Evaluation of User Interaction Concepts for Driver Displays

Analysis of Expert-Based Approaches for Usability Evaluation During Development



Inaugural-Dissertation zur Erlangung der Doktorwürde der Fakultät für Sprach-, Literatur- und Kulturwissenschaften der Universität Regensburg

Vorgelegt von

Lukas Lamm

aus

Freilassing

2022

Erstgutachter: Prof. Dr. Christian Wolff

Zweitgutachter: Prof. Dr. Michael Burmester

Dedicated to the memory of Dieter Lamm.

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Dr. Christian Wolff, who did not hesitate to supervise my dissertation in collaboration with the Mercedes-Benz AG, and who constantly supported, adviced, and encouraged me throughout the entire process. I also would like to thank Prof. Dr. Michael Burmester for taking the position as second assessor of my thesis as well as his helpful comments and motivating discussions.

Of course, I would also like to thank my two supervisors at the Mercedes-Benz AG, Christina Ibrom and Dr. Tim Schlüsener, for their support in carrying out my empirical studies as well as for the opportunity to gain valuable professional experience during my employment as doctoral student in the team *IC/HUD UI-Concepts & Assistance*. Many thanks to the whole team for supporting me during my dissertation, for every warm word as well as professional discussions. I have learned so much in the past years, and for this opportunity I am truly grateful.

I would like to thank my mother, who always supported me, and who probably never imagined that I would write a doctoral thesis. Without her persistent but always cordial nature, and the freedom granted in my personal development, this dissertation would probably not have come about.

Finally, I would like to thank Michaela, who always was understanding and encouraging for this project. I am grateful for your love and support, thank you <3

Abstract

This dissertation deals with the application of expert-based usability evaluation methods in vehicles. A deficit in this area of research could be shown by an exploratory literature review. In two comparative case studies the suitability of two concrete methods was systematically investigated. For the comparison, the methods cognitive walkthrough and guideline review were applied for the usability evaluation of different parts of an in-vehicle information system comparing the results with those of parallel usability tests. In order to satisfy the special context of use of an in-vehicle information system, established definitions of usability as well as their classification in the user-centered design process and the specific implications for the environment in the vehicle were considered. The two methods applied were then prepared specifically for the context of use in the vehicle. While no clear recommendation for the cognitive walkthrough in the context of use of a vehicle was found, the guideline review could achieve satisfactory results. Furthermore, the additional introduction of a new metric, based on the System Usability Scale (SUS), allows a comparison between individual versions of a system in the course of a project and thus closes a gap in the field of expert-based usability evaluation methods.

Zusammenfassung

Diese Dissertation beschäftigt sich mit dem Einsatz von expertenbasierten Usability-Evaluationsmethoden im Fahrzeug. Ein Defizit in diesem Bereich der Forschung konnte durch eine explorative Literaturstudie gezeigt werden. In zwei vergleichenden Fallstudien wurde die Eignung von zwei konkreten Methoden systematisch untersucht. Für den Vergleich wurden die Methoden Cognitive Walkthrough und Guideline Review für die Usability-Evaluation von unterschiedlichen Teilbereichen eines Fahrerinformationssystems angewendet und die Ergebnisse mit denen eines jeweils parallelen Usability-Tests verglichen. Um dem speziellen Nutzungskontext eines Fahrerinformationssystems gerecht zu werden, wurden etablierte Definitionen von Usability, sowie deren Einordnung in einen benutzerorientierten Gestaltungsprozess und die spezifischen Implikationen für die Umgebung im Fahrzeug betrachtet. Daraufhin wurden die beiden angewandten Methoden speziell für den Nutzungskontext im Fahrzeug aufbereitet. Während sich für die Methode des Cognitive Walkthrough keine eindeutige Empfehlung für den Benutzungskontext *Fahrzeug* herauskristallisierte, konnten durch den Einsatz des Guideline Review zufriedenstellende Ergebnisse erzielt werden. Die gleichzeitige Einführung einer neuen Metrik, angelehnt an den System Usability Scale (SUS) erlaubt darüber hinaus einen Vergleich zwischen einzelnen Versionen eines Systems im Projektverlauf und schließt damit eine Lücke im Bereich expertenbasierter Usability-Evaluationsmethoden.

Contents

1. Introduction **1**

1.1. Evolution of Driver Information Displays 1
1.2. Aims and Objectives 3
1.3. Related Publications 4
1.4. Outline of the Thesis 4

2. Usability Engineering **7**

2.1. Definitions of Usability 7
2.1.1. Brian Shackel 7
2.1.2. Jakob Nielsen 8
2.1.3. Donald Norman 10
2.1.4. Ben Shneiderman 11
2.1.5. Usability in International Standards 11
2.2. Usability Factors 13
2.3. Human-Centered Design Process 14
2.4. Evaluation Methods 17
2.4.1. Surveys 18
2.4.2. Interviews and Focus Groups 23
2.4.3. Expert-Based Testing 25
2.4.4. User-based Testing 30
2.4.5. Measuring Human Performance 35
2.4.6. Model-Based Testing 37
2.4.7. Automated Testing 39
2.5. Usability for In-Vehicle Information Systems 40
2.5.1. The Specific Context of Use 40
2.5.2. Design Space 45
2.5.3. Driver Distraction 49

3. Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Information Systems **55**

3.1. Related Work 55
3.2. Sources 57

3.3. Classification Schema . . . 59
3.4. Approach . . . 61
3.4.1. Graph Database . . . 62
3.4.2. Network Analysis . . . 63
3.4.3. Information Visualization . . . 70
3.5. Results . . . 74
3.5.1. Who are the authors? . . . 77
3.5.2. How do these authors collaborate? . . . 79
3.5.3. Which interface types are of interest? . . . 83
3.5.4. Which evaluation methods are applied? . . . 84
3.5.5. Similarity Between Publications . . . 87
3.6. General Discussion . . . 89
3.7. Conclusions . . . 91

4. Expert Reviews for Automotive User Interfaces 93
4.1. Background . . . 93
4.2. Related Work . . . 96
4.3. Who is an Expert? . . . 100
4.4. What is a Usability Problem? . . . 102
4.5. Measuring the Effectiveness of Usability Evaluation Methods . . . 106
4.6. Summary . . . 109

5. Case Study: A Cognitive Walkthrough of a Driver Display Context Menu 111
5.1. Background . . . 111
5.1.1. The Cognitive Walkthrough Method . . . 111
5.1.2. Cognitive Walkthrough in Dual Task Environments . . . 114
5.2. Object of Investigation - The Context Menu Concept . . . 115
5.3. Experiment 1 - User Study . . . 116
5.3.1. Method . . . 116
5.3.2. Results . . . 120
5.3.3. Discussion . . . 126
5.4. Experiment 2 - Expert Review . . . 127
5.4.1. Method . . . 128
5.4.2. Results . . . 130
5.4.3. Discussion . . . 134
5.5. Comparison . . . 135
5.5.1. Task Performance . . . 135
5.5.2. Usability Problems . . . 136
5.5.3. Measuring Effectiveness . . . 140
5.5.4. Discussion . . . 142

6. Case Study: A Guideline Review of a Driver Display Screen Configuration Application **145**
6.1. Background . . . 145
6.1.1. Guideline Review Method . . . 145
6.1.2. Literature Review for Existing Guidelines . . . 146
6.2. Object of Investigation - A Customizable Driver Display . . . 148
6.3. Experiment 1 - User Study . . . 149
6.3.1. Method . . . 149
6.3.2. Results . . . 153
6.3.3. Discussion . . . 158
6.4. Experiment 2 - Expert Review . . . 160
6.4.1. Method . . . 160
6.4.2. Results . . . 163
6.4.3. Discussion . . . 169
6.5. Comparison . . . 170
6.5.1. System Usability Scale and Guideline Conformity Score . . . 170
6.5.2. Usability Problems . . . 171
6.5.3. Measuring Effectiveness . . . 175
6.5.4. Discussion . . . 177

7. Summary & Outlook **181**
7.1. Main Contributions and Discussion . . . 181
7.1.1. Usability Engineering in an Automotive Context . . . 182
7.1.2. Literature Review of the Current State of the Art . . . 183
7.1.3. Expert Reviews in Automotive . . . 185
7.1.4. Case Study: Cognitive Walkthrough . . . 185
7.1.5. Case Study: Guideline Review . . . 187
7.2. Outlook and Future Work . . . 188
7.2.1. Expanding the Scope of Literature Review . . . 188
7.2.2. Consecutive Studies . . . 189
7.2.3. Toward a Toolbox for Usability Evaluation of In-Vehicle Information Systems . . . 190

References **191**

Appendices **215**

A. Exploratory Literature Review **217**

B. Case Study: Cognitive Walkthrough **221**
B.1. User Study . . . 221

B.2. Cognitive Walkthrough 241

C. Case Study: Guideline Review **253**
C.1. Literature Review for Existing Guidelines 253
C.1.1. General HMI Guidelines 253
C.1.2. Website Usability Guidelines 255
C.1.3. Ergonomic Criteria 259
C.1.4. Readability 263
C.1.5. Accessibility 264
C.1.6. Motivation 264
C.1.7. Persuasion 265
C.1.8. Situation Awareness 266
C.1.9. Automotive HMI Guidelines 268
C.2. User Study 277
C.3. Guideline Review 286

List of Figures

1.1. In-vehicle instrumentation: input and output trends 2

2.1. Attributes of system acceptability 9
2.2. Overview of ISO Standards on Usability 12
2.3. The double diamond design process 15
2.4. The human-centered design process 15
2.5. Dimensions of users' experience 32
2.6. Estimated sample sizes to detect event of interest at least once 34
2.7. Concepts of GOMS model 38
2.8. Taxonomy of usability evaluation automation 39
2.9. Distribution of primary, secondary, and tertiary tasks 41
2.10. Representation of multimodal human-machine interaction loop 42
2.11. Interaction between driver and in-vehicle information system (IVIS) . . . 43
2.12. Division of driver's interaction environment 46
2.13. Graphical representation of the design space for driver-based automotive user interfaces 48
2.14. Three-dimensional representation of the structure of multiple resources 50

3.1. Literature review graph schema 60
3.2. SQL database schema for literature review 62
3.3. *COLLABORATE* relationship between authors 64
3.4. Reference model for visualization 71
3.5. Different network frequency distributions. Illustrating the ratios of papers per author, collaborators per author, and involved authors and institutions per paper. 76
3.6. Community graphs for Gary Burnett from the University of Nottingham, Manfred Tscheligi from the University of Salzburg and Bryan Reimer and Bruce Mehler from the Massachusetts Institute of Technology, the node size represents the PageRank. The IDs correspond to the IDs allocated by the *Louvain* algorithm. 81
3.7. Comparison of community detection algorithms output 82
3.8. Investigated interfaces 83
3.9. Frequency of investigated interfaces and used methods. 84

3.10. Method usage per interface type 85
3.11. Similarity relationships between usability evaluation methods (UEMs) . 87
3.12. Clusters of similar publications 88
3.13. Top 3 clusters of similar publications 88

4.1. Dimensions of experts 102
4.2. Model of usability problem components 104
4.3. The Usability Problem Taxonomy 105
4.4. The Usability Problem Classifier 106

5.1. Scribbles of the context menu concepts 116
5.2. Experimental setup and position of displays 118
5.3. Task performance distribution for both conditions 120
5.4. Task rating distribution for both conditions 122
5.5. Number of usability problems. 123
5.6. User study Usability Problem Classifier (UPC) distribution 125
5.7. Task performance distribution for both conditions 130
5.8. Number of usability problems. 131
5.9. Cognitive walkthrough (CW) UPC distribution 132
5.10. Comparison of the task performance results 136
5.11. Comparison of the distribution among the categories of the UPC 140

6.1. Scribbles of the individualization screen 149
6.2. Experimental setup and position of displays 150
6.3. Task performance distribution for both conditions 154
6.4. Distribution of ratings and System Usability Scale (SUS) scores 155
6.5. Number of usability problems from the user study. 156
6.6. Usability problem classification distribution 157
6.7. Distribution of guideline weights 162
6.8. The Guideline Compliance Scale (GCS) and the number of guideline violations per variant. 164
6.9. Number of usability problems from the expert review. 166
6.10. Usability problem classification distribution 167
6.11. Comparison of the SUS and GCS 171
6.12. Comparison of the distribution among the categories of the UPC 174

A.1. Hierarchical cluster analysis of community detection algorithms results. 220

List of Tables

2.1. Existing survey tools and standardized usability questionnaires. 20
2.2. Summary of the MOT-Technique 29
2.3. Symbolic representation used in the graphical design space 47

3.1. List of sources for literature review. 58
3.2. Usage frequency of used usability evaluation methods (UEMs). 63
3.3. A Data Table for the paper entity of the literature review data. 72
3.4. A Data Table for the author entity of the literature review data. 72
3.5. A Data Table for the institution entity of the literature review data. . . . 72
3.6. Network statistics summary 75
3.7. Top 5 most productive authors with their affiliation and the respective number of publications. 77
3.8. Top 5 institutions with highest number of publications, the respective number of authors affiliated, and the publications per author ratio. . . . 78
3.9. Top 5 communities 82

4.1. Keystroke-level model (KLM) operation times 94

5.1. Differences in task performance 121
5.2. Differences between task ratings 123
5.3. Distribution of usability problems among UPC task component categories. 138
5.4. Distribution of usability problems among UPC object component categories. 139
5.5. Benefit-cost ratios for user-based and expert-based approaches. 142

6.1. Wilcoxon signed rank test for task performance 154
6.2. Intraclass correlation between the four experts for GCS. 164
6.3. Distribution of usability problems among UPC task component categories. 173
6.4. Distribution of usability problems among UPC object component categories. 174
6.5. Benefit-cost ratios for user-based and expert-based approaches. 176

A.1. List of communities 217

B.1. Results of Shapiro-Wilk test for normality 221
B.2. Results of correlation analysis 221

B.3. List of identified usability problems 222
B.4. List of identified usability problems 241

C.1. Consolidated automotive human-machine interaction (HMI) guidelines 268
C.2. Results of Shapiro-Wilk test for normality and correlation analysis 277
C.3. List of identified usability problems 278
C.4. Guideline review template 286
C.5. List of identified usability problems 289

List of Abbreviations

AAM Alliance of Automobile Manufacturers.

ACM Association for Computing Machinery.

ACT-R Adaptive Control of Thought-Rational.

ASQ After-Scenario Questionnaire.

ATT Attractiveness.

CAN Controller Area Network.

CPA critical path analysis.

CSUQ Computer System Usability Questionnaire.

CUE Components of User Experience.

CW cognitive walkthrough.

DALI Driving Activity Load Index.

ECG electrocardiogram.

EEG electroencephalogram.

EGR exhaust gas recirculation.

FFT Fast Fourier Transformation.

GCS Guideline Compliance Scale.

GOMS goals, operators, methods, and selection rules.

HCI human-computer interaction.

HE heuristic evaluation.

HEE Human Efficiency Evaluator.

HMI human-machine interaction.

HQ-I hedonic quality – identity.

HQ-S hedonic quality – stimulation.

HTA hierarchical task analysis.

ICC intraclass correlation coefficient.

IEEE Institute of Electrical and Electronics Engineering.

IVIS in-vehicle information system.

KLM keystroke-level model.

KPI key performance indicators.

LCT lane change task.

MCH modified Cooper Harper scale.

meCUE modular evaluation of key Components of User Experience.

MOT metaphors of human thinking.

MTM methods time measurement.

NASA-TLX NASA Task Load Index.

NHTSA National Highway Traffic Safety Administration.

OBD on-board diagnostics.

PQ pragmatic quality.

PSQ Printer-Scenario Questionnaire.

PSSUQ Post-Study System Usability Questionnaire.

QUESI Questionnaire for the Subjective consequences of Intuitive use.

QUIS Questionnaire for User Interaction Satisfaction.

RSME Rating Scale Mental Effort.

SAE Society of Automotive Engineers.

SHERPA systematic human error reduction and prediction approach.

SUMI Software Usability Measurement Inventory.

SUS System Usability Scale.

SWAT Subjective Workload Assessment Technique.

TA think aloud.

TCD transcranial Doppler sonography.

TSOT total shutter open time.

TTT total task time.

UEM usability evaluation method.

UEQ User Experience Questionnaire.

UMUX Usability Metric for User Experience.

UPC Usability Problem Classifier.

UPT Usability Problem Taxonomy.

WIMP windows, icons, mouse, and pointer.

1. Introduction

January 29, 1886, when Carl Benz handed in the patent for his "Fahrzeug mit Gasmotorenbetrieb" (Benz & Co., 1886), is considered as the date of birth of the automobile. With the Model T in 1908, Henry Ford introduced the first mass produced car, which until in 1972 the Volkswagen Beetle took over leadership was the most sold automobile in the world ("Ford Model T," 2018). Already since these early days of vehicle development, ergonomic factors played an important role. Besides the aspect of accessibility, the need for feedback and information grew (Harvey & Stanton, 2013).

Nowadays, the long-established companies in automotive industry also have to face new challenges as well as new players on the market. Apple Senior Vice President of Operations Jeff Williams called the car "the ultimate mobile device" (Snyder, 2015). Or as Damiani et al. (2009) cites "MOTOROLA", the "current production cars contain more computing power than was used to send the Apollo spacecraft to the moon". While consumer electronics technology is evolving, also the offer for information and entertainment in modern cars is growing. Automobiles contain features for media consumption (e.g. radio, audio, etc.) or personal communication through several connectivity options for smartphones. With these growing capabilities, also the workload and the sources of distraction for the driver are increasing. Although the development of advanced driver assistance systems (ADAS) eases the driving task, but one should not lose sight of this primary task (Kern, 2012).

A well-balanced display strategy, as well as sophisticated evaluation procedures can minimize driver distraction and the risk of accidents. The present dissertation therefore aims to investigate suitable evaluation methods for driver information displays in the specific context of use while driving.

1.1. Evolution of Driver Information Displays

Since the very early phase of the automobile, where HMI consisted of steering wheel, accelerator, and brake, it has been a major issue to get information about the operated vehicle status. As Damiani et al. (2009) state, this is also the decisive aspect for the development of the instrument cluster, which at that time displayed basic information

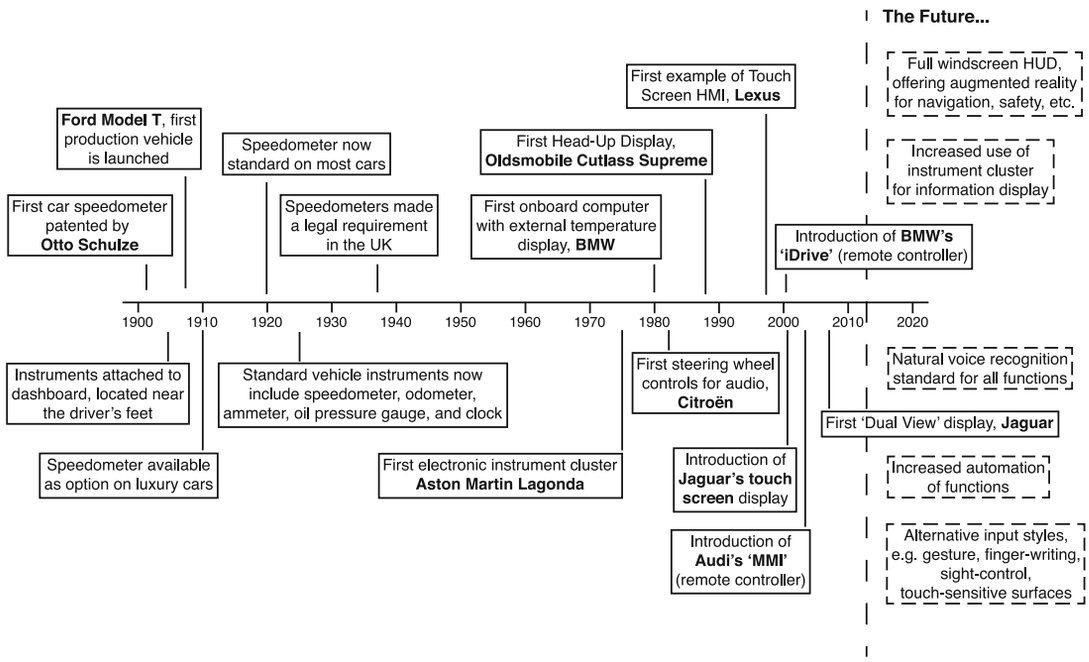


Figure 1.1.: In-vehicle instrumentation: input and output trends. The figure is based on Harvey and Stanton (2013, p. 3).

about speed or fuel level. Harvey and Stanton (2013) describe the process of connecting the driver and the vehicle as instrumentation. Figure 1.1 shows a timeline of instrumentation with different milestones during the development. The first effort of instrumentation was the speedometer developed by Otto Schulze around 1902, which was available as option on luxury cars (Ludvigsen, 1997) in the early years. Later, in the mid-twenties the speedometers got standard (Ludvigsen, 1997) and also a legal requirement in the late-thirties (Harvey & Stanton, 2013). The period during World War II encouraged the effort to measure human factors and design products (also automobiles) with the user in mind, especially through the data available from people enlisting in the armed forces (Ludvigsen, 1997).

The first electronic instrument cluster with all-digital mainly numeric values was introduced by Aston Martin in their Model Lagonda in 1976 ("Aston Martin Lagonda," 2018; Ludvigsen, 1997). Later around 1980 BMW equipped automobiles with onboard computers including for example an external temperature display (Harvey & Stanton, 2013). In the 1980s car manufacturers began to replace roll counters for odometers with LCD displays ("100 Jahre Tachometer. Tempomesser Kam Nur Langsam Auf Touren." 2002) and Citroën introduced steering wheel controls for audio (Harvey & Stanton, 2013). Later in the 1980s, Oldsmobile as the first manufacturer equipped their model Cutclass Supreme with a head-up display (Harvey & Stanton, 2013).

Harvey and Stanton (2013) also list some milestones regarding the development of mainly navigation and infotainment systems, like the introduction of remote controllers like iDrive by BMW or MMI by Audi. Besides these their timeline contains also future developments, like full windscreen head-up displays, augmented reality or the increase of instrument cluster as information display medium. Furthermore they predict alternative input styles like gesture, touch or gaze input to spread (Harvey & Stanton, 2013). While the complexity of in-vehicle information systems (IVIS) is rapidly increasing, the need for user-centered design is also rising. As the following quotation points out, usability is often left standing during the design of human-computer interaction for in-vehicle information systems:

> "However, remember that in the automotive vertical market, user-centered design has sometimes taken a back seat to technology and marketing, or has focused on a narrower range of concerns." (Marcus, 2004)

1.2. Aims and Objectives

The recurring theme of this dissertation is the evaluation of human-computer interaction for in-vehicle information systems through the application of expert-based usability evaluation methods. In a first step the term usability and its implications for the design of in-vehicle information systems (IVIS) is examined. In order to get an overview of the current state of the art and the research landscape of usability engineering methods in the automotive domain, a systematic literature review is performed. To account for the identified need for research on expert-based usability evaluation methods for IVIS, the term *Expert Review* is clarified in the context of this dissertation. In two case studies different approaches of expert reviews are compared to user-based testing regarding effectiveness and efficiency through validity, thoroughness, reliability, and a benefit-cost ratio. These steps should help expressing the importance of a user-centered perspective in the design of human-computer interaction for in-vehicle information systems.

To achieve this higher-level goal, a number of subordinate objectives will have to be met. The following research objectives will be addressed in this dissertation:

- **Usability evaluation for in-vehicle information systems**: Examine several definitions of usability and the corresponding influencing factors; describe different categories of usability evaluation methods and their relation to the user-centered design process; identify the specific context of use for the interaction with in-vehicle information system (cf. chapter 2).

- **Literature review on evaluation of in-vehicle information system**: Capture the state of the art and the need for research on expert-based usability evaluation methods for in-vehicle information systems (cf. chapter 3).
- **Expert reviews for automotive user interfaces**: Analyze framework conditions and the scope of expert-based usability evaluation methods for in-vehicle information systems; identify suitable measures of effectiveness of usability evaluation methods (cf. chapter 4).
- **Case studies of expert reviews methods**: Design and conduct evaluation studies to compare expert-based evaluation techniques with user-based testing methods as a baseline; application of the cognitive walkthrough technique to evaluate context menus in a driver display (cf. chapter 5); application of the guideline review method to evaluate a screen configurator for a driver display (cf. chapter 6).
- **Implications and limitations for the usability evaluation of in-vehicle information systems**: Derive implications and limitations from the consolidated results of the case studies for the application of expert-based usability evaluation methods for in-vehicle information systems; discussion of impacts and future work on expert-based usability evaluation for in-vehicle information systems (cf. chapter 7).

1.3. Related Publications

The following articles/papers covering specific aspects described in this thesis were published in the course of the work on this thesis:

- Lamm, L., & Wolff, C. (2019). Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Systems. *11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19)*
- Lamm, L. (2019). Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Systems
- Lamm, L., & Wolff, C. (2021). GCS: A Quick and Dirty Guideline Compliance Scale. *Journal of Usability Studies*, *16*(3), 179–202

1.4. Outline of the Thesis

Ch. 1: Introduction The introductory chapter gives an overview of the context of this work and the evolution of driver information displays over the years. The chapter also presents the aims of this work and the research objectives.

Ch. 2: Usability Engineering This chapter gives an overview of different definitions of usability and the user-centered design process. The examination of different categories of usability evaluation methods may be seen as the basis for discussion of different evaluation methods in subsequent chapters. The chapter concludes with a detailed consideration of the specific context of use for in-vehicle information systems.

Ch. 3: Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Information Systems In this chapter, after a contemplation of the related work on systematic literature review, the approach of the presented literature review is outlined in detail. Besides the used data for the literature review, the procedure using a graph database and the applied techniques of network analysis are clarified. The chapter concludes with a comprehensive presentation of the results as well as a discussion of the implications for this work.

Ch. 4: Expert Reviews for Automotive User Interfaces To delimit the scope of expert review techniques applied in the case studies in chapters 5 and 6, this chapter gives an overview of related work on the comparison of user-based and expert-based usability evaluation methods. Furthermore, the terms expert and usability problem are examined and specified regarding the already mentioned case studies. Finally, the chapter introduces different measures for comparing usability evaluation methods, focusing mainly on the performance regarding the detection of usability problems.

Ch. 5: Case Study: A Cognitive Walkthrough of a Driver Display Context Menu The case study compares the results of an empirical usability test and a cognitive walkthrough regarding the effectiveness of the latter usability evaluation method. For the measurement of effectiveness through validity, thoroughness, reliability, and a benefit-cost ratio, the results of the usability test are used as a baseline. While the results indicate that the cognitive walkthrough is suited to identify specific task-related usability problems with high validity, the technique is limited regarding the dimensions of a user interface and fails to place the investigated feature inside the bigger picture of the incorporating system.

Ch. 6: Case Study: A Guideline Review of a Driver Display Screen Configuration Application The chapter describes the second case study to compare the results of an empirical usability study with those of an expert review applying the guideline review technique. The study shows that the expert-based approach produced similar results when comparing the System Usability Scale (SUS) from the user study with the introduced Guideline Compliance Scale (GCS) from the guideline review. The effectiveness measures validity and thoroughness for the guideline

review are on a medium level, while the reliability shows a strong effect compared to the results from the empirical study as a baseline. The measure of person-hours per usability problem shows a higher benefit-cost ratio for the guideline review. In summary, the guideline review is suitable in early development phases, for example to narrow the variants for a following user study, and is able to identify the most severe problems in an interface.

Ch. 7: Summary & Outlook The chapter summarizes and discusses the main contributions of this dissertation to the body of knowledge regarding expert-based usability evaluation for in-vehicle information system. Besides the discussion of several definitions of usability, how to assess usability, and the integration of usability evaluation methods in the user-centered design process, the chapter summarizes the examination of the specific context of use inside a vehicle and discusses key insights from the exploratory literature review in chapter 3 as well as the two case studies in chapters 5 and 6. Finally, the chapter presents an outlook on future work extending the scope of this dissertation.

2. Usability Engineering

When discussing usability evaluation methods for driver display concepts, first of all the term usability has to be elaborated. This chapter introduces several definitions of usability throughout the literature as well as the activities of the user-centered design process. As this dissertation investigates usability evaluation methods, the chapter gives an overview of available evaluation methods. Furthermore, the implications of the specific context of use inside a vehicle are discussed.

2.1. Definitions of Usability

> "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." (International Organization for Standardization, 2018, p. 6)

Besides the definition from the ISO standard 9241 — where mainly Nigel Bevan signed responsible for — several authors in the research field contributed significant work to the definition of usability. The following subsections present some of the most significant contributions to the field of human-computer interaction (HCI) and usability by introducing the most important authors individually.

2.1.1. Brian Shackel

As Dillon (2009) states, "Brian Shackel is considered by many to be the grandfather of the field of human-computer interaction" (Dillon, 2009, p. 367). Furthermore, the paper "Ergonomics for a Computer" (Shackel, 1959) is considered to be the first paper on HCI. Shackel (1997) himself, with his paper "Human-Computer Interaction—Whence and Whither?" in the *Journal of the American Society for Information Science*, looked back to over 40 years of work in HCI. The review of this very paper by Grudin (2009) summarizes that "Human-Computer Interaction—Whence and Whither?" not only set out his intentions and contributions, but also the demand for research and application of HCI.

Shackel (1997) also created a first breakdown of users who range from first engineers and non-specialists trained to use computers over professionals to consumers (Grudin, 2009). As Harvey and Stanton (2013, p. 20) summarize Brian Shackel defined the factors "effectiveness, learnability, flexibility, and attitude" (Harvey & Stanton, 2013, p. 20), to describe a usable system — which later got picked up and extended by Stanton and Baber (1992) under the acronym LEAF. The key factor therefore is "acceptable performance", which should be reached within a specified time (learnability), with acceptable effort (attitude) for a defined amount of users (effectiveness), whereby the system "should be able to deal with a range of tasks beyond those first specified" (flexibility) (Stanton & Baber, 1992, p. 154). This is also supported by his statement that usability can be defined by the interaction between the user, the task, and the environment (Shackel, 1986). Further, Shackel proposed to operationalize usability and use quantitative methods to assess the performance in the four mentioned dimensions of usability (Harvey & Stanton, 2013, p. 20).

2.1.2. Jakob Nielsen

The factor of learnability was also described by Jakob Nielsen as one of his five components of usability. In total they are "*Learnability*", "*Efficiency*", "*Memorability*", "*Errors*", and "*Satisfaction*". He describes the term learnability as a fundamental attribute of usability, as in most cases the learning of a specific system is the first experience users have with the system. Therefore the system should be easy to learn for novice users, but one should also consider the level of proficiency — e.g. when upgrading to new releases of a software, as well as the general knowledge about using computers and different learning strategies (Nielsen, 1993, p. 26). Efficiency is described by Nielsen (1993) as the level of performance for experienced users. To measure this performance one needs of course experienced users, for which the system has to be available for some time. Another possibility is to allow the test users to use the system for some time before the initial test and hence familiarize with it. Another key principle for usability is the memorability. It describes the case for casual users, which don't have to learn a system from scratch but need to remember how to use it. Nielsen (1993) also explains two possibilities to measure the memorability of a specific interface — either a standard user test with casual users or a memorability test after a test session. According to Nielsen (1993) a system should also be constructed to allow the user to make as few errors as possible. Thereby he defines an error as "any action that does not accomplish the desired goal" (Nielsen, 1993, p. 32). But one has to differentiate between so called "catastrophic" (Nielsen, 1993, p. 33) errors and errors that can be immediately corrected by the user and therefore only have an effect on the users task completion time. As the final attribute of usability Nielsen (1993) explains the "subjective satisfaction" (Nielsen, 1993, p. 33).

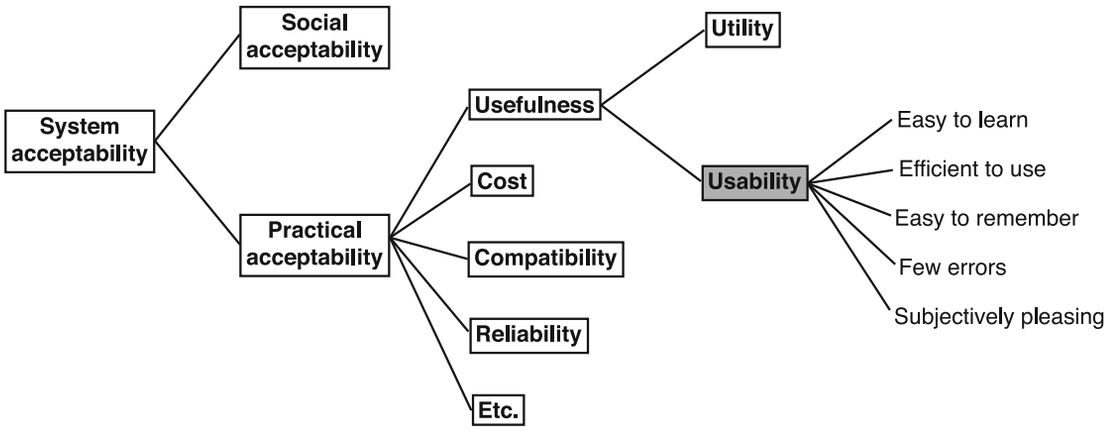


Figure 2.1.: Attributes of system acceptability. The figure is based on Nielsen (1993).

As Figure 2.1 illustrates, usability is considered as part of the larger issue of system acceptability. The key question is whether the system meets all the requirements and needs brought to it by the users and other potential stakeholders. Traditional categories include cost, compatibility, reliability, and usefulness (Nielsen, 1993). But as Bevan (1995) points out all of these dimensions has to be taken into account. For example a system that is easy to use but useless will not sell in most of the cases.

