due to the study participation).	2.6 (0.8); 3	Reformulate for clarification: - Details should be kept to a minimum, referring to the "rental car" problem (see Albers, Radlmayr, et al., 2020 or Forster, Hergeth, Naujoks, Krems, & Keinath, 2019).
10 Procedure	Conduct experiments with a maximum study duration of 2.5 hr (or include breaks).	- Several participants comment that they were overwhelmed and overloaded with information in an unknown domain (amplified due to tension due to the study participation).	2.2 (1.17); 3	Discard: - The recommendation is deemed too generic. Only recommendations specific to usability assessments should be included.

# Category	Preliminary recommendation [optional comment]	Supporting observation	Rating [M (SD); Med]	Proposed changes by experts
11 Procedure	Allow participants room to express their feelings and thoughts on automated driving in general before requesting feedback on the HMI concepts.	- Many participants express their excitement for automated driving. Participants even answer questions about the HMI design with general statements about automated driving. If participants were allowed to express their feelings and thoughts on automated driving in general, it might increase the ability to rate the HMI concept detached from the context of the study.	2.6 (0.49); 3	Reformulate for clarification: - The question should be posed before obtaining the qualitative feedback on the HMI. The discussion ends undecided whether the question should be posted even before obtaining the quantitative feedback or possibly in an additional question before the test drive. - Future research should be initiated on the potential effect of the attitude toward automated driving on quantitative assessments of HMIs for ADS.
12 Sample	Conduct tests within one culture. [HMIs should be validated in different cultural settings before final approval.]	- The study <i>Exp_Culture</i> confirms relative validity between the two samples from Germany and the United States. Furthermore, participants from Germany and the United States rate <i>Controllability/Dependability</i> and <i>Ease of Use</i> (Hinderks et al., 2019) among the most important factors. Only second-order factors and design aspects show differences between the samples, implying differing preferences (→ facet satisfaction). - The implementation of the experiment in only one culture is more efficient.	0.8 (1.72); 1	Discard: - The empirical data is deemed insufficient to formulate this generic recommendation.
13 Sample	Apply sample sizes of at least 20 participants per subsample. Include some buffer regarding the planned statistical tests.	- The effects in the standardized questionnaires are too small to be detected with the available sample sizes. The database of other metrics such as eye-tracking is reduced due to crashes, wrong preconditions (e.g., wrong LoA at scenario start), missing suitability of participants (e.g., bifocals), and insufficient data quality (glare, heavy lids, ...).	2.4 (1.2); 3	Discard: - The recommendation is deemed too generic. Only recommendations specific to usability assessments should be included.

# Category	Preliminary recommendation [optional comment]	Supporting observation	Rating [<i>M (SD); Med</i>]	Proposed changes by experts
14 Sample	Recruit participants who are naïve regarding HMIs for ADS and cover a balanced age and gender distribution.	<ul style="list-style-type: none"> - Participants must be naïve to investigate the intuitive use during first contact interactions. - Female and older participants seem to be more insecure. Additionally, some older participants comment that they struggle with reading small (italic) fonts and report using reading glasses, but not necessarily glasses for driving. 	1.6 (1.02); 2	<p>Reformulate to shift the focus:</p> <ul style="list-style-type: none"> - A target user should be defined, and the characteristics should be included in the recruitment criteria, ensuring that the sample represents the target user. - The term “naïve” should be avoided to preserve the relevance of the recommendation: It is expected that in future settings, automated driving might already be more prevalent. It is deemed possible that depending on the research focus, the user group of interest might have prior experience with automated driving, which can be both a benefit as well as a handicap (“wrong” mental model).
15 Sample	Inquire about the participants' driving experience (general and regarding ADAS) and prior knowledge of automated driving. If possible, inquire after the experiment whether participants have transferred knowledge and experience from other areas or domains to their interaction with the HMI.	<ul style="list-style-type: none"> - Single participants comment unprompted that they transferred prior knowledge of experiences with Tesla vehicles to the interaction with the HMI concepts. Other participants comment that they associated L2 with blue and L3 with green due to the colors on the instruction sheet. 	2.4 (1.2); 3	<p>Shorten and simplify:</p> <ul style="list-style-type: none"> - The first sentence is deemed to be a repetition of Recommendation 14. - It should be inquired after the experiment whether participants were reminded of something similar when they experienced the HMI.

9.3 Derivation of Final Recommendations

This section presents the formulation of the final set of recommendations regarding methods to assess the usability of HMIs for L3 ADS. A preface and a set of 12 recommendations are derived. This section presents the preface and recommendations individually, supplemented with a brief explanation. The explanations consider learnings from the expert workshop, supplemented by learnings from the literature. Additionally, the recommendations are assigned to the terms included in the definition. An overview of the final recommendations is presented in Table 9.2.

The preface is formulated as follows:

The recommendations refer to experiments aiming to assess the relative and intuitive usability of HMIs for L3 ADS that are still in a prototype state. That is, the experiences focus on usability during first contact interactions and focus on comparisons of different HMI concepts. Experiments following these recommendations may help to identify flaws in the design of specific HMI components when comparing different design approaches. The experiments do not claim to provide answers to absolute usability ratings. Moreover, the recommendations do not cover validations of HMIs integrated into series vehicles for the final approval for market launch.

Safety and ethical considerations are always to be prioritized over compliance with recommendations during the planning of the experiment. Additionally, compliance with recommendations should be weighed against the compatibility with the experiment's research question(s).

According to the ISO standard 9241-11, usability is defined as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” (ISO, 2018a, p. 2). The recommendations are assigned to the terms included in the definition.

The preface specifies the preliminary preface in several aspects. The goal of the extension is to narrow down the goals and the scope of the recommendations. Thus, the preface is extended, clearly stating four aspects. First, usability is narrowed down to relative and intuitive usability. Based on the scope of the thesis, no recommendations could be derived for the usability of prolonged use or absolute usability assessments. The latter is identified as a research gap that needs addressing in future research efforts. These efforts may build upon the findings of this thesis. The recommendations refer to relative usability assessments when comparing different HMI concepts. Second, safety and ethical considerations are prioritized over compliance with recommendations. Third, the recommendations serve as a decision aid that must not interfere with the research question of the planned experiment. Thus, the compliance of the study design with the research question is to be prioritized over the compliance with recommendations. Fourth, the definition of usability provided by the ISO

standard 9241-11 (ISO, 2018a) that underlies all recommendations is stated in the preface. In addition to the four aspects, the preface adds the information that all recommendations are assigned to terms of the stated definition of usability to provide a structure.

Table 9.2 Overview of the final recommendations refined after the expert workshop.

#	Assigned terms of the ISO standard 9241-11 (ISO, 2018a)	Recommendation
1	Effectiveness, efficiency, and satisfaction	A within-subject design enables participants to compare different HMI concepts even though the underlying technology is novel and unknown. Thus, a within-subject design is advantageous if the focus is on satisfaction. A between-subject design may test fully naïve participants to assess intuitive usability. Thus, a between-subject design is advantageous if the focus is on effectiveness or efficiency.
2	Specified goals	Define a set of test cases that covers the input channel(s), output channel(s), and dialog logic (Bengler et al., 2020) of the HMI concepts. To evaluate the output channel(s), principles such as the NHTSA minimum requirements (NHTSA, 2017) should be applied.
3	Effectiveness, efficiency, and satisfaction	Define a set of metrics covering all three facets of usability as defined in the ISO standard 9241-11 (ISO, 2018a).
4	Effectiveness, efficiency, and satisfaction	Collect qualitative as well as quantitative data and observational as well as self-reported data. Pay regard to descriptive data analysis as well as inferential data analysis.
5	Specified goals; effectiveness, efficiency, and satisfaction	Be aware of the potential discrepancy between objective and subjective performance measurements.
6	Specified context of use	Conduct user studies in driving simulators instead of field studies.
7	Specified context of use	Simulate an environment that provides a realistic setting for the context of interest, this includes, for example, curves, other traffic participants, or higher speeds.
8	Specified context of use	Test the first contact interaction with the HMI concepts or their specific elements of interest.
9	Specified goals; specified context of use	Provide instructions that are as short as possible, providing only the necessary extent of details (determined in pilot tests).
10	Specified users	Encourage participants to express their feelings and thoughts on automated driving in general, e.g., through an unstructured interview, before inquiring qualitative feedback on the HMI concepts.
11	Specified users	Define a target user (group) and recruit a representative sample according to its characteristics.
12	Specified users	Inquire after the experiment whether the participants associate the HMI (interaction) with something they have experienced before. This provides insights into mental models and knowledge transfer.

9.3.1 Recommendation 1

A within-subject design enables participants to compare different HMI concepts even though the underlying technology is novel and unknown. Thus, a within-subject design is advantageous if the focus is on satisfaction. A between-subject design may test fully naïve participants to assess intuitive usability. Thus, a

between-subject design is advantageous if the focus is on effectiveness or efficiency.

This recommendation is assigned to the three facets of usability effectiveness, efficiency, and satisfaction listed in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 2 (Table 9.1). Following the experts' conclusions, the recommendation is rephrased and slightly toned down. Instead of giving clear recommendations for a study design, the advantages are pointed out. The recommendation is based on the empirical findings in this thesis, underlining the advantage of a between-subject design when testing effectiveness and efficiency and the disadvantage of this design when testing satisfaction. The findings align with the experts' experiences with both within- and between-subject designs in the context of HMIs for ADS.

9.3.2 Recommendation 2

Define a set of test cases that covers the input channel(s), output channel(s), and dialog logic (Bengler et al., 2020) of the HMI concepts. To evaluate the output channel(s), principles such as the NHTSA minimum requirements (NHTSA, 2017) should be applied.

This recommendation is assigned to the specified goals stated in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 3 (Table 9.1). Following the experts' conclusions, the recommendation discards the part on transitions and refers to the framework for HMIs by Bengler et al. (2020) instead. This reference ensures the coverage of all elements of the HMI concept(s). The empirical findings in *Exp_Testing-Environment* and *Exp_Culture* show that single test cases (especially TC10 & TC12) reveal differences between the HMI concepts regarding their output channels. In contrast, other test cases (especially TC3) indicate problems with the control logic. Principles facilitate the selection of test cases and evaluation criteria. While this thesis applies the NHTSA minimum requirements (NHTSA, 2017), other principles (e.g., Mendoza et al., 2022) may be applied based on the research focus.

9.3.3 Recommendation 3

Define a set of metrics covering all three facets of usability as defined in the ISO standard 9241-11 (ISO, 2018a).

This recommendation is assigned to the three facets of usability effectiveness, efficiency, and satisfaction listed in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 4 (Table 9.1). Following the experts' conclusions, the only changes are the removal of the parts on the type of metrics ("observational and self-reported"), the transitions, and the minimum requirements (NHTSA, 2017). The recommendation is supported by the empirical findings in *Exp_Testing-Environment* and *Exp_Culture* that link the different metrics to single facets of usability. Furthermore, the ISO standard 9241-11 (ISO, 2018a, p. 25) states that a single facet of usability cannot fully represent overall usability. Therefore, the standard demands the combination of metrics by covering at least one metric per usability facet. Testing

only single facets of usability has been criticized in previous works by Hornbæk (2006, p. 84) and Frøkjær et al. (2000).

9.3.4 Recommendation 4

Collect qualitative as well as quantitative data and observational as well as self-reported data. Pay regard to descriptive data analysis as well as inferential data analysis.

This recommendation is assigned to the three facets of usability effectiveness, efficiency, and satisfaction listed in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 5 (Table 9.1). Following the experts' conclusions, the recommendation is rephrased, providing a more systematic description of the data types and distinguishing it from the data analysis. In *Exp_Testing-Environment* and *Exp_Culture*, a wide range of metrics is presented that cover differing usability aspects. The analysis of the qualitative metric of control paths, for example, could identify success rates, types of errors, and strategies through descriptive analysis. At the same time, the analysis of TOTs confirms differences between the HMI concepts through statistical tests. Regarding the equal treatment of descriptive and inferential statistical tests, Russell and Grove (2020) strongly recommend not relying on statistical significance when interpreting results in human factors research on automated vehicles. The inclusion of different data types is supported by literature emphasizing the importance of collecting self-reported and observational data. Multiple studies report that results obtained from observational data do not correspond to results obtained from self-reported data (e.g., Drew et al., 2018; Herman, 1996; Large et al., 2019). The ISO standard 9241-11 states that missing feedback might contribute to a discrepancy between observational and self-reported data since users might not be aware of not having successfully completed a task (ISO, 2018a, p. 25). This issue is addressed in the following recommendation.

9.3.5 Recommendation 5

Be aware of the potential discrepancy between objective and subjective performance measurements.

This recommendation is assigned to the specified goals and the three facets of usability effectiveness, efficiency, and satisfaction listed in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 6 (Table 9.1). Following the experts' conclusions, the first part of the preliminary recommendation is discarded as it is included in the preface already. Furthermore, the recommendation is toned down, merely drawing attention to the potential discrepancy without recommending to provide feedback. The recommendation is based on the empirical findings in *Exp_Testing-Environment* and *Exp_Culture*, where participants fail to report problems in the interaction, presumably because they overlooked their non-compliance, for example, with instructed LoAs. This discrepancy between objective and subjective performance measurements has been the subject of previous research bodies (e.g., Drew et al., 2018; Forster, 2020) and the ISO standard 9241-11 (ISO, 2018a, p. 25).

9.3.6 Recommendation 6

Conduct user studies in driving simulators instead of field studies.

This recommendation is assigned to the specified context of use stated in the ISO standard 9241-11 (ISO, 2018a). It is based on the first part of the preliminary recommendation 7 (Table 9.1). Following the experts' conclusions, the two aspects covered in the preliminary recommendation are differentiated by creating two separate recommendations. Chapter 6 concludes with the assumption of the relative validity of testing environments. Furthermore, the data quality, the degree of standardization, and the overall efficiency are considered superior in the driving simulator experiment compared to the test track experiment. Additionally, the driving simulator experiment is more risk-free than the test track experiment. Results suggest that differences identified in the driving simulator are more extreme in test track conditions, which implies that driving simulators are a less sensitive testing environment. In contrast, Purucker et al. (2018) report more conservative results for controllability assessments from the driving simulator experiment compared to the test track experiment in the form of situations rated as more dangerous in the simulator than on the test track. Wynne et al. (2019) argue that perceived risk is lower in driving simulators compared to real-world settings (Bella, 2008; McAvoy et al., 2007), thus facilitating riskier behavior in driving simulators compared to similar settings in real-world settings. This leads to the conclusion that usability problems identified in driving simulator settings will likely be more extreme in real-world settings. Thus, user studies on HMI concepts should first be conducted in driving simulator conditions that pose risk-free, resource- and cost-efficient testing environments (Caird & Horrey, 2011, Table 5.1). After the refinement of the HMI concept(s), follow-up user studies can be carried out in field studies.

9.3.7 Recommendation 7

Simulate an environment that provides a realistic setting for the context of interest, this includes, for example, curves, other traffic participants, or higher speeds.

This recommendation is assigned to the specified context of use stated in the ISO standard 9241-11 (ISO, 2018a). It is based on the second part of the preliminary recommendation 7 (Table 9.1). Following the experts' conclusions, the two aspects covered in the preliminary recommendation are differentiated by creating two separate recommendations. The call for realistic settings is based on participants' comments throughout all three experiments presented in *Exp_Testing-Environment* and *Exp_Culture*. These align with the ISO standard's definition of usability, which emphasizes the significant role of the context of use in the usability assessment (ISO, 2018a).

9.3.8 Recommendation 8

Test the first contact interaction with the HMI concepts or their specific elements of interest.

This recommendation is assigned to the specified context of use stated in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 8 (Table 9.1). Following

the experts' conclusions, the second part of the preliminary recommendation is discarded as it targets a different aspect, which is now covered in the following recommendation. Furthermore, the rephrasing specifies that the first contact interaction refers to the research subject: the HMI in total or a specific HMI element. Interactions with the ADS in general or elements of the HMI that are not focused on are accepted, for example, during a familiarization drive. While acknowledging the importance of usability testing in long-term usage to minimize effects such as disuse, misuse, and abuse (Parasuraman & Riley, 1997), the preface limits the scope of the recommendations to testing intuitive usability. HMI concepts for L3 ADS should demonstrate intuitive usability before usability in prolonged use is targeted. By testing the first contact without prior information, the extreme, but relevant use case of using the car without consulting the manual beforehand ("rental car scenario", see Albers, Radlmayr, et al., 2020 or Forster, Hergeth, Naujoks, Krems, & Keinath, 2019) is covered.

9.3.9 Recommendation 9

Provide instructions that are as short as possible, providing only the necessary amount of details (determined in pilot tests).

This recommendation is assigned to the specified goals and the specified context of use stated in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 9 (Table 9.1). Following the experts' conclusions, the second part of the preliminary recommendation is discarded as it is considered too vague. Instead, the recommendation states that details should be kept to a minimum. Like the previous recommendation, this recommendation refers to the use case described as the "rental car scenario" (see Albers, Radlmayr, et al., 2020 or Forster, Hergeth, Naujoks, Krems, & Keinath, 2019).

9.3.10 Recommendation 10

Encourage participants to express their feelings and thoughts on automated driving in general, e.g., through an unstructured interview, before inquiring qualitative feedback on the HMI concepts.

This recommendation is assigned to the specified users referred to in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 11 (Table 9.1). Following the experts' conclusions, the preliminary recommendation is rephrased through the specification of the timing of the feedback request. The empirical findings in *Exp_Testing-Environment* and *Exp_Culture* suggest that participants' enthusiasm for the topic of automated driving masks potential usability problems. By allowing participants to express their enthusiasm (and other thoughts) before qualitative feedback is requested, the previously described effect is expected to be lessened. Furthermore, the participants' feedback might provide valuable insights that otherwise would be overlooked. The findings typically relate to the UX of automated driving in general (especially the anticipated use), which can be integrated into the further development of HMIs (and ADS functions).

9.3.11 Recommendation 11

Define a target user (group) and recruit a representative sample according to its characteristics.

This recommendation is assigned to the specified users referred to in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 14 (Table 9.1). Following the experts' conclusions, the preliminary recommendation is rephrased by discarding the specific user samples' characteristics regarding age and gender distribution and their naivety. The experts argue that naivety with HMIs for L3 ADS may lose its relevance as ADS become common in vehicles. They acknowledge the empirical findings in *Exp_Testing-Environment* and *Exp_Culture*, indicating gender- and age-specific differences. Nonetheless, they propose referring to a target user or a group of target users through the research questions, e.g., with naivety regarding HMIs for L3 ADS. By recruiting a representative sample, the potential effects of specific characteristics like age and gender are covered.

9.3.12 Recommendation 12

Inquire after the experiment whether the participants associate the HMI (interaction) with something they have experienced before. This provides insights into mental models and knowledge transfer.

This recommendation is assigned to the specified users referred to in the ISO standard 9241-11 (ISO, 2018a). It is based on the preliminary recommendation 15 (Table 9.1). Following the experts' conclusions, the first part of the preliminary recommendation is discarded as it overlaps with the previous recommendation. The second part is rephrased using a more general formulation of the information to be obtained. The experts acknowledge the empirical findings in *Exp_Testing-Environment* and *Exp_Culture*, indicating that single participants transfer knowledge from similar systems or other information sources. Thus, they propose asking whether participants experience associations when interacting with the HMI concept(s).

9.4 Discussion and Outlook

This chapter presents the preliminary recommendations regarding methods to assess the usability of HMIs for L3 ADS formulated by the author of this thesis and the subsequent discussion of these recommendations during an expert workshop. The recommendations are supplemented with observations and comments to facilitate the recommendations' understanding and motivation. After the presentation of the workshop results, the final recommendations are supplemented with a short explanation. The explanation comprises (empirical) findings that partially seem contradictory. By integrating the findings, the explanation intends to facilitate the understanding of deriving the recommendations.

During the expert workshop, two main problems are identified. First, the preface setting the goals and scope of the recommendations is considered too short, leaving room for interpretation. Thus, clarifying the recommendations' goals and limitations formed a recurring theme during the discussion. Second, several recommendations are criticized for being

imprecise, covering several aspects simultaneously. This prolonged and aggravated the discussion since the aspects had to be identified and differentiated before the discussion of the individual aspects could take place.