Jakob Nielsen also introduced the term "'discount usability engineering'" (Nielsen, 1989b). Nielsen (1993) considers that performing even some usability engineering during the development of a product is better than omitting usability because the application of the best methods would be too expensive or needs a strong theoretical background. Therefore his discount method uses the four techniques "User and task observation", "Scenarios", "Simplified thinking aloud", and "Heuristic evaluation" (Nielsen, 1993, p. 17). Scenarios in this case describe strictly reduced prototypes to only match a specific situation or task with a fixed path. According to his definition it describes "a single individual *user*" "using a specific set of computer *facilities*" "to achieve a specific *outcome*" "under specified *circumstances*" (Nielsen, 1990b, p. 315). Later he added the aspect of "over a certain *time interval*" (Nielsen, 1993) to his definition. The think aloud (TA) technique basically means that a single test user should verbalize his thoughts while using the investigated system. This allows to get insights into the mental concepts and predictions users have towards a computer system (Nielsen, 1993). The method of heuristic evaluation is described in detail in section 2.4.

2.1.3. Donald Norman

Among many others, Norman (1983) acknowledged the need for a definition of usability with an essential set of "principles from which to derive the manner of the interaction between person and computer" (Norman, 1983, p. 1). In addition, Norman (1983) appeals to the precision of the above principles, since "statements that proclaim 'Consider the user' are valid, but worthless" (Norman, 1983, p. 1). As Harvey and Stanton (2013, p. 22) cite Norman and Draper (1986), he "banished" the term "user friendliness" from his book to move away from vague concepts of usability. Furthermore, Norman's focus on the user's perspective played an important role in order to shift the definition of usability from the product to the user (Harvey & Stanton, 2013, p. 22). In his best-selling (Durham, 1998) book *The Design of Everyday Things* which was revised in 2013, Norman (2013, p. 8) even mentions the "ever-increasing complexity of the automobile dashboard" as one of these everyday things leading to frustration. Besides that, the automobile is also used as an example to describe the term experience with examples like "the sensation of power during acceleration, their ease of control while shifting or steering, or the wonderful feel of the knobs and switches on the instrument" (Norman, 2013, p. 10).

Norman (2013) lists the following six fundamental principles of interaction: *Affordances*, *Signifiers*, *Constraints*, *Mappings*, *Feedback*, and *Conceptual model*. The term affordance is described as a relationship between a physical object and a person or in other words: "But affordance is not a property. An affordance is a relationship." (Norman, 2013, p. 11). Therefore affordance is not only determined by the qualities of a product or an object, but also by the abilities of the person that is interacting. Because the term affordance describes a relationship rather than a property, Norman (2013) introduces the concept of signifiers. While affordances determine what actions are possible, signifiers communicate where the action should take place (Norman, 2013, p. 14). While the term constraints is separated into four classes — physical, cultural, semantic, and logical constraints — and should explain to the user why he can't perform a specific action (Norman, 2013, pp. 123 ff.), the mapping describes "the relationship between the elements of two sets of things" (Norman, 2013, p. 20). Natural mappings often follow from the principles of perception like groupings or proximity (Norman, 2013, p. 22). Another principle Norman (2013, p. 23) states in his book *The Design of Everyday Things* is the principle of feedback. When it comes to feedback, it should be immediate, informative, and dosed appropriately as well as planned and prioritized. The last principle of conceptual models describes simplified explanations of how something works (Norman, 2013, p. 25). As conceptual models "reside in the minds of the people" (Norman, 2013, p. 26), these may differ regarding the persons background or experience. As conceptual models are constructed by the user, they are often erroneous and lead to issues when using the system. In order to help the user constructing correct conceptual models the principles

of signifiers, affordances, constraints, and mappings should be applied to compose a understandable perceived structure (Norman, 2013, pp. 26 ff.).

2.1.4. Ben Shneiderman

Shneiderman et al. (2018, p. 33) also highlight that the term "'user friendliness'" missed the goal of explaining the key attribute of effective interfaces. Instead a product should create a positive experience and "generate positive feelings of success, competence, and mastery" (Shneiderman et al., 2018, p. 33). They also accredit an important role to the context of use. While life-critical systems first and foremost should be highly reliable and effective, whereby a long learning curve is acceptable, systems intended for office, home and entertainment should yield ease of learning, low error rates, and subjective satisfaction (Shneiderman et al., 2018, p. 35). The key concept of "universal usability" (Shneiderman, 2000, p. 57) is to overcome three main challenges. One challenge represents the variety of different technology, while the others face the actual user. Meeting the needs of users with different skills, knowledge, age, gender, disabilities, culture, income, and many other characteristics as well as the gap between the users knowledge and the state of knowledge needed, represent the other two big challenges (Shneiderman, 2000). This concept is also included in *Shneiderman's Eight Golden Rules of Interface Design*, which also list consistency, shortcuts, informative feedback, and other design guidelines (Shneiderman et al., 2018, pp. 95 ff. Harvey & Stanton, 2013, p. 23).

2.1.5. Usability in International Standards

As Part 11 of ISO 9241 defines, there are several criteria through which usability is determined. Burmester et al. (2008) consider the different aspects of the quotation at the beginning of this section in detail with regard to the specific usage scenario inside a car. Therefore the term product or also called interactive system, derived from DIN EN ISO 13407 which is replaced by part 210 of DIN EN ISO 9241 since January 2011 (International Organization for Standardization, 2011a) and reissued with part 210 of DIN EN ISO 9241 in July 2019 (International Organization for Standardization, 2019b), is defined as a "combination of hardware and/or software and/or services that receives input from, and communicates output to, users" (International Organization for Standardization, 2019b, p. 6). Figure 2.2 shows an overview of the relationship of major sources of guidance for user-system interaction from ISO 9241-110 (International Organization for Standardization, 2019b, p. 7) According to the definition, this includes thinking about the intended usage of computers, any input device, like keyboard and mouse or buttons

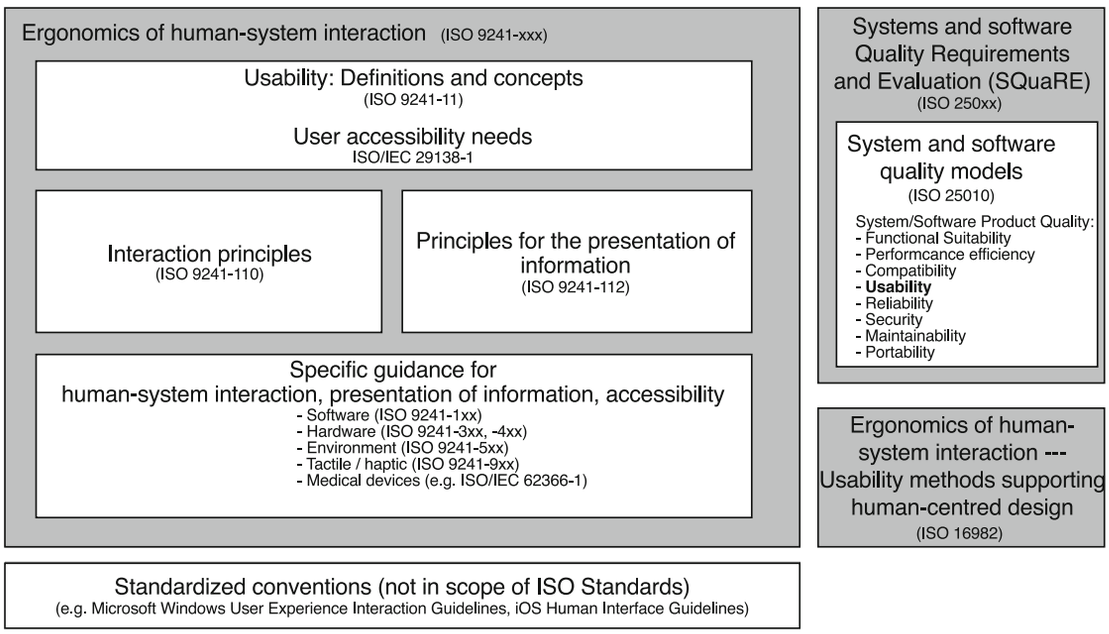


Figure 2.2.: Overview of ISO Standards on Usability. The figure is based on ISO 9241-110 (International Organization for Standardization, 2019b, p. 7).

and knobs, and output devices, like display or loudspeakers, just as virtual control elements (Burmester et al., 2008). The goal or defined as "intended outcome" (International Organization for Standardization, 2019b, p. 6) is often specified from outside in a work environment. In a leisure situation the user mostly defines his own goals — e.g. setting the air-condition to a comfortable temperature (Burmester et al., 2008). When it comes to effectiveness the International Organization for Standardization (2019b) means accuracy and completeness, which first has to be possible through the features of the system and can also be affected by an abstruse user interface (Burmester et al., 2008). Whereas efficiency describes the relation between spent resources and the degree of effectiveness. Measures for efficiency can thereby be the time spent for a specific task, or the number of interaction steps. ISO 9241 describes satisfaction as "freedom from discomfort and positive attitudes towards the use of the product" (International Organization for Standardization, 2011a, p. 7) which was changed in the second edition into the "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (International Organization for Standardization, 2019b, p. 3). This is usually achieved by reaching the goals in an effective and efficient way (Burmester et al., 2008). With the last aspect, the context of use, ISO 9241 addresses one of the main misconceptions when it comes to usability. As Burmester et al. (2008) cite Bevan (1995) "... there is no such thing as a 'usable product' or 'unusable product'. For instance a product which is unusable

by inexperienced users may be quite usable by trained users" (Bevan, 1995, p. 352). Therefore the usability of a product can only be determined by the specific context of use. This includes the technical, physical, social, and organizational environment in which a task with the system is performed (Bevan, 1995; Burmester et al., 2008).

Bevan (2001) summarizes standardization issues for the specific topic of human-computer interaction. Besides the already mentioned definition in the series of standards DIN EN ISO 9241, he also cites ISO/IEC 9126, which defines usability as "a set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users" (International Organization for Standardization, 1991, p. 3). In the first part of the replacing ISO/IEC 9126-1, there is a slightly different definition of usability: "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions"(International Organization for Standardization, 2001, p. 9). Since ISO/IEC 9126 got replaced by ISO/IEC 25010 it is necessary to mention the so called "quality in use model" (International Organization for Standardization, 2011b, p. 3), which extends the model from ISO/IEC 9126-1 and determines quality in use through effectiveness, efficiency, satisfaction, freedom from risk, and context coverage (International Organization for Standardization, 2011b).

2.2. Usability Factors

Considering the different approaches for defining usability in the previous section 2.1, it seems reasonable to suppose that a universal definition of usability cannot be achieved. Harvey (2011) argue that the issue of context is too important to create a single definition of usability. While Nielsen (1993) and Noel et al. (2005) highlight the importance of memorability as an important factor of usability, Harvey (2011) refer to the specific context of use. The standard ISO/IEC 9126 describes three different factors determining the usability of a product — understandability, learnability, and operability (International Organization for Standardization, 1991). However, Nielsen (1993) lists learnability, efficiency, user retention over time, error rate, and satisfaction as factors determining the usability of a product. Furthermore, Bevan (2001) states that "a product has no intrinsic usability, only a capability to be used in a particular context" (Bevan, 2001, p. 537). Analyzing different definitions in the literature, Harvey (2011) derive the following list of high-level usability factors:

- Effectiveness
- Efficiency
- Satisfaction
- Learnability

- Memorability
- Flexibility
- Perceived usefulness
- Task match
- Task characteristics
- User criteria

These factors serve as indications that need to be reflected to the specific context of use of a particular system. Therefore, Harvey (2011) designate them as "loosely defined" (Harvey, 2011, p. 28) starting point to derive key performance indicators (KPIs). The already mentioned context of use, specific for the interaction with in-vehicle information systems, is discussed in detail in section 2.5.1.

2.3. Human-Centered Design Process

As the several definitions of usability in section 2.1 lift out, the user is the central element when it comes to the design of products. To integrate the users needs as early as possible, the design of a system has to follow a defined process. This design process, illustrated in Figure 2.3, is divided in two phases with two main steps in each phase. The first phase focuses on generating an understanding of the problem, while the second phase represents the development of solutions. Within one of these phases the first main step is to explore different alternatives, which is followed by the second step of consolidation. These steps itself follow the user-centered design process (Butz & Krüger, 2017).

The user-centered design process is not strictly linear. As Figure 2.4 shows, it is rather intended as an iterative process which handles switches between different phases. In contrast to the ISO 13407, ISO 9241 part 210 introduces the target group of stakeholders which are are not necessarily considered as users. Therefore, ISO 9241 uses the term human-centered design rather than user-centered design. According to the International Organization for Standardization (2019b) the human-centered design process should include four design activities.

At first one has to gain a profound understanding of the context in which the system is used. This is influenced by different characteristics of the users and their tasks, as well as the environment in which these are performed and the available resources. This phase of the cycle serves to understand the context of use by gathering information on the current usage situation and analysis of existing or similar systems (International Organization for Standardization, 2019b). In a next step the user requirements need to be specified. Therefore these requirements should be created as explicit statements in relation to the before specified context of use. These should also include requirements

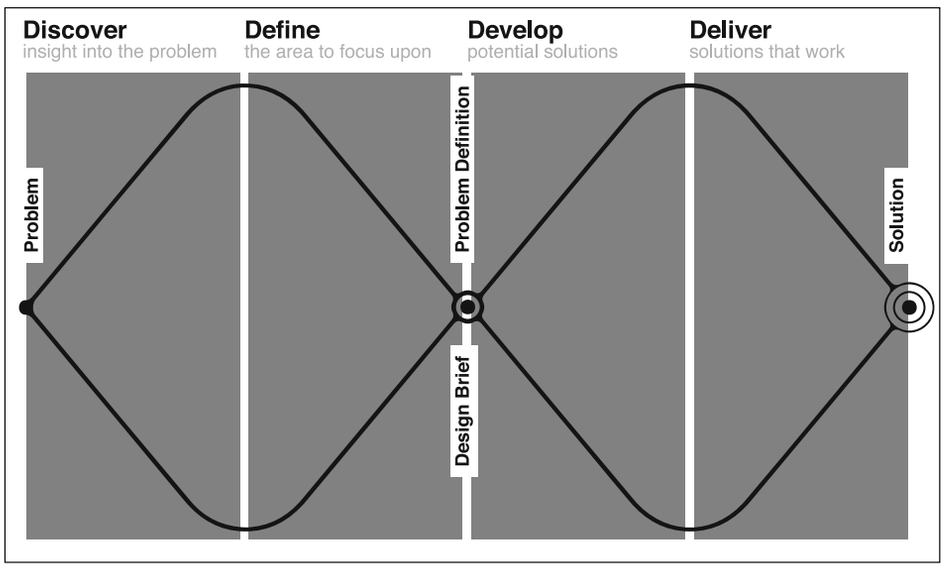


Figure 2.3.: The double diamond design process. The figure is based on Design Council UK (2005).

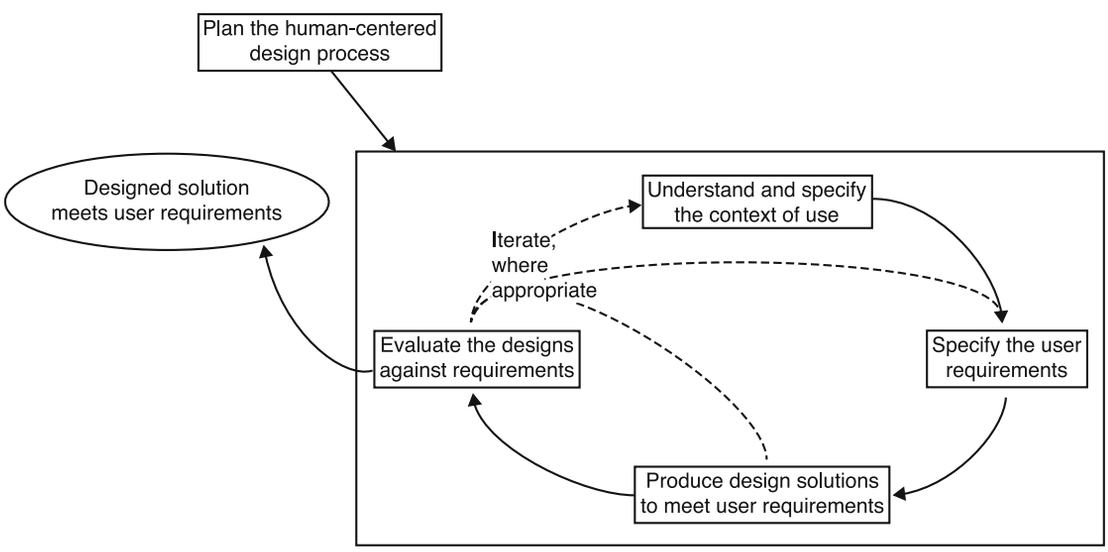


Figure 2.4.: The human-centered design process. The figure is based on ISO9241-210 (International Organization for Standardization, 2019b).

derived from user needs and the context of use, requirements arising from knowledge, standards and guidelines as well as usability requirements, and those derived from organizational requirements (International Organization for Standardization, 2019b).

During the production of design solutions the user tasks, the user-system interaction, and the user interface are designed to meet the previously specified requirements (International Organization for Standardization, 2019b). Part 110 of the standards series ISO 9241 describes a set of principles which should be taken into account. The system should be suitable for the task, self-descriptive, and conform with user expectations as well as controllable and suitable for learning. Furthermore, a system should be tolerant towards errors and suitable for individualization (International Organization for Standardization, 2008). In a newer draft of ISO 9241-110, the principle of individualization was merged into the principle of controllability (International Organization for Standardization, 2019a). In the production phase on the one hand the results received from evaluation activities or feedback should be considered. On the other hand the produced design solutions has to be communicated to those responsible for implementation through documentation, prototypes or exchange via experts in human-centered design.

The last phase in the cycle of human-centered design focuses on the evaluation of the design solutions created in the previous step. ISO 9241 states that "even at the earliest stages in the project, design concepts should be evaluated to obtain a better understanding of user needs" (International Organization for Standardization, 2019b, p. 20). Therefore it is not always necessary to involve users in the evaluation process. The used methods should rather be adjusted to the specific circumstances. The standard differentiates two main approaches in user-centered evaluation, the user-based testing, and the inspection-based evaluation. While inspection-based techniques are ideally performed by usability experts based on usability and accessibility guidelines, user-based testing directly involves users performing tasks with the system. Against the popular opinion these tests can be undertaken at any stage of the project. In earlier stages the product can be evaluated using models, scenarios or sketches. The more the product advances, the more functional and visual prototypes can be used. In later stages the product can also be evaluated using field validation in real environments. According to the International Organization for Standardization (2019b) this phase also includes follow-up evaluation where performance and usage data is collected and analyzed regarding the user needs and requirements.

2.4. Evaluation Methods

Since this thesis investigates the application of evaluation methods for user interfaces in the specific context of use of driver information displays, the evaluation step inside the user-centered design process is of particular interest. Part 210 of ISO 9241 differentiates the two main categories inspection-based and user-based methods (International Organization for Standardization, 2019b), whereas ISO 16982 distinguishes methods with and without direct involvement of users (International Organization for Standardization, 2002). Furthermore, the standard lists the following methods supporting human-centered design (International Organization for Standardization, 2002, p. 5):

- Observation of users
- Performance-related measurements
- Critical incidents analysis
- Questionnaires
- Interviews
- Thinking aloud
- Collaborative design and evaluation
- Creativity methods
- Document-based methods
- Model-based approaches
- Expert evaluation
- Automated evaluation

Comparing different textbooks in the domain of HCI regarding the categorization of evaluation methods for user-centered design, one would register some similarities. When looking at the preparation of research, Shneiderman et al. (2018, p. 168) summarize to consider the following characteristics:

- Stage of design (early, middle, late)
- Novelty of project (well defined versus exploratory)
- Number of expected users
- Criticality of the interface (for example, life-critical medical system versus museum-exhibit support)
- Costs of product and finances allocated for testing
- Time available
- Experience of the design and evaluation team
- Environment where interface is used

Lazar et al. (2017, p. 27) provide a more abstract categorization regarding the type of behavioral research with typical methods assigned. For descriptive investigations, where the focus is on describing a situation or a set of events, typical methods would be observations, field studies, focus groups, and interviews. When the focus is on the identification of relations between multiple variables, observations, field studies, and surveys are typically used. In experimental research, where one wants to identify causes of a situation or a set of events, the suggested method would be controlled experiments.

While Shneiderman et al. (2018, Chapter 5) distinguish between *Expert Reviews* and *Usability Testing* in laboratories, Lazar et al. (2017, pp. 267 ff.) list three categories of usability testing: "expert-based testing, automated testing, and user based-testing". Kuhn (2000) distinguishes two "lines of practice" (Kuhn, 2000, p. 309), the user-driven design that is characterized through direct involvement of users in the design process, and the user-informed design using field studies or interviews to gather information about the context of use. Butz and Krüger (2017) differentiate three dimensions of evaluation methods. They distinguish between formative versus summative, quantitative versus qualitative, and analytical versus empirical methods. While formative methods appear before the decision process, summative methods summarize the result of the development process. Quantitative methods lead to numerical results, which can be task times or error rates, whereas qualitative methods deliver statements focusing on detailed aspects rather than quantifiable results. In the last dimension analytical methods investigate a system through analysis of the same, while empirical methods focus on the results rather than where they come from (Butz & Krüger, 2017). The following section gives a brief overview of different evaluation methods, taking multiple sources into account.

2.4.1. Surveys

According to Shneiderman et al. (2018, p. 187) surveys constitute a good companion for usability tests and expert reviews, as the often large number of responses appear powerful when compared with the relatively small number of participants used for usability tests. It may seem easy to put some items together and run a survey for a specific question, but it should be well planned. It is appropriate to get your survey form reviewed by colleagues and test it with a small number of participants. Also the statistical analysis should be prepared before the survey is distributed, as "direct activities are more successful than unplanned statistics-gathering expeditions" (Shneiderman et al., 2018, p. 187).

Moreover, Lazar et al. (2017, p. 105) state that surveys are commonly used although they are not the most appropriate method, but the easiest. Shneiderman et al. (2018, p. 187) and Lazar et al. (2017, p. 105) also warn of using a poor prepared survey and distributing it to an inappropriate sample, which leads to erroneous results. Unlike in interviews or focus groups there usually is no investigator present, which eliminates the chance of follow-up questions or more detailed questions. Another drawback Lazar et al. (2017, p. 106) draw attention to, is the possibility of biased data, as for example the answers to past behavior can be underestimated or overestimated by different users. Another key challenge for surveys is to define the sample and thus the potential survey respondents. As Lazar et al. (2017, p. 113) state, the goal in HCI research is not to achieve

a population estimate through probabilistic sampling. Instead users are recruited in a nonprobabilistic manner, because often there is not a well-defined population or a list of people who meet a specific criterion.

There exists a variety of different survey tools and standardized usability questionnaires, as shown in Table 2.1. According to Sauro and Lewis (2016) citing Nunnally (1978) standardized measurements offer many advantages. Besides objectivity through the possibility of independently verifying measurements of other researchers, studies are easier to replicate, when using standardized questionnaires. Another advantage is quantification and the application of statistical methods to get a better understanding of the results. Already developed standardized questionnaires can be reused, therefore the expenses of designing a specific questionnaire for every study can be omitted. It is also easier to communicate, when using standardized measurements, and it promotes the generalization of results (Nunnally, 1978; Sauro & Lewis, 2016).

ASQ, PSQ, PSSUQ, and CSUQ

For example J. R. Lewis (1995) introduces four different questionnaires to measure user satisfaction. The After-Scenario Questionnaire (ASQ) as the title suggests is designed to measure satisfaction after every single task in a scenario-based usability study and is quite similar to the earlier version, the Printer-Scenario Questionnaire (PSQ). The Post-Study System Usability Questionnaire (PSSUQ) and the Computer System Usability Questionnaire (CSUQ) are both longer than the previously mentioned questionnaires and are intended to be completed once, at the end of a usability study. As PSSUQ turned out to have limited generalizability it was slightly revised, which resulted in the CSUQ. Both consist of 19 items with the subscales *System Usefulness*, *Information Quality*, and *Interface Quality* (J. R. Lewis, 1995).

QUIS

The Questionnaire for User Interaction Satisfaction (QUIS) is a post-study questionnaire developed by researchers from the Human-Computer Interaction Lab at the University of Maryland at College Park (Sauro & Lewis, 2016). It contains 27 items in the five subscales overall reactions to the software, screen, terminology and system information, learning, and system capabilities (Chin et al., 1988). In their study Chin et al. (1988) compared familiar software products — a liked and disliked software, a command line system, and a menu driven application — with 150 participants. The questionnaire has been shown as reliable (Cronbach's $\alpha = .94$) and a factor analysis revealed a correspondence between the section learning and terminology with the latent factors, whereas the section of system capabilities divide in two different factors (Chin et al., 1988).

SUMI

Another questionnaire of the category of post-study questionnaires is the Software Usability Measurement Inventory (SUMI) with five subscales for efficiency, affect, helpfulness, control, and learnability and 10 items each. The reliability of the SUMI was calculated with Cronbach's alpha ($\alpha = .94$) over 1000 completed questionnaires from 150

Questionnaire	Usage	Citations
After-Scenario Questionnaire (ASQ)	post-task	J. R. Lewis (1995)
Printer-Scenario Questionnaire (PSQ)	post-task	J. R. Lewis (1995)
Post-Study System Usability Questionnaire (PSSUQ)	post-study	J. R. Lewis (1995)
Computer System Usability Questionnaire (CSUQ)	post-study	J. R. Lewis (1995)
Questionnaire for User Interaction Satisfaction (QUIS)	post-study	Chin et al. (1988)
Software Usability Measurement Inventory (SUMI)	post-study	Kirakowski and Corbett (1993) and McSweeney (1992)
Interface Consistency Testing Questionnaire (ICTQ)	post-study	Ozok and Salvendy (2001)
Purdue Usability Testing Questionnaire (PUTQ)	post-study	Lin et al. (1997)
System Usability Scale (SUS)	post-study	Brooke (1996)
User Experience Questionnaire (UEQ)	post-study	Laugwitz et al. (2006)
AttrakDiff 2	post-study	Hassenzahl et al. (2003, 2008)
Modular evaluation of key Components of User Experience (meCUE)	post-study	Minge et al. (2017)
Questionnaire for the Subjective consequences of Intuitive use (QUESI)	post-study	Naumann and Hurtienne (2010)
INTUI	post-study	Ullrich and Diefenbach (2010)
Usability Metric for User Experience (UMUX)	post-study	Finstad (2010b)
Usability Metric for User Experience (UMUX)-LITE	post-study	J. R. Lewis et al. (2013)

Table 2.1.: Existing survey tools and standardized usability questionnaires.

systems. An additional database of results allows to compare the own results with those of similar products and tasks (Kirakowski & Corbett, 1993; Sauro & Lewis, 2016).

SUS

The System Usability Scale (SUS) was developed by Brooke (1996) to satisfy the need for "'quick and dirty' methods to allow low cost assessments of usability" (Brooke, 1996, p. 189). Despite this harsh self-description a study of a collection of usability studies shows that the SUS constituted 43 % of post-test questionnaire usage (Sauro & Lewis, 2009). The questionnaire was developed through selection of 10 out of 50 potential items on a 5-point scale according to the responses of 20 test persons. The selected items describe those with the most extreme responses and therefore the strongest discrimination between the investigated systems. Bangor et al. (2008) investigated data of the SUS of 2324 individual surveys and proved the reliability with Cronbach's alpha, $\alpha = .91$, as well as J. R. Lewis and Sauro (2009) with a value of $\alpha = .92$ using 324 cases.

AttrakDiff2

A questionnaire which goes beyond measuring effective and efficient task completion is the AttrakDiff2. Besides pragmatic product quality it addresses aspects further than usefulness and usability, which measures if the users need for stimulation and identity is fulfilled. Another aspect measured is the attractiveness and thus the global positive-negative evaluation of a product. The questionnaire is divided in the dimensions pragmatic quality (PQ), hedonic quality – stimulation (HQ-S), hedonic quality – identity (HQ-I), Attractiveness (ATT) with 28 items in total. The selected items represent pairs of adjectives, selected out of 133 items developed through an expert workshop and a pilot study including further items of the original version, the AttrakDiff1 (cf. Burmester et al., 2002; Hassenzahl et al., 2000; Hassenzahl, 2001). With a second study Hassenzahl et al. (2003) measured the consistency within (α(HQ-S) = .76–.90; α(HQ-I) = .73–.83; α(PQ) = .83–.85) and between the different scales, which suggests construct validity (Hassenzahl et al., 2003, 2008).

meCUE

Minge et al. (2017) introduced the questionnaire called modular evaluation of key Components of User Experience (meCUE), based on a UX framework, the Components of User Experience (CUE) (Thüring & Mahlke, 2007). The CUE model concepts are integrated into the structure of the meCUE as four different modules. The first module consists of items regarding the product perceptions separated into instrumental product perceptions including usefulness and usability and non-instrumental product perceptions including visual aesthetics, status, and commitment. The items of the second module of user emotions are separated into positive and negative emotions, while the third module consequences of usage includes items regarding the product loyalty as well as the intention to use. The last module added was the single-item of global attractiveness. In order to determine convergent validity, the meCUE was compared with subscales of different validated questionnaires like the AttrakDiff and the User Experience Questionnaire

(UEQ) for instrumental product perceptions and global attractiveness, the VisAWI for visual aesthetics, and the PANAS for emotions (Minge et al., 2017).

QUESI The Questionnaire for the Subjective consequences of Intuitive use (QUESI) was developed to "assess the subjective consequences of intuitive use" (Naumann & Hurtienne, 2010, p. 401). Based on the conclusion that *intuitivity* is a property of human information processes rather than a product feature, the following definition for intuitive use is used as a starting point for the construction of the questionnaire:

> "A technical system is, in the context of a certain task, intuitively usable while the particular user is able to interact effectively, not-consciously using previous knowledge." (Naumann et al., 2007, p. 129)

Mohs et al. (2006) derive criteria for the assessment of intuitive use from this definition that include low subjective mental workload, high perceived achievement of goals, low perceived effort of learning, high familiarity, and low perceived error rate. These serve as subscales for the 14 items of the questionnaire which shows acceptable reliability measured by Cronbach's alpha with $\alpha = .90$ for the overall questionnaire and values between $\alpha = .78$ and $\alpha = .92$ for the different subscales.

INTUI Ullrich and Diefenbach (2010) performed a literature review to collect characteristics of intuitive interaction. From this review the authors derived the five components of intuitive decision making: effortlessness, attention, gut feeling, verbalizability, and magical experience. The 32 items gathered through a workshop formulated as paired contradictory terms were applied in a pilot study in order to perform a factor analysis and to decrease the number of items. As a result the components effortlessness and attention turned out as too similar, therefore attention was removed as a separate scale from the final set of components. The number of items was reduced to 16 for the final INTUI questionnaire which was tested for reliability with satisfying results (*Effortlessness*: $\alpha = .94$; *Gut Feeling*: $\alpha = .68$; *Verbalizability*: $\alpha = .72$; *Magical Experience*: $\alpha = .79$) (Ullrich & Diefenbach, 2010).

UMUX The Usability Metric for User Experience (UMUX) was developed at Intel based on a deconstruction of the SUS. While investigations show that seven-point Likert scales outperform five-point Likert scales in terms of reliability, accuracy, and ease of use (Cox III, 1980; Diefenbach et al., 1993), respondents more likely provide non-integer interpolations in five-point scales than seven-point scales (Finstad, 2010a). Furthermore, Finstad (2010b) aims to more closely conform the survey instrument to the definition of usability used in ISO 9241-11 (International Organization for Standardization, 1998). The resulting questionnaire contains four items addressing the components effectiveness, satisfaction, efficiency, and overall usability on a seven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" (Finstad, 2010b). In a survey study Finstad (2010b)

measured internal reliability through Cronbach's alpha with $\alpha = .94$ for the UMUX and $\alpha = .97$ for the SUS. In further investigations, comparing UMUX with SUS, a significant overall correlation of $r = .96, p < .001$ as well as quite similar results of t-tests with a correlation of $r = .89, p < .01$ could be shown. In a later study J. R. Lewis et al. (2013) introduced a two-item questionnaire based on the UMUX, the UMUX-LITE. Although the investigation could deliver adequate results when compared with the SUS, the authors don't recommend the sole use of UMUX-LITE (J. R. Lewis et al., 2013).

> "Many aspects of usability can best be studied by simply asking users. . . . Questionnaires and interviews are also useful methods for studying how users use systems and what features they particularly like or dislike." (Nielsen, 1993, p. 209)

As Jakob Nielsen summarizes, questionnaires are a useful method to gain insight into user satisfaction and also possible pains with interactive systems. Later in the chapter, where the above quotation is taken from, he also addresses the similarities of questionnaires and interviews "since both involve asking users a set of questions and recording their answers" (Nielsen, 1993, p. 210), as well as the differences and key advantages and disadvantages of each method. While interviews are discussed in detail in section 2.4.2, questionnaires suit especially for asking a large group of users with mostly closed questions, as many users don't feel comfortable with answering open-ended questions or may write hard to interpret statements (Nielsen, 1993).

2.4.2. Interviews and Focus Groups

As already mentioned in section 2.4.1, questionnaires and interviews are quite similar, according to Nielsen (1993). Unlike surveys, interviews suit well for exploratory studies and offer possibilities for follow-up questions or further explanation, if a question is hard to understand. Interviews are also more flexible, as the questions can be adapted to the situation, but often require more resources, especially for large groups of users (Nielsen, 1993). Thus, conducting interviews can be quite cost-intensive and time-consuming, but often provides specific and constructive suggestions (Shneiderman et al., 2018, p. 194).

Lazar et al. (2017, p. 187) outline the key difference through the results of both methods, as "surveys often end up being broad but not deep". Whereas the approach of interviews and focus groups is to go deep but not broad through direct conversations with fewer participants and the flexibility regarding structure and the level of detail. On the other hand an interviewer is faced with the challenges of conducting an interview, e.g. "managing potentially unbounded discussions . . . , listening carefully, taking notes, trying to decide which comments to pursue with further questions, and trying to

understand nonverbal reactions" (Lazar et al., 2017, p. 188). Besides the demanding phase of data collection, interviews with many open-ended questions are much more difficult to analyze compared to surveys with many closed questions, additionally using a particular software. These limitations are also reflecting in the number of participants, as interviews with many participants need much more resources and effort than a survey (Lazar et al., 2017, p. 188).