In addition to these content-related points of criticism, further limitations of the expert workshop should be mentioned. All participating experts are experienced in conducting user tests in the field of HMIs for ADS. Since all experts have worked at the Chair of Ergonomics, the experiences overlap, and their methodological approaches are likely similar. Thus, the resulting variance in the workshop results is assumed to be limited. Furthermore, the expert workshop aims to discuss the set of preliminary recommendations. Therefore, the experts could not formulate recommendations unaffected by the set of recommendations formulated by the author of this thesis. A brainstorming session of the experts without prior input in the form of preliminary recommendations could have provided valuable results, potentially providing a broader picture of recommendations regarding methods to assess the usability of HMIs for L3 ADS. Nonetheless, the expert workshop has generated valuable knowledge within the scope of this chapter. The results form the basis for formulating the final set of recommendations.

A few aspects need to be mentioned regarding formulating the final set of recommendations. The scope of the recommendations is limited. Four of the preliminary recommendations are discarded in the final set of recommendations. Two of these recommendations (preliminary recommendations 10 & 13; Table 9.1) are deemed not specific enough for the scope of the recommendations. The preliminary recommendation 1 (Table 9.1) covers the definition of usability, which is now included in the preface. The preliminary recommendation 12 (Table 9.1) suggests only conducting tests within one culture. Considering the limitations of the two studies—foremost the testing of two Western cultures only—the experts conclude that no general recommendation regarding the cultural effects should be made. The knowledge of cultural effects gained in this thesis is not translated into the formulation of a recommendation. Despite this experts' perception, the findings in *Exp_Culture* and *Survey_Culture* provide valuable findings that are considered in the following chapter.

Finally, the experts identify research gaps for the potential effect of the attitude toward automated driving on quantitative assessments of HMIs or ADS and for developing methods and thresholds for absolute usability assessments.

10 Conclusion

This chapter summarizes and reflects the learnings and limitations of this thesis. The results of the chapters Chapter 4 to Chapter 9 are discussed individually in their respective chapters. Thus, this chapter focuses on the general discussion. The first section briefly presents the answers to the five research questions posed in Chapter 3. Afterward, the limitations of this thesis and the generalizability of results are reflected. Following, fields for future research efforts are identified. The thesis concludes with the formulation of key messages.

10.1 Answers to the Research Questions

This thesis strives to answer the questions on the effect of the testing environment and the users' cultural background on the usability assessment of HMIs for L3 ADS. In the following, the main findings of this thesis are presented and briefly discussed alongside the research questions.

10.1.1 RQ₁: Based on Common Research Methods and Findings, what is the Best Practice Advice for an Experimental Design for Assessing the Usability of HMIs for L3 ADS?

The answer to research question RQ₁ is addressed in Chapter 4 and Chapter 5. In the first step, a systematic literature review is conducted comprising 16 study and theoretical articles. The articles are analyzed regarding the six study characteristics: definition of usability, testing environment, sample characteristics, test cases, dependent variables, and conditions of use. The review reveals that data on the selected study characteristics is unavailable in some articles. The most striking observation is that four articles do not define usability. Nonetheless, a best practice advice is developed based on the review's results.

The best practice advice recommends defining and operationalizing usability in the context of HMIs for L3 ADS by combining the definition provided by the ISO standard 9241-11 (ISO, 2018a) and the NHTSA minimum requirements (NHTSA, 2017). Regarding the testing environment, the best practice advice recommends using high-fidelity driving simulators. For early prototypes, the value of desktop methods is acknowledged. The best practice advice further recommends conducting tests with potential end users. The sample characteristics are supposed to represent the potential user regarding the distribution of characteristics such as age, gender, prior experience, or affiliation with technical devices. The best practice advice recommends focusing on transitions between LoAs, the availability of LoAs, and non-critical scenarios for selecting test cases. The database for the usability assessment is recommended to include both observational and self-reported metrics. The observational data are further specified in collecting visual behavior and interaction performance data. The advice recommends applying the *SUS*, short interviews, and supplementing standardized questionnaires for self-reported data. Finally, the best practice advice recommends providing only general information on the ADS and testing the first contact interaction for the conditions of use.

In the second step, the best practice advice is transcribed into a study design of the validation studies *Exp_Testing-Environment* and *Exp_Culture*. Considering the challenges of

the varying test settings, the applied definition of usability, the targeted sample characteristics, the test cases, the study procedure, the selection of dependent variables, and their analysis are drafted. Furthermore, two HMI concepts are developed that vary in their compliance with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019).

10.1.2 RQ₂: Which Effect has the Testing Environment on Metrics for Assessing the Usability of HMIs for L3 ADS?

In Chapter 6, the validation study *Exp_Testing-Environment* is presented. A static driving simulator experiment is compared to an experiment conducted in an instrumented vehicle on a test track to answer research question RQ₂. The experimental design follows the general method presented in Chapter 5 and includes several observational and self-reported metrics to assess usability. The study sample includes $N = 113$ participants ($n_{Sim_GER} = 52$ & $n_{TT_GER} = 61$) that experience either an HMI for L3 ADS that is highly compliant (*HC-HMI*) with guidelines for HMI design (Naujoks, Wiedemann, et al., 2019) or deliberately violates these guidelines (*LC-HMI*). After receiving instruction on the study procedure and the ADS functionalities, the participants experience a set of 12 test cases covering continuous rides in L0, L2, and L3, changes of the availabilities of LoAs, and transitions between LoAs triggered by the experimenter or the ADS. Based on the literature, two hypotheses are formulated regarding the effect of the testing environment. The hypotheses expect relative validity for the static driving simulator compared to the test track but no absolute validity.

The results show that no differences between the testing environments are identified in most cases. In single cases, an effect of the testing environment is observed, showing that differences between HMI concepts are more extreme in the test track testing environment compared to the simulator. Thus, relative validity is confirmed. While several metrics show absolute validity, overall absolute validity is rejected. The study concludes that problems with HMI concepts identified in driving simulator environments will likely be more pronounced in test track environments. Based on the findings, driving simulators are deemed a valid tool to assess the usability of HMIs for L3 ADS.

10.1.3 RQ₃: Which Effect has the Users' Cultural Background on Metrics for Assessing the Usability of HMIs for L3 ADS?

Chapter 7 presents the validation study *Exp_Culture*. Two test track experiments are conducted. One experiment is conducted with a sample of German participants and one with a sample of U.S.-American participants. The total study sample includes $N = 103$ participants ($n_{TT_GER} = 61$ & $n_{TT_USA} = 42$). The data of *TT_GER* is reused from the validation study *Exp_Testing-Environment*. The experimental design of *Exp_Culture* is identical to the experimental design of *Exp_Testing-Environment* and follows the general method presented in Chapter 5. Based on the literature, two hypotheses are formulated regarding the effect of the testing environment. The hypotheses expect relative and absolute validity to be demonstrated for the observational metrics, while both forms of validity are not expected for the self-reported metrics.

Most metrics show no differences, but equivalence is confirmed for only about a third of the metrics. There are no interactions between the factors *Exp* and *HMI*. Additionally, the inferential statistical tests and the descriptive analyses for observational and self-reported do

not show systematic differences between the experimental conditions. Only small differences in the metric *Final interview* regarding user preferences are detected. Concluding, relative validity is shown for observational data and self-reported data. The database is inconclusive to conclude on absolute validity for observational or self-reported data. Based on the study's results, it is believed that absolute validity should not be anticipated in potential future studies. Further research conducted with more diverse cultures is recommended. Due to relative validity, the study recommends conducting usability assessments for HMIs for L3 ADS within one culture of the Western industrialized world with limitations if the focus is on the facet satisfaction.

10.1.4 RQ4: Which Effect has the Users' Cultural Background on the Subjective Importance Rating of Usability Factors in the Context of HMIs for L3 ADS?

In Chapter 8, the validation study *Survey_Culture* is presented. Data are collected in an extra survey during the experiment *TT_USA*, which is part of the validation study *Exp_Culture*. Furthermore, two online surveys are conducted with participants from Germany and the United States. The total study sample includes $N = 110$ participants ($n_{TT_USA} = 42$; $n_{non_USA} = 30$ & $n_{non_GER} = 38$). The survey first inquires information on the cultural values of the participants (Hofstede & Minkov, 2013b) and afterward inquires participants to rate their subjective importance of 25 usability factors (Hinderks et al., 2019) in the provided context of HMIs for L3 ADS.³⁴ Based on the literature, three hypotheses, including eight usability factors, expect directional differences between the U.S.-American and German samples.

The results on cultural values show that the U.S.-American sample recruited in Germany for *TT_USA* represents U.S.-American citizens regarding its cultural values. The results show the same or similar trends for differences between the U.S.-American and German samples. This aligns with the reference data provided by Hofstede Insights (2023). The study data considerably deviates from the reference data in two dimensions. Possible influences could be attributed to differences in the sample, such as age.

Regarding the subjective importance ratings of the usability factors, no considerable or systematic effect of the users' cultural background is identified. Therefore, all three hypotheses are rejected. The usability factors *Controllability/Dependability* and *Ease of Use* (Hinderks et al., 2019) are rated among the most important factors in all samples. To sum up, the study provides no evidence that there is an effect of the subjective importance rating of usability factors in the context of HMIs for L3 ADS between cultures. Based on previous findings by other researchers, the transferability of results to more diverse cultures is, however, expected to be limited.

10.1.5 RQ5: Which Methods are recommended for Assessing the Usability of HMIs for L3 ADS?

In Chapter 9, the results of an expert workshop are presented. A preliminary preface and a set of 15 recommendations is formulated derived from the empirical findings obtained throughout the thesis. The preface and the recommendations are discussed in an expert

³⁴ Due to a technical error, data of only 24 usability factors (Hinderks et al., 2019) are available.

workshop with five employees of the Chair of Ergonomics. The workshop's results are transcribed into the formulation of a preface and 12 final recommendations (see Table 9.2).

The preface sets the goals and scope of the recommendations. The recommendations focus on the relative and intuitive usability of HMIs for L3 ADS that are still in a prototype state. Furthermore, safety and ethical aspects as well as the compatibility with the research focus should be prioritized over the recommendations. Finally, the preface presents the definition of usability according to the ISO standard 9241-11 (ISO, 2018a, p. 2). The recommendations are assigned to the terms included in the definition. The set of recommendations does not claim to be an extensive checklist when planning experiments in this research field. Instead, the recommendations serve as a decision aid in the planning process that may be amended in future research efforts. Among others, the recommendations cover the selection of the test cases, the dependent variables, and the data analysis. Furthermore, the recommendations provide orientation in selecting the study design and testing environment, as well as the definition of the study procedure and relevant study sample.

Following the results of the expert workshop, the knowledge gained on cultural effects is regarded to be too inconclusive. Therefore, the learnings are not included in the final set of recommendations.

10.2 Reflections on Limitations and Generalizability

This section discusses reflections regarding this thesis' procedure and the conclusions' generalizability.

10.2.1 Limitations

The validation studies *Exp_Testing-Environment* and *Exp_Culture* build the center of the empirical data basis reported in this thesis. The experimental design of the studies is derived from the best practice advice based on the literature review presented in Chapter 4. The analysis of six chosen study characteristics (e.g., definition of usability) proves difficult since the information on these study characteristics is limited. Following the Guidelines for Safeguarding Good Research Practice (Deutsche Forschungsgemeinschaft e.V., 2022), research should be extensively documented, ensuring transparency and the ability to repeat research. Due to this main limitation of analyzing the status quo, an influence on the derived experimental design of the validation studies *Exp_Testing-Environment* and *Exp_Culture* cannot be ruled out. This thesis advocates the documentation of all relevant study characteristics, enabling the comparison of research results and the use of synergies in developing HMIs for ADS. In the context of usability research of HMIs for ADS, these study characteristics could be derived from the terms included in the definition of usability provided by the ISO standard 9241-11 (ISO, 2018a). That is, the system product, system, or service; the specified users; the specified goals; the metrics for effectiveness, efficiency, and satisfaction; and, most importantly, the context of use should be described.

Two HMI concepts are developed as research subjects in the validation studies *Exp_Testing-Environment* and *Exp_Culture*. A heuristic evaluation with experts yields considerable differences in the HMI concepts that exceed the differences identified in the user tests. This observation supports the general recommendation of conducting usability tests with

experts before conducting resource-intensive user tests (Dumas & Salzman, 2006, p. 133; Naujoks, Hergeth, et al., 2019). Experts are familiar with guidelines and norms as well as common pitfalls in design. Thus, fundamental usability problems can efficiently be identified in this stage of HMI development. User studies may, instead, focus on discovering usability problems and attitudes specific to certain user groups, for example, depending on the users' mental models or naivety.

In validation studies *Exp_Testing-Environment* and *Exp_Culture*, usability is treated as a relative measure depending on the two HMI concepts. It is explicitly not the aim of this thesis to investigate methods assessing absolute usability. While this is a limitation of this thesis, comparing different HMI concepts or components of HMI concepts is a common use case in research. Hence, the assessment of relative usability is legitimate. Nevertheless, the findings of this thesis may build a basis for future research efforts on methods for absolute usability assessments or the validation of existing approaches (e.g., Bangor et al., 2009). An advantage of testing relative usability is that potential differences in response behavior can be ignored since all research subjects are affected equally. In cross-cultural research, response biases are known (e.g., Baumgartner & Steenkamp, 2001; Moss & Vijayendra, 2020). The results of the validation study *Survey_Culture* show a systematic offset in the importance rating, increasing the relevance of the relative importance rankings when interpreting the results.

While the resulting usability assessments are of the highest importance to practitioners, the effects of the testing environment and the users' cultural background on the research method, that is, the data collection phase, play an important role, too. As described before, culture affects the response behavior (Baumgartner & Steenkamp, 2001; Moss & Vijayendra, 2020). Also, the match between the participants' and their experimenter's cultural backgrounds influences the collected data (Vatrapu & Pérez-Quiñones, 2006). Regarding testing environments, factors such as immersion or perceived risk are potential confounding factors (Ranney, 2011). Hence, for a comprehensive investigation of the effects of testing environment and users' cultural background on usability assessments, more attention should have been paid to the potential effects during the data collection phase.

10.2.2 Generalizability

The thesis focuses on HMIs for L3 ADS since L3 entails a paradigm change in the role of the human in the car compared to lower LoAs. This thesis includes LoAs L0, L2, and L3, and the test cases cover different aspects, such as continuous rides, transitions, or changes in the availability of LoAs. No systematic differences are observed regarding individual LoAs. Therefore, no differences regarding the validity of this thesis' results are expected, even though use cases might differ when other LoAs are investigated.

Regarding testing environments, including other types of driving simulators (e.g., low-fidelity driving) or other settings such as real-world tests and field studies, future research is needed to validate the results of this thesis. Single observational metrics observe more extreme differences between the HMI concepts in the test track testing environment compared to the driving simulator experiment. A potential reason is the higher workload, possibly induced by a more complex setting with more distractions and sensory impressions (Purucker et al., 2018). This effect might be expressed more pronounced in real-world or field test settings. A similar concern may be expressed regarding the complexity of the test case scenarios. The scenarios in

this thesis are simple. A potential interaction between the workload induced through scenario complexity and the workload induced through the testing environment affecting the usability assessment cannot be ruled out.

Due to the COVID-19 pandemic, only two Western industrialized countries could be compared. As discussed before, the generalizability of the results to other, more different cultures is expected to be limited. Specific HMI components, such as the implementation of avatars, which are common in Asian countries, might amplify such differences.

Finally, the thesis' focus on usability is narrow, ignoring other constructs to assess the quality of the HMI design. The terms usability and related constructs are presented in Section 2.1. Since they are closely related, the results sections of the validation studies *Exp_Testing-Environment* and *Exp_Culture* partially report results on the constructs trust, acceptance, UX, and workload. No divergent trends for these constructs are observed in the results. Therefore, generalizability is plausible, and the thesis' results may be applied to future research, e.g., on UX of HMIs for ADS. Nevertheless, validation studies, including a comprehensive set of metrics to assess the respective constructs, would be needed to confirm generalizability.

10.3 Future Work

Validation studies for driving simulators demonstrate at least relative validity in most results. However, possibly due to reasons such as a lack of visual details or perceived risk, several studies could not confirm the validity of driving simulators. This stresses the importance of continuous research efforts to check the validity of driving simulators, especially for new research fields like automated driving. Therefore, future research is recommended to investigate the validity of other driving simulators. Considering the recent market launch of vehicles equipped with L3 ADS, addressing the validity of driving simulators compared to real-world settings is now feasible and poses a promising research topic.

Likewise, deepening the insights into the effects of this thesis' other main research subject—culture—is imperative. Due to globalization, the blending of different cultures and the worldwide marketing of products are ubiquitous, stressing the importance of developing HMIs that consider potential cultural differences. Following the literature and the conclusions of this thesis, differences regarding the interaction quality with HMIs for ADS may be expected if cultural differences are more pronounced. Thus, future research may supplement this thesis' work. Additionally, future research may take up the challenge of tying cultural values to usability assessments. The validation study *Survey_Culture* could not confirm hypotheses regarding the subjective importance rating of usability factors based on cultural values. This, however, is important for predicting cultural differences and consequently developing tailored HMI designs.

In addition to cross-cultural research, the research methods of different cultures and regions pose a challenge in itself. Albers, Grabbe, et al. (2020) conducted a workshop with international researchers and practitioners in automotive HMIs. The workshop reveals that research approaches vary across regions and research institutions. Future research to standardize research methods are indicated.

Finally, establishing thresholds and measures for absolute usability ratings for HMIs for ADS poses an onerous but worthy endeavor. This thesis' scope is limited to relative usability assessments that do not allow for a conclusion on the absolute usability. In the broad field of usability research, single attempts for absolute usability ratings exist, such as the *SUS* categories introduced by Bangor et al. (2009). These methodological thresholds are highly dependent on the field of research and the context of applying the measures. Therefore, transferring or establishing criteria for absolute usability to the field of HMIs for ADS requires diligent and exhaustive research efforts. Advances in this field promise the enhancement of comparability of studies on HMIs for ADS and thus promote the use of synergies.

10.4 Key Messages

The thesis closes with the formulation of five key messages:

1. Improvements in documenting research regarding relevant study characteristics are needed to facilitate the use of synergies or at least ensure comparability and transparency of the research. For usability research, this thesis recommends reporting details of the experimental design on each component of the definition provided by the ISO standard 9241-11 (ISO, 2018a).
2. In line with the literature (Dumas & Salzman, 2006, p. 133; Naujoks, Hergeth, et al., 2019), this thesis suggests conducting expert reviews before user tests in the development process of HMIs for ADS. A heuristic evaluation conducted with experts confirmed considerable differences in the compliance of two HMI concepts with guidelines for HMI design. In contrast, in the user studies conducted with the two HMI concepts, the identified differences are not extreme and overlap only partially with the differences identified by the experts.
3. The thesis' findings support literature (Frøkjær et al., 2000; Hornbæk, 2006; ISO, 2018a) advocating testing all facets of usability when assessing usability. Furthermore, a mix of observational and self-reported metrics is recommended. Finally, metrics and test cases need to be selected considering all components of the HMI concept, for example, testing the control logic in first contact use to identify problems with the toggle logic.
4. Findings of the validation study *Exp_Testing-Environment* indicate that high-fidelity driving simulators offer relative validity when researching HMIs for ADS. Problems identified in driving simulators are expected to be more pronounced in more complex, for example, real-world settings.
5. Based on the validation studies *Exp_Culture* and *Survey_Culture*, differences in usability assessments between Western industrialized countries are expected to be minor. Relative validity is confirmed, and differences in the usability facet satisfaction are small, only expressed in qualitative data. In the scope of this thesis, no conclusion can be drawn regarding the effects of users' cultural backgrounds where cultural differences are more prominent.