Another challenge the interview method as well as surveys are facing regards the data collection itself, as it is often decoupled from the investigated system and the tasks a user has to perform. Participants in both methods report experiences from their memory, which can lead to different answers compared to those from users actually using the system in the designated context of use (Lazar et al., 2017, pp. 200 ff.). Through this also before mentioned flexibility the interview method offers a variety of properties to refine the study design.

When planning an interview, for example the selection of participants plays an important role. Besides different groups of users, possible interviewees might also include project stakeholders or specific experts and key informants, which are also invited repeatedly (Lazar et al., 2017, pp. 196 ff.). Another property to consider is the interview strategy. While fully structured interviews do not offer the possibility for questions apart from the predefined script the fact that every participant is asked exactly the same question facilitates the data analysis. In semistructured interviews, the interviewer also sticks to a prepared script, but this strategy enables discussions with the participant and the ability of follow-up questions to participants comments. Whereas an unstructured interview is often only based on a list of topics or questions to avoid the discussion from slowing down or stalling. Lazar et al. (2017, p. 199) state that "interviews with less structure require more skill to conduct", while more structure leads to more comparable but less deep insights.

Besides the level of structure interviews can also differ in focusing. This describes the way of tying the interview to a specific context of a problem or technology. A special form of focused interview is the contextual inquiry. Instead of asking a participant to envision or remember a specific interface, the contextual inquiry aims to observe the user how he would actually use the system. During this observation the participant would describe his work and the individual steps taken to reach his goal including his intention behind these steps. Through discussion with the interviewee the interviewer obtains a profound insight into the user's understanding. Participants can reveal potential for optimization unconsciously by mentioning steps they haven't thought about when only describing their work from memory. As in the example from Davis and Rebelsky (2007) where students are asked to describe how to make a peanut butter and jelly sandwich

and most participants would forget the step of removing the foil seal under the lid of the unopened jar (Beyer & Holtzblatt, 1998; Holtzblatt & Beyer, 2017; Lazar et al., 2017).

An alternative to the contextual inquiry, where the participant provides the context, is to use external aids to inspire the participant. These aids can range from non-technical objects, like a set of photos to ask the participant about his approach when organizing photos to elicit feedback and promote engagement, to functional prototypes in form of paper prototypes or even high-fidelity software prototypes "to explore possibilities and understand needs and practices regarding technology use" (Lazar et al., 2017, p. 203).

Interview techniques can also be applied for a group of users at the same time. This allows insights into a broad range of opinions. As Fern (1983, p. 121) cites Davidson (1975): "A focused group interview is a qualitative tool for collecting information in which a number of respondents simultaneously discuss a given topic under the guidance or [*sic*] a moderator." Furthermore the often interactive and balancing discussion between the subjects can spawn issues that might not been identified during a one-to-one interview (Lazar et al., 2017, p. 204; Shneiderman et al., 2018, p. 144) and can also lead to more representative opinions (Kuhn, 2000). But this group dynamics can also lead to monopolies of individual viewpoints or even confrontations between the participants. As an interviewer one should take care of this and also consider for the selection of focus group participants (Lazar et al., 2017, p. 205). As Nielsen (1993) states focus groups can be applied in very early phases of design to assess user needs, as well as when the product has been in use for some time to gather feedback about it.

2.4.3. Expert-Based Testing

Whereas questionnaires are completed by potential users and interviews as well as focus groups are hold with actual users, expert-based testing completely omits the actual user perspective and focuses on structured inspections by experts in interface design. The goal is to identify obvious interface flaws from the interface design perspective, rather than finding deeper, task-related interface flaws (Lazar et al., 2017, p. 268). Shneiderman et al. (2018, pp. 171 ff.) list several methods from the category of expert-reviews.

Heuristic Evaluation

The method of heuristic evaluation (HE) developed by Nielsen and Molich (1990) aims to find usability problems through inspections by multiple evaluators. The individual evaluators therefore inspect the interface each by themselves to reduce bias and often according to certain rules. For example Nielsen (1990c, 1993) himself introduces ten principles that can be used for heuristic evaluation (Nielsen, 1993, p. 20):

- Simple and natural dialogue
- Speak the users' language

- Minimize the users' memory load
- Consistency
- Feedback
- Clearly marked exits
- Shortcuts
- Good error messages
- Prevent errors
- Help and documentation

He also performed an analysis to compare different sets of usability principles and identify those with the broadest explanatory coverage according to actual usability problems. From this analysis he derives two top ten lists of usability heuristics, one that covers all usability problems in the investigated database and on that covers the serious usability problems reported (Nielsen, 1994a). Another set of heuristics for interface design is the list of 8 Golden Rules of Interface Design (Shneiderman et al., 2018, pp. 95 ff.):

- Strive for consistency
- Cater to universal usability
- Offer informative feedback
- Design dialogs to yield closure
- Prevent errors
- Permit easy reversal of actions
- Support internal locus of control
- Reduce short-term memory load

Besides these proposed principles, Nielsen (1994a) investigated several usability heuristics (Apple Computer, 1992; Carroll & Rosson, 1992; Holcomb & Tharp, 1989; Holcomb & Tharp, 1991; Polson & Lewis, 1990; Smith et al., 1982) in a factor analysis in order to determine what heuristics best explain actual usability problems from a database containing usability problems from a variety of projects. The highest coverage of usability problems was provided by a set of nine heuristics: visibility of system status, match between system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, and helping users recognize, diagnose, and recover from errors (Nielsen, 1994a).

To conduct a session of a HE, an evaluator would go through the system several times and test the dialog elements against the predefined heuristics. The evaluator is thereby not limited to the predefined set of heuristics, rather he can raise violations of additional usability principles that may be relevant for the specific system or dialog element. Based

on the domain knowledge of the evaluators they can be assisted by domain experts or supplied with a typical usage scenario. The result of a HE is a list of usability problems regarding the violated usability principle. Thereby the evaluators do not necessarily provide solutions to fix the mentioned problems, rather they point out more obvious usability problems (Nielsen, 1993).

Guideline Review

Guideline reviews are similar to heuristic evaluation, as the reviewers also check the conformance of a system with specific guidelines. Compared to heuristics, which contain around ten items, organizational or other guideline documents are often a much larger collection of design instructions and recommendations. Therefore a guideline review might take more time than a heuristic evaluation for mastering the guidelines as well as the review itself (Shneiderman et al., 2018, p. 172).

The ISO 16982 standard lists the category of document-based methods that incorporate similar methods to guideline review. Document-based analysis describes the procedure of an usability specialist using existing documents like checklists to evaluate an interface. The available documents for testing typically include style guides from the software provider or the company the software is used in, as well as handbooks containing ergonomic recommendations or standards (International Organization for Standardization, 2002).

Consistency Inspection

Another more detailed inspection method is the consistency inspection, where experts check the consistency of a system or interface inside a family of interfaces. Tested topics therefore include consistency of terminology, fonts, color schemes, layout, and input and output formats. Investigated objects in a consistency inspection go beyond the interface itself, as documentation, training material, and online help also have to be consistent across the investigated family (Lazar et al., 2017, p. 269; Shneiderman et al., 2018, p. 172).

Grudin (1989) differentiates between three different types of consistency to consider. The internal consistency with a user interface itself can be determined by graphical layout, naming and wording, dialog forms, as well as several other factors, including domain specific dimensions. Another type is the external consistency of an interface with features of other interfaces. This includes design patterns the user is already familiar with from other interfaces. In some situations these two types of consistency can be in competition with each other. The third type of consistency — the "*correspondence of interface features to familiar features of the world beyond computing*" (Grudin, 1989, p. 1165) — describes the usage of metaphors or analogies to real world object inside an interface.

Cognitive Walkthrough

The method of cognitive walkthrough (CW) focuses on the ease of learning of investigated interfaces (Wharton et al., 1994). C. H. Lewis et al. (1990) initially introduced a methodology based on the *CE+* theory of exploratory learning (Polson & Lewis, 1990) to generate a list of questions regarding the user interface of an interactive system. The

questions focus on aspects of an interface that support the problem-solving and learning process. Positive answers therefore indicate that the interface will be easily learned (C. H. Lewis et al., 1990).

Besides the investigation of learning an interface by exploratory browsing, the CW method also suits for interfaces that need training. During an inspection the evaluator simulates a user going through the system to solve a specific task (Shneiderman et al., 2018, p. 172). Wharton et al. (1994) propose the following four questions, the evaluator should ask with respect to the evaluated interface (Wharton et al., 1994, p. 9):

- Will the user try to achieve the right effect?
- Will the user notice that the correct action is available?
- Will the user associate the correct action with the effect they are trying to achieve?
- If the correct action is performed, will the user see that progress is being made toward solution of their task?

Before the actual investigation there are some conditions to commit to. These include a definition of the user, e.g. containing his background and level of knowledge, the tasks that should be analyzed as well as the single actions needed to accomplish each task. Last but not least a definition of the investigated interface is needed, e.g. an implemented prototype or even a document-based description in earlier stages of development (Wharton et al., 1994).

In most cases a CW session needs to be recorded in some way (e.g. recording on video, electronically, or using flip charts) to capture the critical information for an interface. While the focus of a CW is on identifying problems with an interface, the context and therefore the explanation of a problem can also support the identification of solutions to fix the problem (Wharton et al., 1994).

Metaphors of Human Thinking

With their method of metaphors of human thinking (MOT) Frøkjær and Hornbæk (2002) present five metaphors capturing several aspects of human thinking regarding consistency and information scent. Besides the usage as traits when designing user interfaces, the application for evaluation of user interfaces is also suggested. Table 2.2 summarizes the metaphors used for the technique, as well as their implications for user interfaces including some key questions and examples. These key questions should be considered during the inspection of a user interface.

> "Deciding between these techniques requires careful consideration of the goals of the evaluation, the kinds of insights sought, and the resources available. We believe that heuristic evaluation and usability testing draw much of their strength from the skilled UI professionals who use them." (Jeffries et al., 1991, p. 124)

Metaphor of Human Thinking	Implications for User Interfaces	Key Questions/Examples
Habit formation is like a landscape eroded by water.	Support of existing habits and, when necessary, development of new ones.	Are existing habits supported? Can effective new habits be developed? Is the interface predictable?
Thinking as a stream of thought.	Users' thinking should be supported by recognizability, stability and continuity.	Do the system make visible and easily accessible the important task objects and actions? Does the user interface make the system transparent or is attention drawn to non-task related information? Does the system help users to resume interrupted tasks? Is the appearance and content of the system similar to the situation when it was last used?
Awareness as a jumping octopus.	Support users' associations with effective means of focusing within a stable context.	Do users associate interface elements with the actions and objects they represent? Can words in the interface be expected to create useful associations for the user? Is the graphical layout and organization helping the user to group tasks?
Utterances as splashes over water.	Support changing and incomplete utterances.	Are alternative ways of expressing the same information available? Are system interpretations of user input made clear? Does the system make a wider interpretation of user input than the user intends or is aware of?
Knowing as a site of buildings.	Users should not have to rely on complete or accurate knowledge—design for incompleteness.	Can the system be used without knowing every detail of it? Do more complex tasks build on the knowledge users have acquired from simpler tasks? Is feedback given to ensure correct interpretations?

Table 2.2.: Summary of the MOT-Technique. The Five Metaphors, Their Implications for User Interfaces, and Examples of Questions to be Asked During Usability Inspection (Frøkjær & Hornbæk, 2008).

As Jeffries et al. (1991) state it is not a simple decision for one or another method. They compared four different techniques, among them heuristic evaluation, software guidelines, cognitive walkthrough, and usability testing. While the guidelines method and the cognitive walkthrough technique can be used by software engineers, heuristic evaluation as well as usability testing must be conducted by professionals in the domain of user interfaces. In their study they also discovered, that methods used by experts are not able to find all kinds of usability problems. For example problems that only occur accidentally through operations are hard to identify for expert-based usability methods (Jeffries et al., 1991). In a comparative study Frøkjær and Hornbæk (2008) compare MOT with HE, CW, and the TA technique, with satisfying results when it comes to problem detection.

Expert reviews can be conducted in several phases during the development process. They suit well to review consistency across interfaces or even multiple systems of a family (Shneiderman et al., 2018, pp. 173 ff.) as well as to find obvious interface flaws (Lazar et al., 2017, p. 268) and assess an interface regarding specific guidelines or heuristics. It helps a lot to place the expert in a similar situation and a realistic work environment during the review as the intended user would experience it (Shneiderman et al., 2018, p. 173). No matter how realistic the situation is during a review, expert reviews cannot replace testing with real users. As Lazar et al. (2017, p. 268) state:

> "Interface experts are experts in interfaces but they are typically not experts in the tasks to be performed within a certain interface. Conversely, representative users are typically experts in performing the tasks but are not experts in interface design."

2.4.4. User-based Testing

As Lazar et al. (2017, p. 271) describe: "user-based testing is what most people mean when they refer to usability testing". In contrast to testing with interface experts (see section 2.4.3, the representative users are experts in the specific domain the interface is designed for. Therefore users are able to find task-related issues in an interface, that interface experts may overlook (Lazar et al., 2017, p. 268). The term usability testing summarizes several types of methods for different stages of design and also different purposes of the test.

Mockups, Prototypes, and Simulations

In early project phases paper mockups or low-fidelity prototypes can be assessed regarding aspects like wording, layout, and sequencing. Therefore a moderator simulates the different interface screens and states through different pages. The subject is asked to perform typical tasks with the paper prototype. The later in the project, the fidelity of

these prototypes increases with the help of specific software tools (Shneiderman et al., 2018, pp. 148–149).

A special method for prototyping is the Wizard of Oz experiment. During a simulation, the user perceives the interaction as with an actual application, but in reality the user is interacting with the wizard, who provides the system responses (Lazar et al., 2017, p. 294). In order to find out about the application-specific linguistic characteristics for a natural language interface, Dahlbäck et al. (1993) suggest Wizard of Oz studies. Therefore they developed a simulation environment for the specific use case of natural language interfaces. Another application for the Wizard of Oz method is shown by Gould et al. (1983), who introduce an experiment for a "'listening typewriter'" (Gould et al., 1983, p. 295), a natural language input method for word processing.

Discount Usability Testing

Another approach to usability testing is described in the method of "discount usability engineering" (Nielsen, 1989b, 1990a, 1994b). Nielsen (1993) defines the term through the four principles *user and task observation*, *scenarios*, *simplified thinking aloud*, and *heuristic evaluation*. The method of he) is already presented in section 2.4.3. The technique of think aloud (TA) invites users to express what they are doing and thinking about the interface while performing the task. It comes in different variants, concurrent or retrospective to the actual test session (Shneiderman et al., 2018, p. 181), and can differ in complexity. With simplified thinking aloud Nielsen (1993) wants to delimit his lighter variant through simple note taking against more complex variants of the technique conducted by psychologists, with recordings on video tape and therefore time-consuming data analysis. When using the TA technique, it should be considered that the technique can interfere with a measurement of task times, as verbalizing their thoughts produces cognitive load for the users. Furthermore, it can cause spurious data in eye tracking experiments (Shneiderman et al., 2018, p. 182).

The principle of scenarios relates to the kind of prototypes used for usability testing. A scenario in this sense describes a previously planned path, the user can follow and therefore reduces the functionality and the number of features to a minimum. This allows frequent changes to the scenario as well as to the prototype with minimal effort and produces quick and frequent feedback from real users (Nielsen, 1993). The key argument in favor of discount usability engineering is to overcome the intimidating image of usability engineering, and at least "having *some* usability engineering work performed, even though the methods . . . may not always be the absolutely 'best' method and will not necessarily give perfect results" (Nielsen, 1993, p. 17).

Various Types of Usability Testing

Other variations belonging to the group of user-based testing are for example competitive usability testing, where a new interface is compared to previous versions or similar products from competitors, mostly with a within-subject study design, or A/B testing to

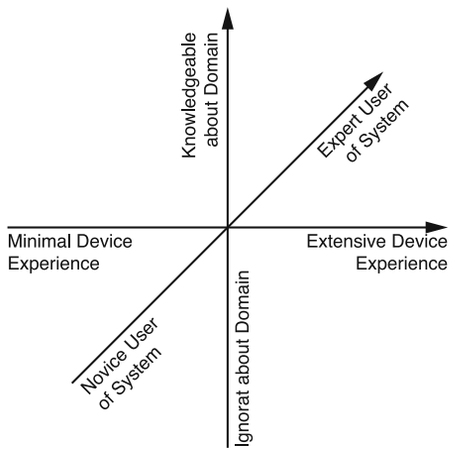


Figure 2.5.: Dimensions of users' experience. The figure is based on Nielsen (1993).

investigate differences between different interface designs. The latter uses a between-subjects study design with two groups and randomly assigned subjects. Further examples include field testing in realistic work environments or more naturalistic environments, or remote usability testing, where the subject and the moderator conduct the test and communicate mostly over web-based applications. On the one hand this widens the range of available participants and increases the possibilities to get users with diverse backgrounds, but on the other hand it limits control over user behavior as well as observation abilities (Shneiderman et al., 2018, pp. 182–186).

Since the movement toward usability testing, organizations and companies start to establish usability laboratories. A traditional usability lab consists of two rooms for the participant and the observer, separated by a one-way mirror (Shneiderman et al., 2018, pp. 175 ff.). Further equipment typically includes a computer or other device where the user performs the intended task, as well as microphones and cameras to record what the user is doing. In most cases the moderator uses several monitors to observe the user's screen and face as well as to control the experiment (Lazar et al., 2017, p. 276).

Categories of Users

Nielsen (1993) claims that an important aspect of usability engineering is to know the actual user. As an analysis of 92 comparisons of hypertext systems regarding which aspects have the largest usability effects shows, much effects are influenced by the individual differences between users (Nielsen, 1989a). Nielsen (1993) therefore differentiates three dimensions of users' experience, illustrated in Figure 2.5. Within this "'user cube'" (Nielsen, 1993, p. 43) these are: experience with the system, with the device (originally "with computers in general" (Nielsen, 1993, p. 43)), and with the task domain.

In traditional web usability the focus is typically on increasing the ease of learning. Usability tests therefore are often performed with novice users. As Nielsen (2000) states this focus in history switched from time to time between novice and expert users. In the late 1980s and early 1990s much research was conducted in the field of expert user performance, e.g. famous case studies for reconfiguration of command keys for systems of American telephone companies, or the design of the "'launch abort' button" (Nielsen, 2000) for the space program launch control center.

Sample Sizes for Usability Studies

When testing with users, during the development of the study design at some point, the question of *How many test users?* will arise. Whereas increasing the sample size for a survey study through an online questionnaire is entailed with relatively low incremental costs compared to the fixed costs, it takes much more resources for moderated usability studies. The fixed costs of a survey are typically allocated to tasks and material in advance of the actual study. Because of the different cost structure of usability studies, the consideration of sample sizes for this type of study is necessary (Sauro & Lewis, 2016).

Sauro and Lewis (2016) differentiate between summative and formative studies regarding sample size estimation. For summative evaluations they suggest an exemplary approach using the formula for the test statistic of the one-sample t-test. By determining the desired level of precision, as well as the level of confidence, the sample size estimate can by calculated through multiple iterations. Unlike summative studies, formative studies do not aim to obtain measurements, but rather have the goal of discovering problems that users have with an interface. For this type of studies Sauro and Lewis (2016) use a probabilistic model to estimate the sample sizes for different probabilities of the detection of a problem. Figure 2.6 shows, the estimated sample size in relation to the likelihood of detecting the problem at least once for different probabilities of problem occurrence. For example for a lowest probability of problem occurrence of interest of .25 and a minimum number of detections of 1 with a cumulative likelihood of discovery of 90 %, the estimated sample size is nine participants (Sauro & Lewis, 2016). Nielsen (1994b) suggests a number of participants between three and five for his approach of "discount usability engineering" to achieve a maximum benefit-cost ratio. In a previous article Nielsen and Landauer (1993) developed a mathematical model to plan the amount of evaluation needed with respect to cost/benefit ratios by analyzing different case studies. According to their results they suggest 20 test users for large projects, 15 test users for medium-large projects, and seven test users for small projects (Nielsen & Landauer, 1993). In an online article Nielsen (2012) also emphasizes to test with five users referring to the approach of "discount usability engineering" as with this number one gets almost close to the maximum benefit-cost ratio for user testing. In this article he also claims to differentiate between types of research. While his number of five is valid for qualitative user testing he suggests at least 20 participants for quantitative studies to get statistically

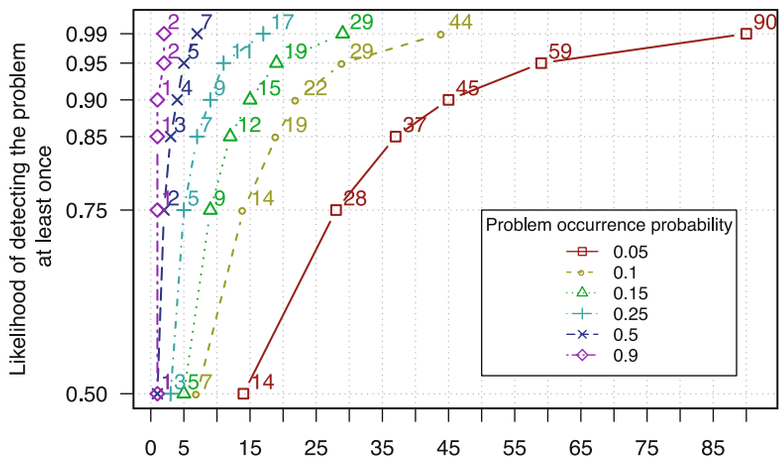


Figure 2.6.: Estimated sample sizes to detect event of interest at least once. The figure is based on Sauro and Lewis (2016).

significant results, 15 users for card sorting experiments and 39 users for eye tracking studies to get stable heatmaps. He argues for small tests with a dropping return on invest when using more than five users and suggests to better invest in additional studies than more participants for a single study (Nielsen, 2012).

Lazar et al. (2017, pp. 275 ff.) cite different authors to emphasize the debate about the sufficient number of users for a usability study. According to Virzi (1992) "five users will find approximately 80 % of usability problems in an interface" (as cited in Lazar et al., 2017, p. 275). The problem with this statement is that it assumes that the total number of usability problems for a user interface is known (Lazar et al., 2017, p. 275). Lindgaard and Chattratichart (2007) investigate the relationship between the proportion of problems found and the number of users by analyzing studies from test teams participating in a workshop at the 2003 CHI conference. As a result they found no correlation between the number of users and the proportion of usability problems found. In their study they also refer to Spool and Schroeder (2001) who discovered through the study of four web sites with 49 users that five users found only 35 % of all the problems and therefore are not enough. Another study conducted by Faulkner (2003) reports that five users reveal only 55 % of the usability problems. Hwang and Salvendy (2010) suggest 10 ± 2 participants to reach 80 % discovery rate, based on their meta analysis of usability evaluation studies from different digital libraries and offline sources and Schmettow (2012) even doubts to discover 80 % of the problems with only 10 users.

"The reality is that most usability testing will never uncover all, or even most, of the usability flaws. And even if all of the flaws were uncovered, most of them will never be fixed" (Lazar et al., 2017, p. 276). As the literature shows, there is still a debate about the appropriate sample size for usability studies. As Lazar et al. (2017, p. 276) summarize,

the goal should be to reveal the most important problems with an interface and actually fix them. There is no value in finding all usability issues without considering if they can be actually fixed (Wixon, 2003).

> "User testing with real user is the most fundamental usability method and is in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the concrete interface being tested." (Nielsen, 1993, p. 165)

Nielsen (1993) describes usability testing as irreplaceable, but also mentions to pay attention to the two main issues of reliability and validity when testing with users. While reliability describes the characteristic of getting the same results when using the same study design, validity shows that the study actually obtains relevant results for the usability of a product. An important decision during the planning of a usability test is between formative and summative evaluation. While a formative evaluation has the goal to improve the interface in an iterative design process and learn about the specific problematic aspects, the focus in a summative evaluation is on the assessment of the overall quality of a system. After the development of a comprehensive test plan including the test goals and its basic conditions, as well as information about the users, the collected data, and the test budget — the developed procedure should be tested with a few pilot subjects (Nielsen, 1993).

2.4.5. Measuring Human Performance

Besides surveys, interviews, and user studies the involvement of users can produce several types of data. These measurements of the human are able to gain significant insights into the ways users interact with interfaces.

Task Performance

A rather obvious type of measurement when looking at usability studies involving users is the measurement of task performance. Lazar et al. (2017, p. 288) differentiate between the two performance measurements task performance and time performance. Task performance describes the correctness, thus how many task were completed correctly and how many tasks were completed unsuccessfully. On the other hand time performance describes the task duration or the time people spent before giving up on a task. Sauro and Lewis (2016) distinguishes three ways of measuring task duration: task completion time, time until failure, and total time on task. Other quantitative measurements in the domain of task performance are error rates, average time to recover from errors, or the number of visits of a specific screen or page.

J. S. Dumas and Loring (2008) indicate the trade-off between the measurement of task performance and the think aloud (TA) technique as a users task times would increase by

expressing their thoughts. A possible solution to the problem could be the application of retrospective TA rather than concurrent TA. As the name implies the users interpret problems after attempting a series of tasks, optionally supported by a video recording of the exercise (Frøkjær & Hornbæk, 2005; Lazar et al., 2017).

Mouse Movements and Keyboard Input

In order to see how users control computer systems it can be helpful to track the user's mouse movements and keyboard interactions. For example traditional log files for websites often only give an insight in the different pages visited. As an HCI practitioner a more detailed view onto the journey of your users can be of specific interest. With an approach like the one Atterer et al. (2006) propose, not only the visited pages of a website got logged. Additionally, techniques like these can give an insight into the areas users tried to click unsuccessfully, the user's scroll events, mouse movements, and much more. Unfortunately, the application for mobile devices or modern interaction techniques like gesture or voice input is not possible without major adaption. Another downside is that mouse and keyboard interactions can't tell where the user is actually looking at.

Eye Tracking

This vacuum of knowledge can be filled by eye tracking techniques. Tracking the orientation of the fovea allows to draw paths of the user's gaze behavior on interactive interfaces. Besides stationary eye tracking systems for computer displays, there exist head-mounted systems for the application in mobile environments (Lazar et al., 2017, pp. 370 ff.).

In general, two different movement categories are distinguished. The *saccades* which describe "rapid eye movements used in repositioning the fovea to a new location in the visual environment" (Duchowski, 2017, p. 40). On the other hand the eye movements when focusing a new area of interest is called *fixation*. Duchowski (2017) adds two more movement categories to the list: *pursuits* describing the visual tracking of a moving targets and *nystagmus* which describes "a smooth pursuit movement interspersed with saccades invoked to compensate for the retinal movement of the target" (Duchowski, 2017, p. 45).

In addition to mouse and keyboard interaction, eye movements can gather relevant information for the evaluation of human-computer interaction of interactive systems. The technique allows to answer questions like where users looked before interacting with input elements or menu items. Furthermore, it allows to get information about which areas of an interface attract the users' attention (Lazar et al., 2017, pp. 371 ff.).

Measuring Workload

In order to create designs that are easy to use the assessment of the user's mental workload can help to understand which parts of the interaction create high mental demands on users. For example the two questionnaires Subjective Workload Assessment Technique (SWAT) (Reid & Nygren, 1988) and NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988) were developed to measure mental workload. The most widely used

NASA-TLX (Lazar et al., 2017, p. 315) consists of six scales addressing questions of mental, physical, and temporal demand as well as performance, effort, and frustration (Hart & Staveland, 1988). Since the development, the tool has been used in several different environments than originally intended (aviation crew complement) and grew to a benchmark for newly developed measures (Hart, 2006). But like other surveys, tools like SWAT and NASA-TLX are attributed by the shortcomings of fallible human memory. As users rate the scales after they accomplished a task the measurements can be inconsistent (Lazar et al., 2017, p. 376).

To overcome these problems several approaches are used to measure mental workload. For example through eye tracking the pupil diameter is measured which increases during stress or frustration (Barreto et al., 2008; Jiang et al., 2014; Klingner et al., 2008). Tokuda et al. (2009), Tokuda et al. (2011) and Tokuda et al. (2011) investigate the relation between saccadic intrusions as well as microsaccades and the mental workload to overcome the shortcoming of brightness-sensitive pupil diameter in vehicle driving environments. Other approaches use electrocardiogram (ECG) or electroencephalogram (EEG) to measure workload during the performance of the task. Another possible technique to assess mental workload is the transcranial Doppler sonography (TCD) which uncovers changing blood flow velocity through ultrasound pulses (G. Matthews et al., 2015).

> "Many HCI questions involve digging deeper than the level of individual tasks." (Lazar et al., 2017, p. 396)

Besides measuring gaze direction and workload through eye tracking and other physiological measurements, the human body holds several other processes that can be measured. Looking at parts of the human body it stands out that human bodies are constantly moving. New technologies, also in the consumer market like Nintendo Wii remote or Microsoft's Kinect, allow tracking of motion and position. Other possible measurements of the human body include physiological data like skin conductivity, blood flow, and respiration rate. As these technologies advance, they can provide new ways for studying how users interact with computers (Lazar et al., 2017, pp. 396 ff.).

2.4.6. Model-Based Testing

ISO 16982 defines model-based approaches as "use of models which are abstract representations of the evaluated product to allow the prediction of the user's performance" (International Organization for Standardization, 2002, p. 5). For example, goals, operators, methods, and selection rules (GOMS) is a model to predict how users use a system. As Figure 2.7 shows it consists of specific goals which can be achieved by one or more methods that determined by the selection rules. The operators thereby describe

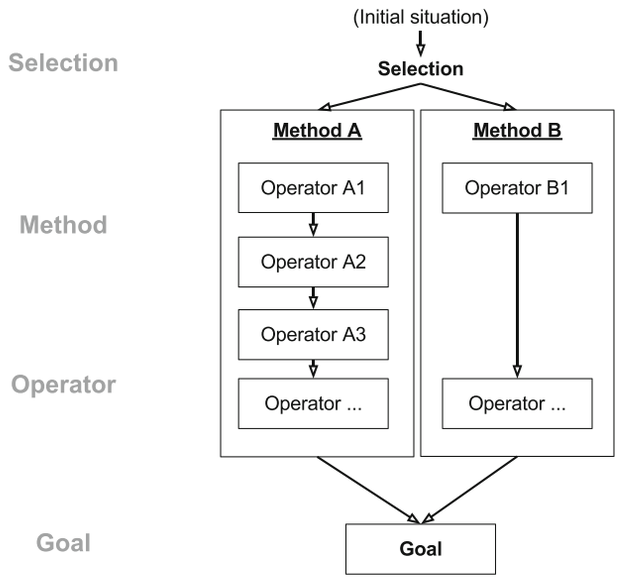


Figure 2.7.: Concepts of GOMS model. The figure is created by Olli Savolainen under the Creative Commons Attribution-Share Alike 3.0 Unported license.

perceptual, motor, or cognitive actions composed to the involved methods (Card et al., 1983) and are assigned a specific execution time. Therefore, the model allows to predict how a user will use a system for the specified tasks.

The first model-based approach of the GOMS family and also a streamlined variation of GOMS is the keystroke-level model (KLM). It uses four different "physical-motor operators" (Card et al., 1980) for keystroking, pointing, homing and drawing. Besides these, a mental operator for mental preparation by the user, and a response operator for the system. To predict the execution time of a task, as a series of operators, each operator has assigned an execution time either measured through experiments or defined through a formula. The summed operators execution times then gives the total execution time of the investigated task (Card et al., 1980).

Another model also used for research in HCI is the Adaptive Control of Thought-Rational (ACT-R). ACT-R describes a cognitive architecture consisting of multiple modules addressing different cognitive mechanisms and associated with distinct cortical regions. Through this, ACT-R aims — inspired by cognitive neuroscience — to define the basic operations that enable the human mind (Anderson et al., 2004). Byrne (2003) investigates several other cognitive architectures besides ACT-R with regard to their application in HCI. Therefore, he defines cognitive architectures as "a broad theory of human cognition based on a wide selection of human experimental data, and implemented as a running computer simulation program" (Byrne, 2003). This allows to simulate tasks with respect to cognitive processes like attention, memory, problem solving, decision making, and

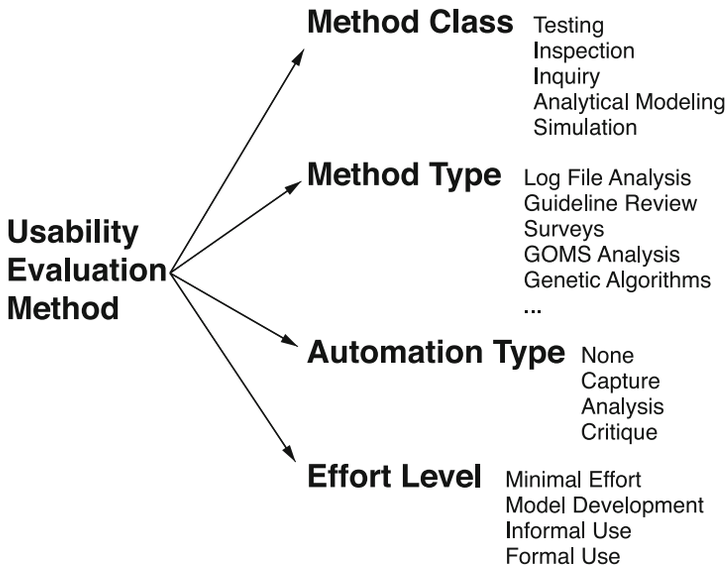


Figure 2.8.: Summary of the usability evaluation methods taxonomy by Ivory and Hearst (2001).

learning. A major strength of cognitive architectures within the field of HCI is the ability to produce execution times, error rates, and even learning curves.

2.4.7. Automated Testing

When talking about automation of usability evaluation, Ivory and Hearst (2001) differentiate between applications following the windows, icons, mouse, and pointer (WIMP) paradigm and web interfaces with mostly limited functionality. Furthermore, they cite Balbo (1995) stating four approaches to automation:

- *Nonautomatic*: methods "performed by human factors specialists";
- *Automatic Capture*: methods that "rely on software facilities to record relevant information about the user and the system, such as visual data, speech acts, keyboard and mouse actions";
- *Automatic Analysis*: methods that are "able to identify usability problems automatically"; and
- *Automatic Critic*: methods that "not only point out difficulties but propose improvements."