11 References

- Abed, F. (1991). Cultural influences on visual scanning patterns. *Journal of Cross-Cultural Psychology*, 22(4), 525–534. <https://doi.org/10.1177/0022022191224006>
- Albers, D., Grabbe, N., Janetzko, D., & Bengler, K. (2020). Saluton! How do you evaluate usability? – Virtual Workshop on Usability Assessments of Automated Driving Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 109–112). ACM. <https://doi.org/10.1145/3409251.3411737>
- Albers, D., Radlmayr, J., Grabbe, N., Hergeth, S., Naujoks, F., Forster, Y., Keinath, A., & Bengler, K. (2021). Human-machine interfaces for automated driving: Development of an experimental design for evaluating usability. In N. L. Black, W. P. Neumann, & I. Noy (Eds.), *Lecture Notes in Networks and Systems: Vol. 221, Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021): Volume III: Sector Based Ergonomics* (1st ed., pp. 541–551). Springer Cham. https://doi.org/10.1007/978-3-030-74608-7_66
- Albers, D., Radlmayr, J., Loew, A., Hergeth, S., Naujoks, F., Keinath, A., & Bengler, K. (2020). Usability evaluation - Advances in experimental design in the context of automated driving human-machine interfaces. *Information*, 11(5), 240. <https://doi.org/10.3390/info11050240>
- Albert, M., Lange, A., Schmidt, A., Wimmer, M., & Bengler, K. (2015). Automated driving – Assessment of interaction concepts under real driving conditions. *Procedia Manufacturing*, 3, 2832–2839. <https://doi.org/10.1016/j.promfg.2015.07.767>
- Amazon Mechanical Turk, Inc. (2023). *Amazon Mechanical Turk*. <https://www.mturk.com/>
- Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3), 114–123. <https://doi.org/10.5555/2835587.2835589>
- Barber, W., & Badre, A. (1998). Culturability: The merging of culture and usability. In *Proceedings of the 4th Conference on Human Factors and the Web*.
- Barnum, C. M. (2021). *Usability testing essentials: Ready, set... test!* (2nd edition). Morgan Kaufmann Elsevier.
- Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Bella, F. (2008). Driving simulator for speed research on two-lane rural roads. *Accident Analysis and Prevention*, 40(3), 1078–1087. <https://doi.org/10.1016/j.aap.2007.10.015>
- Bellem, H., Klüver, M., Schrauf, M., Schöner, H.-P., Hecht, H., & Krems, J. F. (2017). Can we study autonomous driving comfort in moving-base driving simulators? A validation study. *Human Factors*, 59(3), 442–456. <https://doi.org/10.1177/0018720816682647>
- Bengler, K., Mattes, S., Hamm, O., & Hensel, M. (2010). Lane change test: Preliminary results of a multi-laboratory calibration study. In G. Rupp (Ed.), *Performance metrics for assessing driver distraction: The quest for improved road safety* (1st ed., pp. 243–253). SAE International.
- Bengler, K., Rettenmaier, M., Fritz, N., & Feierle, A. (2020). From HMI to HMIs: Towards an HMI framework for automated driving. *Information*, 11(2), 61. <https://doi.org/10.3390/info11020061>
- Bevan, N., Kirakowski, J., & Maissel, J. (1991). What is usability? In *Proceedings of the 4th International Conference on HCI*, Stuttgart.
- Blaauw, G. J. (1982). Driving experience and task demands in simulator and instrumented car: A validation study. *Human Factors*, 24(4), 473–486. <https://doi.org/10.1177/001872088202400408>

- Blana, E. (1996). Driving simulator validation studies: A literature review. *ITS Working Paper*, 480.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler: Mit 242 Tabellen* (6., vollst. überarb. und aktualisierte Aufl.). Springer-Lehrbuch. Springer Medizin.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler* (4th ed.). Springer-Verlag.
- Brandes, R., Lang, F., & Schmidt, R. F. (2019). *Physiologie des Menschen: mit Pathophysiologie* (32nd ed.). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-56468-4>
- Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, L. I. McClelland, & B. A. Weerdmeester (Eds.), *Usability evaluation in industry* (1st ed., pp. 189–194). CRC Press.
- Bruder, R., Abendroth, B., & Landau, K. (2007). Zum Nutzen von Fahrversuchen für die Gestaltung. In R. Bruder & H. Winner (Eds.), *3. Darmstädter Kolloquium* (79-95). ergonomia.
- Bubb, H., Bengler, K., Grünen, R. E., & Vollrath, M. (Eds.). (2015). *Automobilergonomie*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-8348-2297-0>
- Bundesanstalt für Straßenwesen. (2021). *Selbstfahrende Autos – assistiert, automatisiert oder autonom?* [Nr.: 06/2021]. <https://www.bast.de/DE/Presse/Mitteilungen/2021/06-2021.html>
- Caird, J. K., & Horrey, W. J. (2011). Twelve practical and twelve practical and useful questions about driving simulation. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology*. CRC Press.
- Caldwell, A. R. (2022). *Exploring equivalence testing with the updated TOSTER r package*. <https://doi.org/10.31234/osf.io/ty8de>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <https://doi.org/10.1037/h0040950>
- Campbell, J. L., Richard, C. M., Brown, J. L., & McCallum, M. (2007). *Crash warning system interfaces: Human factors insights and lessons learned: Final report*. DOT HS 810 697. Washington. National Highway Traffic Safety Administration.
- Chau, P. Y. K., Cole, M., Massey, A. P., Montoya-Weiss, M., & O'Keefe, R. M. (2002). Cultural differences in the online behavior of consumers. *Communications of the ACM*, 45(10), 138–143. <https://doi.org/10.1145/570907.570911>
- Chetty, M., Buckhalter, C., Best, M., Grinter, R. E., & Guzdial, M. (2007). *Description of computer science higher education in Sub-Saharan Africa: Initial explorations*. Technical Report GIT-GVU-07-14. College of Computing, Georgia Institute of Technology.
- Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, 32(4), 347–361. <https://doi.org/10.1007/s10869-016-9458-5>
- Child, J. (1981). Culture, contingency and capitalism in the cross-national study of organizations. *Research in Organizational Behavior*, 3, 303–356.
- Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In E. Soloway, D. Frye, & S. Sheppard (Eds.), *Human factors in computing systems: Chi '88 Conference proceedings, May 15 - 19, 1988, Washington, D.C* (pp. 213–218). Association for Computing Machinery. <https://doi.org/10.1145/57167.57203>
- Christensen, R. H. B. (2022). *ordinal - regression models for ordinal data* (Version 2022.11-16) [Computer software]. <https://CRAN.R-project.org/package=ordinal>

- Clercq, K. de, Dietrich, A., Núñez Velasco, J. P., Winter, J. de, & Happee, R. (2019). External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Human Factors*, 61(8), 1353–1370. <https://doi.org/10.1177/0018720819836343>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Courage, C., & Baxter, K. (2005). *Understanding your users: A practical guide to user requirements methods, tools, and techniques*. Elsevier. <https://doi.org/10.1016/B978-1-55860-935-8.X5029-5>
- Del Re, A. C. (2013). *compute.es: Compute effect sizes* [Computer software]. <https://cran.r-project.org/package=compute.es>
- Deutsche Forschungsgemeinschaft e.V. (2022). *Guidelines for Safeguarding Good Research Practice: Code of Conduct*. Bonn.
- Dey, D., Matviienko, A., Berger, M., Pfleging, B., Martens, M., & Terken, J. (2021). Communicating the intention of an automated vehicle to pedestrians: The contributions of eHMI and vehicle behavior. *It - Information Technology*, 63(2), 123–141. <https://doi.org/10.1515/itit-2020-0025>
- DIN (2003). *Grund- und Sicherheitsregeln für die Mensch-Maschine-Schnittstelle, Kennzeichnung: Codierungsgrundsätze für Anzeigengeräte und Bedienteile* (DIN EN 60073:2003-05).
- Douglas, I., & Liu, Z. (2011). *Global usability*. Springer London. <https://doi.org/10.1007/978-0-85729-304-6>
- Drew, M. R., Falcone, B., & Baccus, W. L. (2018). What does the System Usability Scale (SUS) measure? In A. Marcus & W. Wang (Eds.), *Lecture Notes in Computer Science: Vol. 10918, Design, User Experience, and Usability: Theory and Practice: 7th International Conference, DUXU 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I* (pp. 356–366). Springer International Publishing. https://doi.org/10.1007/978-3-319-91797-9_25
- Dumas, J. S., & Salzman, M. C. (2006). Usability assessment methods. *Reviews of Human Factors and Ergonomics*, 2(1), 109–140. <https://doi.org/10.1177/1557234X0600200105>
- Edelmann, A., Stümper, S., & Petzoldt, T. (2021). Cross-cultural differences in the acceptance of decisions of automated vehicles. *Applied Ergonomics*, 92, 103346. <https://doi.org/10.1016/j.apergo.2020.103346>
- Ergoneers Group. (2022). *D-Lab* (Version 3.60.9179.0) [Computer software].
- Faas, S. M., & Baumann, M. (2019). Light-based external human machine interface: Color evaluation for self-driving vehicle and pedestrian interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, pp. 1232–1236). Sage Publications, Incorporated. <https://doi.org/10.1177/1071181319631049>
- Feierle, A., Danner, S., Steininger, S., & Bengler, K. (2020). Information needs and visual attention during urban, highly automated driving - An investigation of potential influencing factors. *Information*, 11(2), 62. <https://doi.org/10.3390/info11020062>
- Feldhüller, A., Härtwig, N., Kurpiers, C., Hernandez, J. M., & Bengler, K. (2019). Effect on mode awareness when changing from conditionally to partially automated driving. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics* (pp. 314–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_34
- Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5), 323–327. <https://doi.org/10.1016/j.intcom.2010.04.004>

- Fischer, O., & Al-Issa, A. (2012). In for a surprise piloting the Arab version of the VSM 94. *International Journal of Intercultural Relations*, 36(5), 737–742. <https://doi.org/10.1016/j.ijintrel.2012.04.007>
- Fors, C., Ahlström, C., & Anund, A. (2013). *Simulator validation with respect to driver sleepiness and subjective experiences: Final report of the project SleepEYE II, part 1*. Virtual Prototyping and Assessment by Simulation.
- Forster, Y. (2020). *Preference versus performance in automated driving: A challenge for method development* [Doctoral thesis]. Chemnitz University of Technology, Chemnitz.
- Forster, Y., Hergeth, S., Naujoks, F., Beggiato, M., Krems, J. F., & Keinath, A. (2019). Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, 62, 599–614. <https://doi.org/10.1016/j.trf.2019.02.013>
- Forster, Y., Hergeth, S., Naujoks, F., Krems, J., & Keinath, A. (2019). User education in automated driving: Owner's manual and interactive tutorial support mental model formation and human-automation interaction. *Information*, 10(4), 143. <https://doi.org/10.3390/info10040143>
- Forster, Y., Hergeth, S., Naujoks, F., Krems, J. F., & Keinath, A. (2020a). Empirical validation of a checklist for heuristic evaluation of automated vehicle HMIs. In N. A. Stanton (Ed.), *Advances in Human Factors of Transportation: Proceedings of the AHFE 2019 International Conference on Human Factors in Transportation, July 24-28, 2019, Washington D.C., USA* (Vol. 964, pp. 3–14). Springer International Publishing. https://doi.org/10.1007/978-3-030-20503-4_1
- Forster, Y., Hergeth, S., Naujoks, F., Krems, J. F., & Keinath, A. (2020b). Self-report measures for the assessment of human-machine interfaces in automated driving. *Cognition, Technology & Work*, 22(4), 703–720. <https://doi.org/10.1007/s10111-019-00599-8>
- Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (Version 3) [Computer software]. Sage. Thousand Oaks, CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- François, M., Osiurak, F., Fort, A., Crave, P., & Navarro, J. (2017). Automotive HMI design and participatory user involvement: Review and perspectives. *Ergonomics*, 60(4), 541–552. <https://doi.org/10.1080/00140139.2016.1188218>
- Frandsen-Thorlacius, O., Hornbæk, K., Hertzum, M., & Clemmensen, T. (2009). Non-universal usability? A survey of how usability is understood by Chinese and Danish users. In *Chi '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: April 4 - 9, 2009 in Boston, USA* (pp. 41–50). ACM Press. <https://doi.org/10.1145/1518701.1518708>
- Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 345–352). ACM. <https://doi.org/10.1145/332040.332455>
- Fuest, T., Feierle, A., Schmidt, E., & Bengler, K. (2020). Effects of marking automated vehicles on human drivers on highways. *Information*, 11(6), 286. <https://doi.org/10.3390/info11060286>
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51. <https://doi.org/10.2307/30036519>

- Gong, Z., Ma, J., Zhang, Q., Ding, Y., & Liu, L. (2020). Automotive HMI guidelines for China based on culture dimensions interpretation. In C. Stephanidis, V. G. Duffy, N. Streitz, S. Konomi, & H. Krömker (Eds.), *Lecture Notes in Computer Science. HCI International 2020 – Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments: 22nd HCI International Conference, HCII 2020* (pp. 96–110). Springer International Publishing. https://doi.org/10.1007/978-3-030-59987-4_8
- Goodenough, R. (2010). *The geometric field of view and speed perception in a driving simulator* [Master's thesis]. Clemson University. http://tigerprints.clemson.edu/cgi/viewcontent.cgi?article=1978&context=all_theses
- Götze, M. (2018). *Entwicklung und Evaluation eines integrativen MMI Gesamtkonzeptes zur Handlungsunterstützung für den urbanen Verkehr* [Doctoral thesis]. Technical University of Munich, Garching.
- Graefe, J. (2021). *Realfahrzeugstudie zur Evaluation zweier Bedienkonzepte für das automatisierte Fahren* [Master's thesis]. Technical University of Munich, Garching.
- Green, P., Levison, W., Paelke, G., & Serafin, C. (1994). *Suggested human factors design guidelines for driver information systems*. Technical Report UMTRI-93-21 (FHWA-RD-94-087). The University of Michigan Transportation Research Institute.
- Heimgärtner, R. (2007). Towards Cultural Adaptability in Driver Information and -Assistance Systems. In N. Aykin (Chair), *Usability and Internationalization. Global and Local User Interfaces: Second International Conference on Usability and Internationalization, UI-HCII 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part II*.
- Hergeth, S., Lorenz, L [Lutz], Krems, J. F., & Toenert, L. (2015). Effects of take-over requests and cultural background on automation trust in highly automated driving. In *Driving Assessment Conference 2015*. Symposium conducted at the meeting of University of Iowa, Iowa City, Salt Lake City, Utah, USA.
- Herman, L. (1996). Towards effective usability evaluation in Asia: Cross-cultural differences. In *Proceedings Sixth Australian Conference on Computer-Human Interaction* (pp. 135–136). IEEE. <https://doi.org/10.1109/OZCHI.1996.559999>
- Herzberg, F., Mausner, B., & Snyderman, B. B. (1967). *The motivation to work* (2nd ed.). Wiley.
- Hinderks, A., Winter, D., Thomaschewski, J., & Schrepp, M. (2019). Applicability of user experience and usability questionnaires. *Journal of Universal Computer Science*, 25(13), 1717–1735. <https://doi.org/10.3217/jucs-025-13-1717>
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (Second edition). Sage Publications. <https://doi.org/010498>
- Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1014>
- Hofstede, G. (2013). Replicating and extending cross-national value studies: Rewards and pitfalls – An example from Middle East Studies. *Insights*, 13(2), 5–7.
- Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival* (Revised and expanded third edition). McGraw-Hill. <http://www.loc.gov/catdir/enhancements/fy1009/2010010437-b.html>
- Hofstede, G., & Minkov, M. (2013a). *VSM 2013: Values survey module 2013 manual*. Geert Hofstede BV.
- Hofstede, G., & Minkov, M. (2013b). *VSM 2013: Values survey module 2013 questionnaire*. English language version. Geert Hofstede BV.

- Hofstede Insights. (2023). *Country Comparison Tool* [Germany vs. United States]. <https://www.hofstede-insights.com/country-comparison-tool?countries=germany%2Cunited+states>
- Honold, P. (1999). Learning how to use a cellular phone: Comparison between German and Chinese users. *Technical Communication*, 46(2), 196–205.
- Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2), 79–102. <https://doi.org/10.1016/j.ijhcs.2005.06.002>
- Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65(1), 19–51. <https://doi.org/10.2307/2657288>
- ISO (2009). *Straßenfahrzeuge – Ergonomische Aspekte von Fahrerinformations- und Assistenzsystemen: Anforderungen und Bewertungsmethoden der visuellen Informationsdarstellung im Fahrzeug* (DIN EN ISO 15008:2009). Berlin.
- ISO (2012). *Road vehicles - Ergonomic aspects of transport information and control systems - Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems* (14198). Geneva.
- ISO (2018a). *Ergonomics of human-system interaction: Part 11: Usability: Definitions and concepts* (ISO 9241-11). Geneva.
- ISO (2018b). *Road vehicles - Measurement of driver visual behaviour with respect to transport information and control systems* (ISO/CD 15007:2018(E)).
- Jahoda, G. (1984). Do we need a concept of culture? *Journal of Cross-Cultural Psychology*, 15(2), 139–151. <https://doi.org/10.1177/0022002184015002003>
- Jamson, H. (2001). Image characteristics and their effect on driving simulator validity. In *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 190–195). University of Iowa, Iowa City. <https://doi.org/10.17077/drivingassessment.1036>
- Jones, P. S., Lee, J. W., Philips, L. R., Zhang, X. W., & Jaceldo, K. B. (2001). An adaptation of Brislin's translation model for cross-cultural research. *Nursing Research*, 50(5), 300–304. <https://doi.org/10.1097/00006199-200109000-00008>
- Kaptein, N. A., Theeuwes, J., & van der Horst, R. (1996). Driving simulator validity: Some considerations. *Transportation Research Record: Journal of the Transportation Research Board*, 1550(1), 30–36. <https://doi.org/10.1177/0361198196155000105>
- Kassambra, A. (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (Version 0.7.2) [Computer software]. <https://CRAN.R-project.org/package=rstatix>
- Kelsch, J., Dziennus, M., Schieben, A., Schömig, N., Wiedemann, K., Merat, N., Louw, T., Madigan, R., Kountouriotis, G., Aust, M. L., Söderman, M., Johansson, E., DLR Team, BMW Team, Continental Team, CRF Team, Daimler Team, Ford Team, IKA Team, . . . VW Team. (2017). *Final functional human factors recommendations*. AdaptIVe Deliverable D3.3. AdaptIVe Consortium.
- Khan, T., Pitts, M., & Williams, M. (2016). Cross-cultural differences in automotive HMI design: a comparative study between UK and Indian users' design preferences. *Journal of Usability Studies*, 11(2), 45–65.
- Khan, T., & Williams, M. (2014). A study of cultural influence in automotive HMI: Measuring correlation between culture and HMI usability. *SAE Int. J. Passeng. Cars – Electron. Electr. Syst.*, 7(2), 430–439. <https://doi.org/10.4271/2014-01-0263>
- Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry*. Taylor & Francis.