Based on several lacks in the classification of Balbo (1995), Ivory and Hearst (2001) suggest a more precise taxonomy for classifying UEMs, shown in Figure 2.8. Through this taxonomy they outline the state of the art of automating different usability engineering methods with respect to the method class, the automation type, the method type, and

the effort level. For all of the given method classes *Testing*, *Inspection*, *Inquiry*, *Analytical Modeling*, and *Simulation*, the three different types of automation *Capture*, *Analysis*, and *Critique* are investigated. While the automation type *Capture* describes the automatic recording of usability data, *Analysis* is defined as software automatically identifying usability problems, and *Critique* as software that automates analysis and suggests improvements.

Taking the method class of inspection methods as an example, Ivory and Hearst (2001) list the HyperCard-based documenting of a cognitive walkthrough by Rieman et al. (1991) with a formal level of effort. For the automation type *Analysis*, Ivory and Hearst (2001) discuss several automation approaches for the method type guideline review for applications following the WIMP paradigm (Mahajan & Shneiderman, 1997; Parush et al., 1998) as well as for user interfaces in the web (Faraday, 2000; Scholtz et al., 1998; Theng & Marsden, 1998; Thimbleby, 1997).

2.5. Usability for In-Vehicle Information Systems

> "When drivers interact with in-vehicle information and communication systems (telematics devices) that have visual-manual interfaces there is the potential for distraction of the driver from the driving task." (Alliance of Automobile Manufacturers, 2006, p. 6)

As the quotation of the Alliance of Automobile Manufacturers (AAM) implies, driving a car is a complex task. In order to reduce distraction the AAM suggests specific guidelines that support the product development of so called telematic devices by providing criteria and evaluation procedures (Alliance of Automobile Manufacturers, 2006). Also the Commission of the European Communities provides recommendations for the design of human-machine interfaces for in-vehicle information and communication systems, "considering that the driver's primary task is to control the vehicle safely in a complex and dynamic traffic environment" (Commission of the European Communities, 2008, p. 3). Besides these, there exist several guidelines (Bhise, 2002; Japan Automobile Manufacturers Association, 2004; Kroon et al., 2016; M. L. Matthews et al., 2001; National Highway Traffic Safety Administration, 2013; Stevens & Cynk, 2011) to assess usability for in-vehicle information systems with regard to the driving task.

2.5.1. The Specific Context of Use

As Harvey (2011) state, in order to develop a definition of usability for human-computer interaction (HCI) for in-vehicle information systems (IVIS), the preliminary step would

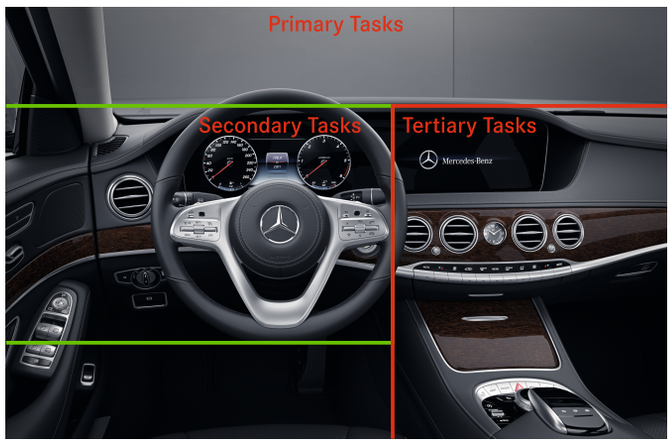


Figure 2.9.: Distribution of primary, secondary, and tertiary tasks. The figure is based on Tönnis et al. (2006).

be to define the specific context of use inside the vehicle. Tönnis et al. (2006) cite Geiser (1985) who describes interactive tasks in the vehicle in three different classes. *Primary tasks* are related to actual control of the vehicle, regarding heading and speed of the car. *Secondary tasks* are tasks that increase safety while driving and are mandatory, e.g. activating the turn signal or windshield wiper, whereas entertainment and information functions fall within the category of *tertiary tasks*. The different placement of information for different task categories in traditional vehicles is illustrated in Figure 2.9. Through a review of the literature, Harvey (2011) derive several "context-of-use factors" (Harvey, 2011, p. 30) as the main issues influencing the usability of IVIS.

Dual Task Environment

One of the most important influencing factors is the concept of a dual task environment which describes that the interaction with the IVIS is subsidiary alongside the primary driving task (Burnett, 2000; Lansdown et al., 2002). Conflicts between the performance of the driving task and the interaction with the IVIS can be indicated by the measure of efficiency (Harvey, 2011). A higher effort focusing on the IVIS reduces the amount of effort that is left for the driving task which leads to degradation of driving performance and potential risks to safety (Endsley, 1995; M. L. Matthews et al., 2001). If the secondary task is taking attention away from the primary task it is called driver distraction which is much more important for the usability than distraction in other contexts. As Green (2012) disagrees with the former CEO of Sun Microsystems who once said: "A car is nothing more than a Java technology-enabled browser with tires" (Kayl, 2000) and highlights that he "knows of no one who has ever been killed as a consequence of operating a computer at their desk, but the loss of life associated with crashes arising from normal motor vehicle operation is huge" (Green, 2012, p. 750) Another quotation from Jordan (1998) also illustrates the importance of dual task interference:

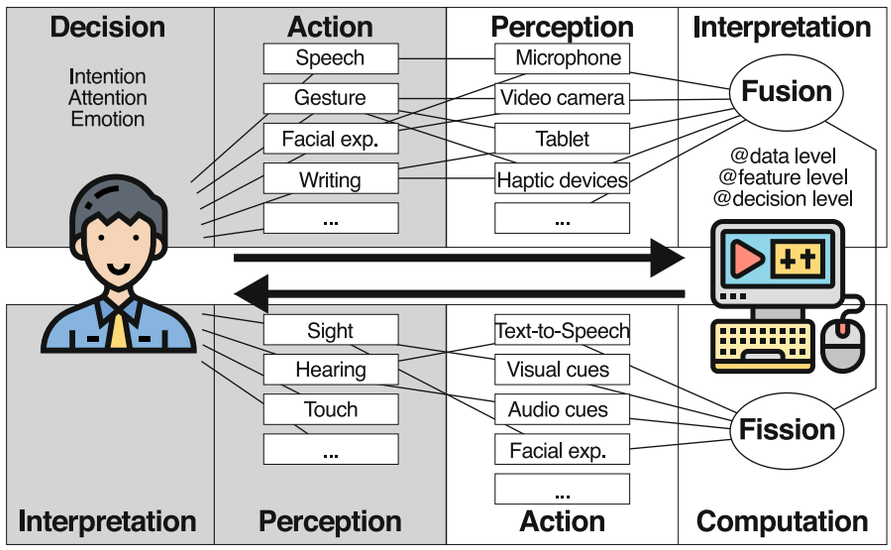


Figure 2.10.: A representation of the multimodal human-machine interaction loop by B. Dumas et al. (2009).

> "Whilst lack of usability in a video cassette recorder (VCR) may result in the user recording the wrong television programme, lack of usability in a car stereo may put lives at risk by distracting driver's attention from the road." (Jordan, 1998, p. 2)

Multimodal Interaction

A principle that correlates with the dual task environment of driving a car is the principle of multimodal interactions. Based on the action cycle by Norman (1988), B. Dumas et al. (2009, pp. 9 ff.) present a model of "multimodal man-machine communication". As Figure 2.10 shows, the communication can be divided in four different stages. In the decision state, the message content is prepared either consciously for an intention, or unconsciously for attentional content or emotions. The following action state servers to select the communication means to transmit the message. While in the perception state the message information is received by the multimodal system, in the interpretation state the system will try to give these information some meaning. After that, the system will respond with action according to the business logic defined by the developer in the computation state. The answer from the system is then generated using the most relevant modalities based on the context of use as well as the user's profile and transmitted in the action state. The user receives the answer in the perception state and processes it in the interpretation state (B. Dumas et al., 2009).

As summarized by Harvey and Stanton (2013, pp. 44–45) user inputs in the car can be made via physical movement like pushing a button or turning a knob or verbal using speech commands. Outputs of the IVIS can be made through visual, auditory, or physical modes. This transfer of information between driver and IVIS is illustrated in

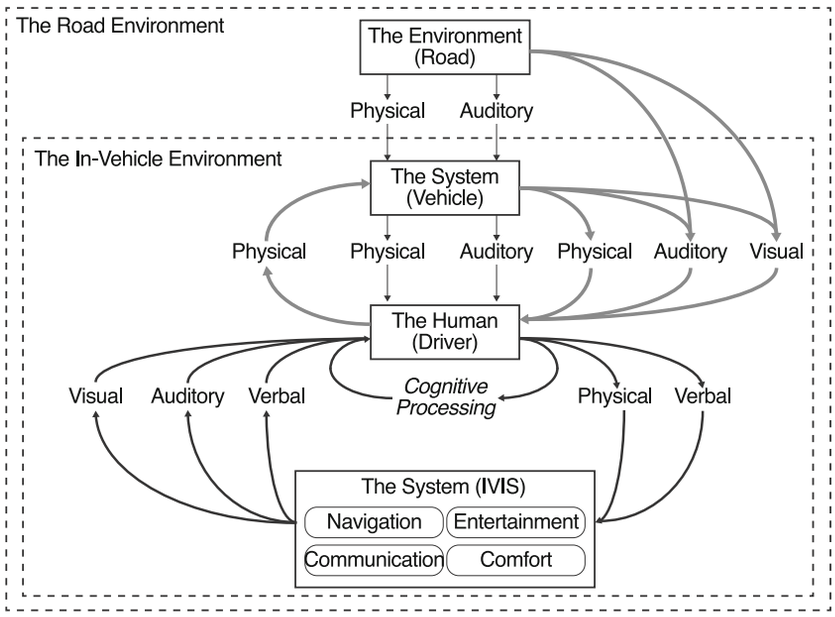


Figure 2.11.: Interaction between driver and IVIS. The figure is based on Harvey and Stanton (2013, p. 38).

Figure 2.11. Although the visual channel is already used for the driving task, the visual mode is the most common mode to present information form the IVIS to the driver while driving (Agah, 2000; Bach et al., 2008; Haslegrave, 1993; Sivak, 1996; Wierwille, 1993). Moreover, auditory tasks can occur simultaneously with visual tasks with minimal interference between the two information-processing channels (Fang et al., 2006; Wickens, 2002) and are relatively underused in driving tasks (Harvey & Stanton, 2013, p. 44) for which reason the spare capacity can be used for secondary in-vehicle tasks (Hulse et al., 1998). Physical interactions like haptic feedback or vibrations to alert the driver are not commonly used to gather information while driving.

Environmental Conditions

Another issue when designing for IVIS is the external vehicle environment. While the driving task is influenced by environmental factors such as visibility, road alignment, road marking, road signs and signals, road surfaces and curve radii, camber angles, and the existence of other road users (Fuller, 2005), the environment also influences the secondary task — the interaction with an IVIS. Harvey (2011) gives an example of lighting conditions that should be taken into account. Several guidelines and standards state that the operation of IVIS must not be adversely affected by glare either caused by sunlight or a bright display at night (Alliance of Automobile Manufacturers, 2006; Commission of the European Communities, 2008; International Organization for Standardization, 2017; Japan Automobile Manufacturers Association, 2004; Stevens et al., 2002). Harvey (2011) suggest to evaluate the usability through comparison of results on effectiveness of user

operations under varying conditions. The aim is to achieve a high level of effectiveness with only little variation between the different conditions (Harvey, 2011).

> "... the device should be designed to counter any adverse effects resulting from external environmental conditions." (Harvey, 2011, p. 32)

Range of Users

A definition of the potential user group is essential for the definition of the context of use. The potential user group for IVIS contains drivers and passengers and therefore a diverse range of physical, intellectual, and perceptual characteristics (Harvey, 2011). Commission of the European Communities (2008) highlight the need of taking these characteristics into account in several principles under the headline *Verification/applicable methods*: "Verification requires assessment and judgement taking into account the system's functionality and the intended user groups" (Commission of the European Communities, 2008). According to Baldwin (2002) and Herriotts (2005) with an increasing age of older drivers the likelihood of degradation of physiological, sensory, cognitive, and motor abilities increases. Stevens et al. (2002) also highlight elder drivers as special user group, while emphasizing the possible influence of different attitudes, emotional states, or the reason for a trip. The driving performance differs between experienced drivers and those with less driving experience. Furthermore, driving performance is not consistent, but is influenced by fatigue, stress, or the influence of alcohol and drugs (Stevens et al., 2002). In addition, the interaction with IVIS is not limited to the driver. While the interaction of passengers with a system in the car is not as critical as for the driver, possible conflicts between passenger and driver have to be considered as well as the operation of a system by the passenger. During evaluation the effectiveness and efficiency has to be tested ideally with the full range of potential users (Harvey, 2011).

Training Provision

As new users of IVIS mostly do not have time to read through instruction manuals before using the device for the first time (Commission of the European Communities, 2008) the learnability of such systems is an important usability factor. Keeping the need for complicated instructions or training as minimal as possible offers advantages that should be considered during the design of IVIS (Stevens et al., 2002). While learnability can be assessed by measuring the time taken to reach an acceptable level of performance (Harvey, 2011), Fastrez and Haué (2008) suggest measuring the initial effectiveness and efficiency as indication for usability of the IVIS when used for the first time (Harvey, 2011).

Frequency of Use

The frequency with which the investigated system is used holds as an important factor influencing the context of use. As Harvey (2011) state in the example of memorability, which will be more important for systems that are used infrequently than for systems that are often used and therefore facilitate information retention. Furthermore, an IVIS can provide functions that are rarely used although the driver uses the car on a daily basis (Commission of the European Communities, 2008). Another use case where

frequency of use plays an important role is the interaction with an IVIS in a rental car. In most cases the driver interacts with the system for the first time (Noel et al., 2005), whereby the usability factor of learnability is brought to the foreground.

An important usability factor, also addressed by Nielsen (1993) — the satisfaction — relates to the frequency of use. Hix and Hartson (1993) differentiate between short-term satisfaction which describes the satisfaction after initial use and is therefore important to ensure that the user will use the device frequently, and long-term satisfaction to maintain the frequent use of the product or system.

Uptake

As already stated at the beginning of this section, the interaction with an IVIS is not necessary to safely operate the vehicle. But Harvey (2011) argue that it can be chosen by the driver to enhance the driving experience. The most important factors influencing uptake of an IVIS are therefore satisfaction and perceived usefulness. While satisfaction cannot be measured as an isolated parameter and rather has to be considered as a trade-off between various usability factors, it is difficult to measure the perceived usefulness of a product before the actual product is released. Therefore, the likely behavior of real users can only be predicted through subjective evaluation (Harvey, 2011).

2.5.2. Design Space

After specifying the context of use while interacting with an IVIS, it is important to consider the actual space for which the system is designed. An approach to capture the area of interaction inside a car in a structured way is the design space introduced by Kern and Schmidt (2009). In their design space for driver-based user interfaces, input and output devices that can differ in placement and modality for in-car use are further investigated. Besides a detailed graphical representation of the design space, they have developed an abstract view for classification and compare different arrangements with regard to the three classes of the driving task. They have evaluated their approach with regard to completeness through the classification of more than 100 cars (Kern & Schmidt, 2009).

Through an analysis of 706 photographs of cockpits of 117 different car models from 35 different manufacturers, Kern (2012, p. 44) identified several different input modalities. In a first step the input modalities were assigned to one of the three classes defined by Tönnis et al. (2006) listed at the beginning of section 2.5.1. Furthermore, eight different input possibilities were identified. Buttons are separated into soft and mechanical buttons, as well as knobs which are differentiated between continuous and discrete knobs. Besides these, Kern (2012, p. 45) also includes sliders, pedals, and stalk controls in the group of traditional input modalities in cars. Modern multifunctional steering

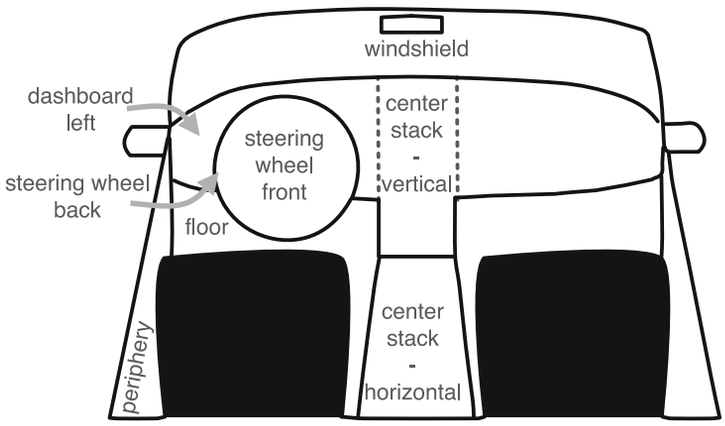


Figure 2.12.: Division of driver's interaction environment. The figure is cited from Kern and Schmidt (2009).

wheels often use thumbwheels, while modern infotainment systems are often controlled by a multifunctional controller in the center stack that can be turned, pressed, and often shifted in different directions. Besides new interaction techniques like speech or gesture recognition through microphones respectively cameras are also included, an input modality on the border of the output modalities is represented by the touchscreen which combines input and output in a single device (Kern, 2012, p. 46).

The output modalities in the design space are separated in the three classes visual, auditive, and haptic. While the class of auditive output modalities only contains loudspeakers, the haptic class includes vibrating steering wheels as well as mechanical knobs. Indicator lamps included in the class of visual output modalities can either be a simple light above a description or shaped where the symbol indicates the meaning. Furthermore, the display category ranges from analog displays like the classic speedometer to digital displays to present specific information, multifunctional displays often used for infotainment and comfort functions, and even head-up displays where the information is projected directly in the driver's field of sight (Kern, 2012, p. 48).

Regarding the positioning of input and output devices, Kern and Schmidt (2009) differentiate the positions visualized in Figure 2.12. The center stack of the vehicle cockpit is divided in vertical and horizontal, and the steering wheel is divided in front and back. The windshield as interaction area is mostly used for head-up displays, while the dashboard traditionally covers displays to visualize driving related functions like a speedometer, a revmeter, and fuel level indicator. The pedals for gas, brake, and clutch are most commonly placed in the front area near to the floor of the vehicle. Besides these, there are interaction devices in the peripheral area of a vehicle like a microphone that cannot be located precisely (Kern, 2012, pp. 49–50).

Input/Output Modality	Primary	Secondary	Tertiary
button			
slider			
knob (d=discrete, c=continuous)			
lever			
thumbwheel			
pedal			
multifunctional-knob			
warning and indicator lamps			
display {f ∈ [a=analog, d = digital, m = multifunctional, h=HUD]}			
loudspeakers			
speech recognition			

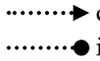
 direct connection between input and output device
 indirect connection between input device and output domain

Table 2.3.: Symbolic representation used in the graphical design space. The table is cited from Kern (2012, p. 51).

Kern (2012, pp. 50 ff.) created two different graphical representations of their design space for driver-based automotive user interfaces considering the findings regarding input and output modalities and the positioning in the cockpit. The more abstract representation notates existing interaction possibilities as triples representing the task classes primary, secondary, and tertiary. These are placed in a matrix of position and modality. As an example Kern (2012, p. 53) present the triple *(0-2,3,0)* in the column for the left hand input to the backside of the steering wheel. This indicates that the driver can interact with zero to two primary devices and with three secondary devices on the backside of the steering wheel with his left hand. This representation is used when describing or categorizing a set of cars (Kern, 2012, p. 53).

A more detailed graphical representation of a single car cockpit can be constructed by using the symbols in Table 2.3. The different input and output modalities are represented through different shapes, while the color indicates the three classes of tasks. The numbers inside the symbols indicate the number of occurrences in the investigated vehicle. The list of input and output modalities is designed to be further extended in cases where additional modalities are needed (Kern, 2012, p. 52). Figure 2.13 shows an example with

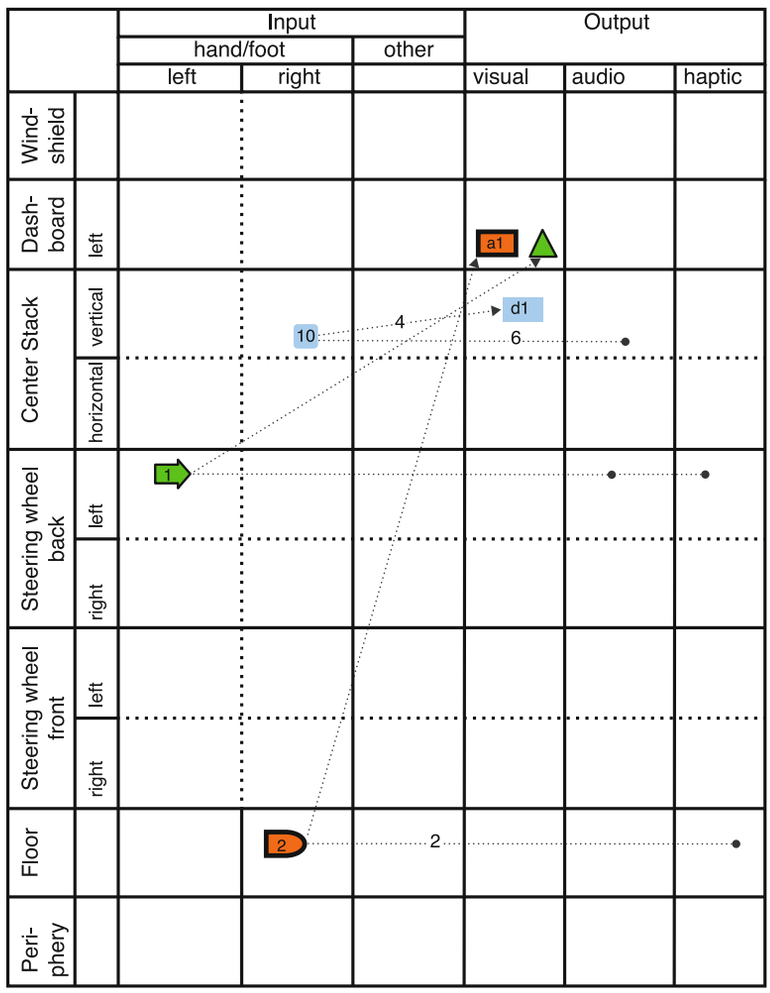


Figure 2.13.: Graphical representation of the design space for driver-based automotive user interfaces by Kern and Schmidt (2009). The figure is cited from Kern (2012, p. 52).

in this case two pedals on the floor likely for gas and brake. As both pedals provide direct haptic feedback, an indirect connection points to the column for haptic output in the floor area. Furthermore, the pedals deliver output to the analog display on the left side of the dashboard — the speedometer reacts when accelerating by a step on the gas pedal (Kern, 2012, p. 53).

> "The main purpose of the design space of driver-based automotive user interfaces is to provide a common basis for analyzing existing layouts and for generating new ideas." (Kern, 2012, p. 62)

As main purpose, Kern (2012) lists the supply of a "common ground" (Kern, 2012, p. 62) for discussion and documentation of user interfaces in an automotive context. Besides a historical analysis and an analysis of photos from the 2007 International Motor Show in Frankfurt, her dissertation presents different prototypes for automotive user interfaces, where the design space was used for exploration (Kern, 2012).

2.5.3. Driver Distraction

As already mentioned in section 2.5.1, the driving scenario is divided in tasks relevant for operating the vehicle safely through traffic and secondary or tertiary functions used for supporting the driver or increasing the comfort while driving. The topic of driver distraction addresses sources and reasons of distracted drivers from the driving task and possible solutions to identified problems.

Definitions of Driver Distraction

The National Highway Traffic Safety Administration (NHTSA) defines the term driver distraction as "a specific type of inattention that occurs when drivers divert their attention away from the driving task to focus on other activity" (National Highway Traffic Safety Administration, 2013, p. 24819). Thereby, the distraction is categorized into visual, manual, and cognitive distraction. While visual distraction requires the driver to look away from the roadway, manual distraction requires the driver to take a hand off the steering wheel, and cognitive distraction diverts the driver's mental attention away from the driving task (National Highway Traffic Safety Administration, 2013). Young et al. (2007) use the term driver distraction in a more specific context. They only speak of driver distraction when the driving performance is compromised because "drivers are no longer able to adequately divide their attention between the driving and secondary tasks and maintain driving performance at a satisfactory level" (Young et al., 2007, p. 380). The authors also cite Treat (1980), who states that "driver distraction occurs when a driver is delayed in the recognition of information needed to safely accomplish the driving task because some event, activity, object or person within or outside the vehicle compelled or tended to induce the driver's shifting attention away from the driving task". Furthermore, Streff and Spradlin (2000, p. 3) use the term driver distraction "as a shift of attention away from stimuli critical to safe driving toward stimuli that are not related to safe driving".

According to Young et al. (2007) driver distraction can be categorized into four types — visual, auditory, biomechanical (physical), and cognitive distraction. While the visual and auditory distraction types mean a shift of focus of visual respectively auditory attention to other signals than the road environment, physical distraction describes removing one or both hands from the steering wheel. Cognitive distraction occurs when the driver's attention is influenced by absorbing thoughts in a way that safe driving

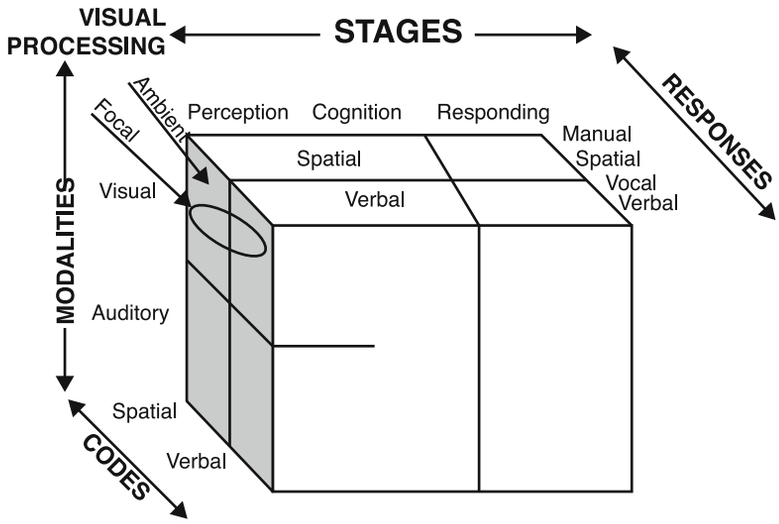


Figure 2.14.: Three-dimensional representation of the structure of multiple resources. The fourth dimension (visual processing) is nested within visual resources. The figure is cited from Wickens (2002).

is no longer possible (Young et al., 2007). The specified dual-task environment of the context of use inside a vehicle (cf. section 2.5.1) leads to interference between similar cognitive, perceptual, or motor resources (Wickens, 2002). In the four-dimensional multiple resource model, Wickens (2002) describes the four dimensions that "account for variance in time-sharing performance" (Wickens, 2002, p. 163). These dimensions — processing stages, perceptual modalities, visual channels, and processing codes — are shown in Figure 2.14. Wickens (2002) states that the concurrent demand of a level of one dimension by two tasks lead to more interference, than demanding separate levels on the given dimension. In detail, the stages perception and cognition involve the same resources which are functionally separated from the resources used for selection and execution of responses. This is supported by the observation that varying the difficulty of responding in a task does not affect perceptual and cognitive demanding tasks. Respectively, increasing the perceptual and cognitive difficulty does not influence the response-related performance significantly. Besides this, splitting attention between the visual and auditory perceptual modalities is better than between two visual or two auditory channels (Wickens, 2002).

Young et al. (2007) derive a higher dual-task interference for tasks that have visual inputs and require manual response compared to auditory or cognitive distraction from the multiple resource theory of Wickens (2002), when looking at primarily visual-spatial-manual tasks of driving. This theory is also supported by different studies in the literature (Hurwitz & Wheatley, 2002; Tijerina et al., 1998). Furthermore, Young et al. (2007) highlight the fact that dual-task interference is not negligible when two concurrent

tasks demand different resources. This is supported by a set of studies addressing mobile phone usage while driving (Haigney et al., 2000; R. Matthews et al., 2003; Patten et al., 2004; Strayer & Drew, 2004; Treffner & Barrett, 2004) which involves auditory inputs and speech responses compared to the visual, spatial, and manual demands of driving (Young et al., 2007).

Driver Analysis and Driver Modeling

To better understand the behavior of drivers in this dual-task environment, researchers perform studies analyzing driver behavior through driver monitoring. According to Islinger, Köhler, and Wolff (2011a) most of the research utilizes information from camera images (Devi & Bajaj, 2008; Itoh et al., 2005; Lee et al., 2008) with a few exceptions using analog speed graphs (Rygula, 2009). Islinger, Köhler, and Wolff (2011a) propose another approach for a driver analyzing system consisting of three modules, driver preferences, driving style analysis, and driver state. In order to gather relevant data for these modules, the authors limit themselves to using data from the standardized CAN bus system since most of the information available from the CAN bus is independent of the vehicle platform. Besides a method for driver state detection using Fast Fourier Transformation (FFT) of steering wheel angle (Islinger, Köhler, & Ludwig, 2011), Islinger, Köhler, and Wolff (2011b) propose a user model for an automotive environment. The generalized architecture allows for driver state detection using different physiological states like hunger and thirst, as well as psychological states like happiness, sadness, anger, anxiety, nervousness, relaxed, bored, stress, attentiveness, and drowsiness. The suggested human model for the automotive environment not only includes physical characteristics of each state and the respective measurable values, but also the measurability with the necessary input data (Islinger, Köhler, & Wolff, 2011b).

Measures of Driver Distraction

There exists a set of methods to measure driver distraction, separated by Wu (2009) into physiological and performance-based measures. While physiological measurements use eye tracking (Hurtado & Chiasson, 2016; Trösterer et al., 2014), the visual occlusion technique, or other advanced measures like EEG (G. Matthews et al., 2015) and ECG (Engström et al., 2005; Manseer & Riener, 2014; G. Matthews et al., 2015), the performance-based measures focus on driving performance through deviation from the driving lane (Hurwitz & Wheatley, 2002) or by analyzing GPS traces (Vhaduri et al., 2014).

Other measures of driving performance that give information about driver distraction are longitudinal control, lateral control, event detection and reaction time, gap acceptance, and subjective mental workload. One of the most common measures for longitudinal control is speed, in detail the variation in driving speed which has been shown to be higher when using a mobile phone (Burns et al., 2002; Green et al., 1993; Reed & Green, 1999). Furthermore, studies show that speed is often reduced when talking on a mobile phone (Burns et al., 2002; Haigney et al., 2000), operating a navigation system (Srinivasan

& Jovanis, 1997), or a CD player (Jenness et al., 2002). Besides lowering the vehicle speed, drivers tend to increase the headway when interacting with secondary tasks. Studies show that this observation applies particularly to secondary tasks with mainly visual demands (Greenberg et al., 2003; Hjälmdahl & Várhelyi, 2004; Östlund et al., 2004). In addition to speed and headway, measures of lateral control like lane keeping and steering are applied to quantify driver distraction. Östlund et al. (2004) show that particularly visual demanding secondary tasks lead to problems in lateral control, while Engström et al. (2005) only measured lane deviation in driving simulator studies — their field study results do not indicate similar results. Östlund et al. (2004) also analyze different measures related to steering wheel movements — among them reversal rate, angle variation, and rapid steering wheel turnings — indicating "sluggish steering wheel movements" (Östlund et al., 2004, p. 155) for visual demanding tasks. In a comparison study of different route guidance types — ranging from route and guidance map display with and without supplementary voice guidance to paper map and textual direction list — it could be shown that visual presentations influence lane deviations rather than voice guidance (Dingus et al., 1995). Tijerina et al. (1998) also show that destination entry for route guidance systems with mainly visual-manual demands results in a greater number of lane exceedences as well as longer eyes-off-road-ahead times when compared with voice input.

As the risk of being involved in a car crash is also related to the detection of external events and the respective reaction time of the driver, these measures can act as indicators for driver distraction (Young et al., 2009). For example, several studies show that phone usage, either handheld or hands-free, can increase the time needed for drivers to react to hazards or common road events like traffic light changes by up to 30 % (Brookhuis et al., 1991; Burns et al., 2002; Östlund et al., 2004; Strayer & Johnston, 2001).

Another way of measuring driver distraction is the assessment of mental workload through standardized questionnaires, as already mentioned in section 2.4.5. Young et al. (2009) list different scales to measure workload, among them the NASA-TLX,[1] the SWAT,[2] the modified Cooper Harper scale (MCH),[3] and the Rating Scale Mental Effort (RSME).[4] While the MCH and the RSME focus on a single dimension — difficulty level (Wierwille & Casali, 1983) or respectively invested effort (Zijlstra, 1993) — the SWAT uses three load dimensions — time, mental effort, and psychological stress — with the respective levels low, medium, and high (Reid et al., 1989). The NASA-TLX scales from very low (0 points) to very high (100 points) in 5-point steps for mental, physical, and temporal demand as well as performance, effort, and frustration. To create an overall task load index,

[1] Hart and Staveland, 1988.
[2] Reid and Nygren, 1988.
[3] Wierwille and Casali, 1983.
[4] Zijlstra, 1993.

the scores of these six scales are weighted according to the relevance for the specific study and thus combined to a single mean weighted workload score (Hart & Staveland, 1988). It is also used as a basis for the Driving Activity Load Index (DALI), which is developed particularly for the operation of vehicles. The Driving Activity Load Index (DALI) uses the same basic concept of "six pre-defined factors, followed by a weighting procedure" (Pauzié, 2008, p. 316). The workload dimension used are: effort of attention, visual, auditory, and temporal demand, interference and situational stress (Pauzié, 2008).

Compared to Wu (2009), Pettitt (2008) uses a different type of separation between existing methods to measure driver distraction into user trial methods and non-user trial methods. His focus during the review of the literature on measures of driver distraction is on the measurement of visual distraction, as it could be argued as the most concerning, but not exclusive source of distraction (Pettitt, 2008). As the most commonly used interaction with IVIS is mostly visual-manual demanding, visual distraction should be attributed an important factor to safety (Klauer et al., 2006).