- Kluckhohn, C. (1959). The study of culture. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and method* (2nd ed., pp. 86–101). Stanford University Press.
- Knappe, G., Keinath, A., Bengler, K., & Meinecke, C. (2007). Driving simulator as an evaluation tool—assessment of the influence of field of view and secondary tasks on lane keeping and steering performance. In *20th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*. Symposium conducted at the meeting of National Highway Traffic Safety Administration, Lyon, France.
- Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, 66, 18–31. <https://doi.org/10.1016/j.apergo.2017.07.006>
- Kos, J. (2020). *Implementierung eines Human-Machine-Interfaces für automatisiertes Fahren in unity zur Nutzung in Realfahrtversuchen* [Bachelor's thesis]. Technical University of Munich, Garching.
- Krause, M., Yilmaz, L., & Bengler, K. (2014). Comparison of real and simulated driving for a static driving simulator. In T. Ahram, W. Karwowski, & T. Marek (Eds.), *Advances in human factors and ergonomics 2014 / ed. by Neville Stanton, Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014* (pp. 29–40). AHFE Conference.
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Large, D. R., Burnett, G., Crundall, E., Lawson, G., Skrypchuk, L., & Mouzakitis, A. (2019). Evaluating secondary input devices to support an automotive touchscreen HMI: A cross-cultural simulator study conducted in the UK and China. *Applied Ergonomics*, 78, 184–196. <https://doi.org/10.1016/j.apergo.2019.03.005>
- Large, D. R., Burnett, G., & Mohd-Hasni, Y. (2017). Capturing cultural differences between UK and Malaysian drivers to inform the design of in-vehicle navigation systems. *International Journal of Automotive Engineering*, 8(3), 112–119. https://doi.org/10.20485/jsaeijae.8.3_112
- Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *Lecture Notes in Computer Science: Vol. 5298, HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008, Proceedings* (1st ed., pp. 63–76). Springer Berlin. https://doi.org/10.1007/978-3-540-89350-9_6
- Lee, S. H., & Eom, H. (2015). Design of driver-vehicle interface to reduce mode confusion for adaptive cruise control systems. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 67–71). ACM. <https://doi.org/10.1145/2809730.2809757>
- Lee, Y. M., Madigan, R., Garcia, J., Tomlinson, A., Solernou, A., Romano, R., Markkula, G., Merat, N., & Uttley, J. (2019). Understanding the messages conveyed by automated vehicles. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 134–143). ACM. <https://doi.org/10.1145/3342197.3344546>
- Lesch, M. F., Rau, P.-L. P., Zhao, Z., & Liu, C. (2009). A cross-cultural comparison of perceived hazard in response to warning components and configurations: Us vs. China. *Applied Ergonomics*, 40(5), 953–961. <https://doi.org/10.1016/j.apergo.2009.02.004>

- Lewis, C. (1982). *Using the 'thinking-aloud' method in cognitive interface design*. RC 9265 (#40713). Yorktown Heights, NY. IBM Thomas J. Watson Research Center.
- Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *IET Intelligent Transport Systems*, 36(2), 368–378. <https://doi.org/10.1177/001872089403600215>
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1), 57–78. <https://doi.org/10.1080/10447319509526110>
- Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4. edition, pp. 1267–1312). Wiley.
- Lewis, J. R. (2019). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human–Computer Interaction*, 35(15), 1404–1419. <https://doi.org/10.1080/10447318.2018.1533152>
- Lindgren, A., Chen, F., Jordan, P. W., & Zhang, H. (2008). Requirements for the design of advanced driver assistance systems – The differences between Swedish and Chinese drivers. *International Journal of Design*, 2(2), 41–54.
- Liu, P., Jiang, Z., Li, T., Wang, G., Wang, R., & Xu, Z. (2021). User experience and usability when the automated driving system fails: Findings from a field experiment. *Accident Analysis and Prevention*, 161. <https://doi.org/10.1016/j.aap.2021.106383>
- Loew, A., Sogemeier, D., Kulessa, S., Forster, Y., Naujoks, F., & Keinath, A. (2022). A global questionnaire? An international comparison of the system usability scale in the context of an infotainment system. In T. Ahram & C. Falcão (Eds.), *AHFE International, Usability and User Experience: Proceedings of the 13th AHFE International Conference on Usability and User Experience* (pp. 224–232). AHFE International. <https://doi.org/10.54941/ahfe1001711>
- Lorenz, L [L.], Kerschbaum, P., Hergeth, S., Gold, C., & Radlmayr, J. (2015). Der Fahrer im Hochautomatisierten Fahrzeug: Vom Dual-Task zum Sequential-Task Paradigma. In *Proceedings of the 7. Tagung Fahrerassistenz*, Muenchen, Germany.
- Mahoney, M. J. (1978). Experimental methods and outcome evaluation. *Journal of Consulting and Clinical Psychology*, 46(4), 660–672. <https://doi.org/10.1037/0022-006X.46.4.660>
- Marcus, A., & Gould, E. W. (2000). Crosscurrents: Cultural dimensions and global web user-interface design. *Interactions*, 7(4), 32–46.
- The MathWorks Inc. (2022). *MATLAB* (Version 9.13.0.2126072 (R2022b) Update 3) [Computer software]. Natick, Massachusetts, US. <https://www.mathworks.com>
- McAvoy, D. S., Schattler, K. L., & Datta, T. K. (2007). Driving simulator validation for nighttime construction work zone devices. *Transportation Research Record: Journal of the Transportation Research Board*, 2015(1), 55–63. <https://doi.org/10.3141/2015-07>
- Mehler, J., Guo, Z., Zhang, A., & Rau, P.-L. P. (2021). Quick buttons on map-based human machine interface in vehicles is better or not: A Cross-cultural comparative study between Chinese and Germans. In M. Rauterberg (Ed.), *Lecture Notes in Computer Science: Vol. 12795, HCII 2021: Culture and Computing. Design Thinking and Cultural Computing* (pp. 432–449). Springer International Publishing. https://doi.org/10.1007/978-3-030-77431-8_27
- Melcher, V., Rauh, S., Diederichs, F., Widlroither, H., & Bauer, W. (2015). Take-over requests for automated driving. *Procedia Manufacturing*, 3, 2867–2873. <https://doi.org/10.1016/j.promfg.2015.07.788>
- Mendoza, L., Pauzié, A., & Mathis, L.-A. (2022). *D 6.4: ESoP on HMI for AVs*. Drive2theFuture (815001).
- Minkov, M., & Hofstede, G. (2013). *Cross-cultural analysis: The science and art of comparing the world's modern societies and their cultures*. Sage.

- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Mooshofer, N. (2020). *Probandenstudie zur Usability-Bewertung von zwei unterschiedlich kompatiblen SAE L3 Human-Machine-Interfaces* [Master's thesis]. Technical University of Munich, Garching.
- Moss, F., & Vijayendra, B. (2020). *When difference doesn't mean different*. Ipsos Views. Ipsos Knowledge Centre.
- Mourant, R. R., Rengarajan, P., Cox, D., Lin, Y., & Jaeger, B. K. (2007). The effect of driving environments on simulator sickness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, pp. 1232–1236). Sage Publications, Incorporated. <https://doi.org/10.1177/154193120705101838>
- Mullen, N., Charlton, J., Devlin, A., & Bédard, M. (2011). Simulator validity: Behaviors observed on the simulator and on the road. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology*. CRC Press.
- Naujoks, F., Hergeth, S., Wiedemann, K., Schömig, N., Forster, Y., & Keinath, A. (2019). Test procedure for evaluating the human-machine interface of vehicles with automated driving systems. *Traffic Injury Prevention*, 20(sup1), S146-S151. <https://doi.org/10.1080/15389588.2019.1603374>
- Naujoks, F., Purucker, C., Neukum, A., Wolter, S., & Steiger, R. (2015). Controllability of partially automated driving functions – Does it matter whether drivers are allowed to take their hands off the steering wheel? *Transportation Research Part F: Traffic Psychology and Behaviour*, 35, 185–198. <https://doi.org/10.1016/j.trf.2015.10.022>
- 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>
- Naujoks, F., Wiedemann, K., Schömig, N., Jarosch, O., & Gold, C. (2018). Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX*, 5, 579–592. <https://doi.org/10.1016/j.mex.2018.05.007>
- Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Sage.
- NHTSA. (2013). *Visual–manual NHTSA driver distraction guidelines for in-vehicle electronic devices* (2013-09883).
- NHTSA. (2016). *Federal automated vehicles policy: Accelerating the next revolution in roadway safety* (12507-091216).
- NHTSA. (2017). *Automated driving systems 2.0: A vision for safety* (DOT HS 812 442).
- Niehaus, F., Huang, T.-Y., Prinz, F., Bertleff, S., Voß, G., & Ladwig, S. (2020). Cultural impact on HMI of trucks: A comparative study between German and Japanese. In Aachener Kolloquium (Chair), *29th Aachen Colloquium Sustainable Mobility 2020*, Aachen.
- Nielsen, J. (1993). *Usability engineering*. Kaufmann.
- Nielsen, J. (1994). Usability inspection methods. In *Conference on Human Factors in Computing Systems: Chi 1994, Boston, Massachusetts, USA, April 24-28, 1994, Proceedings* (pp. 413–414). Addison-Wesley.
- Nielsen, J. (2000). Why you only need to test with 5 users. <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>
- Nielsen, J. (2005). *Ten usability heuristics*. ISSN 1548-5552. http://www.useit.com/papers/heuristic/heuristic_list.html

- Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In J. C. Chew & J. Whiteside (Chairs), *CHI90: Conference on Human Factors in Computing*.
- O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance* (pp. 1–49). John Wiley & Sons.
- Orlovska, J., Wickman, C., & Söderberg, R. (2020). Naturalistic driving study for automated driver assistance systems (ADAS) evaluation in the Chinese, Swedish and American markets. *Procedia CIRP*, 93, 1286–1291. <https://doi.org/10.1016/j.procir.2020.04.108>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *IET Intelligent Transport Systems*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parker, P. M. (1997). *Cross cultural statistical encyclopedia of the world: A statistical reference*. Greenwood Press.
- Peterson, M. F., & Smith, P. B. (2008). Social structures and processes in cross-cultural management. In P. B. Smith, M. F. Peterson, & Thomas David C. (Eds.), *The handbook of cross-cultural management research* (pp. 35–58). Sage Publications, Incorporated.
- Piao, J., McDonald, M., Henry, A., Vaa, T., & Tveit, Ø. (2005). An assessment of user acceptance of intelligent speed adaptation systems. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005*. IEEE.
- Poisson, C., Barré, J., Bourmaud, G., & Forzy, J.-F. (2020). Driver Behavior in Conditional Automation: Comparison of Driving Simulator and Wizard of Oz Conditions. In R. Bernhaupt, F. 'Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). ACM. <https://doi.org/10.1145/3334480.3382854>
- Purucker, C., Schneider, N., Rüger, F., & Frey, A. (2018). Validity of research environments – Comparing criticality perceptions across research environments. In K. Bengler, J. Drüke, S. Hoffmann, D. Manstetten, & A. Neukum (Eds.), *UR:BAN human factors in traffic* (pp. 423–446). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15418-9_25
- Quesenberry, W. (2004). Balancing the 5Es: Usability. *Cutter IT Journal*, 17(2), 4–11.
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (Version 4.2.2) [Computer software]. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Radlmayr, J. (2020). *Take-over performance in conditionally automated driving: Effects of the driver state and the human-machine-interface* [Doctoral thesis]. Technical University of Munich, Garching.
- Rahman, M. M., Lesch, M. F., Horrey, W. J., & Strawderman, L. (2017). Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. *Accident Analysis and Prevention*, 108, 361–373. <https://doi.org/10.1016/j.aap.2017.09.011>
- Ranney, T. A. (2011). Psychological fidelity: Perception of risk. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology*. CRC Press.
- Response 3. (2009). *Code of practice for the design and evaluation of ADAS*.
- Rödel, C., Stadler, S., Meschtscherjakov, A., & Tscheligi, M. (2014). Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1–8). ACM. <https://doi.org/10.1145/2667317.2667330>

- Roessger, P. (2003). An international comparison of the usability of driver-information systems: Tools, results and implications. *Journal of Passenger Cars: Electronic and Electrical Systems*, 112(7), 776–779. <https://www.jstor.org/stable/44699740>
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3), 393–404. <https://doi.org/10.5465/amr.1998.926617>
- Russell, S., & Grove, K. (2020). Hf considerations when testing and evaluating ACIVs. In D. L. Fisher, W. J. Horrey, J. D. Lee, & M. A. Regan (Eds.), *Handbook of human factors for automated, connected, and intelligent vehicles* (First edition). CRC Press an imprint of Taylor & Francis Group LLC.
- SAE International (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (J3016).
- Sarodnick, F., & Brau, H. (2016). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung* (3., unveränderte Auflage). Hogrefe.
- Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Elsevier/Morgan Kaufmann.
- Schmidtke, H., & Schulze, P. (1989). Voraussetzungen für eine Systembewertung. In H. Schmidtke (Ed.), *Handbuch der Ergonomie: Mit ergonomischen Konstruktionsrichtlinien und Methoden. Band 5* (2., überarb. und erw. Aufl.). Hanser. D.2.1.1.
- Schneider, S. (2021). *Behavioral validity in virtual reality pedestrian simulators* [Doctoral thesis]. Technical University of Munich, Garching.
- Schrepp, M. (2023). *User experience questionnaire handbook: All you need to know to apply the UEQ successfully in your projects*. <https://www.ueq-online.org/>
- Senserrick, T. M., Brown, T., Quistberg, D. A., Marshall, D., & Winston, F. K. (2007). Validation of simulated assessment of teen driver speed management on rural roads. *Annual Proceedings of the Association for the Advancement of Automotive Medicine*(51), 525–536.
- Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N., & Rosenfeld, R. (2009). Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *International Conference on Information and Communication Technologies and Development (ICTD)* (pp. 447–457). IEEE. <https://doi.org/10.1109/ICTD.2009.5426682>
- Shi, E., Gasser, T. M., Seeck, A., & Auerswald, R. (2020). The principles of operation framework: A comprehensive classification concept for automated driving functions. *SAE International Journal of Connected and Automated Vehicles*, 3(1). <https://doi.org/10.4271/12-03-01-0003>
- Shuttleworth, J. (2019). *SAE Standards News: J3016 automated-driving graphic update: SAE updates J3016 Levels of Automated Driving graphic to reflect evolving standard*. SAE International. <https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic>
- Signorell, A. (2023). *DescTools: Tools for Descriptive Statistics* (Version 0.99.48) [Computer software]. <https://CRAN.R-project.org/package=DescTools>
- Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2023). *afex: Analysis of Factorial Experiments* (Version 1.3) [Computer software]. <https://CRAN.R-project.org/package=afex>
- Singmann, H., & Kellen, D. (2020). An introduction to mixed models for experimental psychology. In D. Spieler & E. Schumacher (Eds.), *Taylor & Francis eBooks. New methods in cognitive psychology* (First published.). Routledge Taylor & Francis Group.

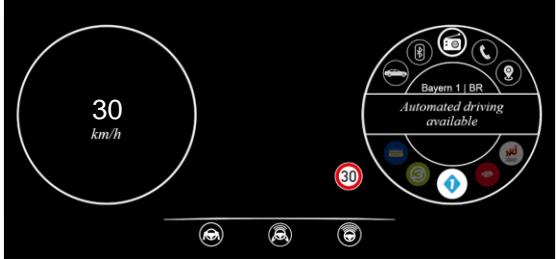
- Sogemeier, D., Forster, Y., Naujoks, F., Krems, J. F., & Keinath, A. (2022). How to map cultural dimensions to usability criteria: Implications for the design of an automotive human-machine interface. In *14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 123–126). ACM. <https://doi.org/10.1145/3544999.3554786>
- Stevens, A., Quimby, A., Board, A., Kersloot, T., & Burns, P. (2002). *Design guidelines for safety of in-vehicle information systems* (PA3721/01). TRL Limited.
- Strle, G., Xing, Y., Miller, E. E., Boyle, L. N., & Sodnik, J. (2021). Take-over time: A cross-cultural study of take-over responses in highly automated driving. *Applied Sciences*, 11(17), 7959. <https://doi.org/10.3390/app11177959>
- Sugiura, E. (2021). *Honda launches world's first level 3 self-driving car: Plan for just 100 lease sales reflects Japanese automaker's cautious approach*. Nikkei Asia. <https://asia.nikkei.com/Business/Automobiles/Honda-launches-world-s-first-level-3-self-driving-car>
- Tacay, C. (2020). *Entwicklung und Evaluierung eines Human-Machine-Interfaces in dem adaptiven Kombiinstrument für das automatisierte Fahren* [Term paper]. Technical University of Munich, Garching.
- Utesch, F. (2014). *Unscharfe Warnungen im Kraftfahrzeug: Eignet sich eine LED-Leiste als Anzeige für Fahrerassistenzsysteme?* [Doctoral thesis]. Technical University of Braunschweig, Braunschweig.
- van Gijssel, A. (2012). *Assisting driver sovereignty: A fail-safe design approach to driver distraction* [Doctoral thesis]. Delft University of Technology, Delft.
- Vatrapu, R., & Pérez-Quiñones, M. A. (2006). Culture and usability evaluation: The effects of culture in structured interviews. *Journal of Usability Studies*, 1(4), 156–170. <https://doi.org/10.5555/2835531.2835533>
- Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Proceedings of the Human Factors Society Annual Meeting*, 34(4), 291–294. <https://doi.org/10.1177/154193129003400411>
- Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2022). *skimr: Compact and flexible summaries of data* (Version 2.1.5) [Computer software]. <https://CRAN.R-project.org/package=skimr>
- Werner, A. (2018). New colours for autonomous driving: An evaluation of chromaticities for the external lighting equipment of autonomous vehicles. *Colour Turn*, 2018(1), Article IV. <https://doi.org/10.25538/tct.v0i1.692>
- Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12). <https://www.jstatsoft.org/v21/i12/> [Computer software].
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* [computer software] (Second edition). *Use R!* Springer. <https://doi.org/10.1007/978-3-319-24277-4>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686> [Computer Software].
- Winter, D., Hinderks, A., Schrepp, M., & Thomaschewski, J. (2017). Welche UX Faktoren sind für mein Produkt wichtig? In S. Hess & H. Fischer (Eds.), *Mensch und Computer 2017 - Usability Professionals*. Gesellschaft für Informatik e.V. <https://doi.org/10.18420/muc2017-up-0002>
- Wynne, R. A., Beanland, V., & Salmon, P. M. (2019). Systematic review of driving simulator validation studies. *Safety Science*, 117, 138–151. <https://doi.org/10.1016/j.ssci.2019.04.004>

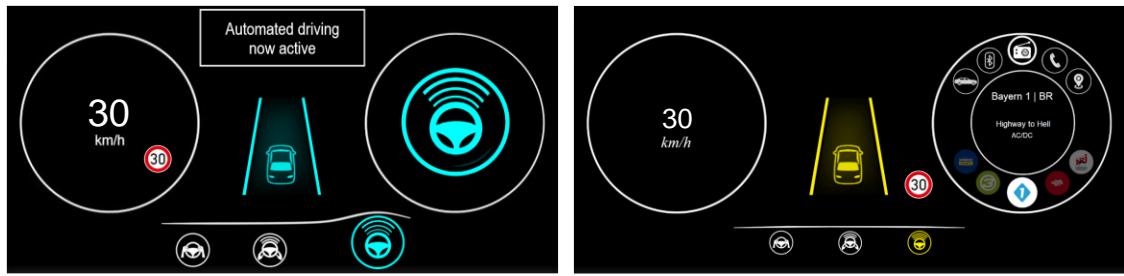
- Yeo, A. W. (2001). Global-software development lifecycle. In J. A. Jacko & A. Sears (Eds.), *CHI letters: Vol. 3,1. Chi '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 104–111). ACM Press. <https://doi.org/10.1145/365024.365060>
- Young, K. L., Rudin-Brown, C. M., Lenné, M. G., & Williamson, A. R. (2012). The implications of cross-regional differences for the design of In-vehicle Information Systems: A comparison of Australian and Chinese drivers. *Applied Ergonomics*, 43(3), 564–573. <https://doi.org/10.1016/j.apergo.2011.09.001>

12 Appendix I

This appendix contains supplemental material to the experimental design of studies *Exp_Testing-Environment* and *Exp_Culture* (Chapter 5).

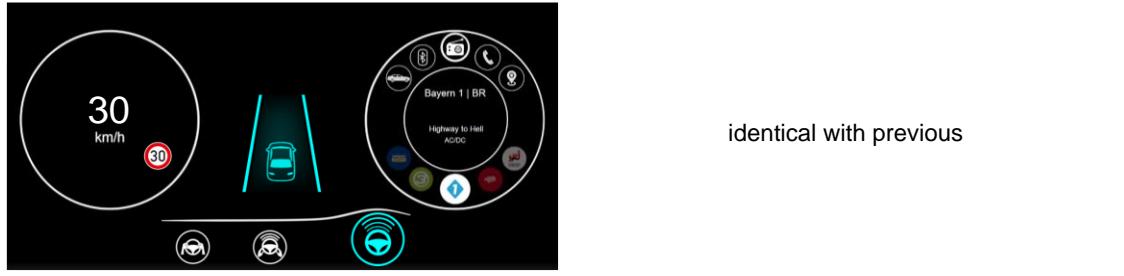
Table 12.1 Excerpts of the HMI concepts *HC-HMI* and *LC-HMI* for the 12 test cases and interaction errors.

<i>HC-HMI</i>	<i>LC-HMI</i>
	
TC1: L0 [-] TC2a: L0 [-] → L0 [L2, L3]	
	
TC2b: L0 [-] → L0 [L2, L3]	
<i>HC-HMI</i> : Pop-Up (notification) disappears after 7 s. <i>LC-HMI</i> : Pop-Up (notification) disappears after 4 s.	
	
TC2c: L0 [-] → L0 [L2, L3] TC3a: L0 [L2, L3] → L3	
	
TC3b: L0 [L2, L3] → L3	
<i>HC-HMI</i> : Pop-Ups (icon & notification) disappear after 7 s.	



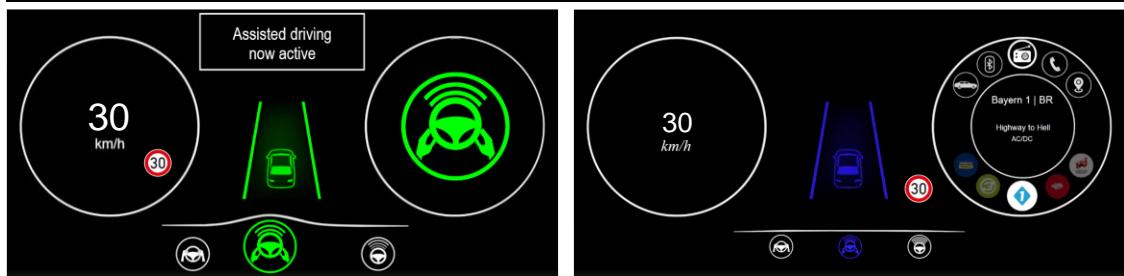
TC3c: L0 [L2, L3] → L3

HC-HMI: Pop-Ups (icon & notification) disappear after 7 s.



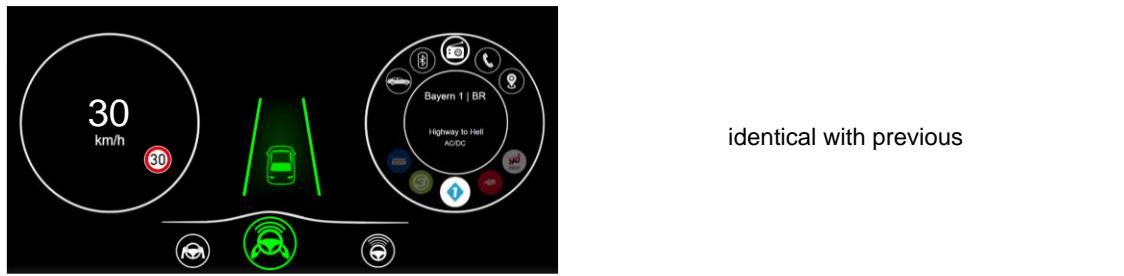
TC3d: L0 [L2, L3] → L3 | TC4: L3 | TC5a: L3 → L2 [L3]

identical with previous



TC5b: L3 → L2 [L3]

HC-HMI: Pop-Ups (icon & notification) disappear after 7 s.



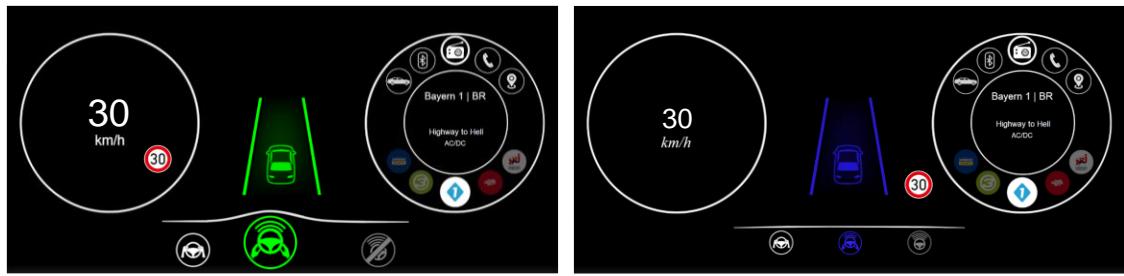
TC5c: L3 → L2 [L3] | TC6a: L2 [L3] → L2 [-] (malfunction)

identical with previous

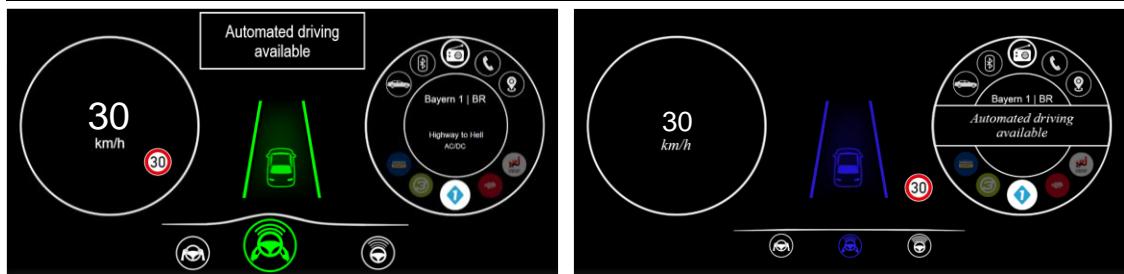


TC6b: L2 [L3] → L2 [-] (malfunction)

HC-HMI: Pop-Up (notification) disappears after 7 s. | LC-HMI: Pop-Up (notification) disappears after 4 s.



TC6c: L2 [L3] → L2 [-] (malfunction) | TC7: L2 [-] | TC8a: L2 [-] → L2 [L3]

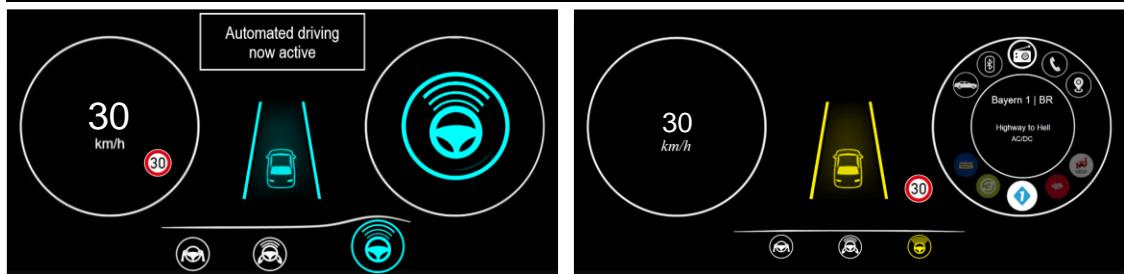


TC8b: L2 [-] → L2 [L3]

HC-HMI: Pop-Up (notification) disappears after 7 s. | LC-HMI: Pop-Up (notification) disappears after 4 s.



TC8c: L2 [-] → L2 [L3] | TC9a: L2 [L3] → L3



TC9b: L2 [L3] → L3

HC-HMI: Pop-Ups (icon & notification) disappear after 7 s. | LC-HMI: Pop-Ups (icon & notification) disappear after 4 s.



identical with previous

TC9c: L2 [L3] → L3 | TC10a: L3 → L0 [-] (ODD end)



TC10b: L3 → L0 [-] (ODD end)

HC-HMI and LC-HMI: Pop-Up (notification) shows for 7 s before TC10c starts.



TC10c: L3 → L0 [-] (ODD end)

HC-HMI and LC-HMI: Pop-Up (notification) shows for 6 s before TC10d starts.

HC-HMI: One box per second disappears indicating a countdown. Yellow LED lights flash on the steering wheel and a warning sound with low criticality is triggered.

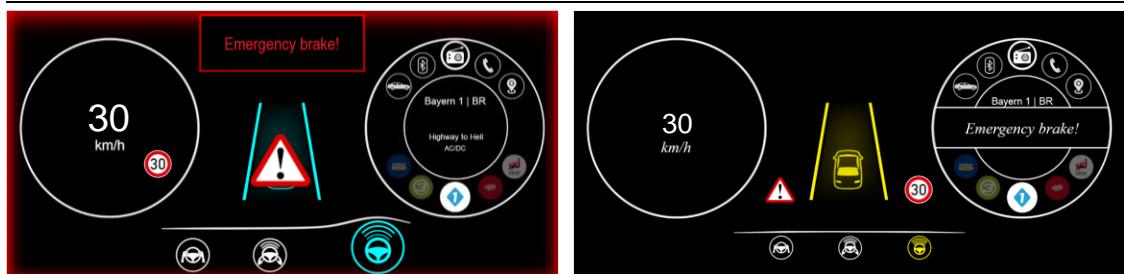


identical with previous

TC10d: L3 → L0 [-] (ODD end)

HC-HMI and LC-HMI: Pop-Up (notification) shows for 7 s before TC10e starts.

HC-HMI: One box per second disappears indicating a countdown. Red LED lights flash on the steering wheel and a warning sound with high criticality is triggered.



TC10e: L3 → L0 [-] (ODD end)

HC-HMI: Red LED lights flash on the steering wheel.



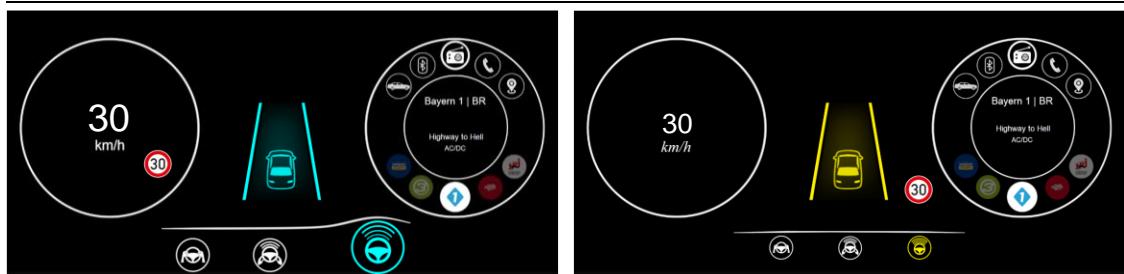
TC10f: L3 → L0 [-] (ODD end)

HC-HMI: Pop-Ups (icon & notification) disappear after 7 s.

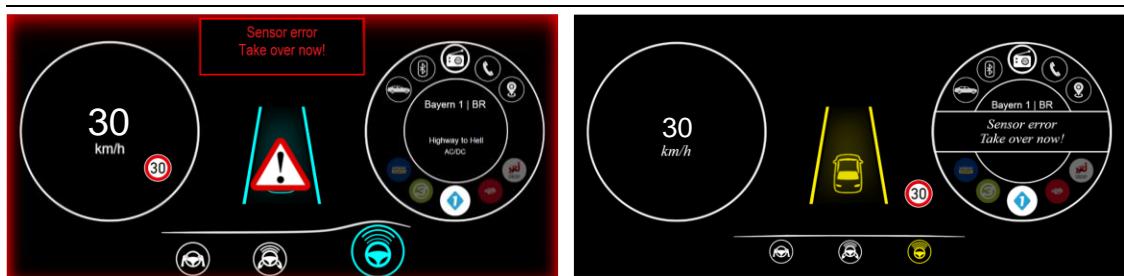


TC10g: L3 → L0 [-] (ODD end) | TC11a: L0 [-] → L3

See TC2b-TC3c for TC11b-TC11e



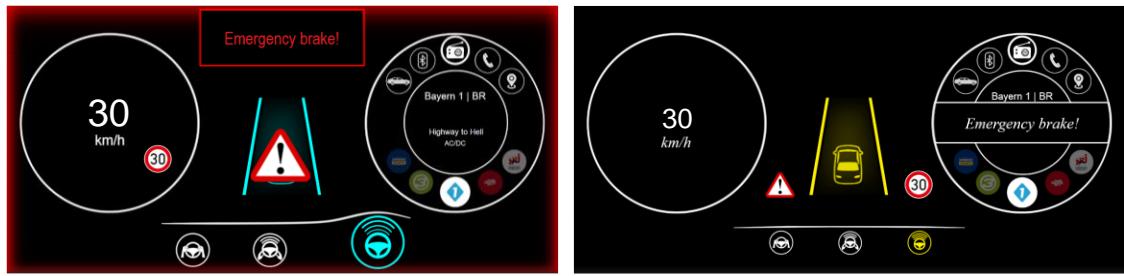
TC11f: L0 [-] → L3 | TC12a: L3 → L0 [-] (malfunction)



TC12b: L3 → L0 [-] (malfunction)

HC-HMI and LC-HMI: Pop-Up (notification) shows for 6 s before TC12c starts.

HC-HMI: Red LED lights flash on the steering wheel and a warning sound with high criticality is triggered.



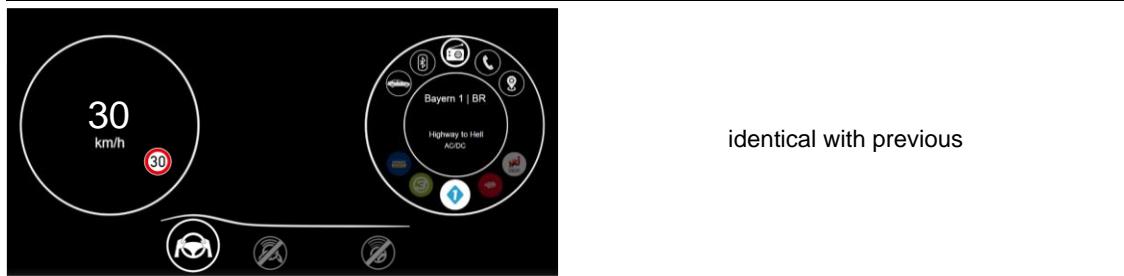
TC12c: L3 → L0 [-] (malfunction)

HC-HMI: Red LED lights flash on the steering wheel.



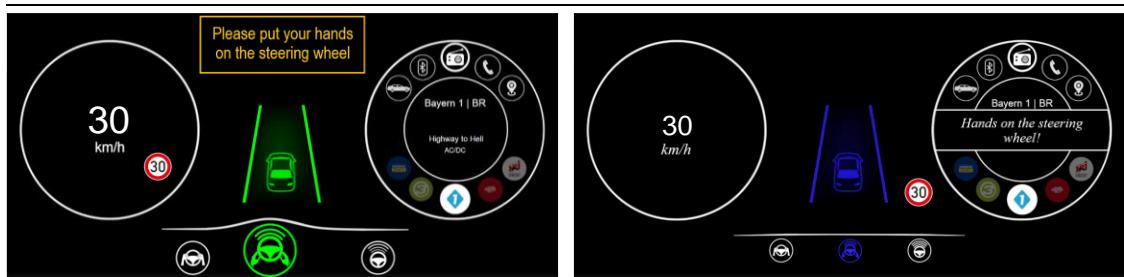
TC12d: L3 → L0 [-] (malfunction)

HC-HMI: Pop-Ups (icon & notification) disappear after 7 s.



TC12e: L3 → L0 [-] (malfunction)

identical with previous



H-off notification in L2 (a)

HC-HMI and LC-HMI: Pop-Up (notification) shows for 7 s before H-off notification in L2 (b) starts.

HC-HMI: Yellow LED lights flash on the steering wheel.



Identical with previous

H-off notification during L2 driving (b)

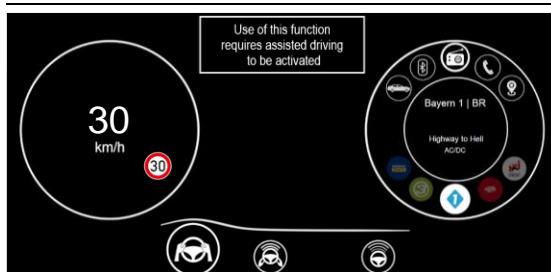
HC-HMI and LC-HMI: Pop-Up (notification) shows for 6 s before H-off notification in L2 (c) starts.

HC-HMI: Red LED lights flash on the steering wheel and a warning sound with low criticality is triggered.



H-off notification during L2 driving (c)

HC-HMI: Red LED lights flash on the steering wheel.



Feeedback for handling error (MOD button has no function in L0)

HC-HMI: Pop-Up (notification) disappears after 7 s.

13 Appendix II

This appendix contains supplemental material to the validation study *Exp_Testing-Environment* (Chapter 6).

Table 13.1 Weather and light conditions in the study *Exp_Testing-Environment*.

Metric	Condition	Proportion [% (n)]			
		Sim_GER-HC (26)*	Sim_GER-LC (26)*	TT_GER-HC (33)	TT_GER-LC (28)
Weather	Sunny, blue sky	0 (0)		6.06 (2)	10.71 (3)
	Lightly clouded	100 (26)		39.39 (13)	39.29 (11)
	Heavily clouded	0 (0)		39.39 (13)	35.71 (10)
	Light rain	0 (0)		15.15 (5)	14.29 (4)
Light	Very bright, blinding	0 (0)		21.21 (7)	25 (7)
	Bright	100 (26)		72.73 (24)	57.14 (16)
	Gloomy, dusky	0 (0)		6.06 (2)	17.86 (5)

Note. The mode values of each metric are indicated in bold.

* In the driving simulator, the weather and light conditions are consistent across all participants.

Table 13.2 Summary table of the descriptive analysis of the metrics on the sociodemographic data for the study *Exp_Testing-Environment* (Section 6.2).

Metric	Statistic / Response	Value / Proportion [% (n)]			
		Sim_GER-HC (26)	Sim_GER-LC (26)	TT_GER-HC (33)	TT_GER-LC (28)
Age*	M	41.92	38	37.55	37.43
	SD	16.9	17.52	14.88	15.12
	Range	19-71	18-73	22-69	20-65
	Age group: 18-24	15.38 (4)**	23.08 (6)	21.21 (7)	28.57 (8)
Gender	Age group: 25-39	30.77 (8)	38.46 (10)	39.39 (13)	32.14 (9)
	Age group: 40-54	26.92 (7)	23.08 (6)	21.21 (7)	17.86 (5)
	Age group: > 54	23.08 (6)	15.38 (4)**	18.18 (6)	21.43 (6)
	Male	61.54 (16)	61.54 (16)	57.58 (19)	64.29 (18)
Need of visual aid	Female	38.46 (10)	38.46 (10)	42.42 (14)	35.71 (10)
	Diverse	0 (0)	0 (0)	0 (0)	0 (0)
	Other / not indicated	0 (0)	0 (0)	0 (0)	0 (0)
	No	76.92 (20)	53.85 (14)	66.67 (22)	78.57 (22)
Color deficiency / color blindness	Yes & currently used	19.23 (5)	34.62 (9)	24.24 (8)	14.29 (4)
	Yes & currently not used	3.85 (1)	11.54 (3)	9.09 (3)	7.14 (2)
	No	100 (26)	92.31 (24)	90.91 (30)	96.43 (27)
Color deficiency / color blindness	Yes, (slight) red-green	0 (0)	7.69 (2)	9.09 (3)	3.57 (1)
	Yes, other	0 (0)	0 (0)	0 (0)	0 (0)

Note. The mode values of each metric are indicated in bold.

* Sim_GER-HC_{TP18} does not provide age information.

** The targeted minimum of five participants per age group (NHTSA, 2013) is not met.

Table 13.3 Summary table of the descriptive analysis of the metrics on the driving background for the study *Exp_Testing-Environment* (Section 6.2).

Metric	Response	Proportion [% (n)]			
		Sim_GER-HC (26)	Sim_GER-LC (26)	TT_GER-HC (33)	TT_GER-LC (28)
<i>Frequency of driving (12 months)</i>	Every day	65.38 (17)	38.46 (10)	30.3 (10)	50 (14)
	Several times a week	19.23 (5)	38.46 (10)	33.33 (11)	25 (7)
	Several times a month	7.69 (2)	19.23 (5)	27.27 (9)	3.57 (1)
	Less than once a month	7.69 (2)	3.85 (1)	6.06 (2)	21.43 (6)
<i>Mileage (12 months)</i>	Never	0 (0)	0 (0)	3.03 (1)	0 (0)
	> 20,000 km	19.23 (5)	3.85 (1)	18.18 (6)	21.43 (6)
	10,001 km-20,000 km	30.77 (8)	34.62 (9)	24.24 (8)	35.71 (10)
	5,001 km-10,000 km	34.62 (9)	42.31 (11)	33.33 (11)	10.71 (3)
<i>No ADAS experience</i>	< 5,000 km	15.38 (4)	19.23 (5)	24.24 (8)	32.14 (9)
	Yes	19.23 (5)	30.77 (8)	18.18 (6)	10.71 (3)
	No	80.77 (21)	69.23 (18)	81.82 (27)	89.29 (25)
<i>Usage frequency (12 months): CC</i>	Several times a day	11.54 (3)	7.69 (2)	6.06 (2)	10.71 (3)
	Every day	19.23 (5)	3.85 (1)	0 (0)	7.14 (2)
	Every week	11.54 (3)	23.08 (6)	15.15 (5)	17.86 (5)
	Every month	15.38 (4)	26.92 (7)	42.42 (14)	35.71 (10)
	Seldom	0 (0)	0 (0)	0 (0)	0 (0)
	Never	19.23 (5)	7.69 (2)	15.15 (5)	17.86 (5)
<i>Usage frequency (12 months): ACC</i>	No prior experience	23.08 (6)	30.77 (8)	21.21 (7)	10.71 (3)
	Several times a day	0 (0)	0 (0)	3.03 (1)	7.14 (2)
	Every day	7.69 (2)	0 (0)	0 (0)	3.57 (1)
	Every week	11.54 (3)	7.69 (2)	0 (0)	3.57 (1)
	Every month	11.54 (3)	11.54 (3)	15.15 (5)	14.29 (4)
	Seldom	0 (0)	0 (0)	0 (0)	0 (0)
<i>Usage frequency (12 months): LKA</i>	Never	3.85 (1)	3.85 (1)	27.27 (9)	3.57 (1)
	No prior experience	65.38 (17)	76.92 (20)	54.55 (18)	67.86 (19)
	Several times a day	0 (0)	0 (0)	0 (0)	7.14 (2)
	Every day	23.08 (6)	7.69 (2)	6.06 (2)	3.57 (1)
	Every week	7.69 (2)	3.85 (1)	0 (0)	3.57 (1)
	Every month	7.69 (2)	11.54 (3)	18.18 (6)	17.86 (5)
<i>Prior knowledge in the field of automated driving</i>	Seldom	0 (0)	0 (0)	0 (0)	0 (0)
	Never	19.23 (5)	15.38 (4)	27.27 (9)	21.43 (6)
	No prior experience	42.31 (11)	61.54 (16)	48.48 (16)	46.43 (13)
	4: expert	0 (0)	0 (0)	9.09 (3)	3.57 (1)
3	3	3.85 (1)	11.54 (3)	9.09 (3)	3.57 (1)
	2	23.08 (6)	30.77 (8)	15.15 (5)	17.86 (5)
	1	34.62 (9)	23.08 (6)	24.24 (8)	28.57 (8)
0: no prior knowledge		38.46 (10)	34.62 (9)	42.42 (14)	46.43 (13)

Note. The mode values of each metric are indicated in bold.