The list of user trial methods by Pettitt (2008) for example contains the 15-second rule, which is also standardized by the Society of Automotive Engineers (SAE) as *Static Method* in the recommended practice *J2365*. The rule describes that task with a total completion time of more than 15 seconds in a stationary vehicle should not be accessible by the driver while driving. The task completion time is therefore measured in 30 trials (three trials for each of ten participants) and averaged through the logarithmic mean (Society of Automotive Engineers, 2016). Pettitt (2008) argues that despite the simple application of the 15-second rule, it misses the sensitive assessment of potential distraction.

Another user trial method is the occlusion technique where the gaze behavior while driving a vehicle is simulated through occlusion goggles. These goggles are designed in a way that allow to block the participants vision frequently during the test trial which mimics the interruption of a task by glances to the road environment (Pettitt, 2008). A so called *chunkable* task, which describes that a task could be completed effectively through a series of short glances, is considered as easy to resume and therefore more acceptable to perform while driving. Two measurements often raised by studies that apply the occlusion technique are the total task time (TTT) as a measure for task duration, and the total shutter open time (TSOT) as a measure for the visual time required to complete a task (Stevens et al., 2004). The ratio of total shutter open time (TSOT) to total task time (TTT) is described by Frank et al. (2002) as the chunkability index of a task ($R = TSOT_{mean}/TTT_{unocc\,mean}$).

The category of user trial techniques according to Pettitt (2008) contains also standardized methods like the lane change task (LCT) by Mattes (2003) and the Peripheral Detection Task (cf. Patten et al., 2004) as well as simulator and on-road studies. However, the

non-user trial methods involve qualitative methods and mostly model-based approaches. Hereby, models from the GOMS family — especially KLM — or cognitive behavior models are used to predict task times as well as eyes-off-road times. Further techniques involve the application of software to automatically generate user models, like keystrokes for KLM (Pettitt, 2008).

Driver Distraction Guidelines

Besides specific techniques to measure driver distraction, several industrial groups (cf. Alliance of Automobile Manufacturers, 2006; Bhise, 2002; Japan Automobile Manufacturers Association, 2004; Kroon et al., 2016; Society of Automotive Engineers, 2001; Stevens & Cynk, 2011) and governmental organizations (cf. Commission of the European Communities, 2008; National Highway Traffic Safety Administration, 2013) published guidelines for the design of HMI for IVIS. These guidelines involve different principle categories for installation, information presentation, interaction, system behavior, and information about the system. For the case study in chapter 6, these guidelines were reviewed and analyzed regarding common principles. This analysis is included in the appendix section C.1.9.

3. Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Information Systems

Parts of the following chapter have already been published in Lamm and Wolff (2019) and Lamm (2019). The following chapter is a revised and extended version of the published conference paper. As the previous chapter 2.4 shows, the range of methods to evaluate user interfaces is diverse. To get a revealing insight into the research landscape of human-computer interaction (HCI) for in-vehicle information systems (IVIS) — especially the used usability evaluation methods (UEMs) — an exploratory literature review was performed. The following sections explain in detail the investigated sources and the criteria for relevant literature (c.f. section 3.2), as well as the underlying classification schema (c.f. section 3.3) and the exploratory approach for the literature review (c.f. section 3.4).

3.1. Related Work

Because literature reviews give a valuable insight in the research landscape of specific subdomains, they are a welcome method when looking at a specific topic. Apart from the specific topic of HCI for IVIS, the actual use of different methods is of further interest as well. V. Böhm and Wolff (2013, 2014) analyze 55 empirical studies in an intercultural context regarding the employed methods. They refer to the classes introduced in ISO 16982 (International Organization for Standardization, 2002) to classify different UEMs. These classes are listed in section 3.3, where the schema used during this approach is discussed in detail. The most commonly used methods are interviews, questionnaires, and the thinking aloud technique, followed by observations of users, performance measurement, and creativity methods. The classification of ISO 16982 is also used as a basis for the survey conducted by Roche et al. (2014). They have distributed a questionnaire about the knowledge, usage, and context of specific methods from the aforementioned classes as well as the participants' profile to 139 professionals in the domain of HCI, of which 98 participants are selected because of their access to users

and their experience in usability methods. They split their results in methods with and without end users, whereby the most commonly used methods are user testing, observation, and interviews, respectively heuristic, document, and expert analysis.

When looking at research literature in the form of a literature review, not only the actual publication, but also the connections between different researchers can be of interest. While there are controversies about scientific authorship in different research areas (De S. Price, 1981; Edge, 1979; Katz & Martin, 1997; Shapiro, 1994; Van Raan, 1998), co-authorship analysis of scientific work provides an intense topic for research. Besides the investigation of multilateral international co-authorship (De Lange & Glänzel, 1997; Glänzel & De Lange, 1997; Van Raan, 1998), recent research focuses on measures of co-authorship networks and social network analysis (Abbasi et al., 2011; Di Caro et al., 2012; Ioannidis, 2008; Liu et al., 2005; Lu & Feng, 2009; Solomon, 2009). The collection of discussion papers published by the Alexander von Humboldt-Stiftung (2009) gives an overview of the publication behavior in different scientific disciplines. As Schuh (2009) summarizes, the different disciplines follow different principles to declare authorship, ranging from alphabetical order or "'equal contribution'" to "'first-last-author-emphasis'" (Schuh, 2009, p. 8) and honorary authorship.

In terms of research performance, Toutkoushian et al. (2003) summarize different performance indicators used in academic research. Besides the absolute number of publications in journals and summative indexes constructed from publication counts observed in the literature, they report different measures to rate research institutions among themselves. In order to create similar ratings from publication estimates from the Institute of Scientific Inquiry, the research performance is measured through the total number of publications and a calculated ratio of publications to full-time faculty (Toutkoushian et al., 2003).

In the domain of co-authorship analysis, Liu et al. (2005) investigated the field of digital libraries through social network analysis. Therefore, they constructed several co-authorship networks with different graph structures for the past ACM, IEEE and joint ACM/IEEE digital library conferences from data of the computer science bibliography *DBLP*.[1] Besides different metrics like component size analysis and centrality, they have also applied the *PageRank* algorithm (c.f. Brin & Page, 1998; Page et al., 1998), "the ranking mechanism at the heart of Google" (Liu et al., 2005, p. 1468) to the data. As *PageRank* is designed for search engine results, Liu et al. (2005) propose a modification of the algorithm, called *AuthorRank* which is especially suited for the weighted and directional network of co-authorship. A similar analysis is performed by Mezzanzanica et al. (2018) for the data of the *DBLP* in their project *GraphDBLP*. While the focus in Liu et al. (2005) is on the analysis of co-authorship, Mezzanzanica et al. (2018) study semantic keyword similarities and social network analysis. On the data of the *DBLP*, they perform several

[1] Digital Bibliography & Library Project (https://dblp.org/)

queries regarding the profiling and comparison of author publication records as well as social network analysis over the whole research community.

Another publication also studying collaboration in different research areas (biomedical research, physics, and computer science) shows that randomly chosen pairs of scientists within a research community "are typically separated by only a short path of intermediate acquaintances" (Newman, 2001, p. 404), also known as the "small-world problem" (Milgram, 1967, p. 62). Based on the concept of the "Erdös number" (Newman, 2001, p. 405) known from the mathematics community, the author measures distances between authors. The study has shown that in all different communities the average distance between scientists varies logarithmically with the size of the considered community. Furthermore Newman (2001) found that the investigated research communities are highly clustered, because the probability that two scientists collaborated is higher if both have a common collaborator.

3.2. Sources

In order to identify potential sources for literature in the research area of HCI for IVIS, the literature analysis software *SciVal*[2] of the scientific publisher Elsevier was used to screen different fields of research. *SciVal* offers different predefined so called research areas, where the areas *Human Factors and Ergonomics* and *Usability/User Experience* were analyzed regarding relevant conferences and journals. Furthermore, a self-defined research area using the keywords *human-computer interaction*, *user interfaces*, and *usability evaluation* was analyzed regarding sources of relevant literature. Further reference points were offered by the digital libraries of the Association for Computing Machinery (ACM)[3] and the Institute of Electrical and Electronics Engineering (IEEE),[4] which allow browsing through the organizations publications.

In the literature review, the articles of three conferences and eleven journals on the topic from 2015 to 2017 were captured in the investigation. In addition, three monographs or rather anthologies were screened for relevant chapters or articles. The full list of sources is listed in Table 3.1 along with the number of relevant articles. In order to refine the total number of papers from 4,864 to only include relevant papers for the topic of HCI for IVIS, the following selection criteria were defined (adapted from V. Böhm & Wolff, 2013, 2014):

- An explicit thematic relation to one or more usability evaluation methods.

[2] https://www.scival.com/
[3] https://dl.acm.org/
[4] https://ieeexplore.ieee.org

	Source	Relevant Studies
Conferences	ACM International Conference on Automotive User Interfaces and Interactive Vehicular Applications	93
	Mensch und Computer Workshop Automotive HMI	12
	Conference on Human Factors in Computing	13
Journals	IEEE Transactions on Intelligent Transportation Systems	10
	IEEE Transactions on Human-Machine Systems	7
	International Journal of Human-Computer Studies	2
	International Journal of Human-Computer Interaction	12
	Applied Ergonomics	39
	IET Intelligent Transport Systems	22
	ACM Transactions on Computer-Human Interaction	0
	Interaction Studies	0
	Advances in Human-Computer Interaction	0
	Human-Computer Interaction	0
	Journal of Usability Studies	1
Monographs	Harvey, C., & Stanton, N. A. (2013). *Usability Evaluation for In-Vehicle Systems*. CRC Press	4
	Winner, H., Hakuli, S., Lotz, F., & Singer, C. (Eds.). (2016). *Handbook of Driver Assistance Systems*. Springer International Publishing. http://link.springer.com/10.1007/978-3-319-09840-1	4
	Meixner, G., & Müller, C. (Eds.). (2017). *Automotive User Interfaces*. Springer International Publishing	4

Table 3.1.: List of sources for literature review.

- A practical application of the method in an empirical part of the study.
- The context of use is limited to interaction of the passengers of a vehicle.

From the total amount of 4864 papers, 223 papers — 118 conference papers, 93 journal articles, and 12 monograph/anthology chapters — were identified as relevant. The full list of papers with the corresponding publication source, and the year of publication can be found in the published dataset on *Mendeley Data* (Lamm, 2019).

3.3. Classification Schema

In a further step a classification schema was developed, to systematically organize the data from the articles selected as relevant. For the development of the schema illustrated in Figure 3.1, several meta data was taken into account. Starting from the entity *Paper*, with the main property *title*, as central element, each author is assigned the institutions he was affiliated when authoring the paper including the *name* and the *type*. Institution types are differentiated between *University*, *Research Institute*, *Company*, and *Government Authority* and connected to their country of origin identified by the *name*. On the other hand each paper is connected to the source it is published in, assigned with the source *title* and *type* as well as an additional property *year* when it was published for the relationship between paper and source. The source type corresponds to the differentiation used in Table 3.1. Therefore, the type includes *Monography*, *Journal*, *Conference*, and *Workshop*, whereby the workshop on "Automotive HMI" within the conference "Mensch und Computer" is assigned to the separate type *Workshop*, when compared to Table 3.1.

While the mentioned entities (*Paper*, *Author*, *Institution*, *Country*, and *Source*) of the classification schema can be applied to almost any discipline in research, the presented schema furthermore contains the entities *Interface* and *Method*, specific to HCI. The differentiation for the applied UEMs corresponds to the categories of ISO 16982 (International Organization for Standardization, 2002). According to V. Böhm and Wolff (2014) the standard "appears to be a reliable and distinct source of different categories". The standard lists twelve classes of UEMs:

1. Observation of users
2. Performance-related measurements
3. Critical incidents analysis
4. Questionnaires
5. Interviews
6. Thinking aloud
7. Collaborative design and evaluation
8. Creativity methods

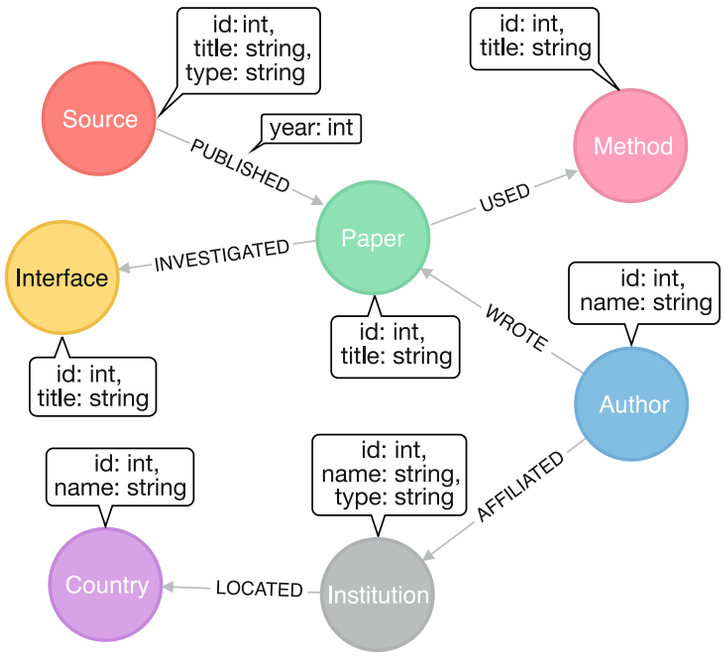


Figure 3.1.: A graph representation of the classification schema used for the literature review.

9. Document-based methods
10. Model-based approaches
11. Expert evaluation
12. Automated evaluation

Furthermore, the methods are divided into two categories depending on the direct involvement of users. While classes 1 to 7 imply direct involvement of users, the other classes imply at most the indirect involvement of users.

The second HCI-specifc entity *Interface* describes the investigated type of interface. Here, the categories are based on the three classes of the driving task by Geiser (1985) and the positioning dimension of the design space by Kern and Schmidt (2009) (cf. section 2.5.1 and 2.5.2) with a stronger differentiation of the input or output modality. Looking at the list of interface types, the different items can mainly be arranged easily in a default cockpit layout of a car:

- Center Stack Display
- Driver Display
- Head-up Display
- Speech Dialog System
- Auditory Display
- Air Gesture

- Steering Wheel
- Seat
- Lighting
- Virtual Reality
- Nomadic Device
- Thermal Interface
- Olfactory Interface
- Pedals

While the selection criteria in section 3.2 requires that each paper applies at least one of the above listed UEMs, the classification also includes papers that do not refer to a specific interface type. For example, more technical-oriented papers presenting a driving simulator implementation or investigating human behavior towards traffic signs in driving situations also include empirical data derived from a usability evaluation method but do not investigate a specific interface type.

3.4. Approach

In order to filter by relevance according to the criteria defined in section 3.2, all 4,864 papers were sighted manually. Papers that did not match the criteria got sorted out, while the relevant papers undergone an intense inspection. Therefore, information about the authors and their affiliations was determined from the front matter of each paper. In some cases further investigation was necessary to classify the type of institution between *University*, *Research Institute*, *Company*, and *Government Authority*. While defining the entity *Source* for each paper was trivial, the entities *Interface* and *Method* needed further analysis of the paper. Since the information about investigated interface types and applied UEMs is typically not included in the meta data of a paper, the content has to be analyzed manually with respect to the methods used in the empirical part of the studies as well as the investigated interface types.

The gathered information was initially stored in a *SQLite*[5] database. Figure 3.2 shows the entity relationship model behind the temporary database structure. The graphical user interface and editing software *DB Browser for SQLite*[6] offers export functionality for the most common exchange formats *SQL*, *JSON*, and *CSV*. The latter is used in the presented approach to import the collected data into the graph database software *Neo4j*.[7]

[5] https://www.sqlite.org/
[6] https://sqlitebrowser.org/
[7] https://neo4j.com/

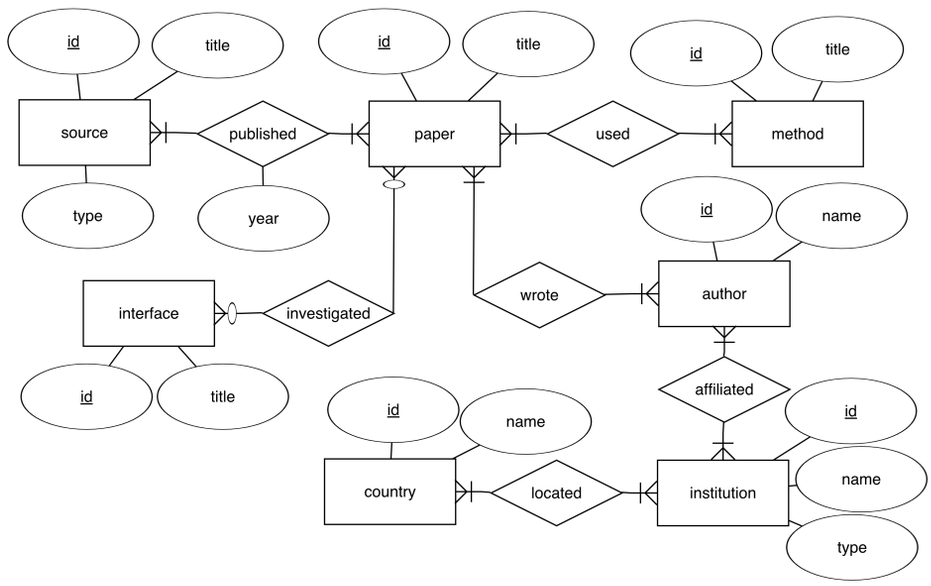


Figure 3.2.: The entity relationship diagram of the temporary used SQLite database.

3.4.1. Graph Database

Because spreadsheet formulas and SQL queries including multiple entities can get quite complex and difficult to manage, the presented approach uses the "property graph model" (Neo4j Inc., 2018c) which *Neo4j* is based on. According to this, the four building blocks of *Neo4j* are *Nodes*, *Relationships*, *Properties*, and *Labels*. Transferring these four principles to the classification schema introduced in Figure 3.1, the nodes are represented by the entities in the schema, while each entity type has a different label related to the entity type associated with it. The node properties (e.g. id, title, and type) describe the different columns an entity has, when looking at the schema from a relational perspective, as a key-value pair. Finally, the connections between different nodes including their attributes represent the relationships. In the property graph model these connections consist of "directed binary relations, which always have a start and an end node" (Mezzanzanica et al., 2018).

In order to query the graph database, Neo4j provides the SQL-inspired query language *Cypher*. The ascii-art syntax of the language is created around the concept of pattern matching (Neo4j Inc., 2018b). Besides subgraphs as results, *Neo4j* can also output tabular data. For example, if the usage of methods is of interest, the following query retrieves a table of methods sorted by usage frequency.

```
MATCH (method:Method)
RETURN method.title AS Method,
```

Method	Count
Performance-related measures	122
Observation of users	114
Questionnaires	103
Interviews	38
Thinking aloud	6
Creativity methods	3
Model-based approaches	2
Expert evaluation	2
Collaborative design and evaluation	2
Critical incidents analysis	0
Document-based methods	0
Automated evaluation	0

Table 3.2.: Usage frequency of used usability evaluation methods (UEMs).

```
    size((method)<-[:USED]-(:Paper)) AS Count
ORDER BY Count DESC
```

In the query, nodes are surrounded by parentheses, e.g. `(method:Method)`, while relationships are represented through arrows (`-->`) with an optional relationship-type in square brackets, e.g. the *USED* relationship in `(method)<-[:USED]-(:Paper)`. With the `MATCH` statement, the query selects all nodes of type *Method* and stores them in the variable `method`. The `RETURN` statement defines the output. In this case the query should return the property *title* of the node and how often the node shares a relationship of type *USED* with nodes of type *Paper*. With the `AS` statement the columns in the output table can be named. The result of this query is printed in Table 3.2. In the presented survey, the user interface *Neo4j Desktop* version 1.1.13 was used together with the *Neo4j* database version 3.5.0.

3.4.2. Network Analysis

While the mentioned application of the *Cypher* query language is suited well for exploratory analysis of connected data, *Neo4j* offers so called *Graph Algorithms* to perform basic network analysis and compute graph metrics (Neo4j Inc., 2018a). As the presented schema not only contains nodes representing actors and relationships as interactions between them, the presented data does not apply to the general definition of social networks by Aggarwal (2011). Following Aggarwal (2011), the presented data is rather associated with more generalized "*information networks*, in which the nodes

could comprise *either* actors or *entities*, and the edges denote the relationships between them" (Aggarwal, 2011, p. 2). The *Neo4j* workbench offers a library (*Graph Algorithms*[8]) to perform several algorithms for data analytics. Unlike typical social networks from the real world, the presented schema is not based on direct connections between the protagonists, like friendship relationships in Facebook. Rather a connection between two authors is established through one or more papers they co-authored. To simplify the data handling and add weighted links for collaboration an additional direct relationship between co-authors is introduced through the following query.

```
MATCH (a1:Author)-[:WROTE]->(p:Paper)<-[:WROTE]-(a2:Author)
WITH a1, a2, count(DISTINCT p) as count
MERGE (a1)-[c1:COLLABORATE]->(a2)
SET c1.count = count
MERGE (a1)<-[c2:COLLABORATE]-(a2)
SET c2.count = count
```

This adds the relationship type *COLLABORATE* with the number of publications as edge property. An example for the authors David Large and Gary Burnett from the University of Nottingham is visualized in Figure 3.3. The social network analytics presented here are based on this co-authorship connection unless specified differently.

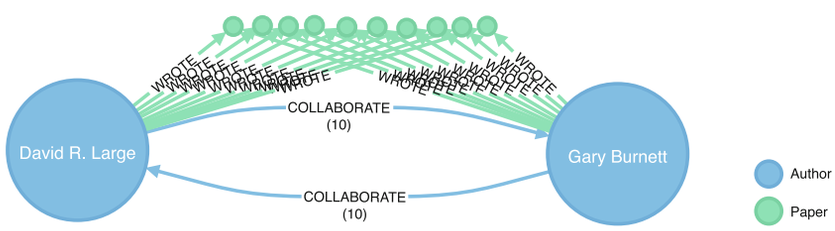


Figure 3.3.: The *COLLABORATE* relationship between two authors contributing to at least one common paper. As the two authors David Large and Gary Burnett from the University of Nottingham have common *WROTE* relationships to ten papers, the *COLLAORATE* relationships get a weight of 10.

Centrality

A concept of analytical research of networks is the concept of centrality. As Freeman (1978) outlines, the concept of centrality is a frequently discussed topic in scientific research. In the presented literature review four different measures of centrality are used. The degree centrality assigns a degree of importance based on the number of connections of each node in the network. Mathematically the degree centrality of a node n_i is defined as (Scott, 2017):

$$C_D(n_i) = \sum_{j=1}^{g} a(n_i, n_j), \tag{3.1}$$

[8] https://neo4j.com/docs/graph-algorithms/3.5/

where g is the total number of nodes and the function $a(n_i, n_j)$ is equal to 1 if n_i and n_j are connected, otherwise $a(n_i, n_j) = 0$. As the degree centrality measure in the presented network describes the direct co-authorship connections of each author, it functions as a measure of general collaboration of an author (Diestel, 2017). In *Neo4j* it is calculated by counting the outgoing *COLLABORATE* relationships of each author:

```
MATCH (a:Author)
RETURN a.id, a.name, size( (a)-[:COLLABORATE]->(:Author) ) AS
    degree
ORDER BY a.id ASC
```

The betweenness centrality on the other side measures how many times each node lies on the shortest path between other nodes. It can be described as (Wasserman & Faust, 1994):

$$C_B(n_i) = \sum_{j<k} \frac{\breve{g}_{jk}(n_i)}{\breve{g}_{jk}}, \tag{3.2}$$

where $\breve{g}_{jk}$ is the number of shortest paths between n_j and n_k and $\breve{g}_{jk}(n_i)$ describes the number of shortest paths between n_j and n_k that contain n_i. It shows which authors act as bridges between other authors in the network and therefore influence the flow around a system (Easley & Kleinberg, 2010). In *Neo4j* it is calculated through the following *Cypher* query:

```
CALL algo.betweenness.stream(
  'Author', 'COLLABORATE', {direction: 'both'}
)
YIELD nodeId, centrality
MATCH (a)
WHERE id(a) = nodeId
RETURN a.id, a.name, centrality as betweenness
ORDER BY betweenness DESC
```

Besides the mentioned centrality measures, closeness centrality measures how close each node is to all other nodes in the network. Closeness is calculated through the sum of shortest paths of a node. Mathematically it is defined as inverse average distance between node n_i and all other nodes in the network (Newman, 2018):

$$C_C(n_i) = \frac{1}{\sum_j d(n_i, n_j)}, \tag{3.3}$$

where the function $d(n_i, n_j)$ returns the length of the shortest path between the nodes n_i and n_j. In social network analysis, closeness is an indicator for nodes that are placed to influence the network most quickly (Freeman, 1978). As the original closeness centrality algorithm is problematic when applying to disconnected graphs, the variant of Marchiori

and Latora (2000) suggests calculating the sum of the inverse distances between node n_i and all other nodes, calculated by harmonic closeness centrality:

$$C_H(n_i) = \sum_{j \neq i} \frac{1}{d(n_i, n_j)}. \tag{3.4}$$

As the presented network contains several disconnected subgraphs, the harmonic closeness centrality is preferred over original closeness centrality. The following *Cypher* query calculates harmonic closeness centrality in *Neo4j*:

```
CALL algo.closeness.harmonic.stream(
  'Author', 'COLLABORATE', {}
)
YIELD nodeId, centrality
MATCH (a)
WHERE id(a) = nodeId
RETURN a.id, a.name, centrality as harmonic
ORDER BY harmonic DESC
```

Unlike degree centrality, the PageRank algorithm not only takes connections of the current node into account, but also includes the importance of a connected node itself (Brin & Page, 1998; Page et al., 1998). Therefore, a high PageRank indicates authors that influence a network beyond their direct collaboration relationships. Mathematically it can described as (Page et al., 1998):

$$R'(n_i) = c \sum_{n_j \in B_i} \frac{R'(n_j)}{N_j} + cE(n_i), \tag{3.5}$$

where $E(n_i)$ describes a vector corresponding to a source of rank, N_j is the number of connections from n_j, B_i is the set of nodes pointing to n_i, and c is used as normalizing factor. The following *Cypher* query calculates the PageRank in *Neo4j*:

```
CALL algo.pageRank.stream(
  'Author', 'COLLABORATE', {weightProperty: 'count'}
)
YIELD nodeId, score
MATCH (a)
WHERE id(a) = nodeId
RETURN a.id, a.name, score as pagerank
ORDER BY pagerank DESC
```

While degree centrality is calculated by simply counting the number of co-authorship connections, *Neo4j* offers the functionality to also calculate betweenness centrality,[9]

[9] https://neo4j.com/docs/graph-algorithms/current/algorithms/betweenness-centrality/

closeness centrality,[10] harmonic centrality,[11] and the PageRank[12] through the *Graph Algorithms* library.

Community Detection

Besides centrality measurements, the *Graph Algorithms* library also includes different functions for community detection in different networks. Because the result of the *Label Propagation*[13] algorithm is not stable across multiple calculations (Raghavan et al., 2007), it is not pursued for this study. The *Triangle Counting* or *Clustering Coefficient*[14] algorithm does not output communities per se, as it calculates the number of triangles a node is member of, together with a local clustering coefficient that depends on the connections of the node's neighbors itself (Schank & Wagner, 2005).

Alongside these mentioned algorithms for community detection, the *Graph Algorithms* library of *Neo4j* implements the *Louvain* method for community detection. The algorithm is originally designed by Blondel et al. (2008) based on the principle of modularity Q, which is defined by Newman (2004) as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{3.6}$$

where c_i is the community to which the node i is assigned, A_{ij} represents the weight of the relationship between i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the relationships attached to node i, the function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \frac{1}{2} \sum_{ij} A_{ij}$ is the number of nodes in the graph. The above Equation 3.6 serves as an indicator of the quality of a partition into communities. Nonzero values indicate deviations from randomness, while a value of zero claims no more within-community relationships than by random (Newman, 2004). In the first phase, the algorithm evaluates the gain modularity for removing i from its community as well as placing it in the community of j and places it in the community for which the gain is maximum and a positive value. In the second phase, the network is restructured according to the communities found during phase one by adjusting the relationship weights by the sum of the relationship weights in the corresponding communities. These phases are iterated "until there are no more changes and a maximum of modularity is attained" (Blondel et al., 2008, p. 4).

The other applied algorithm for community detection — *Connected Components* or *Union Find* — is according to Hopcroft and Tarjan (1973) which cite Shirey (1969) well known. The algorithm basically consists of two functions, a union function and a find function.

[10] https://neo4j.com/docs/graph-algorithms/current/algorithms/closeness-centrality/
[11] https://neo4j.com/docs/graph-algorithms/current/algorithms/harmonic-centrality/
[12] https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/
[13] https://neo4j.com/docs/graph-algorithms/current/algorithms/label-propagation/
[14] https://neo4j.com/docs/graph-algorithms/current/algorithms/triangle-counting-clustering-coefficient/

While the find function serves to determine in which subset a particular node is, the union function joins two nodes into a single subset. For each relationship between two authors, the union function joins the two participating nodes into a component. If the investigated nodes are already in a subset, the different subsets are merged together. Therefore, the algorithm creates groups of connected components, where each node is reachable from any other node. In a pre-investigation the *Union Find* algorithm was applied with different threshold values for the number of common publications of two authors (no threshold /minimum threshold of two common papers). This shows that the total number of identified communities differ extremely between 103 for the no-threshold sample and 580 for the sample with at least two publications as threshold. As for the total number of 661 authors a number of 580 communities does not seem reasonable, the *Union Find* algorithm is applied with no threshold.

As the collaboration connection between two authors in the investigated network is represented as a directed connection with the number of publications as edge weight, one could argue that the application of the *Union Find* algorithm mismatches the requirements. On the other hand, the *Strongly Connected Components*[15] algorithm ignores weighted edges, which seems unproblematic as the weights of the two directed connections between two authors coincide and the preinvestigation showed to not consider the weights at all. The *Strongly Connected Components* algorithm can be implemented in different ways. The first linear-time algorithm was introduced by Tarjan (1972) and is based on the application of the depth-first search. The search begins from a random start node and visits every node of the graph once. During this pass, the algorithm maintains a stack of already visited nodes and assigns an *index* value according to the order it was visited and a *lowlink* value that represents the smallest *index* of any node known to be reachable. The root of a strongly connected component is then determined by each node that has the same value for *index* and *lowlink*, while the component contains all of the nodes above it on the stack (Nuutila & Soisalon-Soininen, 1994).

Because the *COLLABORATE* relationship is represented through two directed edges between the authors with the same weight applied, the output of the *Strongly Connected Components* algorithm matches the output of the *Union Find* algorithm without a weight threshold. This prognostication is also sustained by a hierarchical cluster analysis using Ward's minimum variance method (Ward, 1963), comparing the differences in community detection algorithms. The results are discussed in detail in section 3.5.2.

Similarity

Besides the concept of centrality and the detection of communities, *Neo4j*'s *Graph Algorithms* library offers different node similarity algorithms. In order to find similarities between publications, the *Jaccard Similarity* algorithm, based on the Jaccard similarity

[15] https://neo4j.com/docs/graph-algorithms/current/algorithms/strongly-connected-components/

coefficient (Jaccard, 1902), measures dissimilarities through the division of the common features by the number of features (Niwattanakul et al., 2013) as defined in Equation 3.7.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.7}$$

In this literature review, this method is used to measure similarities between publications regarding the used usability methods and the investigated interface types. The result of the similarity algorithm is then used to calculate clusters by applying the already mentioned *Louvain* algorithm for community detection. The emerging communities from the algorithm contain publications that are similar according to the usage of specific UEMs and the investigation of specific interface types. The following *Cypher* query calculates a relationship labeled *SIMILAR* with the similarity score as edge weight.

```
MATCH (interface)<-[:INVESTIGATED]-(p:Paper)-[:USED]->(method)
WITH {
        item:id(p),
        categories: collect(DISTINCT id(method)) + collect(DISTINCT
            id(interface))
        } as userData
WITH collect(userData) as data
CALL algo.similarity.jaccard(data, {
        similarityCutoff: 0.001,
        write: true,
        writeRelationshipType: "SIMILAR",
        writeProperty: "score" })
YIELD nodes, similarityPairs, write, writeRelationshipType,
   writeProperty, min, max, mean, stdDev, p25, p50, p75, p90, p95,
   p99, p999, p100
RETURN nodes, similarityPairs, min, max, mean, stdDev, p25, p50,
   p75, p90, p95, p99, p999, p100
```

The value for *similarityCutoff* thereby allows to set a threshold for the similarity score. The relationships with a similarity score below this threshold will be ignored. The query moreover outputs general statistics of the calculated similarity measure like minimum, maximum, mean, standard deviation, and several percentiles.

Another node similarity measure — the *Overlap Similarity* — is used to find overlaps of different usability methods. While the *Jaccard Similarity* is calculated through the intersection of two sets divided by the size of the union of the two sets, the *Overlap Similarity* is calculated through the intersection of two sets divided by the size of the smaller of the two sets (Arnaboldi et al., 2016).

$$O(A,B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{3.8}$$

The results of the algorithm can then be used to detect methods that are used in combination. The following *Cypher* query calculates the overlap similarity for UEMs with respect to the usage in the publications of the network.

```
MATCH (p:Paper)-[:USED]->(method)
WITH {item:id(method), categories: collect(DISTINCT id(p))} as
   userData
WITH collect(userData) as data
CALL algo.similarity.overlap.stream(data, { similarityCutoff: 0.0
   })
YIELD item1, item2, count1, count2, intersection, similarity
RETURN algo.getNodeById(item1).id AS from, algo.getNodeById(item2).
   id AS to, intersection, similarity
ORDER BY from, to ASC
```

Besides these measures the *Graph Algorithm* library offers two more similarity measures, the *Euclidian Distance* and the *Cosine Similarity*. As both algorithms use edge weights to calculate similarity these methods are not applicable for the presented network.

3.4.3. Information Visualization

Besides the application of network analysis techniques, the data structure of a graph database offers another benefit — the visualization of the stored information. Card et al. (1999, p. 1) describe the term "information visualization" as the broader application of evolved computers as a medium for graphics. Futher they use the following definition:

> "The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." (Card et al., 1999, p. 7)

In contrast to scientific visualization, Munzner (2008, p. 149) uses the following definition of information visualization: "it's infovis when spatial representation is chosen, and it's scivis when the spatial representation is given". While "scientific visualization researchers deal primarily with three-dimensional physical objects and processes", "information visualization researchers are concerned with abstract phenomena for which there may not be a natural physical reality" (Bederson & Shneiderman, 2003, ch. Preface).

The Basics of Information Visualization Using the Literature Review Data

For information visualization Card et al. (1999, pp. 12–14) differentiate four levels of use — infosphere, information workspace, visual knowledge tools, and visual objects. While the infosphere describes the visualization of information outside the user's environment, the information workspace uses visualization to support the user in organizing information to perform some tasks. The level of visual knowledge tools, where the presented exploratory literature review approach in this chapter belongs to, describes tools that arrange information or use manipulation of information to detect patterns or visual

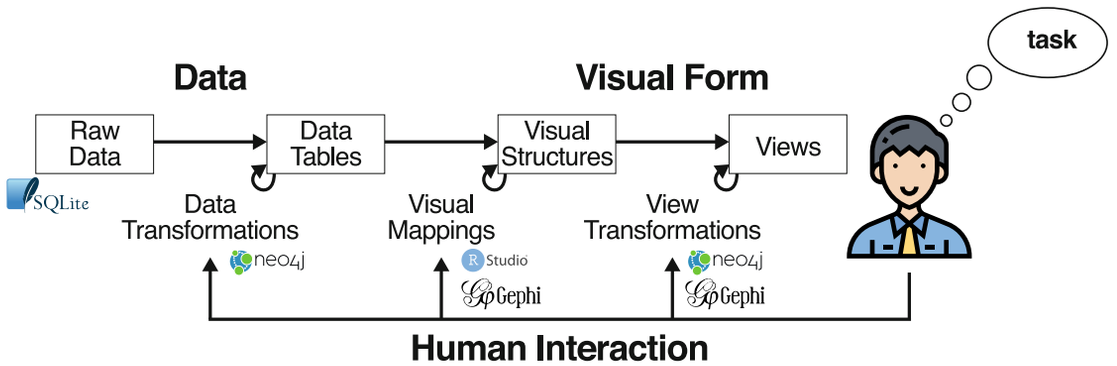


Figure 3.4.: Reference model for visualization. The figure is based on Card et al. (1999, p. 17). The used software and tools for the main steps of the visualization pipeline was added.

calculations. Visual objects use information visualization to enhance virtual physical objects in order to package collections of abstract information (Card et al., 1999, pp. 12–14).

Card et al. (1999, pp. 17 ff.) propose a reference model to describe visualization as the mapping of data to visual form supporting human interaction (see Figure 3.4). Through *Data Transformations* the *Raw Data*, specified as data in some idiosyncratic format, is mapped into relational descriptions of data — the *Data Tables*. These *Data Tables* are extended to include metadata which can lead to new values in the *Data Table* as well as new structure of the *Data Table*. *Visual Mappings* transform *Data Tables* into *Visual Structures* that combine so called "spatial substrates", "marks", and "graphical properties". Through the specification of graphical parameters like position, scaling, and clipping as *View Transformations*, *Views* of the *Visual Structures* are created by the user. The raw data for the literature review in this chapter is derived from information provided through the full text of the selected papers. This data can be transformed to include the relevant information for the literature review, exemplified in the *Data Table* in Table 3.3. The table uses the notation introduced by Card et al. (1999, pp. 17 ff.) using the input variables as rows and the cases as columns. While the first column specifies the variable name, the second column defines the variable type of the variable (N = *Nominal*, O = *Ordinal*, Q = *Quantitative*). As the variables *Author Names*, *Methods*, and *Interfaces* use cardinalities with more than a single possible value, the values are listed in curly braces. Since each author can be affiliated to several institutions, a separate *Data Table* in Table 3.4 is used for the authors. Again, the cardinality that one or more authors can be affiliated with one or more institutions leads to the notation of several values for the variable *Institutions* in curly braces. The data about each institution is stored in a separate *Data Table*, exemplified in Table 3.5. Other than in the tables above, in which the raw data is collected from the full text of the paper, this data is usually collected through research in online search engines.

PaperID	N	1	170	...
Title	N	Visual Distraction Effects of In-Car Text Entry Methods - Comparing Keyboard, Handwriting and Voice Recognition	Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling	...
Source Title	N	International Conference on Automotive User Interfaces and Interactive Vehicular Applications	Applied Ergonomics	...
Source Type	N	Conference	Journal	...
Year	Q_t	2017	2017	
Author Names	N	{Tuomo Kujala, Hilkka Grahn}	{Christian Purucker, Frederik Naujoks, Andy Prill, Alexandra Neukum}	...
Methods	N	{Performance-related measurements, Questionnaires}	{Observation of users, Model-based approaches}	...
Interfaces	N	{Center Stack Display, Speech Dialog System}	{Center Stack Display}	...

Table 3.3.: A Data Table for the paper entity of the literature review data.

AuthorID	N	1	493	...
Name	N	Tuomo Kujala	Francesco Biondi	...
Institutions	N	{University of Jyväskylä}	{Jaguar Land Rover, University of Padova, University of Utah}	...

Table 3.4.: A Data Table for the author entity of the literature review data.

InstitutionID	N	137	50	...
Name	N	University of Jyväskylä	Jaguar Land Rover	...
Type	N	University	Company	...
Country	N	Finland	United Kingdom	...

Table 3.5.: A Data Table for the institution entity of the literature review data.

In terms of *Visual Structures*, Card et al. (1999, pp. 23 ff.) cite Mackinlay (1986), who differentiates spatial substrate, marks, and graphical properties. Variables of a *Data Table* can be encoded using spatial position along several axes (unstructered, nominal, ordinal, or quantitative). All variables used in this literature review are of the nominal type, except the year of publication which is of the quantitative type time (Q_t). The graphs in this chapter mostly use the spatial position of the nodes to make the objects more visually salient. In order to support perception through clustering, gestalt principles like proximity are used. The nodes in the form of points and the edges or relationships in the form of lines represent the marks — "the visible things that occur in space" (Card et al., 1999, p. 28). According to Card et al. (1999, pp. 29 ff.) citing Bertin (1983), graphical properties are separated into spatial and object properties that can be separated according to whether the property is suitable for expressing the extent of a scale or for differentiating marks. While the spatial properties position and size are used to express the extent of a scale, orientation is used to differentiate marks. To express the extent of a scale through object properties the gray scale can be used, while the properties color, texture, and shape belong to the differential category. Card et al. (1999, pp. 29 ff.) list several other graphical properties from the literature like crispness, resolution, transparency, and arrangement as well as the breakdown of color into value, hue, and saturation (MacEachren, 1995, as cited in Card et al., 1999, p. 30). Further visual features listed are: number, line orientation, length, width, size, curvature, terminators, intersection, closure, color, intensity, flicker, direction of motion, binocular luster, stereoscopic depth, 3D depth cues, and lighting direction (Healey et al., 1996, as cited in Card et al., 1999, p. 30).

Regarding the *View Transformations* in Figure 3.4, Card et al. (1999, pp. 31 ff.) differentiate between the three categories location probes, viewpoint controls, and distortions to create different views of the data. While location probes describe operations like visual effects when hovering over elements, viewpoint controls mostly include controls to zoom, pan, or clip the viewpoint. Distortions are used to create views offering focus as well as context or combine overview and detail. For example using graphs, several nodes that are not relevant for the specific question can be shrinked as in Figure 3.3, where the *Paper* nodes are shrinked because the relevant information are the names of both authors and the number of publications they contributed together.

Interaction and Transformation

Card et al. (1999) state that human interaction in the reference model in Figure 3.4 includes techniques for data transformations, visual mappings and view transformations. For the literature review in this chapter, the implementation as a graph database in *Neo4j* offers an integrated user interface — the *Neo4j* Browser. The *Neo4j* Browser allows to create *Data Transformations* through the *Cypher* query language. For example, the addition of the supplementary *COLLABORATE* relationship creates a new derived structure as well as the calculation of similarity relationships for papers and methods. The calculation of centrality measures like the PageRank adds new derived values to the *Author* nodes

of the network. When it comes to *Visual Mappings*, the *Cypher* query language offers an intuitive way to create mappings through the ascii-art syntax. However, the *Neo4j* Browser offers only limited graphical properties to be adjusted like coloring and labelling of nodes and relationships. In order to gain more control over the *Visual Structures*, the desktop software *Gephi*[16] was used with imported data via the *GraphML* interchange format. *Gephi* offers several possibilities to create *Visual Mappings* like mapping the node size of *Author* nodes to the PageRank value (e.g. Figure 3.6), increasing the edge thickness with increasing similarity between *Paper* or *Method* nodes (e.g. Figure 3.11), or using color and proximity to visualize communities of authors (e.g. Figure 3.7) or clusters of similar publications regarding the applied usability evaluation methods and the investigates interface types (e.g. Figure 3.12). As the literature review follows an exploratory approach, the *Neo4j* Browser, but even more *Gephi*, offer possibilities to get detailed information about a displayed node or relationship and includes features to transform the viewpoint and create distortions.

Tools and Visualization Pipeline

This paragraph gives a short overview of the applied software and tools during the visualization pipeline. During data collection from the article full texts, the gathered data was organized using a *SQLite* database with several tables for entities and relationships. In order to create a graph representation, the collected data was exported from the *SQLite* database and imported into the *Neo4j* graph database using a *CSV* format. For *Data Transformations* like network analytics, the *Neo4j* Browser was used with the *Cypher* query language to calculate and export data using the formats *CSV* for tabular data and *GraphML* for graph data. The visualizations using tabular data like column charts in Figure 3.5 were generated using the development environment *R Studio* for the statistic programming language *R* using the *CSV* export data. Whereas the graph visualizations were created using the graph visualization software *Gephi* using the *GraphML* export data.

3.5. Results

The created network of publications in the domain of HCI for IVIS offers several possibilities to explore different aspects like co-authorship relationships, information about the usage of methods, or investigated display types. Therefore, the following section gives an overview of different issues regarding the network.

In total, the graph contains 1142 nodes: 661 authors from 184 different institutions and 31 countries; 223 papers of 17 different sources investigating 14 different interface types with 12 categories of UEMs. Between those nodes exists a total of 5322 connections,

[16] https://gephi.org/

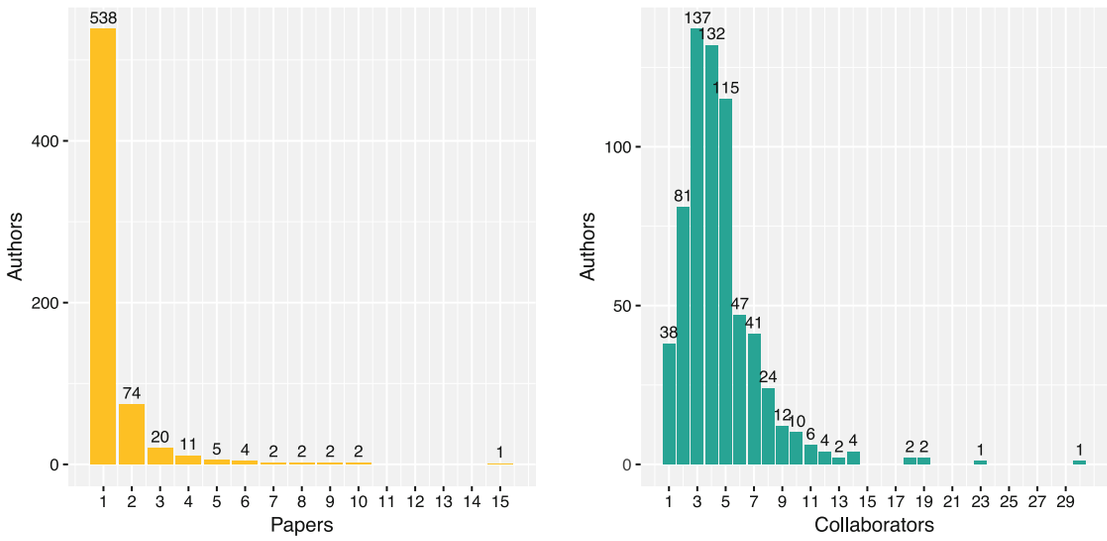
	M	Q_1	Q_2	Q_3	SD	Min	Max
Papers per author	1.39	1	1	1	1.22	1	15
Collaborators per author	4.56	3	4	5	2.84	1	30
Authors per paper	4.13	3	4	5	1.58	1	10
Institutions per paper	1.79	1	2	2	0.89	1	8
Interfaces per paper	1.37	1	1	2	0.69	1	5
Methods per paper	1.76	1	2	2	0.71	1	3

Note. M = mean; Q_1 = first quartile; Q_2 = second quartile/median; Q_3 = third quartile; SD = standard deviation; Min = minimum; Max = maximum.

Table 3.6.: Summary of the network statistics of the investigated graph.

excluding the calculated *COLLABORATE* relationship. As the data in the presented network was gathered manually, the problems described in Newman (2001) regarding the number of authors concerning different notations of a single author does not apply. Authors with the same name could be identified through research regarding their affiliation, and associated with different unambiguous identifiers. Furthermore, slight differences in authors identification (e.g. using different numbers of initials or the full name inconsistently) could eliminated through normalization. The presented network therefore contains 661 distinct authors. The average clustering coefficient between these authors is .506, calculated over the connections through the entity *Paper*, by applying the *Triangle Count* algorithm from the *Graph Algorithms* library. Therefore the authors in the presented network cultivate collaboration on a medium level.

As Table 3.6 shows, on average the authors wrote on about a single paper. The number of authors per paper lies mostly between three and five, with a mean of 4.13. While the standard deviation for the papers per author and the authors per paper ratios deviate on average only by 1.22, respectively 1.58, the number of distinct collaborators per author deviates stronger by 2.84 on average. The average number of authors involved with a paper lies around four, the average number of institutions involved with a paper lies below two. The mean number of investigated interface types and used UEMs per paper amounts to 1.37 for investigated interfaces and 1.76 for used UEMs. Even more information about the structure of the raised network data give the histograms in Figure 3.5. While most of the authors ($n(1) = 538$) only wrote on a single paper in the network, few authors contributed to more than one paper ($n(2) = 74$; $n(3) = 20$; $n(4) = 11$), and the most productive author contributed to 15 papers (cf. fig. 3.5a). However, taking a look at the frequency of authors per paper, only two papers were written by a single author. Most of the papers were authored by three to five authors ($n(3) = 52$; $n(4) = 57$; $n(5) = 44$), and a few were written by less or more authors ($n(2) = 30$; $n(6) = 23$), as shown in Figure 3.5c. This observation is also condensed



(a) Papers per author.

(b) Collaborators per author.

(c) Involved authors per paper.

(d) Involved institutions per paper.

Figure 3.5.: Different network frequency distributions. Illustrating the ratios of papers per author, collaborators per author, and involved authors and institutions per paper.

Author	Institution	Publications
Gary Burnett	University of Nottingham	15
David R. Large	University of Nottingham	10
Manfred Tscheligi	University of Salzburg	10
Alexander Meschtscherjakov	University of Salzburg	9
Bryan Reimer	Massachusetts Institute of Technology	9
Bruce Mehler	Massachusetts Institute of Technology	8
Stephen A. Brewster	University of Glasgow	8
Lee Skrypchuk	Jaguar Land Rover	7
Neville A. Stanton	University of Southampton	7

Table 3.7.: Top 5 most productive authors with their affiliation and the respective number of publications.

by the ratio of collaborators per author in Figure 3.5b, which also lies mostly between three and five. The average is with 4.56 collaborators per author slightly higher than the average authors per paper with 4.13, and also the dispersion is almost twice for the collaborators per author with a standard deviation of 2.84 compared to 1.58 for authors per paper. Looking at the number of distinct institutions involved with a paper (cf. fig. 3.5d), the average number (1.79) amounts to less than half of the average number of authors involved with a paper (4.13), while the standard deviation (0.89) lies a bit above half of the standard deviation of the authors per paper (1.58).

3.5.1. Who are the authors?

Looking at the most productive authors, Table 3.7 lists the nine most productive authors partially sharing the positions one to five with at least seven relevant publications for the presented literature review. According to the data, the most productive author is Gary Burnett with 15 publications, followed by his colleague David Large — both affiliated with the University of Nottingham located in the UK — and Manfred Tscheligi from the University of Salzburg with 10 publications. Furthermore, the University of Salzburg employs Alexander Meschtscherjakov with 9 relevant publications, on the same level as Bryan Reimer from the Massachusetts Institute of Technology. With 8 publications, Bruce Mehler from the Massachusetts Institute of Technology and Stephen Brewster from the University of Glasgow in the UK share position four, followed by Lee Skrypchuck from Jaguar Land Rover and Neville Stanton from the University of Southampton with 7 publications. The full list of authors with the corresponding number of publication can be found in the published dataset on *Mendeley Data* (Lamm, 2019). Looking at the data in more detail shows that all of the ten publications David Large contributed to are

Institution	Publications	Authors affiliated with Institution	Publications per author ratio
University of Nottingham	20	16	1.250
Jaguar Land Rover	13	13	1.000
University of Salzburg	10	15	0.667
Massachusetts Institute of Technology	10	17	0.588
University of Southampton	9	8	1.125
University of Glasgow	8	9	0.889

Table 3.8.: Top 5 institutions with highest number of publications, the respective number of authors affiliated, and the publications per author ratio.

also co-authored by Gary Burnett. The same applies to the nine papers by Alexander Meschtscherjakov that are also co-authored by Manfred Tscheligi as well as the eight publications of Bruce Mehler that are co-authored by Bryan Reimer.

Another research question addresses the origin of the authors in the network. With a total of 192 authors, Germany makes up the largest share, followed by the USA with 129 authors and the UK with 83 authors. Further countries include South Korea (36), China (35), Netherlands (31), Japan (29), and Austria (21). The full list of countries with the corresponding number of publications can be found in the published dataset on *Mendeley Data* (Lamm, 2019).

Considering only the data in Table 3.7, deriving the most productive institutions would result in the University of Nottingham followed by the University of Salzburg and the Massachusetts Institute of Technology. Based on the work of Toutkoushian et al. (2003), the research performance of institutions in this review is measured through the ratio of publications per author affiliated with an institution. As Table 3.8 shows, the most productive institution when considering the number of publications with 20 publications and 16 employed authors is the University of Nottingham, followed by Jaguar Land Rover with a publication per author ratio of 1.000 and the University of Salzburg and the Massachusetts Institute of Technology with 10 publications and a ratio of 0.667 respectively 0.588 publications per author. The positions four and five are claimed by the University of Southampton with 9 publications and a ratio of 1.125 publications per author and the University of Glasgow with 8 and a ratio of 0.889 publications per author. Taking a closer look at the collaboration relationships of the most productive institutions it is apparent that the authors from the two most productive institutions — the University of Nottingham and Jaguar Land Rover — worked together on seven publications, which is therefore the most productive collaboration of two institutions. Another productive relationship is maintained by the University of Nottingham and

the University of Southampton with six joint publications. On the third position ranges the collaboration between the University of Oldenburg and the OFFIS - Institute for Information Technology with five common publications. The full list of institutions with the corresponding number of authors, and publications can be found in the published dataset on *Mendeley Data* (Lamm, 2019).

3.5.2. How do these authors collaborate?

In order to analyze how the authors in the network work together on different publications, the centrality measures in section 3.4.2 are used to identify important authors that act as interfaces between communities. In a further step the graph is analyzed regarding communities of authors in the network. The presented results in the following sections report exemplary highlights.

Centrality Measures

When looking at the list of authors with respect to centrality, it is not surprising that productive authors also have a high degree centrality. With the highest value for degree centrality of 30, Gary Burnett also has the highest number of publications (15). He chaired the Automotive'UI conference in 2015, leads the group of Road Transport within the Human Factors Research Group, and is affiliated as Professor of Transport Human Factors at the Faculty of Engineering at the University of Nottingham. Therefore, 15 of his collaborators are also affiliated with the University of Nottingham. On the other hand he maintains a productive relationship with different authors from Jaguar Land Rover with 7 common publications. Another 2 publications arose in his relationships to different authors from the Virginia Tech. Besides the highest value for degree centrality, Gary Burnett also has the highest value for betweenness centrality (1047.65). This indicates that he is not only strongly interconnected, but also lies on the shortest paths between several other nodes. Therefore, he acts as a bridge between different authors in the graph. The calculated closeness centrality for Gary Burnett is 0.44. Because of the problems appearing for closeness centrality in disconnected subgraphs, the value will not be reported for the following authors. As the highest value of harmonic closeness with 0.06 for Gary Burnett is very low, the authors in the network are generally widely spread across the network. Nevertheless, Gary Burnett as well as other individual authors from very productive communities — like David Large (C_H = 0.05), Lee Skrypchuck (0.05), Manfred Tscheligi (0.05) — rank high with respect to harmonic closeness centrality.

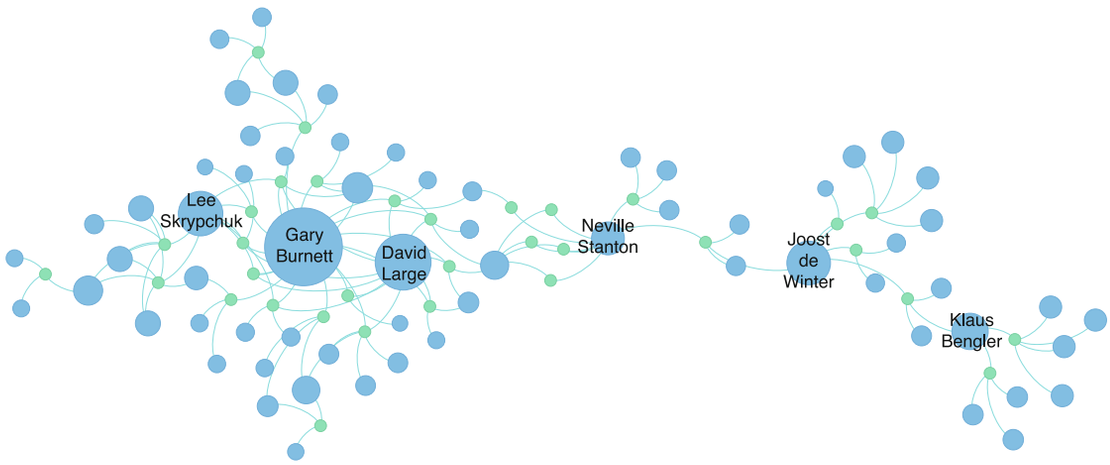
Finally, the PageRank calculation also shows the highest value for Gary Burnett (5.71) which emphasizes his position in the network. On the second place, when ordered by PageRank, is Manfred Tscheligi with a PageRank score of 3.87. He also holds the second highest value for degree centrality (23), while his betweenness centrality value of 302.58 is only the seventh highest value in the network. The authors on places two to

six regarding betweenness centrality are all from the same community as Gary Burnett (c.f. Figure 3.6a), more precisely: Neville Stanton (C_D = 8, C_B = 1003.33, R' = 1.83), Joost de Winter (C_D = 13, C_B = 883, R' = 2.76), Catherine Harvey (C_D = 6, C_B = 520.55, R' = 1.42), Victoria Banks (C_D = 4, C_B = 416.11, R' = 0.62), and Klaus Bengler (C_D = 10, C_B = 411, R' = 2.13). Looking at the PageRank score, positions three and four are occupied by David Large (C_D = 18, C_B = 271.41, R' = 3.82) from the same community as Gary Burnett and Alexander Meschtscherjakov (C_D = 19, C_B = 131.08, R' = 3.33) from the same community as Manfred Tscheligi (c.f. Figure 3.6b). Both are followed by Bryan Reimer (C_D = 19, C_B = 61.63, R' = 3.11) and Bruce Mehler (C_D = 18, C_B = 46.13, R' = 2.89) from the Massachusetts Institute of Technology and also from the same community (c.f. Figure 3.6c).

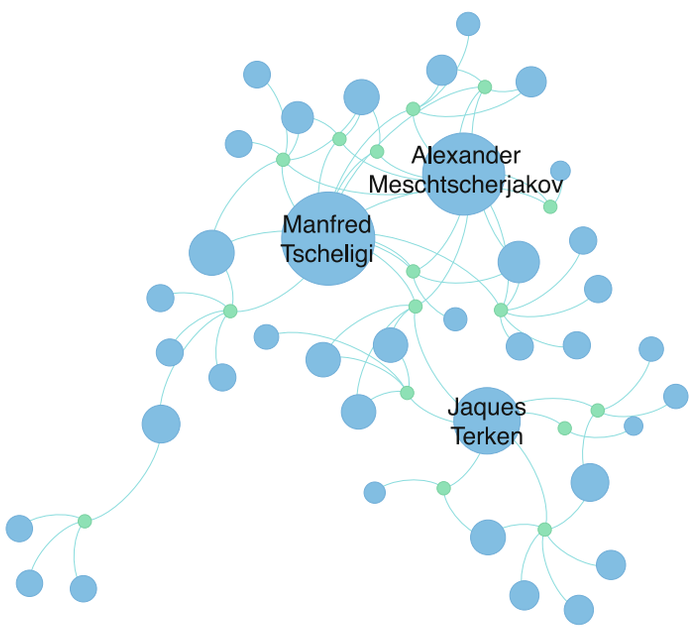
Community Detection

As already mentioned in the previous section, the performed community detection algorithms detected several communities in the network. While the *Louvain* algorithm detected 104 communities, *Union Find* and *Strongly Connected Components* detected 103 communities. As explained above in section 3.4.2 the both algorithms *Union Find* and *Strongly Connected Components* produce the same result.

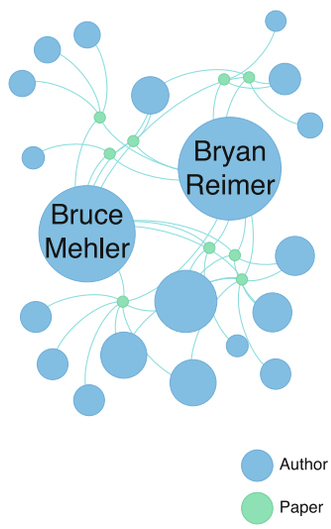
The performed hierarchical cluster analysis shows that all three algorithms detect the same communities — except one difference. While *Union Find* and *Strongly Connected Components* detect the very same communities, the *Louvain* algorithm splits one community which is detected as one single community by *Union Find* and *Strongly Connected Components* as two separate communities. In detail the *Louvain* algorithm splits the community visualized in Figure 3.6a in two separate communities. While the left part of the graph containing Lee Skrypchuk, Gary Burnett, David Large and Neville Stanton belong to one community, the second community starts on the right side with Joost de Winter. The two co-authors of the connecting publication between Neville Stanton and Joost de Winter — Daniël Heikoop and Bart van Arem — belong to the community of Joost de Winter. The difference of the two algorithms *Louvain* and *Strongly Connected Components* is also visualized in Figure 3.7. While the two algorithms *Strongly Connected Compontents* and *Union Find* detect a single large community (the blue highlighted structure in Figure 3.7b), because of the collaboration connection between the two sub-communities, the *Louvain* algorithm divides this large community into two separate communities (the blue and brown highlighted structure in Figure 3.7a). Looking at the dendrogram visualization in Figure A.1 in appendix A, it is clear that the communities with the IDs 3 and 42 detected by the *Louvain* algorithm are united in the community with the ID 426 detected by the *Union Find* as well as the community with the ID 10 detected by the *Strongly Connected Components* algorithm. As Table 3.9 shows, these communities are the largest and one of the fourth largest communities regarding the number of authors. When ignoring the two communities with only a single author, the community with the ID 3 ranks on the third position regarding the publications



(a) Community graph for Gary Burnett from the University of Nottingham (ID 3).

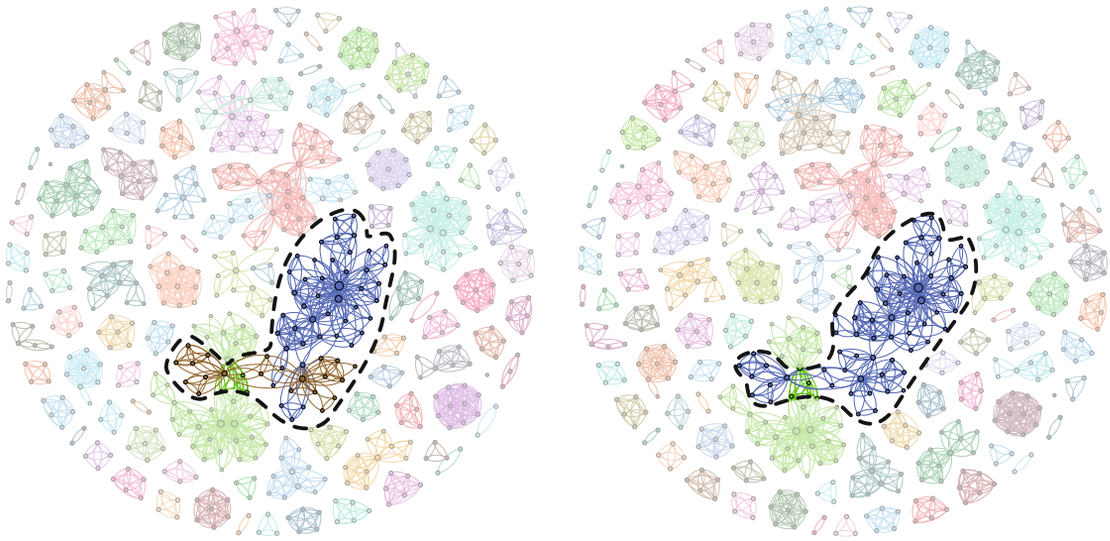


(b) Graph for Manfred Tscheligi (ID 9).



(c) Graph for Bryan Reimer and Bruce Mehler (ID 6).

Figure 3.6.: Community graphs for Gary Burnett from the University of Nottingham, Manfred Tscheligi from the University of Salzburg and Bryan Reimer and Bruce Mehler from the Massachusetts Institute of Technology, the node size represents the PageRank. The IDs correspond to the IDs allocated by the *Louvain* algorithm.



(a) Output of the *Louvain* algorithm for community detection.

(b) Output of the *Strongly Connected Components* algorithm for community detection.

Figure 3.7.: Comparison of the two applied community detection algorithms. The different colors represent different communities, the node size represents the PageRank, and the highlighted area shows the difference in the community detection between the two algorithms.

Community ID	Authors	Publications	Publications per author	Institutions
3	45	27	0.6	11
9	37	16	0.43	7
15	27	7	0.26	7
6	20	9	0.45	4
42	20	7	0.35	6

Table 3.9.: The five largest detected communities by the *Louvain* algorithm regarding the number of authors.

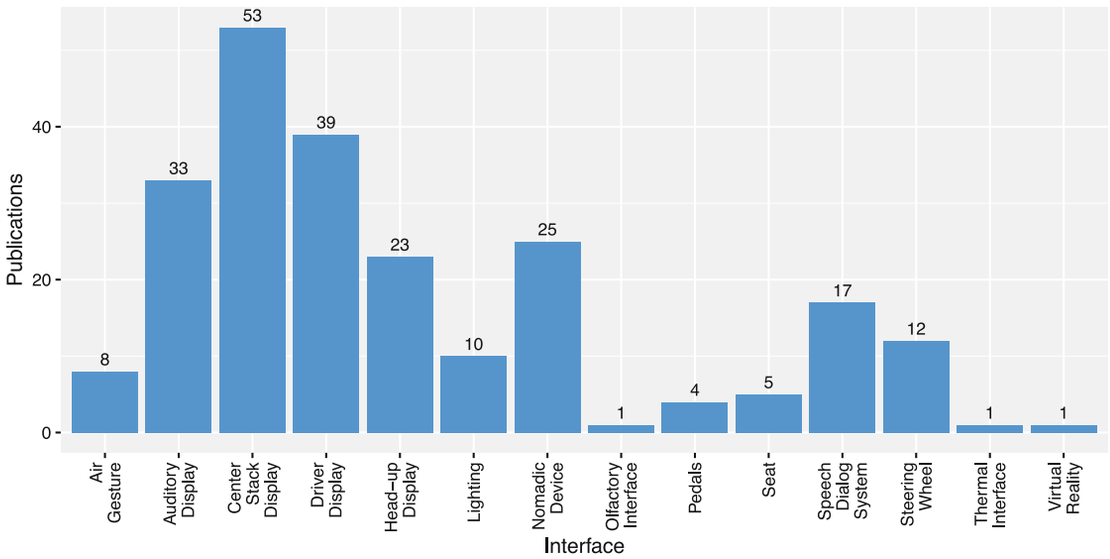
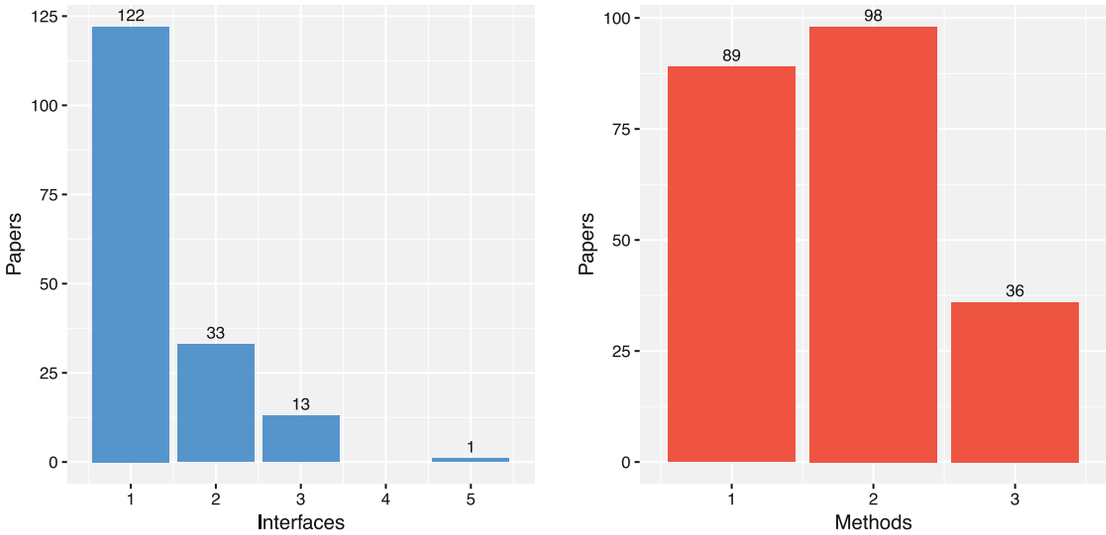


Figure 3.8.: Number of publications for investigated interface types.

per author ratio. The two preceding communities are relatively small with 11 authors in the community with the ID 1 (8 publications, 0.73 publications per author) and 3 authors in the community with the ID 33 (2 publications, 0.67 publications per author), when compared to the community with the ID 3. The second largest community is the community with the ID 9 with 37 authors, 16 publications, 0.43 publications per author, and 7 involved institutions. The average community size when applying the *Louvain* algorithm is 6.36 with an average number of publications of 2.15 and an average ratio of 0.34 publications per author. The full list of detected communities by the *Louvain* algorithm with the corresponding number of authors, institutions, publications, and publications per author ratio can be found in Table A.1 in appendix A.