Figure 13.1 Visualization of the individual driving behavior for the metric *Control path of the first activation* for the study *Exp_Testing-Environment* (Paragraph Control Path of First Activation).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L0 at the start of the instruction (leftmost column signals that L0 is active) are included. Participants using the ideal path are marked with a black box. The sample sizes are as follows: *Sim_GER-HC*: $n = 23$; *Sim_GER-LC*: $n = 23$; *TT_GER-HC*: $n = 27$; & *TT_GER-LC*: $n = 23$.



Figure 13.2 Visualization of the individual driving behavior for the metric *Take-over Path after Rtl* for Rtl_{20s} for the study *Exp_Testing-Environment* (Paragraph *Take-Over Path after Rtl*).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L3 at the start of the Rtl (leftmost column signals that L3 is active) are included. Participants using one action only are marked with a black box. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 32$; & *TT_GER-LC*: $n = 26$.

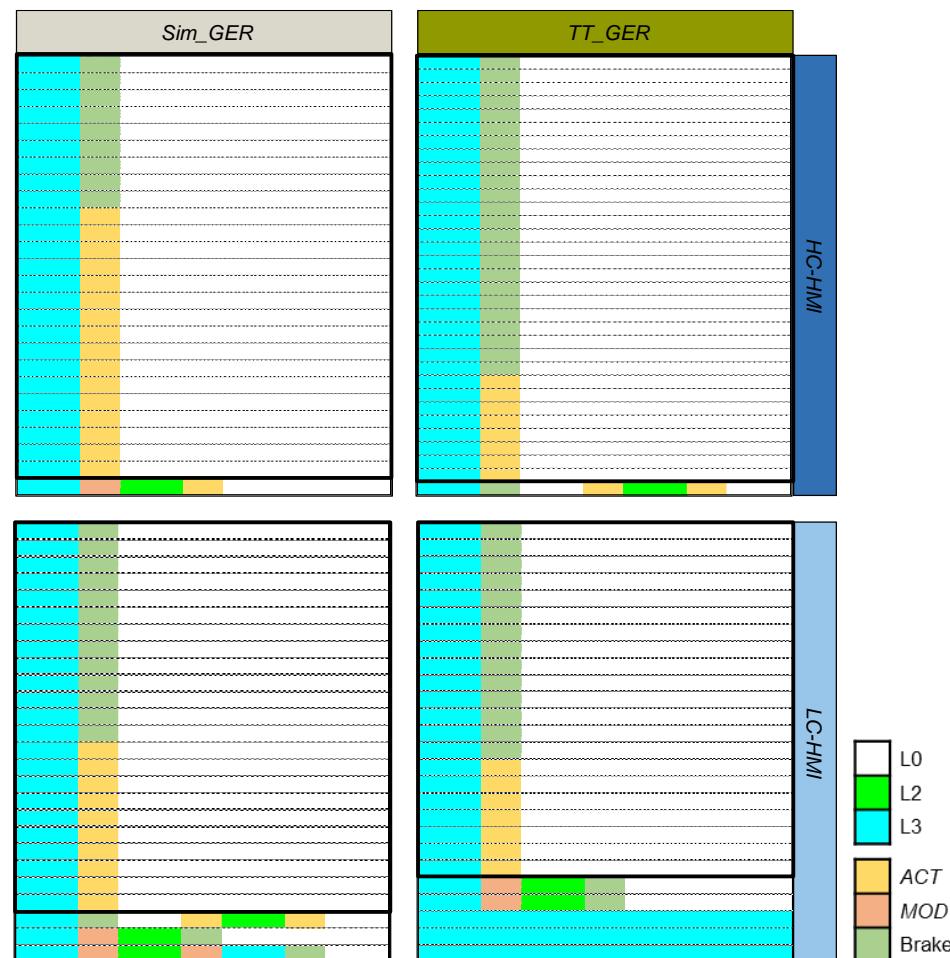


Figure 13.3 Visualization of the individual behavior for the metric *Take-over Path after Rtl* for Rtl_{6s} for the study *Exp_Testing-Environment* (Paragraph *Take-Over Path after Rtl*).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L3 at the start of the Rtl (leftmost column signals that L3 is active) are included. Participants using one action only are marked with a black box. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 33$; & *TT_GER-LC*: $n = 26$.

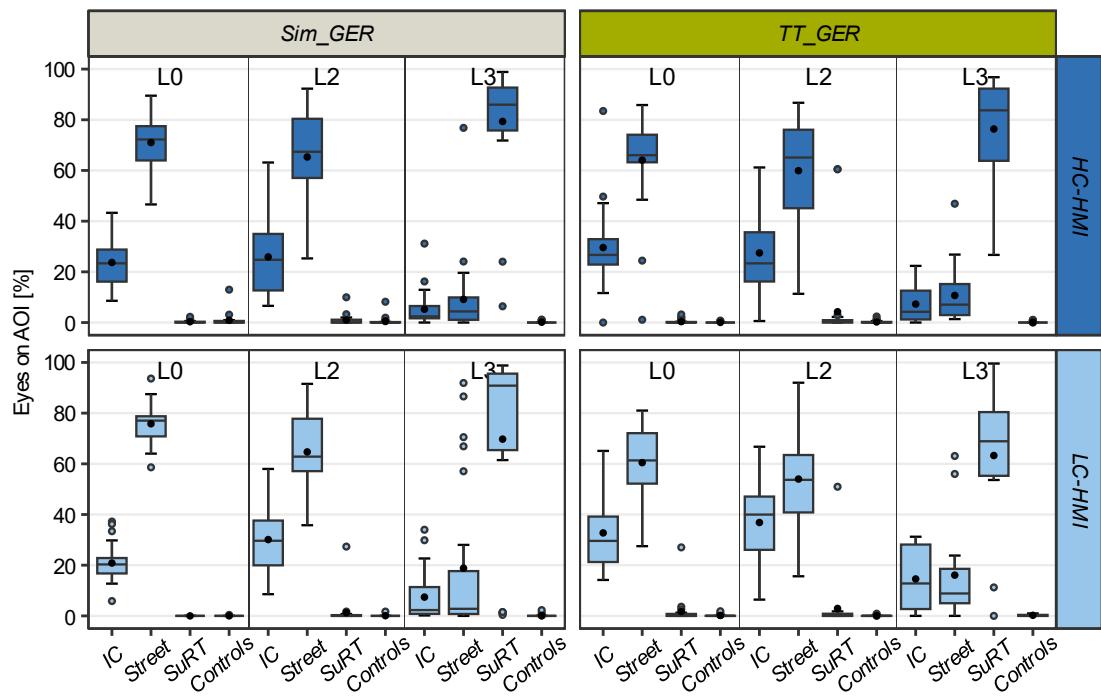


Figure 13.4 Boxplot diagram visualizing the results of the metric *Attention ratio during continuous rides in L0, L2, & L3* for all four AOIs for the study *Exp_Testing-Environment* (Paragraph Attention Ratio during Continuous Rides in L0, L2, & L3).

Note. The sample sizes are as follows: TT_GER-HC: $n = 22$ (L0), $n = 22$ (L2), $n = 22$ (L3); TT_GER-LC: $n = 25$ (L0), $n = 25$ (L2), $n = 25$ (L3); TT_GER-HC: $n = 27$ (L0), $n = 32$ (L2), $n = 27$ (L3); & TT_GER-LC: $n = 22$ (L0), $n = 20$ (L2), $n = 14$ (L3).

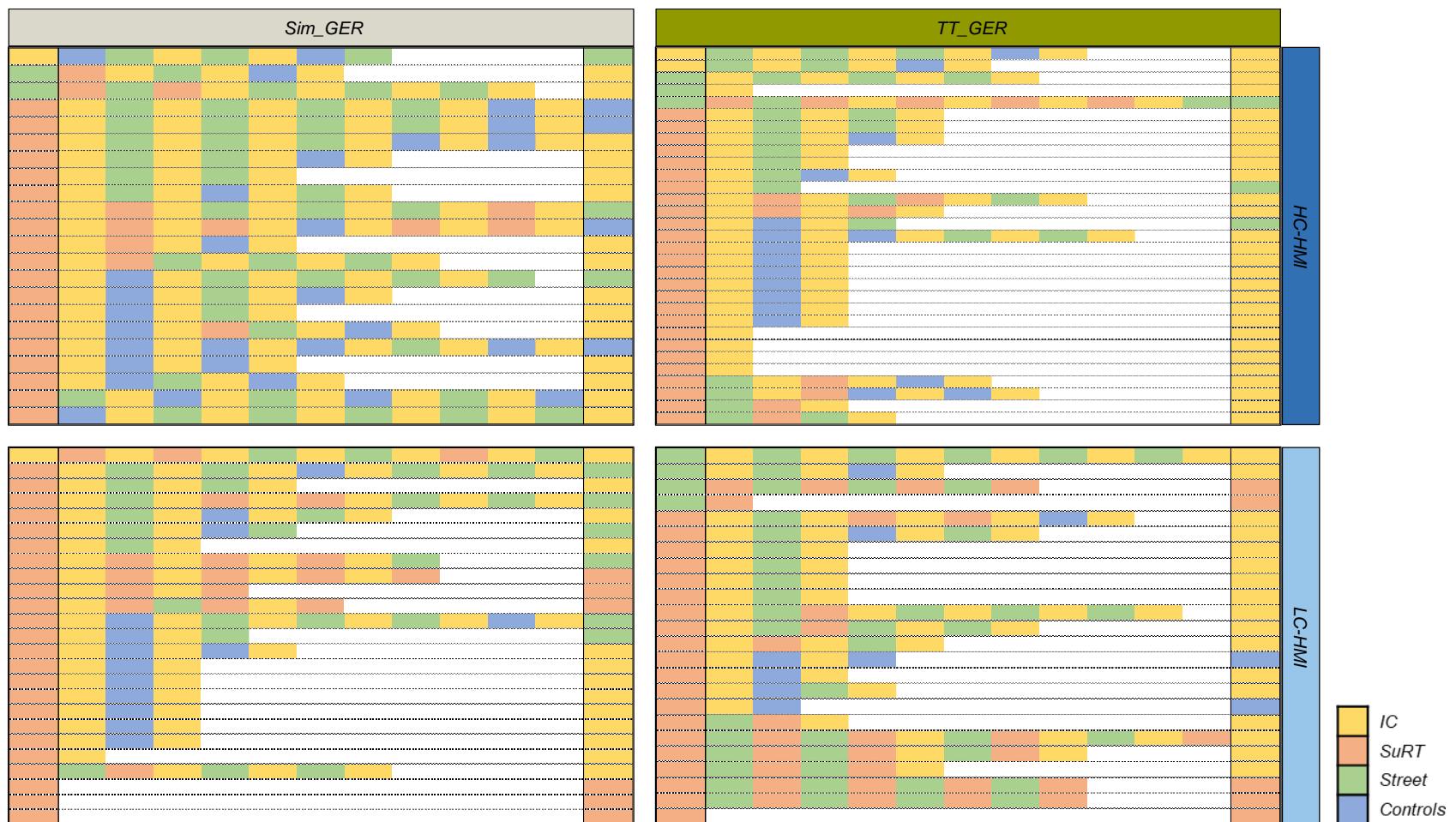


Figure 13.5 Visualization of the individual gaze behavior for the metric *Gaze Behavior during Rtl* for Rtl_{20s} for the study *Exp_Testing-Environment* (Paragraph *Gaze Behavior during Rtl*).

Note. The figure shows active AOIs of the participants between the start and the end of the Rtl. The end of the Rtl is marked by the start of emergency braking maneuver or the transition to L0. Only participants driving L3 at the start of Rtl are included. The sample sizes are as follows: *Sim_GER-HC*: $n = 22$; *Sim_GER-LC*: $n = 25$; *TT_GER-HC*: $n = 31$; & *TT_GER-LC*: $n = 24$.

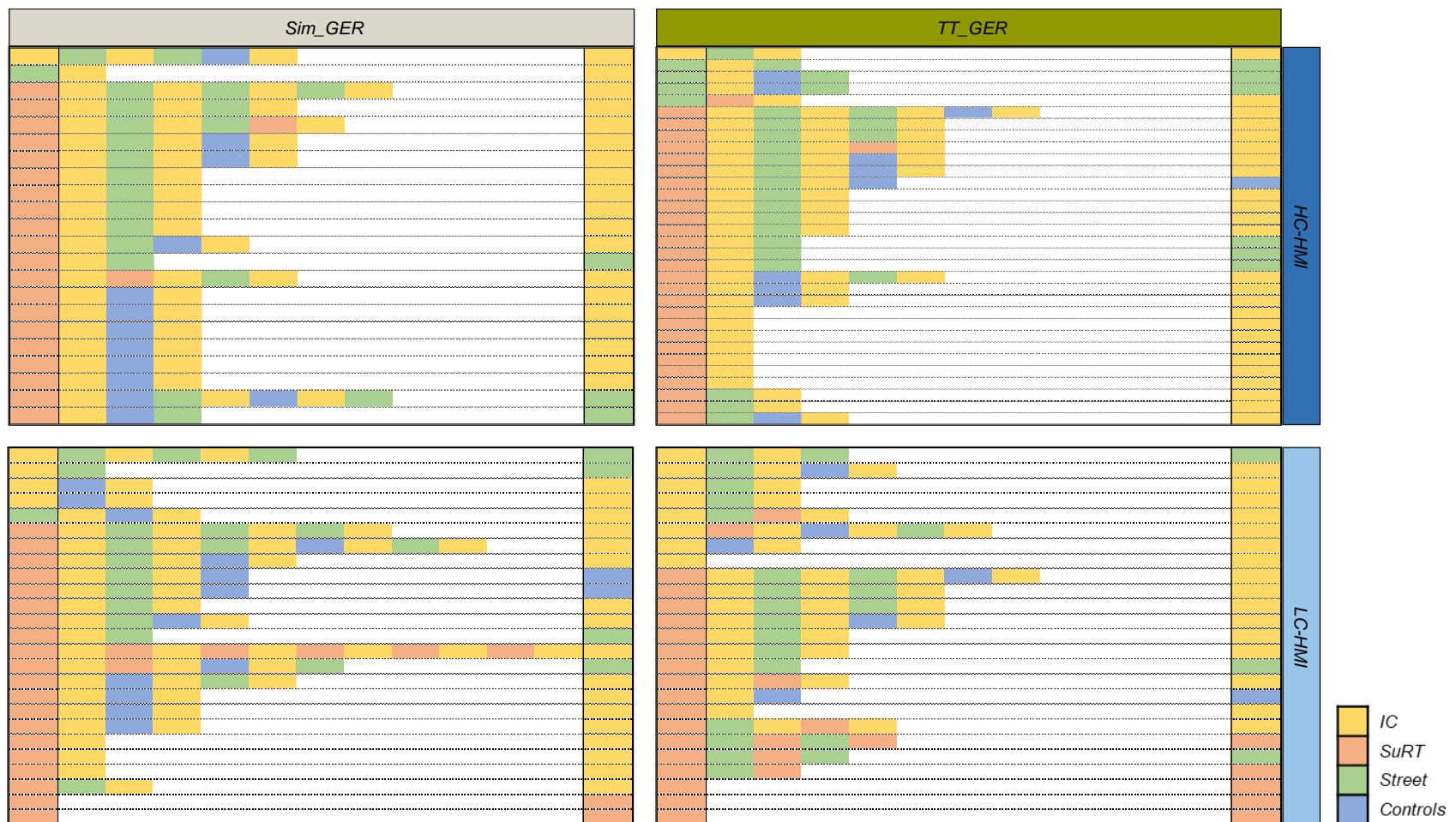


Figure 13.6 Visualization of the individual gaze behavior for the metric *Gaze Behavior during Rtl* for Rtl_{6s} for the study *Exp_Testing-Environment* (Paragraph *Gaze Behavior during Rtl*).

Note. The figure shows active AOIs of the participants between the start and the end of the Rtl. The end of the Rtl is marked by the start of emergency braking maneuver or the transition to L0. Only participants driving L3 at the start of Rtl are included. The sample sizes are as follows: *Sim_GER-HC*: $n = 22$; *Sim_GER-LC*: $n = 25$; *TT_GER-HC*: $n = 32$; & *TT_GER-LC*: $n = 25$.

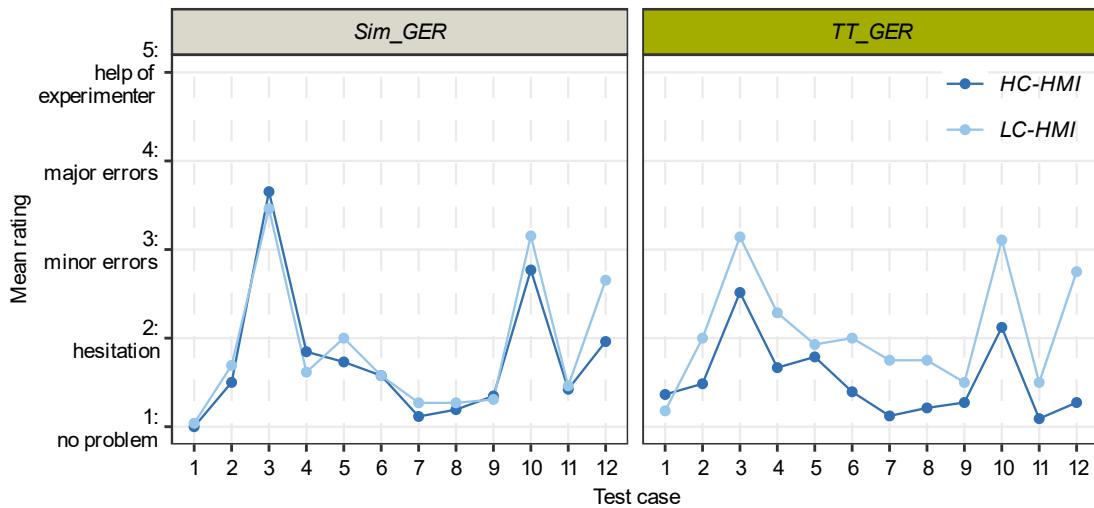


Figure 13.7 Visualization of the mean ratings per test case for the metric *Experimenter rating* for the study *Exp_Testing-Environment* (Subsubsection Experimenter Rating).

Note. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 33$; & *TT_GER-LC*: $n = 28$.

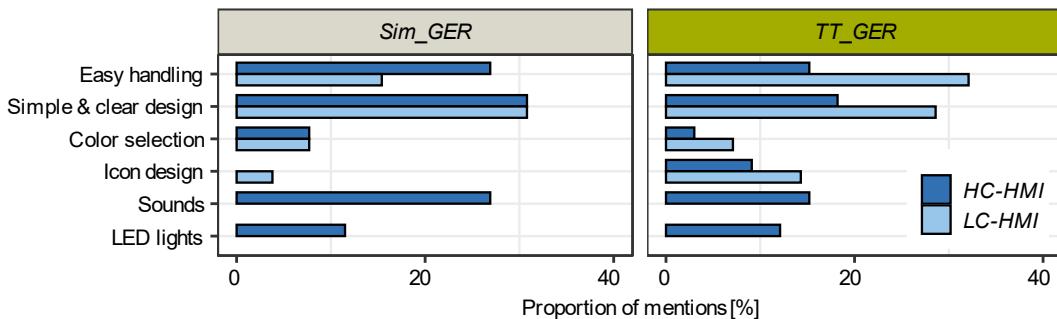


Figure 13.8 Overview of the clustered replies for praised components of the HMI concepts in the metric *Final Interview* for the study *Exp_Testing-Environment* (Subsubsection Final Interview).

Note. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 33$; & *TT_GER-LC*: $n = 28$.