3.5.3. Which interface types are of interest?

Taking a look at the investigated interface types in Figure 3.8 it is noticeable that the most commonly investigated interface type is the *Center Stack Display* with 53 publications (23.77 %). Since most comfort and entertainment features as well as navigation systems are positioned in the *Center Stack Display*, the investigation of this interface type increases with an increase in such functionalities (Kern & Schmidt, 2009). On positions two and three are following *Driver Display*s with 39 publications (17.49 %) and *Auditory Display*s with 33 publications (14.80 %). This can be explained with an increase of devices and functionalities to support the primary driving task (Kern & Schmidt, 2009), which also leads to studies analyzing human-computer interaction with *Driver Display*s. Besides



(a) Investigated interfaces per paper.

(b) Used methods per paper.

Figure 3.9.: Frequency of investigated interfaces and used methods.

this, Kern and Schmidt (2009) already mentioned the opportunities of speech and gesture recognition as well as the windshield area through providing head-up display or spatial audio technologies. Therefore, often investigated interface types include *Nomadic Device*s (25 publications, 11.21 %), *Head-up Display*s (23 publications, 10.31 %), and *Speech Dialog System*s (17 publications, 7.62 %).

As several publications investigate multiple interface technologies, Figure 3.9a shows the distribution of different interface types per paper. While the defined selection criteria in section 3.2 requires the application of at least one UEM, there is no explicit need for the investigation of a specific interface type. Therefore, only a subset of 169 publications are actually connected to an interface type from which 122 papers investigate a single interface type. A number of 33 papers examine two interface types and 13 publications investigate three interface types. Only one paper surveys five different interface types in a single publication, which studies fatigue warning systems with auditory, visual, tactile, and electric stimuli in the interface types *Auditory Interface*, *Driver Display*, *Steering Wheel*, *Seat*, and *Nomadic Device*.

3.5.4. Which evaluation methods are applied?

Besides the distribution of investigated interfaces per publication, the usage of different methods per paper could also be examined. Figure 3.9b shows that together more

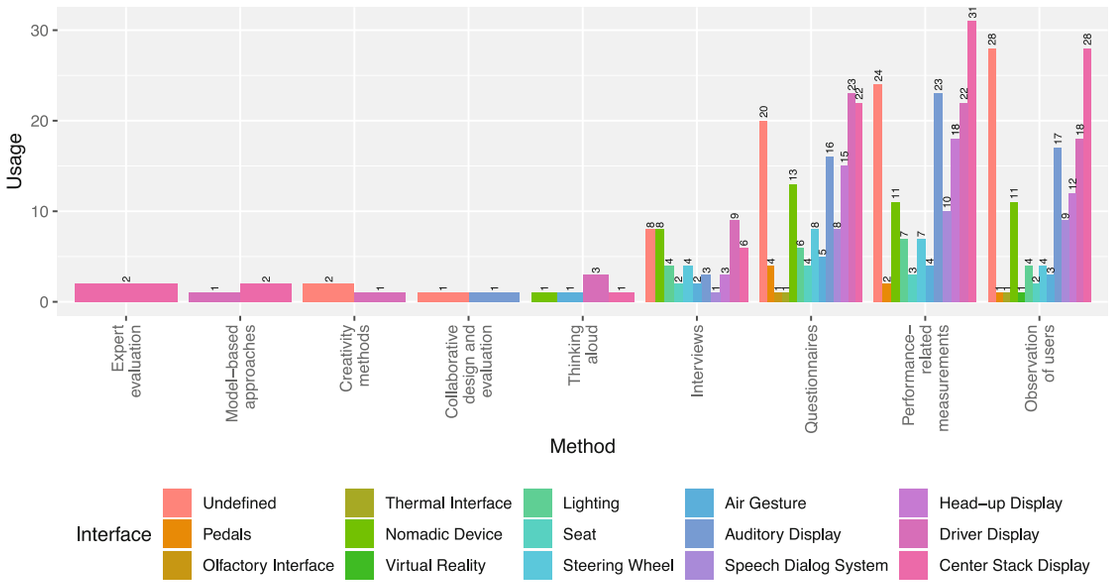


Figure 3.10.: usability evaluation method usage per interface type.

than three quarters of publications consult a single (39.91 %) or two different methods (43.95 %). The remaining 16.14 % of publications use three different UEMs.

In detail the three most commonly used UEMs are performance-related measurements (used by 54.71 % of publications), observation of users (51.12 %), and questionnaires (46.19 %). As Figure 3.10 might lead to the misinterpretation that the method of questionnaires is in total more often used than the observation of users, the absolute frequency for all methods is also listed in the previous section 3.4.1 in Table 3.2.

Figure 3.10 shows the usage of different UEMs for the investigation of different interface types. The overlap for different publications that investigate more than one interface type or consult more than one UEM is ignored in the visualization. For the display type *Center Stack Display* the most commonly used methods are from the category of performance-related measurements (31), followed by observation of users (28), and questionnaires (22). Five publications consult a combination of all of the three mentioned UEMs, eight papers use the combination of performance-related measurements and questionnaires, and a single publication uses the combination of performance-related measurements, questionnaires, and interviews. Looking at the category *Driver Display*, questionnaires are used by 23 publications and performance-related measurements are used by 22 publications, with a number of 13 overlapping publications. While four papers that investigate the interface type *Driver Display* use the combination of performance-related measurements and questionnaires, seven publications also use observation of users together with the mentioned combination. In total 18 publications that investigate the

interface type *Driver Display* consult observation of users as UEM. Besides this, the thinking aloud technique as well as interviews are combined in single publications with the combination of performance-related measurements and questionnaires. The third most commonly investigated interface type *Auditory Interface* is investigated through performance-related measurements in 23 publications, while 17 publications report the usage of observation of users and 16 publications use questionnaires. Here, two publications use a combination of all three mentioned methods, while eight papers use performance-related measurements and questionnaires, and a single paper uses performance-related measurements and questionnaires in combination with interviews.

Besides the three most commonly used methods, interviews are used in several publications to investigate different interface types, like *Driver Display* (9), *Nomadic Device* (8), *Center Stack Display*, *Steering Wheel*, or *Lighting*. The thinking aloud technique is used in three publication to investigate the interface type *Driver Display* as well as in single publications to investigate the respective interface types *Nomadic Device*, *Air Gesture*, and *Center Stack Display*. The two studies using methods from the category of model-based approaches to investigate the interface type *Center Stack Display* refer to applications of models like the keystroke-level model (KLM) to predict eyes-off-road times (Purucker et al., 2017) and critical path analysis (CPA) to predict task times (Harvey & Stanton, 2013, ch. 5). The paper investigating *Driver Display* through a model-based approach applies the SEEV (Saliency, Effort, Expectancy, and Value) model by Wickens et al. (2001) to predict human attention and the Human Efficiency Evaluator (HEE) to create models of attention allocation (Feuerstack et al., 2016).

These findings are also supported by the similarity calculation between different methods through the *Overlap Similarity* algorithm. The results reveal the incidence of the combination of different UEMs. As the similarity graph in Figure 3.11 shows, there are UEMs that are frequently used together. A very striking example is the combination of creativity methods and interviews with a similarity score of 1, as in all three publications that applied creativity methods the authors also applied interviews. Another high correlation exists between thinking aloud and interviews with a similarity score of .667. With a similarity score of .553, several publications using questionnaires also apply performance-related measurements in their research.

Finally the category of methods studied in this dissertation — *Expert evaluation* — is used by two publications to investigate the interface type *Center Stack Display*. One of the two papers investigating the display type *Center Stack Display* through expert evaluation additionally consults the usability evaluation method observation of users and interviews (similarity score .5). The other publication uses expert evaluation in combination with model-based approaches (similarity score .5).

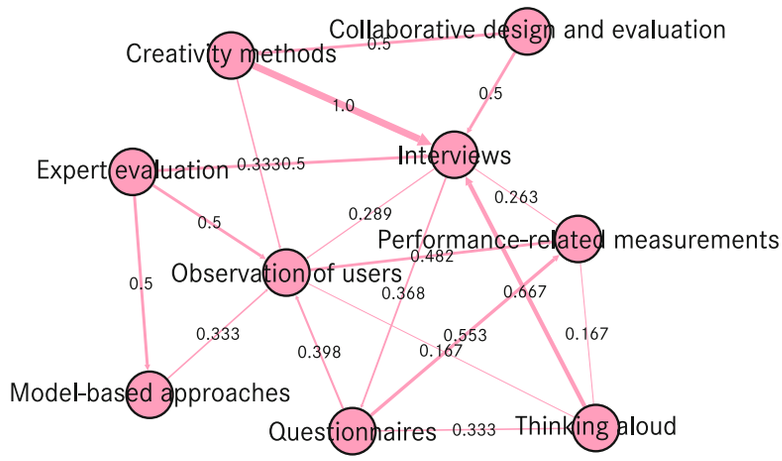


Figure 3.11.: Similarity relationships between UEMs. The edge thickness represents the similarity between the two UEMs.

3.5.5. Similarity Between Publications

The relationships between the nodes of type *Paper* and the types *Interface* and *Method* are used to calculate the similarity between individual publications. As in this literature review only the strongly similar publications are of interest, the threshold for the *Jaccard Similarity* algorithm is set to .75 (the mean for all publications is .301). Figure 3.12 shows the calculated communities of similar publications. The graph visualization shows several smaller disconnected clusters as well as larger groups of publications that are interconnected through one or more publications.

The three largest groups in Figure 3.13 revolve obviously around the most commonly used UEMs performance-related measurements, observation of users, and questionnaires. While in the biggest cluster with respect to the number of contained publications in Figure 3.13a the most commonly investigated interface type is *Center Stack Display*, the second and third biggest clusters in Figure 3.13b and Figure 3.13c mainly address the interface types *Driver Display* and *Head-up Display*. Another finding is that the cluster with ID 17 in Figure 3.13b also contains the UEM thinking aloud and the cluster with ID 30 in Figure 3.13c also contains interviews — with a single publication each. With respect to the results from the above sections 3.5.3 and 3.5.4, the clusters of similar publications support the findings regarding the most commonly investigated interface types and applied UEMs.

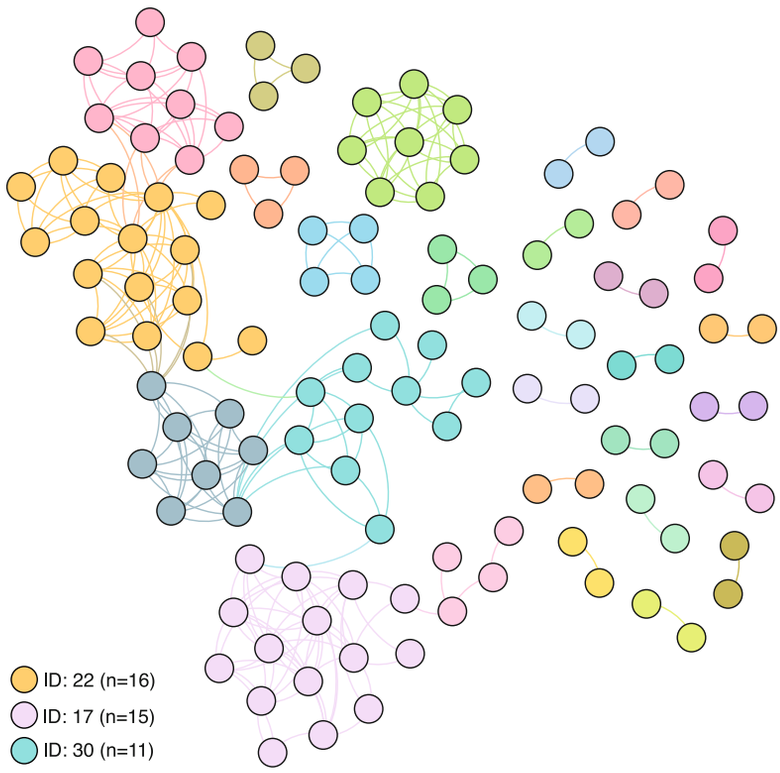
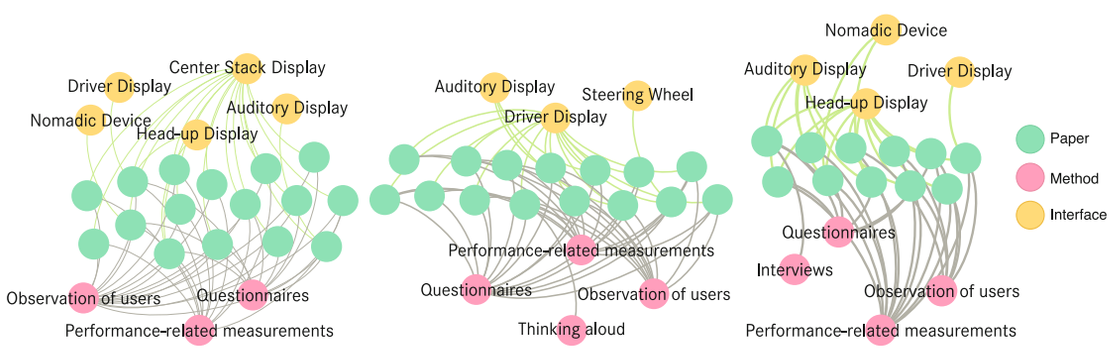


Figure 3.12.: Clusters of similar publications with regard to the applied usability evaluation methods and the investigated interface type. The different colors represent the different clusters detected by the *Louvain* algorithm.



(a) Cluster with ID 22. (b) Cluster with ID 17. (c) Cluster with ID 30.

Figure 3.13.: Top 3 clusters of similar publications with regard to the applied UEMs and the investigated interface type. The different colors represent the different node types.

3.6. General Discussion

The investigated dataset in this chapter shows a snapshot of the current research situation. Therefore, the results should be interpreted with caution but allow the indication of some trends.

The overall network statistics show that most authors contributed to a single publication during the time period of three years. The number of authors per paper ranges mainly between three and five as well as the collaborators per author, and in most publications authors from a single or two different institutions are involved. Compared to the research of Newman (2001) on scientific collaboration in the fields biomedical research, physics, and computer science, the papers per author ratio is higher for the research fields biomedical research and physics. Although Newman (2001) investigates a time period of five years, the papers per author ratio for the research area computer science is only slightly higher. When looking at the numbers of collaborators per author the result is most likely comparable to the discipline of computer science, while from the other research disciplines higher average numbers of collaborators are reported. For the specific research area of high-energy physics the average number of authors per paper is reported higher than in this literature study, while the other research areas of biomedical research, physics and computer science show slightly lower authors per paper ratios.

A look at the most productive authors shows that the colleagues in several institutions work together very closely. While most authors in the network are affiliated with institutions in Germany, the most productive institutions, when looking at the number of publications, are located in the UK, Austria and the USA. Besides finding productive collaborations between individual authors, the results also show relationships between institutions that work together on several publications.

The centrality measures from the field of social network analysis applied to the collaboration relationships between individual authors show detailed insights into the network. Especially the PageRank emphasizes the importance of authors like Gary Burnett from the University of Nottingham and Manfred Tscheligi from the University of Salzburg, who take a key role in their corresponding research community. The betweenness centrality measure shows that the community around Gary Burnett contains many individual authors acting as bridges between smaller sub-communities — including Neville Stanton, Joost de Winter, Catherine Harvey, Victoria Banks, and Klaus Bengler. Therefore, there exists an exchange between the University of Nottingham, the University of Southampton, the Delft University of Technology, and the Technical University of Munich. The results for harmonic closeness centrality show an overall trend of a widespread network.

This finding is also highlighted by the results of the applied community detection algorithms. The average community size of six authors with around two publications on average shows that the network consists of many smaller communities with few exceptions like the communities around Gary Burnett from the University of Nottingham, Manfred Tscheligi from the University of Salzburg, Martin Baumann from the Ulm University, Bryan Reimer from the Massachusetts Institute of Technology, and Joost de Winter from the Delft University of Technology. Except one difference where a single community in the results of the *Strongly Connected Components* algorithm is split into two communities in the results of the *Louvain* algorithm, both applied algorithms deliver similar results. Here Neville Stanton is awarded with the key role of bridging two of the biggest communities in the network.

Since human-machine interaction possibilities inside the car are no longer limited to steering wheel, accelerator, and brake, the number of interface types inside the vehicle is rapidly growing. Publications investigating more traditional interface types including *Center Stack Display*, *Driver Display*, and *Auditory Display* represent a little more than half of the publications in the network. Besides these, the interface types *Nomadic Device*, *Head-up Display*, and *Speech Dialog System* are investigated regularly, while also vehicle-specific interface types like *Steering Wheel*, *Seat*, and *Pedals* or innovative interaction techniques through interface types *Air Gesture* and *Lighting* are of research interest.

From a user-centered perspective it is not striking that the usage of methods with direct involvement of users prevails. Roche et al. (2014) notice a similar result in their survey. While the most commonly used methods are user testing, observation and interview, they also determine that method usage is highly dependent on the academic training, expertise (Coutaz and Balbo, 1994, as refered to in Roche et al., 2014), and the industry of the specific usability professional. Besides exploratory analysis, the application of *Overlap Similarity* supports the findings regarding frequently combined UEMs. All papers using creativity methods also use interviews and more than half of the publications applying thinking aloud also apply interviews. Furthermore, the application of another similarity measure — the *Jaccard Similarity* — allows to reveal similarity between papers based on common method usage or mutual investigated interface types. Analyzing the result data with the *Louvain* algorithm for community detection allows to create clusters of similar publications. Looking at the results it is not surprising that the largest clusters revolve around the three most commonly used UEMs as well as some of the most commonly investigated interface types.

The dataset used in this study is available under the open source license GNU General Public License Version 3 in different formats on *Mendeley Data* (Lamm, 2019). For usage inside the graph database software *Neo4j* the dataset includes a database dump as well as data in the *GraphML* interchange format. Moreover, the dataset provides a dump in

SQL format to use the data in relational databases as well as lists of entities and their connections as *CSV* and *JSON*.

3.7. Conclusions

The aim of the presented literature review in this chapter is to get an overview of the research landscape in human-computer interaction (HCI) for in-vehicle information system (IVIS). Through the exploratory approach using the graph database system *Neo4j*, a network of research collaboration could be obtained which allows analysis from a bird's eye perspective as well as in detail. General network analytics show similar results to the work by Newman (2001) considering the research data from the field of computer science regarding publications per author, collaborators per author, and authors per publication. The centrality measures degree centrality, betweenness centrality, harmonic closeness centrality and PageRank highlighted important authors in the network that play a key role for their corresponding community or act as bridge between different research communities and sub-communities. Besides individual authors, the literature review discovered several closed research communities through the graph algorithms *Louvain* and *Strongly Connected Components*. The information specific to HCI is the usage of different usability evaluation methods (UEMs). While the categories performance-related measures, observation of users, and questionnaires are the most popular for most of the investigated interface types, other techniques are used rarely. The most commonly investigated interface types were *Center Stack Display*, *Driver Display*, and *Auditory Display*. The application of similarity algorithms for combination of methods and similarity between publications regarding used UEMs and investigated interface types supported the findings and allowed detailed insights. The findings of this exploratory study highlight the need for research in the field of UEMs other than techniques with users involved. This is discussed in detail in the following chapter.

4. Expert Reviews for Automotive User Interfaces

As the literature review in the previous chapter shows, the usage of different usability evaluation methods (UEMs) is not evenly distributed. It is not striking that methods with the direct involvement of users are applied preferably, as it is about the *user* as the human, and the *system*, *application*, or *software* as the computer in human-computer interaction (HCI). It is obvious that the involvement of users is unavoidable in a human-centered design process. But there are various arguments for expert-based evaluation methods discussed in this chapter.

4.1. Background

Looking again to the previous chapter, especially in Figure 3.9 in section 3.5.4, not only the category of expert evaluation is kind of underrepresented. Taking the category of model-based approaches as an example, one could argue that they have the potential to measure the effectiveness of interaction not only in front of a computer, but also for the dual-task environment of a car. Studies by Green (1999), Pettitt et al. (2007), Schneegaß et al. (2011), and Purucker et al. (2014) show the extensive examination of model-based approaches for in-vehicle evaluation.

In his conference paper Green (1999) elaborates the already outlined method of the 15-second rule (cf. section 2.5.3), also standardized as recommended practice J2365 by the Society of Automotive Engineers (SAE). Besides glance-behavior studies to estimate the time a typical task interacting with an in-vehicle information system (IVIS) takes, Green pleads for the keystroke-level model (KLM). The model allows to predict execution times of tasks through the identification of a keystroke-level description for each goal and subgoal of the tasks. The overall task-time can then be calculated by adding up the estimated operation times for each keystroke. Table 4.1 lists the operator times for young as well as elder drivers (for ages 60 to 65) derived from a combination of the original operators by Card et al. (1980), specific operators in the automotive context (Manes et al.,

Code	Name	Description	Base Time (s)	Age-Adj. Time (s)
Rn	Reach near	from steering wheel to other parts of the wheel, stalks, or pods	0.31	0.56
Rf	Rach far	from steering wheel to center console	0.45	0.81
C1	Cursor once	press a cursor key once	0.80	1.44
C2	Cursor 2 times or more	time/keystroke for the second and each successive cursor keystroke	0.40	0.72
L1	Letter or space 1	press a letter or space key once	1.00	1.80
L2	Letter or space 2 times or more	time/keystroke for the second and each successive cursor keystroke	0.50	0.90
N1	Number once	press the letter or space key once	0.90	1.44
N2	Number 2 times or more	time/keystroke for the second and each successive number key	0.45	0.81
E	Enter	press the enter key	1.20	2.16
F	Function keys or shift	press the function keys or shift	1.20	2.16
M	Mental	time/mental operation	1.50	2.70
S	Search	search for something on the display	2.30	4.14
Rs	Response time of system-scroll	time to scroll one line	0.00	0.00
Rm	Response time of system-new menu	time for new menu to be painted	0.50	0.50

Table 4.1.: KLM operation times. The table is cited from Green (1999).

1998; Steinfeld et al., 1996), and the methods time measurement (MTM) system by Barnes (1968).

Pettitt et al. (2007) address the concern that visual attention in a dual-task scenario cannot be considered by predicting static task time through KLM. An "extended KLM technique" (Pettitt et al., 2007, p. 1518), using operator values from Card et al. (1983) and Green (1999), is used to additionally predict the total shutter open time (TSOT) as used in the occlusion technique as a measure of visual time required for completing a task. Therefore, the development of a KLM is divided in two stages — *1*) a traditional KLM and *2*) an assessment of the KLM against the vision/no-vision timeline of a trial with the occlusion technique. In a feasibility study (Pettitt et al., 2006) three assumptions of the degree of disruption the occlusion process causes were created (Pettitt et al., 2007, p. 1520):

1. During 1.5 second periods of vision the operator sequence can progress without interruption;
2. An operator that begins in a period of vision can continue into a two second occluded period *providing* it is not specifically associated with vision, for example, reading information from a display;
3. An operator can *only* begin in an occluded period when vision is not required at any point in its duration, for example a keystroke where the finger is already placed on the control in question.

In a comparative study between the predicted results and those observed in occlusion trials with users, strong correlations for TSOT (.93) and static task time (.98) could be measured. Therefore, the extended KLM technique holds the "potential as a useful first-pass design tool" (Pettitt et al., 2007, p. 1523) which delivers similar results to those raised by user trials with the occlusion technique, but without the requirement to run user trials.

Another somewhat underrepresented usability evaluation method, when looking at the research area of IVIS, is the think aloud (TA) technique. As already mentioned in section 2.4.5, the technique interferes with the measurement of human performance, for instance through measuring task completion times, time until failure, or average time to recover from errors. Concluding from the fact that in the literature review in chapter 3 performance-related measurements were the most commonly used methods in the automotive domain, it is not that surprising that TA is not used very often. Furthermore, in some cases the technique is maybe not reported particularly, because of its subjective character.

The category of automated evaluation on the other hand might be underrepresented because of technical and data security issues. While more technically oriented control units in the car, for example systems like an exhaust gas recirculation (EGR) system or the exhaust gas oxygen sensor (lambda probe) are logged and diagnosed with the on-board diagnostics (OBD), data loggers specific to HMI are not standardized in the automotive industry. Besides that, the logging of usage data of IVIS is associated with issues regarding data transfer from the vehicle to analyze it and possible restrictions according to data protection laws in different countries.

While expert review or inspection techniques are used successfully in different domains to reveal usability issues, according to the literature review in chapter 3 the method is not commonly used in the automotive domain. This is surprising since most of the following advantages for heuristic evaluation (HE) listed by Nielsen and Molich (1990) can be transfered to other techniques of the expert evaluation category (Nielsen & Molich, 1990, p. 255):

- It is cheap.
- It is intuitive and it is easy to motivate people to do it.
- It does not require advanced planning.
- It can be used early in the development process.

Furthermore, Sears (1997) summarizes the advantages of inspection techniques in the following quotation (Sears, 1997, p. 213):

> "Inspection-based evaluation techniques are popular because they require less formal training, are quick, can be used throughout the development process, do not require test users, and can result in finding numerous usability problems."

Tory and Möller (2005) even go a step further and state that "formal laboratory user studies might be inappropriate during an exploratory phase of research when clear objectives and variables might not yet be defined" (Tory & Möller, 2005, p. 8). Furthermore, most user interfaces for in-car usage are subject to strict secrecy before market launch which makes it much more complicated to test with real users. To avoid information leaks that expose parts of a newly developed information structure or design blueprints, the automotive industry needs evaluation approaches that do not involve real users. In order to investigate the suitability of expert review methods in this domain, the following sections illustrate related work on the field of expert review methods (cf. section 4.2) and define the scope of expert reviews for this investigation (cf. sections 4.3–4.5). The following chapters 5 and 6 showcase comparative studies between selected examples of expert review methods and applied usability studies with users as a baseline.

4.2. Related Work

When looking at expert reviews there is no way around two of the most popular operators and founders of heuristic evaluation (HE). Nielsen and Molich (1990) performed four different experiments with different user interfaces. While the studies were originally designed to show that there are around three to five non-expert reviewers needed to deliver comparable results to those raised by the authors, they also show applications of the method of HE in different user interfaces of different levels of maturity. In their first experiment 37 computer science students were asked to investigate ten screen dumps of a videotex system based on the heuristics introduced by Molich and Nielsen (1990). The results were then compared to those raised by the authors, which states in 52 usability problems in total. The 37 students found on average 51 % of the known usability problems, but when aggregated a random sample of ten reviewers was able to reveal 97 % of the problems on average. The second study used a written specification

of an information system for customers of a fictional telephone company, especially designed for the experiment. The 77 evaluators were recruited through a contest in a magazine for industrial computer professionals and found on average 38 % of the usability problems and on average 83 % of the issues for an aggregated random sample of ten evaluators. In experiments three and four 34 computer science students evaluated two different voice response information systems. The reviewers revealed on average 26 % (78 % for an aggregated random sample of ten evaluators) of the first and 20 % (71 % for an aggregated random sample of ten evaluators) of the second interface (Nielsen & Molich, 1990).

Frøkjær and Hornbæk (2008) performed three experiments using the metaphors of human thinking (MOT) technique in comparison with different methods — heuristic evaluation (HE), cognitive walkthrough (CW), and think aloud (TA) user testing. The first experiment comparing MOT and HE shows that both methods reveal an equal number of usability issues, while those found by MOT are categorized as "more serious, more complex to repair, and more likely to persist for expert users" (Frøkjær & Hornbæk, 2008, p. 14). The investigated student portal web application got reviewed by 87 computer science students (44 used MOT technique) identifying 12 % very critical and 52 % serious problems with MOT compared to 7 % very critical and 42 % serious problems with HE. In the second experiment — evaluating two e-commerce web sites — MOT was compared with CW since both techniques are derived from a psychological theoretical foundation. The 20 evaluators identified 31 % more usability problems with MOT compared to CW, while the problems found by MOT "had a wider coverage of a reference collection describing important and typical problems with e-commerce web sites" (Frøkjær & Hornbæk, 2008, p. 20). In a final study 58 participants evaluated a natural language interface in the form of a telephone dialog and a phonebook application on a mobile phone using two of the three methods MOT, CW, and TA. The results showed that MOT found more problems than CW and TA, and also TA reveals with 5 % (for the phone application) and 7 % (for the natural language interface) relatively few problems that were not identified by one of the other methods, when compared to MOT with 13 % and 7 %. Furthermore, Frøkjær and Hornbæk (2008) identified that the overlap of problems between techniques differs between the two investigated interfaces. In experiment two and three, the MOT technique was the preferred inspection technique, respectively (Frøkjær & Hornbæk, 2008).

In a study using a combined technique of HE, CW, and usability walkthrough — the so called heuristic walkthrough — Sears (1997) showed that the new technique is able to find more problems than CW and results in fewer false positives than HE. The evaluated system consisted of design documents of a visual e-learning application for rendering algorithms that contained already identified usability issues detected through user testing. None of the three applied evaluation methods — HE, CW, and heuristic

walkthrough — missed serious or intermediate problems already detected by user testing sessions. Applying the measures explained in section 4.5, resulted in HE appearing to be less valid than either of the other two methods, because of the high number of false positives from HE. Thoroughness generally increased with the number of reviewers for each method, but CW was not able to reveal the same number of intermediate and minor usability problems as the other methods. When it comes to reliability, heuristic walkthrough and HE appeared to be more reliable for small numbers of evaluators, while the difference decreased with an increasing number of evaluators (Sears, 1997).

Tory and Möller (2005) used HE to compare different visualization tools. For the specific context of use the authors used heuristics based on standard GUI heuristics, generic visualization tasks, and visualization tasks specific to their investigated tools. The approach focused on the direct comparison of two interfaces for each trial and provides valuable insight into usability problems. However, Tory and Möller (2005) also highlight that expert reviews "should not be used exclusively and should not replace user studies" (Tory & Möller, 2005, p. 11).

A summary of several studies investigating the effectiveness of UEMs is given by Andre (2000, pp. 59–61). For example, Desurvire et al. (1992) compare the problem detection for HE and CW in relation to the problems reported from a laboratory study with users. The expert review methods were conducted by three groups — human factors experts, non-experts, and system's designers — consisting of three evaluators each, while in the user study 18 subjects performed six different tasks. For the investigated telephone-based interface, the HE method performed better than the CW when using experts as evaluators. The experts reported 44 % of the problems uncovered by the laboratory study as well as 31 % of potential problems using the HE method. Applying the CW method the experts only detected 28 % of the problems from the user study, but also 31 % of the potential problems. The other evaluator groups system designers ($f_{HE,CW}$ = 16%) and non-experts ($f_{HE,CW}$ = 8%) performed worse than the experts for both expert review methods.

Another study by Doubleday et al. (1997) compared user testing with HE for the evaluation of an information retrieval interface. While the authors found that the HE method found 86 heuristic errors compared to 38 usability problems reported from the user testing. Yet, 39 % of the usability problems detected by user testing could not be identified through HE. According to the authors this is influenced by the applied heuristics and the evaluators expertise. Therefore, HE often leads to subjective reports and usability problems that are not distinct. Doubleday et al. (1997) argue for a combination of several UEMs to fully assess an interface.

While Karat et al. (1992) identified the most problems as well as a significant number of relatively severe problems through empirical testing which is also reported by Desurvire et al. (1991), Jeffries et al. (1991) observed that the most serious problems were identified through HE. Karat et al. (1992) suggest using empirical usability testing for baseline and checkpoint testing during the development cycle, whereas walkthroughs can draw on its strengths as a cost-effective alternative in early stages of development to support decisions between alternative designs.

An application of several UEMs to evaluate IVIS usability is presented by Harvey and Stanton (2013, ch. 5). The book section describes a case study investigating the methods hierarchical task analysis (HTA), critical path analysis (CPA), systematic human error reduction and prediction approach (SHERPA), Layout Analysis, and HE. The study describes a comparison of two IVIS, a touch screen as well as joystick operated system. As the aim of the study was to explore different analytic UEMs, the result showed that HTA was not useful for comparing IVIS but could be used as a starting point for CPA and SHERPA. While CPA was used as a measure of performance, SHERPA was applied to generate a comprehensive list of potential usability problems. Harvey and Stanton (2013) argue that CPA would require an extension in order to account for the dual task environment of the driving scenario. Furthermore, the SHERPA method lacks an assessment of error frequency and severity, while HE is not suited for comparison of different IVIS. Whereas, HE as well as Layout Analysis score through low training and application times. The authors highlight the "trade-off between subjectivity and focus on context-of-use" (Harvey & Stanton, 2013, p. 101).

Miniukovich et al. (2019) present a comparison study for the application of web readability guidelines collected through a literature review (Miniukovich et al., 2017), using the results of an eye-tracking experiment as the ground-truth for readability. The authors investigated the readability for several web pages and collected the ground-truth readability through eye-tracking data and subjective readability ratings. In a manual guideline evaluation 35 experts from different domains were asked to perform a online review of a subset of the web pages according to 39 readability guidelines. For another automatic evaluation Miniukovich et al. (2019) matched the list of guidelines to different metrics of readability features and text complexity. While the automatic approach was able to highlight several problematic aspects that were not rated consistently by the experts, the automatic evaluation had problems applying guidelines based on text content regarding understanding and interpretation.