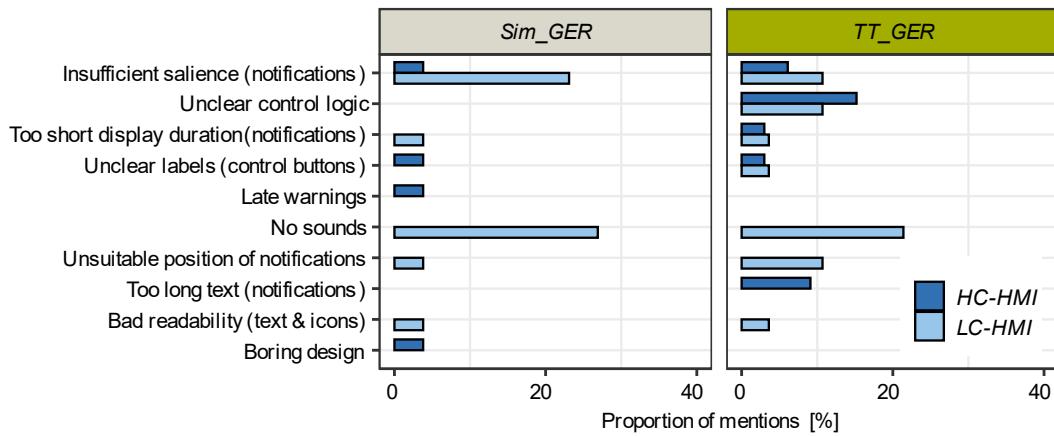


Figure 13.9 Overview of the clustered replies for criticized components of the HMI concepts in the metric *Final Interview* for the study *Exp_Testing-Environment* (Subsubsection Final Interview).

Note. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 33$; & *TT_GER-LC*: $n = 28$.

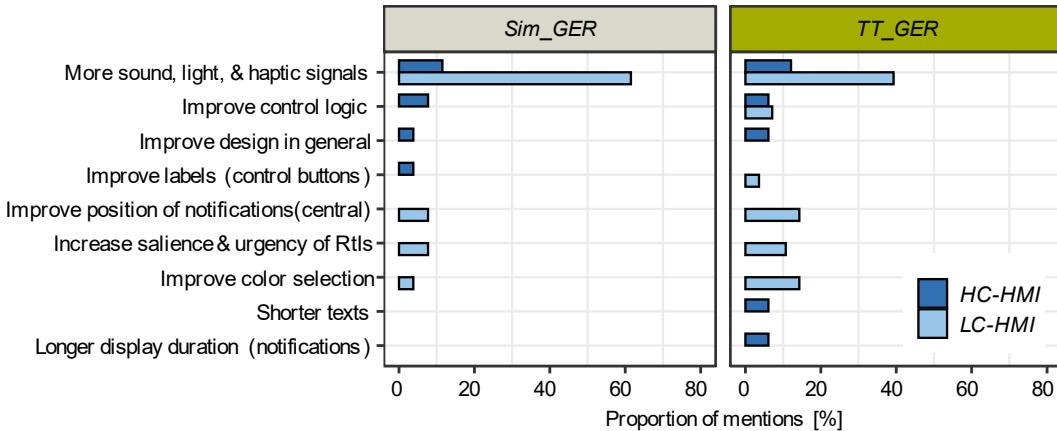


Figure 13.10 Overview of the clustered replies for improvement suggestions for components of the HMI concepts in the metric *Final Interview* for the study *Exp_Testing-Environment* (Subsubsection Final Interview).

Note. The sample sizes are as follows: *Sim_GER-HC*: $n = 26$; *Sim_GER-LC*: $n = 26$; *TT_GER-HC*: $n = 33$; & *TT_GER-LC*: $n = 28$.

14 Appendix III

This appendix contains supplemental material to the validation study *Exp_Culture* (Chapter 7).

Table 14.1 Weather and light conditions in the study *Exp_Culture*.

Metric	Condition	Proportion [% (n)]			
		TT_GER-HC (33)	TT_GER-LC (28)	TT_USA-HC (21)	TT_USA-LC (21)
Weather	Sunny, blue sky	6.06 (2)	10.71 (3)	19.05 (4)	14.29 (3)
	Lightly clouded	39.39 (13)	39.29 (11)	33.33 (7)	38.1 (8)
	Heavily clouded	39.39 (13)	35.71 (10)	47.62 (10)	38.1 (8)
	Light rain	15.15 (5)	14.29 (4)	0 (0)	9.52 (2)
Light	Very bright, blinding	21.21 (7)	25 (7)	42.86 (9)	33.33 (7)
	Bright	72.73 (24)	57.14 (16)	52.38 (11)	52.38 (11)
	Gloomy, dusky	6.06 (2)	17.86 (5)	4.76 (1)	14.29 (3)

Note. The mode values of each metric are indicated in bold.

Table 14.2 Summary table of the descriptive analysis of the metrics on the sociodemographic data for the study *Exp_Culture* (Section 7.2).

Metric	Statistic Response	Value / Proportion [% (n)]			
		TT_GER-HC (33)	TT_GER-LC (28)	TT_USA-HC (21)	TT_USA-LC (21)
	M	37.55	37.43	38.43	37.86
	SD	14.88	15.12	9.7	10.14
	Range	22-69	20-65	21-60	20-59
Age	Age group: 18-24	21.21 (7)	28.57 (8)	9.52 (2)*	9.52 (2)*
	Age group: 25-39	39.39 (13)	32.14 (9)	47.62 (10)	47.62 (10)
	Age group: 40-54	21.21 (7)	17.86 (5)	33.33 (7)	38.1 (8)
	Age group: > 54	18.18 (6)	21.43 (6)	9.52 (2)*	4.76 (1)*
Gender	Male	57.58 (19)	64.29 (18)	57.14 (12)	52.38 (11)
	Female	42.42 (14)	35.71 (10)	42.86 (9)	47.62 (10)
	Diverse	0 (0)	0 (0)	0 (0)	0 (0)
	Other / not indicated	0 (0)	0 (0)	0 (0)	0 (0)
Need of visual aid	No	66.67 (22)	78.57 (22)	57.14 (12)	61.9 (13)
	Yes & currently used	24.24 (8)	14.29 (4)	33.33 (7)	28.57 (6)
	Yes & currently not used	9.09 (3)	7.14 (2)	9.52 (2)	9.52 (2)
Color deficiency / color blindness	No	90.91 (30)	96.43 (27)	95.24 (20)	100 (21)
	Yes, (slight) red-green	9.09 (3)	3.57 (1)	4.76 (1)	0 (0)
	Yes, other	0 (0)	0 (0)	0 (0)	0 (0)

Note. The mode values of each metric are indicated in bold.

* The targeted minimum of five participants per age group (NHTSA, 2013) is not met.

Table 14.3 Summary table of the descriptive analysis of the metrics on the driving background for the study *Exp_Culture* (Section 7.2).

Metric	Response	Proportion [% (n)]			
		TT_GER-HC (33)	TT_GER-LC (28)	TT_USA-HC (21)	TT_USA-LC (21)
Frequency of driving (12 months)	Every day	30.3 (10)	50 (14)	14.29 (3)	19.05 (4)
	Several times a week	33.33 (11)	25 (7)	42.86 (9)	28.57 (6)
	Several times a month	27.27 (9)	3.57 (1)	23.81 (5)	4.76 (1)
	Less than once a month	6.06 (2)	21.43 (6)	19.05 (4)	42.86 (9)
Mileage (12 months)	Never	3.03 (1)	0 (0)	0 (0)	4.76 (1)
	> 20,000 km	18.18 (6)	21.43 (6)	4.76 (1)	4.76 (1)
	10,001 km-20,000 km	24.24 (8)	35.71 (10)	9.52 (2)	19.05 (4)
	5,001 km-10,000 km	33.33 (11)	10.71 (3)	42.86 (9)	19.05 (4)
No ADAS experience	< 5,000 km	24.24 (8)	32.14 (9)	42.86 (9)	57.14 (12)
	Yes	18.18 (6)	10.71 (3)	9.52 (2)	0 (0)
	No	81.82 (27)	89.29 (25)	90.48 (19)	100 (21)
Usage frequency (12 months): CC	Several times a day	6.06 (2)	10.71 (3)	0 (0)	0 (0)
	Every day	0 (0)	7.14 (2)	0 (0)	14.29 (3)
	Every week	15.15 (5)	17.86 (5)	9.52 (2)	14.29 (3)
	Every month	42.42 (14)	35.71 (10)	23.81 (5)	23.81 (5)
	Seldom	0 (0)	0 (0)	57.14 (12)	42.86 (9)
	Never	15.15 (5)	17.86 (5)	0 (0)	4.76 (1)
Usage frequency (12 months): ACC	No prior experience	21.21 (7)	10.71 (3)	9.52 (2)	0 (0)
	Several times a day	3.03 (1)	7.14 (2)	4.76 (1)	0 (0)
	Every day	0 (0)	3.57 (1)	0 (0)	0 (0)
	Every week	0 (0)	3.57 (1)	0 (0)	4.76 (1)
	Every month	15.15 (5)	14.29 (4)	9.52 (2)	9.52 (2)
	Seldom	0 (0)	0 (0)	19.05 (4)	19.05 (4)
Usage frequency (12 months): LKA	Never	27.27 (9)	3.57 (1)	0 (0)	4.76 (1)
	No prior experience	54.55 (18)	67.86 (19)	66.67 (14)	61.9 (13)
	Several times a day	0 (0)	7.14 (2)	4.76 (1)	0 (0)
	Every day	6.06 (2)	3.57 (1)	9.52 (2)	0 (0)
	Every week	0 (0)	3.57 (1)	0 (0)	4.76 (1)
	Every month	18.18 (6)	17.86 (5)	4.76 (1)	9.52 (2)
Prior knowledge in the field of automated driving	Seldom	0 (0)	0 (0)	19.05 (4)	38.1 (8)
	Never	27.27 (9)	21.43 (6)	0 (0)	4.76 (1)
	No prior experience	48.48 (16)	46.43 (13)	61.9 (13)	42.86 (9)
	4: expert	9.09 (3)	3.57 (1)	0 (0)	9.52 (2)
	3	9.09 (3)	3.57 (1)	9.52 (2)	9.52 (2)
	2	15.15 (5)	17.86 (5)	33.33 (7)	23.81 (5)
0: no prior knowledge	1	24.24 (8)	28.57 (8)	33.33 (7)	47.62 (10)
	0: no prior knowledge	42.42 (14)	46.43 (13)	23.81 (5)	9.52 (2)

Note. The mode values of each metric are indicated in bold.

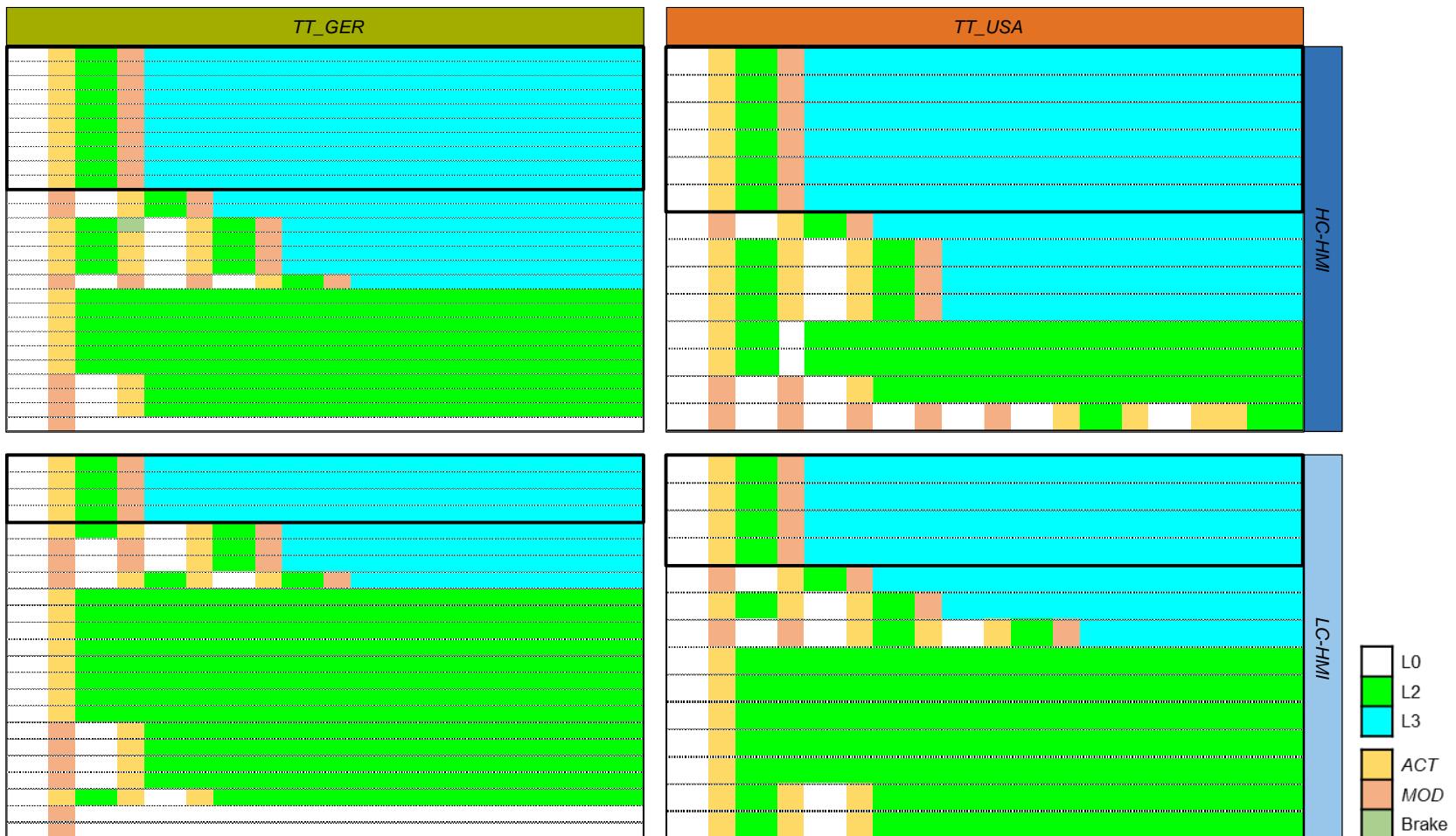


Figure 14.1 Visualization of the individual driving behavior for the metric *Control path of the first activation* for the study *Exp_Testing-Environment* (Paragraph Control Path of First Activation).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L0 at the start of the instruction (leftmost column signals that L0 is active) are included. Participants using the ideal path are marked with a black box. The sample sizes are as follows: TT_GER-HC: $n = 27$; TT_GER-LC: $n = 23$; TT_USA-HC: $n = 14$; & TT_USA-LC: $n = 14$.

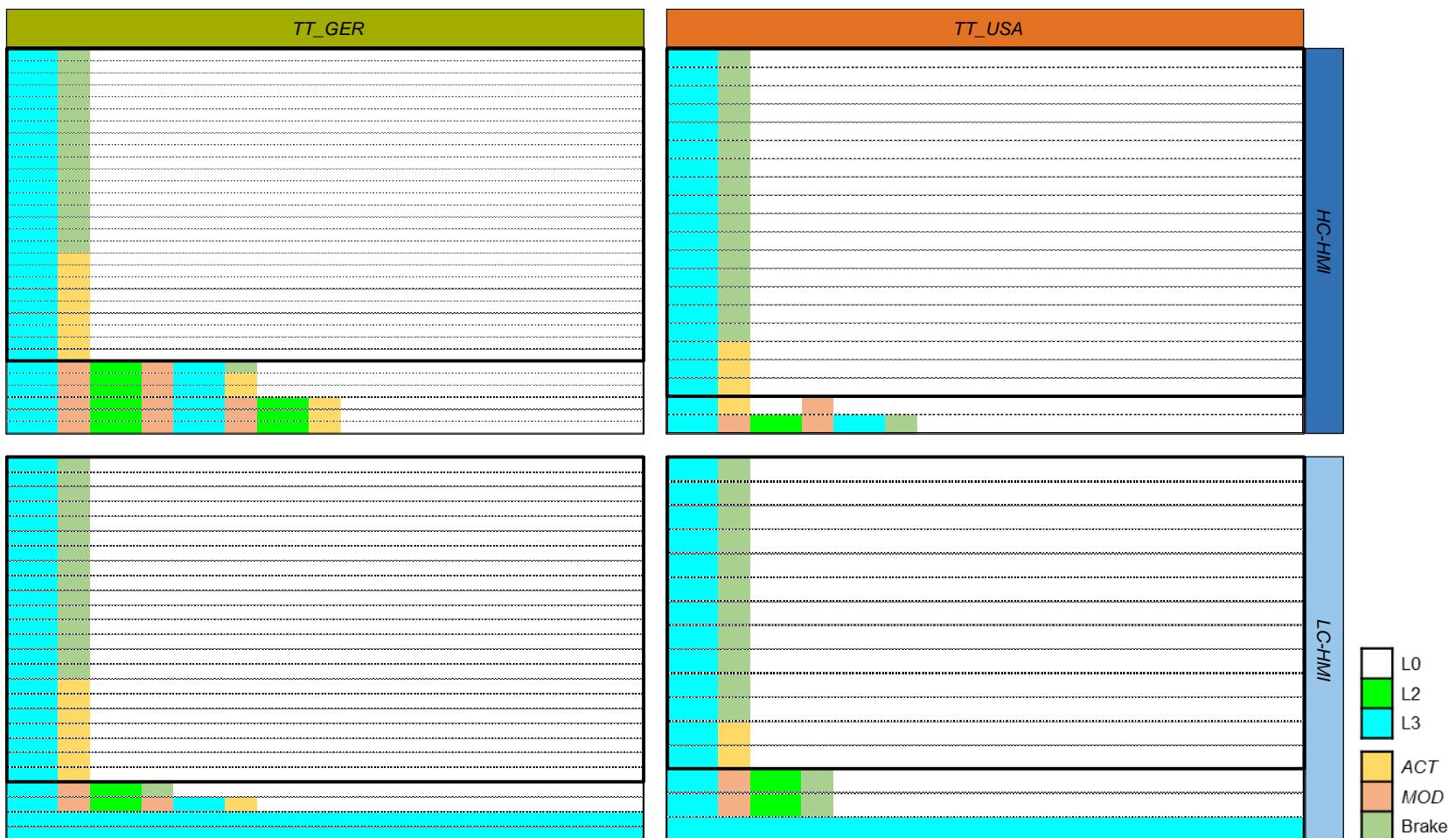


Figure 14.2 Visualization of the individual driving behavior for the metric *Take-over Path after Rtl* for Rtl_{20s} for the study *Exp_Culture* (Paragraph *Take-Over Path after Rtl*).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L3 at the start of the Rtl (leftmost column signals that L3 is active) are included. Participants using the one action only are marked with a black box. The sample sizes are as follows: TT_GER-HC: $n = 32$; TT_GER-LC: $n = 26$; TT_USA-HC: $n = 21$; & TT_USA-LC: $n = 16$.

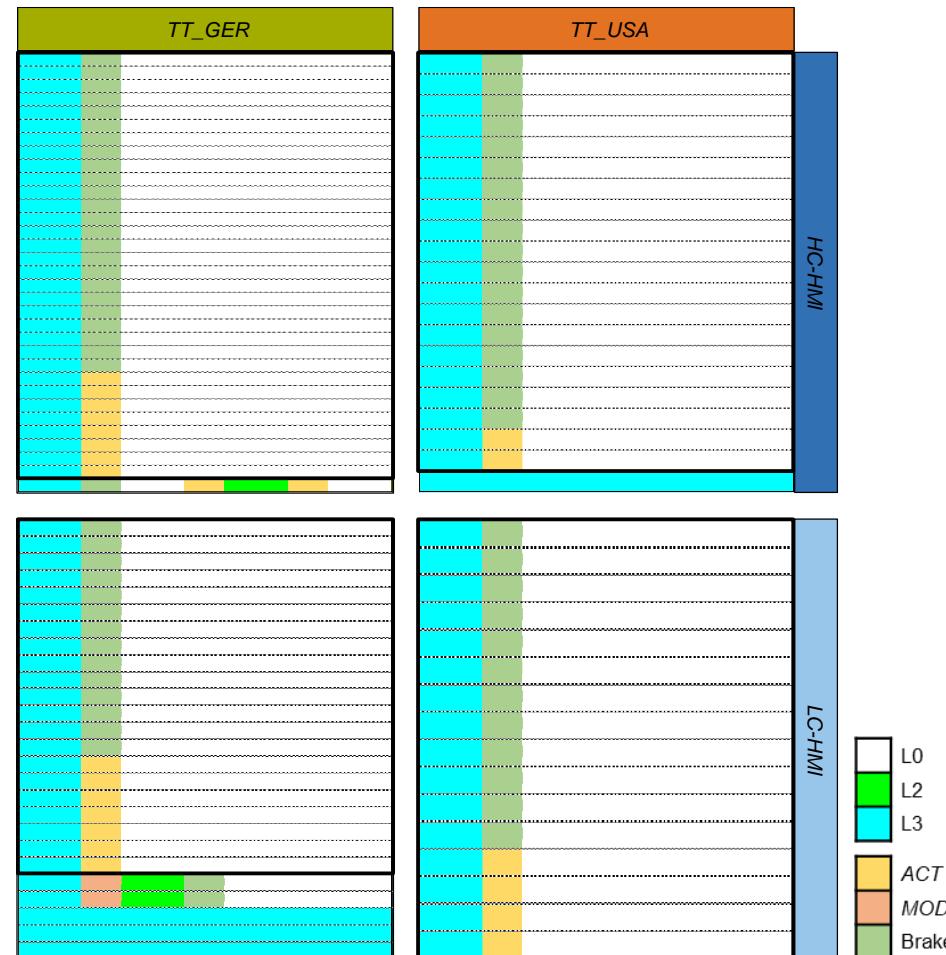


Figure 14.3 Visualization of the individual driving behavior for the metric *Take-over Path after Rtl* for Rtl_{6s} for the study *Exp_Culture* (Paragraph *Take-Over Path after Rtl*).

Note. The figure shows actions of the participants and the resulting LoAs. Only participants driving L3 at the start of the Rtl (leftmost column signals that L3 is active) are included. Participants using one action only are marked with a black box. The sample sizes are as follows: TT_GER-HC: $n = 33$; TT_GER-LC: $n = 26$; TT_USA-HC: $n = 21$; & TT_USA-LC: $n = 16$.

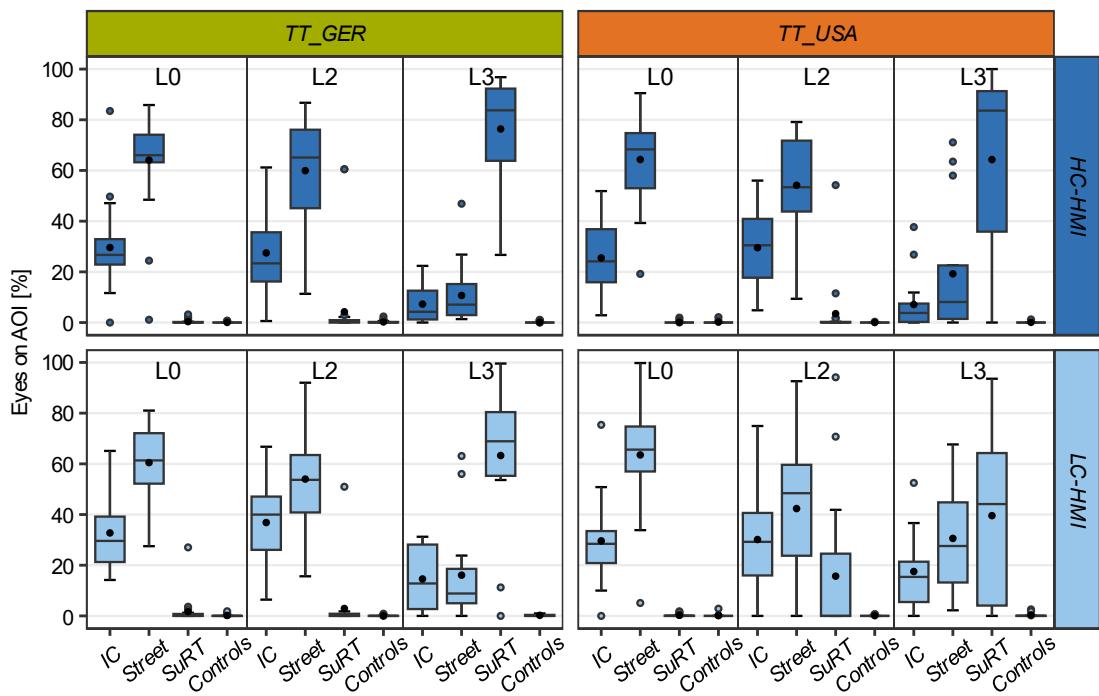


Figure 14.4 Boxplot diagram visualizing the results of the metric *Attention ratio during continuous rides in L0, L2, & L3* for all four AOIs for the study *Exp_Culture* (Paragraph Attention Ratio during Continuous Rides in L0, L2, & L3).

Note. The sample sizes are as follows: *TT_GER-HC*: $n = 27$ (L0), $n = 32$ (L2), $n = 27$ (L3); *TT_GER-LC*: $n = 22$ (L0), $n = 20$ (L2), $n = 14$ (L3); *TT_USA-HC*: $n = 19$ (L0), $n = 20$ (L2), $n = 15$ (L3); & *TT_USA-LC*: $n = 20$ (L0), $n = 18$ (L2), $n = 12$ (L3).

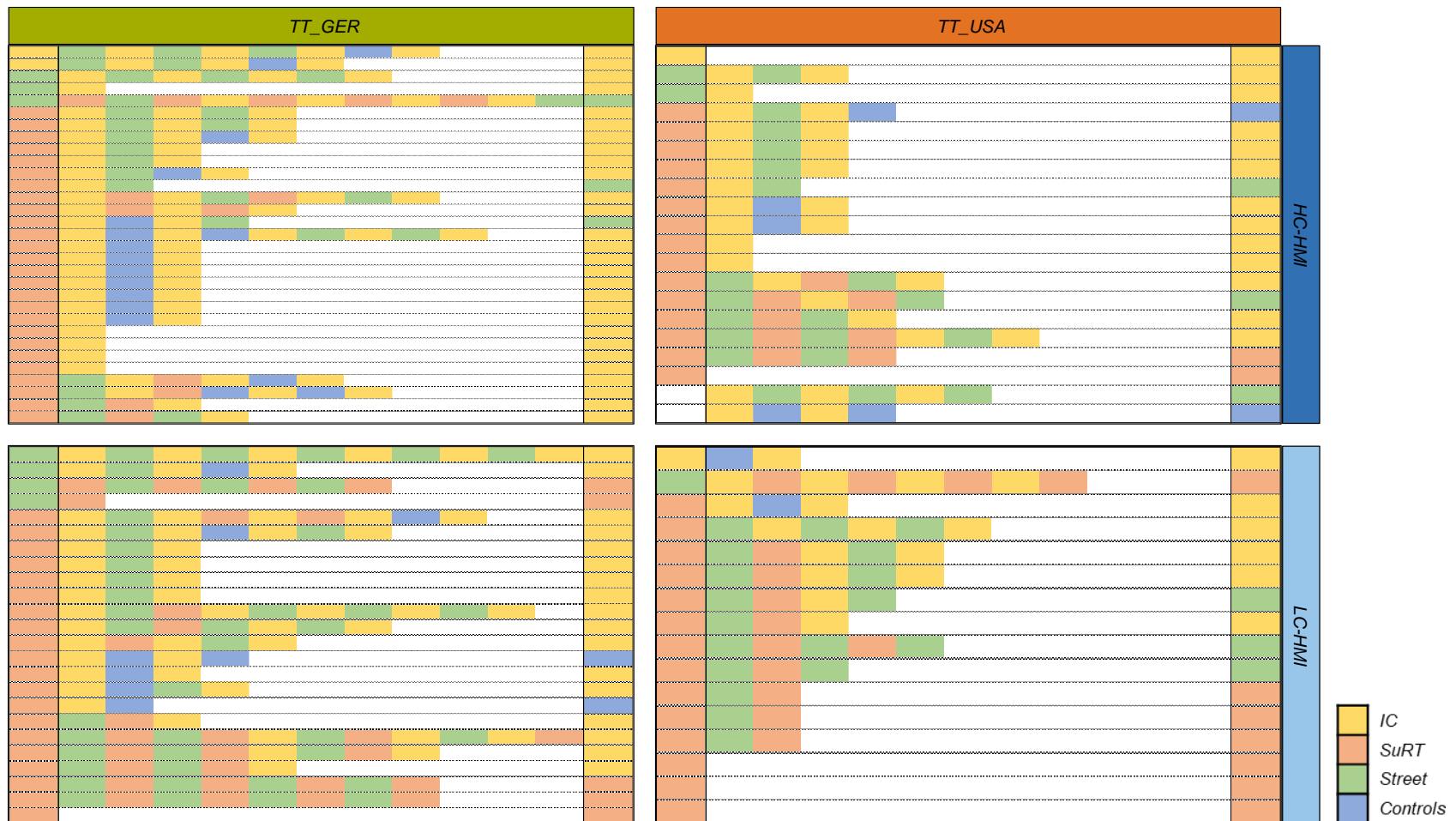


Figure 14.5 Visualization of the individual gaze behavior for the metric *Gaze Behavior during Rtl* for Rtl_{20s} for the study *Exp_Culture* (Paragraph *Gaze Behavior during Rtl*).

Note. The figure shows active AOIs of the participants between the start and the end of the Rtl. The end of the Rtl is marked by the start of emergency braking maneuver or the transition to L0. Only participants driving L3 at the start of Rtl are included. The sample sizes are as follows: TT_GER-HC: $n = 31$; TT_GER-LC: $n = 24$; TT_USA-HC: $n = 20$; & TT_USA-LC: $n = 16$.

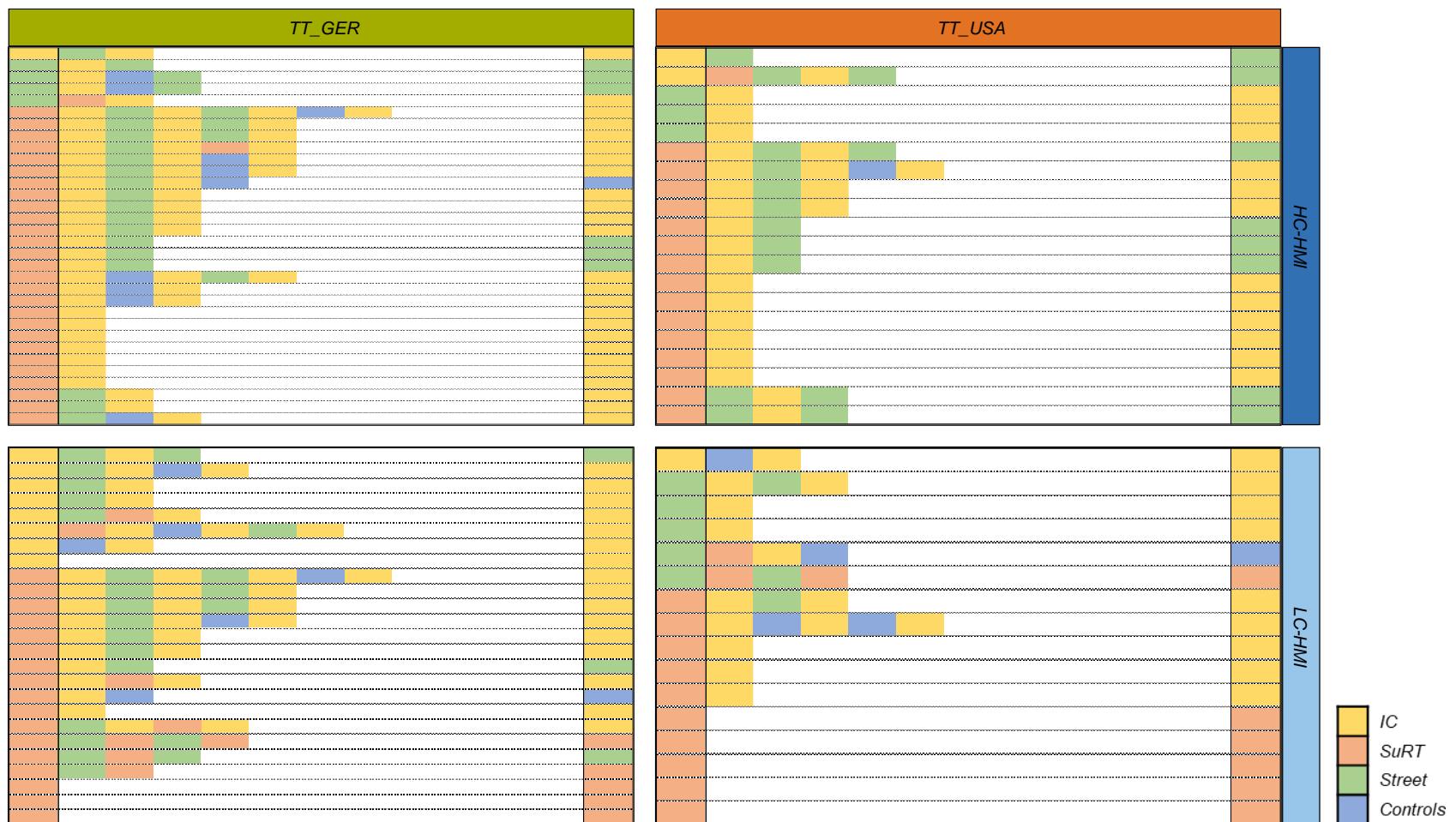


Figure 14.6 Visualization of the individual gaze behavior for the metric *Gaze Behavior during Rtl* for Rtl_{6s} for the study *Exp_Culture* (Paragraph *Gaze Behavior during Rtl*).

Note. The figure shows active AOIs of the participants between the start and the end of the Rtl. The end of the Rtl is marked by the start of emergency braking maneuver or the transition to L0. Only participants driving L3 at the start of Rtl are included. The sample sizes are as follows: TT_GER-HC: $n = 32$; TT_GER-LC: $n = 25$; TT_USA-HC: $n = 20$; & TT_USA-LC: $n = 16$.

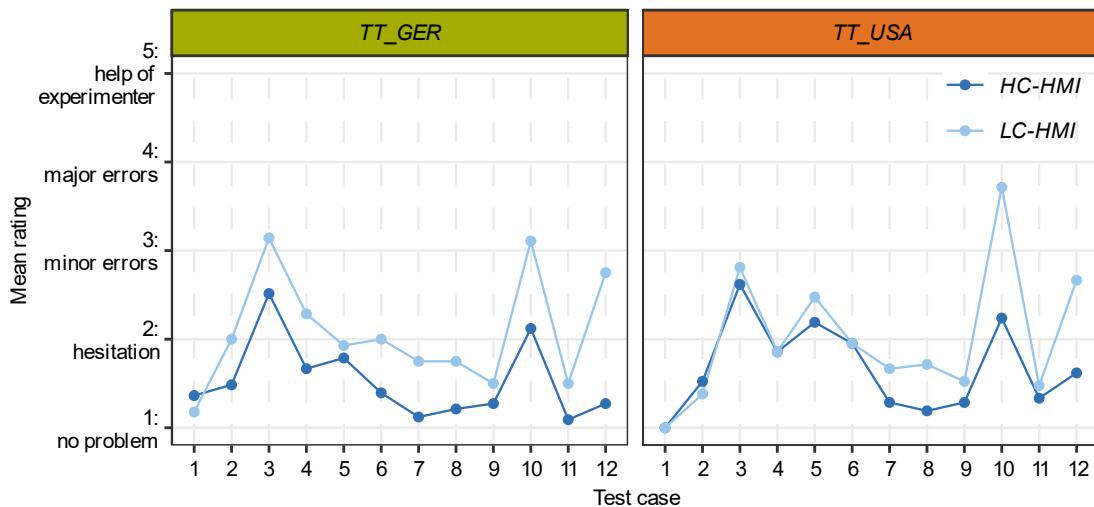


Figure 14.7 Visualization of the mean ratings per test case for the metric *Experimenter rating* for the study *Exp_Culture* (Subsubsection Experimenter Rating).

Note. The sample sizes are as follows: TT_GER-HC: $n = 33$; TT_GER-LC: $n = 28$; TT_USA-HC: $n = 21$; & TT_USA-LC: $n = 21$.

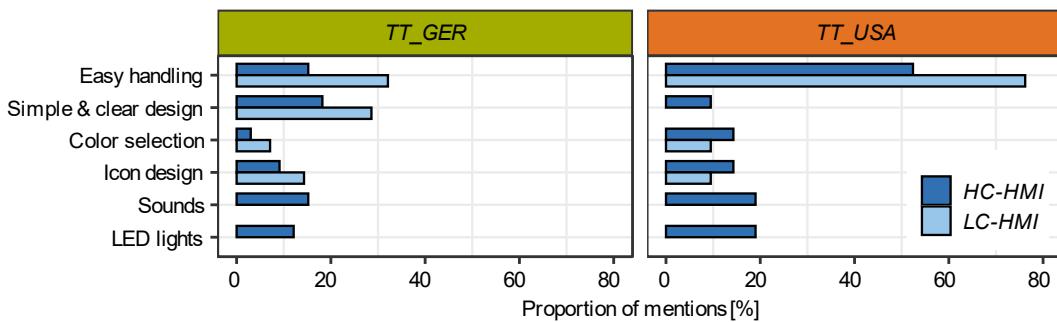


Figure 14.8 Overview of the clustered replies for praised components of the HMI concepts in the metric *Final Interview* for the study *Exp_Culture* (Subsubsection Final Interview).

Note. The sample sizes are as follows: TT_GER-HC: $n = 33$; TT_GER-LC: $n = 28$; TT_USA-HC: $n = 21$; & TT_USA-LC: $n = 21$.

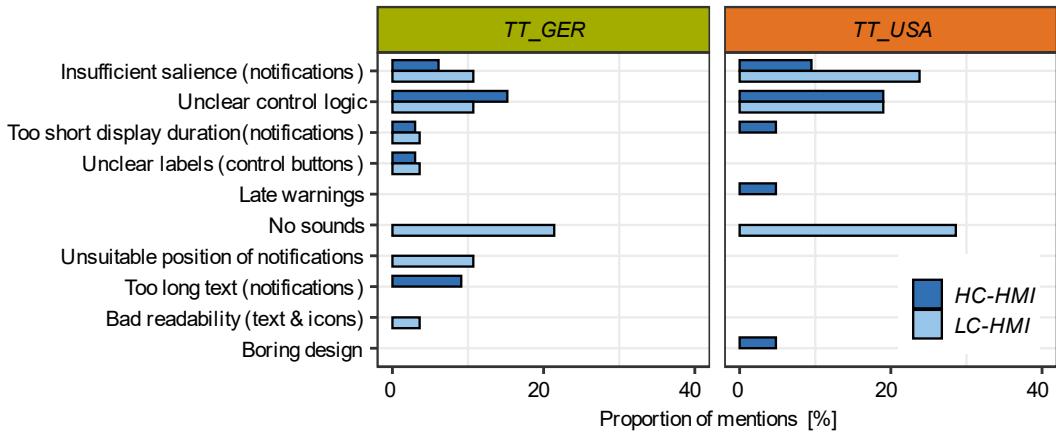


Figure 14.9 Overview of the clustered replies for criticized components of the HMI concepts in the metric *Final Interview* for the study *Exp_Culture* (Subsubsection Final Interview).

Note. The sample sizes are as follows: TT_GER-HC: n = 33; TT_GER-LC: n = 28; TT_USA-HC: n = 21; & TT_USA-LC: n = 21.

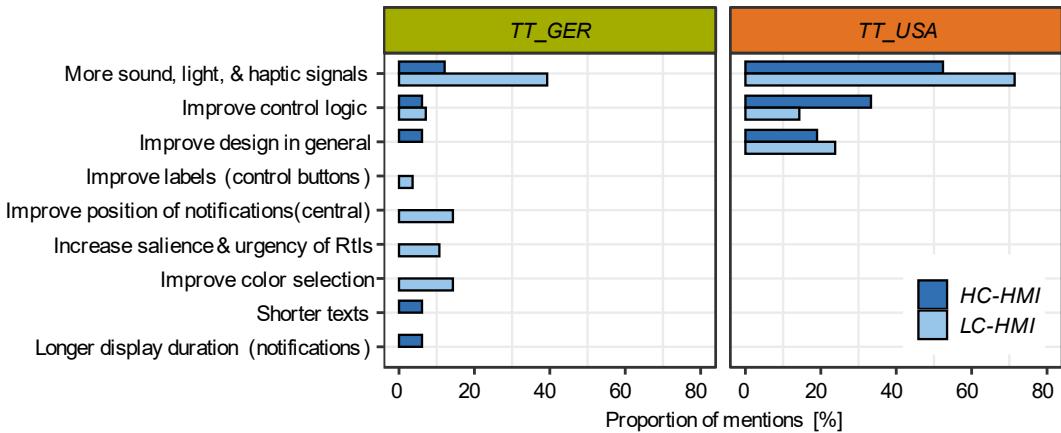


Figure 14.10 Overview of the clustered replies for improvement suggestions for components of the HMI concepts in the metric *Final Interview* for the study *Exp_Culture* (Subsubsection Final Interview).

Note. The sample sizes are as follows: TT_GER-HC: n = 33; TT_GER-LC: n = 28; TT_USA-HC: n = 21; & TT_USA-LC: n = 21.