4.3. Who is an Expert?

In order to set the scope of expert reviews for this dissertation, it is crucial to specify the term *expert*. Tory and Möller (2005) suggest to "choose usability experts with strong communication skills, experience conducting usability inspections, and experience with data display (not just usability)" (Tory & Möller, 2005, p. 10). Equivalent to the selection of users for a user study, the experts for an expert review have to be selected according to specified criteria. Since an expert review cannot simply be transfered to a series of user trials with experts as subject of the experiment, the experts cannot be selected according to the intended user groups.

There are several factors to consider during the selection of reviewers for an expert review. On the one hand a reviewer can be an expert regarding the applied review technique. For example, when applying a heuristic evaluation it is crucial to choose experts with knowledge of the employed heuristics or guidelines and at least some experience in the application of them in a heuristic or guideline review. Harley (2018) highlights this issue in the following quotation:

> "If you design in a vacuum and never see how the target audience interacts with your design, you don't build the kind of UX expertise that's needed to review interfaces." (Harley, 2018)

On the other hand reviewers with expertise in the specific domain and the context of use should be considered. Reviewers with an insight in the way people interact with the system have a profound understanding of the users goals as well as common tasks. Another issue to keep in mind is the so called "fresh perspective" (Harley, 2018) of a reviewer. People involved in the concept or design phase of the reviewed product are already biased and may give restrained feedback.

However, Burghardt (2014, p. 89) states in his dissertation that "a single evaluator is more likely to produce consistent results". Involving several evaluators would lead to more heterogeneous results also described as "evaluator effect" which was investigated in detail by Hertzum and Jacobsen (2001). It describes the phenomenon that multiple evaluators detect different sets of problems when applying the same UEM to the same interface. In order to analyze this phenomenon, eleven studies applying CW, HE and TA were reviewed with respect to the evaluator effect. Hertzum and Jacobsen (2001) report that the average agreement among any two evaluators ranges from 5 % to 65 % without an exception for any of the methods. The suggestion of the authors is to involve at least multiple evaluators in an expert review to overcome some of the issues accompanying with the evaluator effect. Besides this, an explicit analysis of goals of a

usability evaluation as well as a profound selection of tasks to test can help coping the effect (Hertzum & Jacobsen, 2001).

Since the application of UEMs often suffers from a substantial evaluator effect (Hertzum & Jacobsen, 2001), several studies (c.f. Hertzum & Jacobsen, 1999; Jacobsen et al., 1998; Nielsen, 1992) calculate the average detection rate which represents the number of problems detected by a single evaluator by the number of problems detected by all evaluators (Hertzum & Jacobsen, 2001):

$$\text{Detection rate} = \text{Average of } \frac{\mid P_i \mid}{\mid P_{all} \mid} \text{ over all n evaluators} \tag{4.1}$$

where P_i is the set of usability problems detected by a single evaluator, and P_{all} is the number of usability problems detected by all evaluators. According to Hertzum and Jacobsen (2001) this calculation has a weakness regarding the lower bound of the detection rate, as for a single evaluator the detection rate will always be 100 % and for two evaluators the detection rate will be at least 50 %, even when there is no overlap in the problem sets. Furthermore, the detection rate assumes that P_{all}, the number of usability problems detected by all evaluators, is identical to the number of existing usability problems P_n for an interface. However, in most cases an increase in the number of evaluators will also increase the probability to detect problems that would have been missed with a smaller number of evaluators. The calculation of the detection rate will cause a decrease in the individual evaluators performance which would normally represent an increase in the evaluators' collective performance. To overcome these difficulties with the detection rate, Hertzum and Jacobsen (2001) suggest the any-two agreement measure. It describes the number of problems detected by two evaluators commonly divided by the number of problems they detected collectively (Hertzum & Jacobsen, 2001):

$$\text{Any-two agreement} = \text{Average of } \frac{\mid P_i \cap P_j \mid}{\mid P_i \cup P_j \mid} \text{ over all } \tfrac{1}{2}\text{n(n-1) pairs of evaluators} \tag{4.2}$$

where P_i and P_j are the sets of problems detected by evaluator i and evaluator j and n is the number of evaluators. The resulting value ranges from 0 % to 100 % and represents the agreement of evaluators, rather than the effectiveness of the UEM.

As Burghardt (2014, p. 90) performs several reviews with a single evaluator, he refers to the concept of a "double expert" (Nielsen, 1992, p. 376) — an expert that is experienced in the domain the system belongs to as well as the usability evaluation technique. Taking into account the categories of users defined by Nielsen (1993) already mentioned in section 2.4.4, similar dimensions also apply to experts. The dimensions of reviewers expertise in Figure 4.1 adopt the dimension of domain knowledge from Nielsen (1993). The dimensions device and system experience from Nielsens model are unified in the

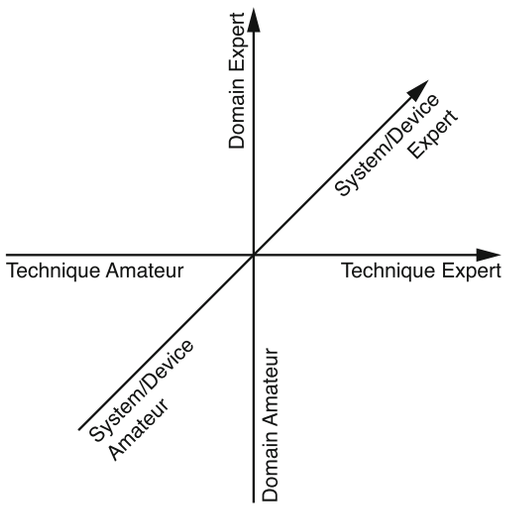


Figure 4.1.: Dimensions of reviewers expertise.

dimension of *System/Device Expert*, and the new dimension for the expertise regarding the review technique is added.

4.4. What is a Usability Problem?

During the discussion of different definitions of the term usability in section 2.1, several principles that constitute usability were discussed. A simple answer to the question, What is a usability problem? would be to look at violations of these principles. Manakhov and Ivanov (2016) mention three specific functions of an explicit definition of a usability problem. On the on side it helps an evaluator who is dealing with many observations to decide which of them are an actual usability problem. It allows to filter out problems that are not in the scope of HCI. On the other side a consistent analysis strategy and description format supports comparability. The problem sets of different studies with different UEMs can be compared when using a consistent definition of what a usability problem is. Finally, a common definition of a usability problem is necessary to guide attention to relevant aspects of data while teaching UEMs (Manakhov & Ivanov, 2016).

As most of the definitions from the literature do not cover all of these mentioned functions, Manakhov and Ivanov (2016) developed requirements of a consistent problem definition. Their requirements include the following features (Manakhov & Ivanov, 2016, p. 3146):

1. Include all HCI phenomena
2. Distinguish usability problems from problems in general
3. Distinguish a problem and its cause

4. Imply a relational position on usability
5. Distinguish a problem and a recommendation.

Taking these into account they review several definitions of the term usability problem. For example, the definition by Jeffries et al. (1991), "'anything that impacts ease of use — from a core dump to a misspelled word'" misses a possibility to filter problems out of the HCI-scope. This together with other definitions (Karat et al., 1992; Mack & Nielsen, 1994; Rubin & Chisnell, 2008) "lead to significant theoretical issues", according to Manakhov and Ivanov (2016). Furthermore, the theoretically plausible definitions of a usability problem (Kahn & Prail, 1994; Lavery et al., 1997; Mack & Montaniz, 1994) are described as not detailed enough or miss a description of certain aspects.

Manakhov and Ivanov (2016) themselves describe the term usability problem with the following definition:

> "A usability problem is a set of negative phenomena, such as user's inability to reach his/her goal, inefficient interaction and/or user's dissatisfaction, caused by a combination of user interface design factors and factors of usage context." (Manakhov & Ivanov, 2016, p. 3146)

A detailed consideration of the term "negative phenomena" results in the three non-overlapping classes from part 11 of ISO 9241 — effectiveness, efficiency, and satisfaction (International Organization for Standardization, 2018, p. 9). The cause of a usability problem, also outlined in Figure 4.2, is a "combination of user interface design factors and factors of a usage context" (Manakhov & Ivanov, 2016, p. 3148). The negative consequences deriving from usability problems are not an immediate component of a usability problem, but "are essential to estimate the severity of the usability problem" (Manakhov & Ivanov, 2016, p. 3148).

When talking about severity, the literature offers several approaches to classify usability problems by their severity. According to Nielsen (1995b), the severity of a usability problem is defined by three factors:

- The frequency with which the problem occurs.
- The impact of the problem if it occurs.
- The persistence of the problem.

Despite his definition of these three components, Nielsen (1995b) suggests the following rating scale to rate the severity of usability problems:

0 = I don't agree that this is a usability problem at all
1 = Cosmetic problem only: need to be fixed unless extra time is available on project
2 = Minor usability problem: fixing this should be given low priority

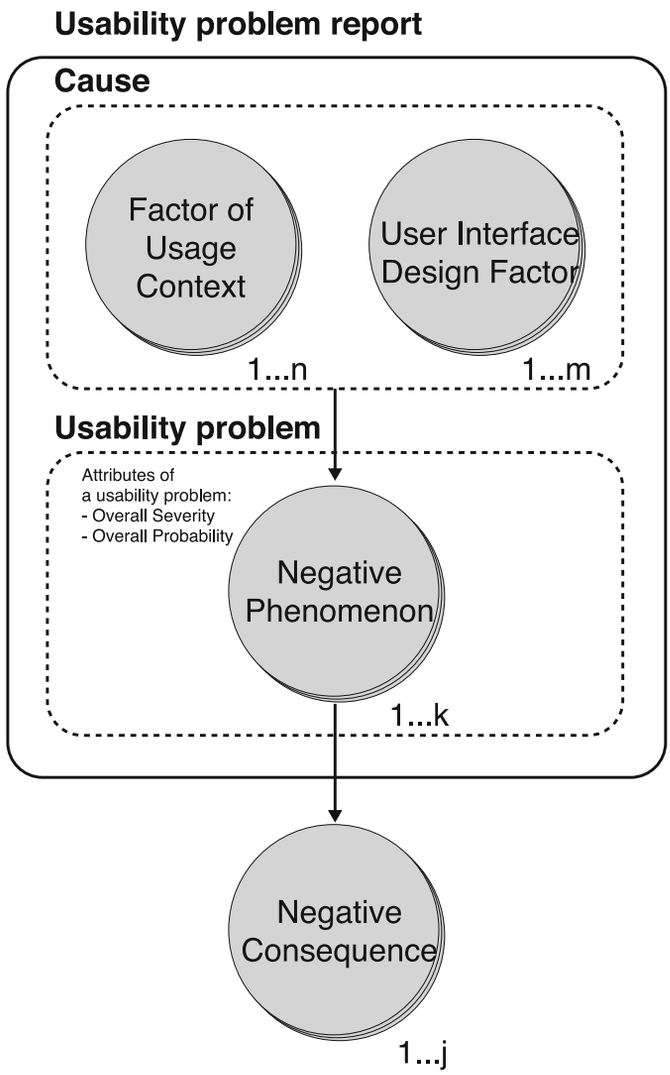


Figure 4.2.: A model of all components of a usability problem. The figure is based on Manakhov and Ivanov (2016).

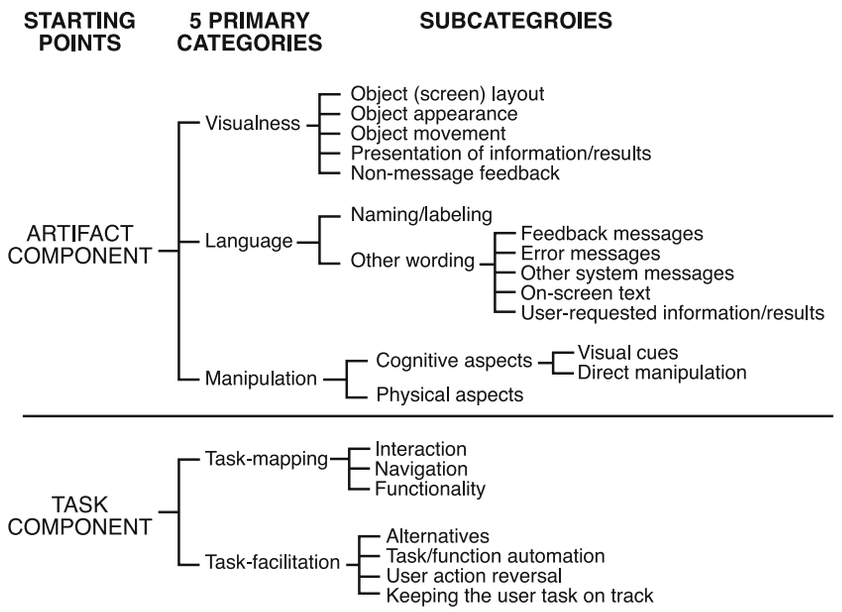


Figure 4.3.: The Usability Problem Taxonomy (UPT). The figure is based on Keenan et al. (1999).

3 = Major usability problem: important to fix, should be given high priority
4 = Usability catastrophe: imperative to fix this before product can be released

Besides other classification schemes based on severity or importance (Desurvire, 1994; J. S. Dumas & Redish, 1999; Rubin & Chisnell, 2008; Wilson & Coyne, 2001), Keenan (1996) introduces the Usability Problem Taxonomy (UPT). The classification schema in Figure 4.3 consists of five categories of usability problems: *Visualness*, *Language*, *Manipulation*, *Task-mapping*, and *Task-facilitation*. These are grouped into an artifact component and a task component, where each usability problem receives a classification in each of the components. A usability problem can be classified either to the deepest available level of subcategories (full classification), a category one or more levels above the deepest level (partial classification), or even no category at all (null classification) (Keenan et al., 1999).

Based on the UPT, van Rens (1997) enhanced the classifier resulting in different version of the Usability Problem Classifier (UPC). As Figure 4.4 shows, the third version of the classifier is extended with the temporal dimension of the usability problem's occurrence. The three categories before, during, and after are followed by the task attributes. The subsequent hierarchy level contains the object components specific to the temporal category (Andre et al., 2000; Andre, 2000).

A consistent definition of the term usability problem is inevitable especially when comparing different techniques regarding the detection of usability problems. For the

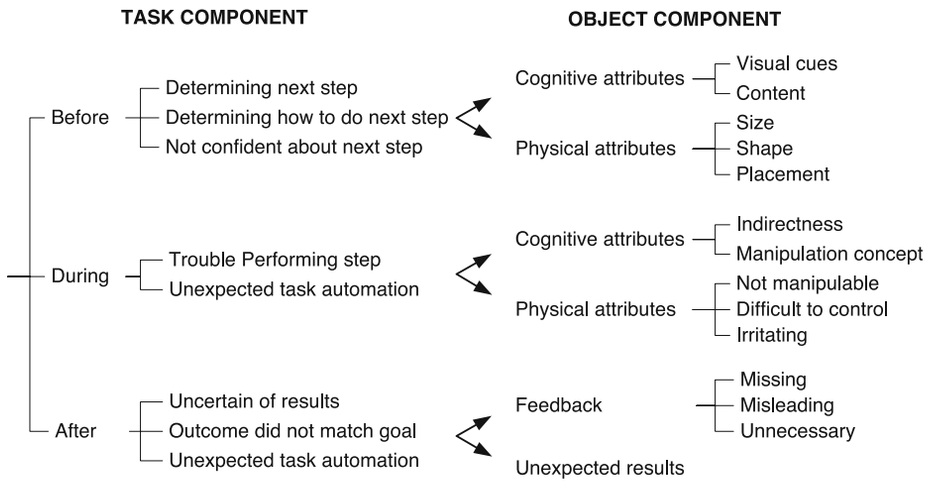


Figure 4.4.: The third version of the Usability Problem Classifier (UPC). The figure is based on Andre (2000) as well as Andre et al. (2000).

following case studies in chapters 5 and 6, the definition by Manakhov and Ivanov (2016) is used as it eliminates weaknesses of other definitions. Furthermore, the usability problems are classified according to the UPC, which is also used in the user action framework (Andre et al., 2001) to classify usability problems. Unlike other severity rating scales, the UPC uses a comprehensible theoretic foundation to classify usability problems based on their temporal occurrence, attributes regarding the task, and object components dealing with the objects interacted with by the user. The structured approach of the UPC aims to support the inspector moving the discussion from the feature itself to the effect on the user in the context of a task (Andre, 2000, p. 88).

4.5. Measuring the Effectiveness of Usability Evaluation Methods

In order to compare the results of expert review techniques to other UEMs, specific measures for effectiveness are used. Andre (2000) summarizes the three shortcomings responsible for the difficulty of comparing UEMs. On the one hand, there are no established standard criteria from the HCI literature, on the other hand the existing definitions, measures, and metrics are not standardized in a way to be used for researchers and practitioners. Moreover, there is no standard process for the evaluation and comparison of UEMs. In their article Gray and Salzman (1998) raised serious concerns with regards to the validity of five popular studies comparing different UEMs. It is questioned, whether the applied measures "really reflect sensitivity to usability" (Gray

& Salzman, 1998, p. 206) and therefore, whether "that what is being compared across UEMs is their ability to assess usability" (Gray & Salzman, 1998, p. 206).

Andre (2000) differentiates two different sets of criteria. While ultimate criteria are usually simple and direct, but nearly impossible to measure, actual criteria can be measured more easily, but can only be predictors of ultimate criteria.

> "The ultimate criterion for UEMs is how well they help inspectors discover problems that impact users in real work contexts." (Andre, 2000, p. 39)

For comparison studies the key concern is to design the correct actual criteria, while minimizing criterion deficiency by not leaving out important aspects as well as criterion contamination by moving the focus away from criteria that fail predicting the ultimate criterion (Andre, 2000).

According to Landauer (1995), user testing is used as a gold standard. Also Andre et al. (1999) propose user testing as the "standard yardstick" (Andre et al., 1999, p. 1090), to which UEMs should be compared. While the optimal setup would be "studying real workers doing real jobs in real environments" (Landauer, 1995, p. 281), in reality researchers have to deal with compromises and approximations with user-based studies in laboratories. Besides taking the usability problem set from user-based testing as part of the actual criterion (Andre, 2000), Sears (1997) suggest a union of all the individual usability problem sets, determined through the different methods compared. He addresses the characteristics validity, thoroughness, and reliability defined by Bastien and Scapin (1995). These three measures are commonly used to evaluate heuristics as mentioned by Lohmüller et al. (2019). Throughout the literature these three measures have been used to measure the effectiveness of UEMs (P. Böhm et al., 2014; V. Böhm et al., 2018; Hartson et al., 2003; Ling & Salvendy, 2005; Lohmüller et al., 2018; Meier et al., 2017; Schmargendorf et al., 2018; Sears, 1997).

The characteristic of validity describes the focus on issues that actually impact users. It is measured as "the ratio of 'real' usability problems identified to all issues identified as usability problems" (Sears, 1997, p. 214), defined in the following Equation 4.3:

$$Validity = \frac{Real\ Problems\ Found}{Issues\ Identified\ as\ Problems} \tag{4.3}$$

Hereby, the term "real problems" is based on the assumption that experts identify some issues as usability problems that are not actually problems. As Sears (1997) compares several expert-based techniques, his definition focuses on validity of these techniques over several evaluators. Another approach to measure validity of so called "inspection methods" is presented by Andre (2000). He defines validity as the ratio of the intersection

of usability problems found by the inspection method and the user test to the problems found only by user testing (Andre, 2000, p. 135):

$$\textit{Validity} = \frac{\mid P \cap A \mid}{\mid P \mid} \tag{4.4}$$

where P is defined as the set of usability problems detected by the inspection method and A represents the set of usability problems identified through user testing.

Besides measuring the validity of a technique, the thoroughness measures if the technique is able to identify all real usability problems in the system. He defines it as "the ratio of real problems that are identified to the number of problems that exist in the system" (Sears, 1997, p. 214):

$$\textit{Thoroughness} = \frac{\textit{Real Problems Found}}{\textit{Real Problems That Exist}} \tag{4.5}$$

As the number of existing problems in a system is almost impossible to determine, he uses an approximation of "the number of real problems identified by all participants plus the problems found during user testing" (Sears, 1997, p. 229). To overcome the issue that different techniques can identify problems with different severity, Andre (2000) suggests to use weighted counts of usability problems, as defined in Equation 4.6:

$$\textit{Weighted Thoroughness (by severity)} = \frac{\sum \mathrm{s}(\mathrm{rpf}_i)}{\sum \mathrm{s}(\mathrm{rpe}_i)} \tag{4.6}$$

where $\mathrm{s}(u)$ is the severity of usability problem u, rpf_i is the ith real problem found by the UEM, and rpe_i is the ith real problem that exists in the system. Equivalent to weighting by severity, the problems could also be weighted by frequency (see Equation 4.7, as often the most important problems are those experienced by a higher number of users (Andre, 2000):

$$\textit{Weighted Thoroughness (by frequency)} = \frac{\sum \mathrm{f}(\mathrm{rpf}_i)}{\sum \mathrm{f}(\mathrm{rpe}_i)} \tag{4.7}$$

where $\mathrm{f}(u)$ is the frequency of a usability problem. Another simple approach to thoroughness of a technique without weighting by severity or frequency is also described by Andre (2000) in the following equation:

$$\textit{Thoroughness} = \frac{\mid P \cap A \mid}{\mid A \mid} \tag{4.8}$$

where P is defined as the set of usability problems detected by the inspection method and A represents the set of usability problems identified through user testing. Furthermore, reliability is an important characteristic of a UEM. According to Sears (1997) it "implies that similar results should be obtained under similar conditions" (Sears, 1997, p. 215). He uses a measure whether different evaluators or rather techniques tend to find similar

numbers of usability problems, defined as "the ratio of the standard deviation of the number of problems found to the average number of problems found" (Sears, 1997, p. 215) in Equations 4.9 and 4.10.

$$R_{Temp} = 1 - \frac{\text{stdev}(\textit{Real Problems Found})}{\text{average}(\textit{Real Problems Found})} \tag{4.9}$$

$$\textit{Reliability} = \text{Maximum}(0, R_{Temp}) \tag{4.10}$$

As Gray and Salzman (1998) point out, the presented measures should not be used exclusively to compare the results of different UEMs. Rather a combination of measures are needed to predict the effectiveness of a UEM accurately.

> "For example, high thoroughness alone allows for inclusion of problems that are not real, and high validity alone allows real problems to be missed." (Andre, 2000, p. 51)

An issue beyond the detection of usability problems is the usability of the UEM itself. Mack and Nielsen (1994) argue for the ease of learning and appliance of the HE as its main motivation. Besides usability, this is also an issue of cost effectiveness, as many development organizations are not able to spend resources for training their personnel (Andre, 2000). Jeffries et al. (1991) propose a benefit-cost ratio calculated by the sum of severity scores related to the time spent on analysis. As the UPC does not provide severity scores, the measure could be slightly adapted to calculate the number of detected problems per person-hour.

4.6. Summary

The chapter discussed different aspects concerning the application of expert review techniques. While several publications address the comparison of user-based and expert-based evaluation methods, these are based on different comprehensions of the definition of experts, usability problems, and comparison metrics. The dimensions of users expertise by Nielsen (1993) can be adopted to classify experts according to their expertise in technique, domain, and system. These criteria for the selection of appropriate experts to evaluate a user interface are used for the case studies in the following chapters. Because of the range of different definitions for the term usability problem, Manakhov and Ivanov (2016) created requirements for a consistent definition and compared different definitions from the literature according to these requirements. For the mentioned case studies, the identified usability problems are classified according to the Usability Problem Classifier (UPC) as it is based on a comprehensible theoretic foundation. While cost effectiveness, ease of learning, and other secondary factors of usability evaluation methods (UEMs)

are difficult to quantify, Sears (1997) and Andre (2000) present reproducible definitions for the factors validity, thoroughness, and reliability. The comparison of results in the case studies of this thesis therefore focus on these quantifiable factors addressing the detection of usability problems.

5. Case Study: A Cognitive Walkthrough of a Driver Display Context Menu

In order to investigate the application of expert-based usability evaluation techniques in an automotive context, this chapter presents a case study comparing the cognitive walkthrough (CW) method with a user-based evaluation approach. The motivation for the work was to understand how the CW performs for the evaluation of an interaction concept in a driver display.

5.1. Background

The following section introduces the CW method and its background in a theory of exploratory learning. The procedure of a CW session is explained in detail presenting the four key questions of the CW technique. In order to adjust the CW to a dual task environment, a fifth question is added regarding the interaction while driving.

5.1.1. The Cognitive Walkthrough Method

As already broached in section 2.4.3, the CW method was originally introduced to evaluate the learning process for user interfaces. Based on the *CE+* theory of exploratory learning (Polson & Lewis, 1990), C. H. Lewis et al. (1990) generated a list of questions regarding the evaluated user interface.

In a first step, the group of users has to be defined. The more specific the description on background experience or technical knowledge, the more revealing the outcome would be. Moreover, the tasks that should be evaluated have to be defined as concrete and realistic as possible. These are then divided into sequences of simple actions (Wharton et al., 1994). For example, if the task is to log in on a website, these actions might include the following (The Interaction Design Foundation, 2017):

- Open browser
- Navigate to site

- Click login button
- Enter user name in user name field
- Enter password in password field
- Click the login button

Finally, for the preparation phase, the interface has to be defined. It must describe the responses to the earlier specified actions to complete the investigated task. While these information is available from an implementation or an interactive prototype, in earlier stages of a project, when no implementation exists, the actions could also be described by text. Regarding the detail of these information it is important to consider the users body of knowledge.

During the analysis phase, each action of the previously defined sequences is used to "tell a credible story as to why the expected users would choose that action" (Wharton et al., 1994, p. 8). These stories are based on the definition of the group of users and their background as well as their problem-solving process which is described by the *CE+* theory by Polson and Lewis (1990). The steps of this process are described briefly in Wharton et al. (1994, p. 8):

> "In brief, that problem-solving process holds that users: (1) start with a rough description of the task they want to accomplish, (2) explore the interface and select actions they think will accomplish the task (or some part of it), (3) observe the interface reactions to see if their actions had the desired effect, and (4) determine what action to take next."

In order to evaluate critical features, this theory is applied to the actions by asking the following four questions (Wharton et al., 1994, p. 9):

- Will the user try to achieve the right effect?
- Will the user notice that the correct action is available?
- Will the user associate the correct action with the effect they are trying to achieve?
- If the correct action is performed, will the user see that progress is being made toward solution of their task?

Will the user try to achieve the right effect?

With this question the cognitive walkthrough can identify flaws regarding the assumptions an interface is making about a user's level of experience or knowledge. On the other hand, discrepancies between the user's expectations of an action and the actual action taken by the user can be revealed. Wharton et al. (1994) use the following example: "Maybe their task is to print a document, but the first thing they have to do is select a printer. Will they know that they should be trying to get a printer selected?" (Wharton et al., 1994, p. 9). Another example described by The Interaction Design Foundation (2017) is a gun that fires by pushing a button on the side, while the trigger is used to

chamber a round. A user that is used to handling guns will try to pull the trigger, in order to fire the gun and will therefore be disappointed by the outcome.

Will the user notice that the correct action is available?

The second question addresses possibly hidden or obscured controls. Opening a file with a triple-click on the icon in a file explorer would be rarely discovered by the users, as a double-click is the familiar interaction. On the one side presenting every available control on the screen can overwhelm the user, on the other side commonly used controls buried in a menu system can be found hardly (The Interaction Design Foundation, 2017; Wharton et al., 1994).

Will the user associate the correct action with the effect they are trying to achieve?

This question is about the user's ability to connect the action to the desired outcome. Poorly labeled action using overly complex language or industry jargon make it hard for users to work out what is needed for the desired outcome. This also applies to complex actions like a key combination (e.g. Ctrl+Alt+Del for the task manager in Microsoft Windows) or combinations of different inputs (The Interaction Design Foundation, 2017).

If the correct action is performed, will the user see that progress is being made toward solution of their task?

The last question identifies missing, badly worded, easy to miss, or ambiguous feedback in an interface. Feedback is the only way to let your users know about their progress toward a task goal. Therefore, the term feedback is also stressed in usability heuristics (c.f. Nielsen, 1993; Shneiderman et al., 2018). Especially in dual task scenarios like in-vehicle interaction, the channel to transmit feedback is of special importance. Since the visual channel is most often blocked for the driving task, feedback should be transmitted through auditive or tactile channels to reduce driver distraction (c.f. Hulse et al., 1998, p. 13).

While dealing with these questions, it is important to document the interface flaws and peculiarities of the system usage. Rowley and Rhoades (1992) suggest to record the evaluation sessions on video with an additional approach for real time event logging synchronized with the video timer. This allows to review the sessions after the walkthrough is carried out and recap the assessor's issues and comments. Rieman et al. (1991) use a software tool on a computer to reduce overhead of printed task descriptions and questions for each action. Other forms include group visible materials like flip charts or overheads (Wharton et al., 1992) or paper-based forms (C. H. Lewis et al., 1991; Wharton et al., 1992; Wharton, 1992). For each of the defined actions of a task, Wharton et al. (1994) suggest to craft success or failure stories based on the user's background knowledge and the problem-solving process. These stories enriched by credibility arguments serve as a documentation of interface flaws and usability problems.

An exemplary application of the cognitive walkthrough method for IVIS could be found in a study by Mitsopoulos-Rubens et al. (2011). In the first phase of a combined study setup together with empirical user testing, the authors evaluate three different concepts for a music selection device using screen mockups on a computer together

with a rotary control. The four experts entrusted with the evaluation worked through a list of pre-defined song selection tasks documenting issues and navigation errors. In a follow-up meeting with all evaluators, the findings were discussed along with suggestions for design improvements. While Mitsopoulos-Rubens et al. (2011) simulated a driving task during the user testing, the dual-task environment was ignored in the cognitive walkthrough since the results served as a basis for interface improvements before the actual user test.

An earlier application of the cognitive walkthrough for a in-car navigation system by Curzon et al. (2002) showed that during a driving scenario users interact with the navigation system indirectly. When performing a road navigation task "seeing whether an action is available is generally outside the control of the navigation system itself" (Curzon et al., 2002, p. 40). Trying to overcome these limitations, the following section discusses an extension of the cognitive walkthrough method for automotive user interfaces.

5.1.2. Cognitive Walkthrough in Dual Task Environments

As discussed in section 2.5.1 — at least from the driver's perspective — a car is a dual task environment. The primary driving task therefore needs to be considered while performing a cognitive walkthrough for an IVIS. Basically, there are two different approaches regarding the cognitive walkthrough. On the one hand, it is possible to extend the four mentioned question with specific questions focusing on the dual task scenario of the driving task. Another possible solution is the integration of the importance of the dual task scenario in the briefing of the assessors.

In order to apply the cognitive walkthrough method more effectively regarding the coverage of the specific context of use, an additional question is added to the set of questions. The added question reads as follows: *Will the user be able to perform the action without being distracted from the driving task?* In more detail, the product should allow the user to interact with the system while driving. Therefore, the known guidelines on HCI for IVIS should be met. The following examples are presented to the evaluators to support dealing with the added question:

- At least one hand should remain on the steering wheel.
- The relevant information can be captured with just a few glances.
- The system does not demand time-critical interactions.
- The system does not impede the controls of the original driving task.

However, the definition of the group of users and their background knowledge also includes a concise description of the context of use, that should be considered by the assessors.

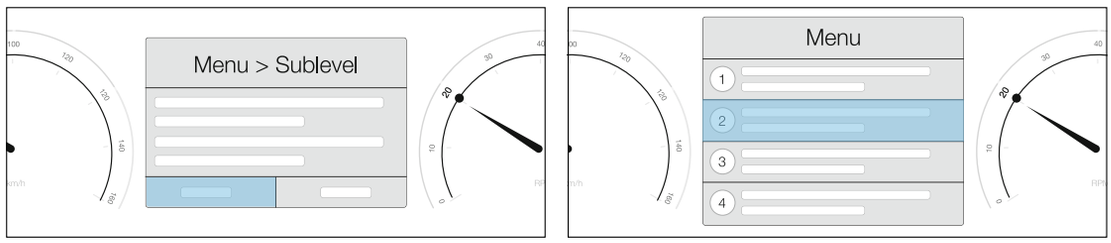
Since the given extension is not based on the theoretical foundation of the cognitive walkthrough, the *CE+* theory of exploratory learning, it actually does not fulfill the purpose of the method. Furthermore, the additional question has received little attention in its application. The investigators tended to consider the question a hindering factor and, based on their experience, reviewed the tasks autonomously in the context of the driving task. For this reasons, and because the extension has led to unsatisfactory results and unnecessary delay in implementation, it will not be discussed in later sections of this chapter.

5.2. Object of Investigation - The Context Menu Concept

Several factors play a role in the selection of the object of investigation. Important criteria are the maturity of the concept and its scope. As already explained in chapter 4, work from other disciplines has shown that expert-based methods can best demonstrate their strengths in early phases. For the cognitive walkthrough, only a subarea of an application (the context menu) was deliberately selected and thus also a reduction to certain forms of interaction. In addition, of course, there is a required practical interest to explore the concepts within a larger research project and the availability of the experts in the domain and method.

The investigated user interface concept shows two different approaches for a context menu in a driver display. The interface concept contains thematic screens for the instrument cluster display which is operated via a multi-functional controller on the steering wheel. The controller allows swipe operation in four directions with an additional button press operation for confirmation. While the up and down operations are used to select menu items and the additional button press to confirm a menu item, a separate key is used as a global back key.

Figure 5.1 shows scribbles demonstrating the differences. The first approach in Figure 5.1a shows the context-sensitive variation. The menu items in the context menu are limited to those relevant for the current visible screen. As an example, if the current screen shows the trip odometer the context menu shows an option to reset the kilometers, while a navigational screen shows options for the current route or for planning a route. On the other hand, Figure 5.1b shows the global option variation. Opening the context menu shows an overview of all available options for the entire instrument cluster system, grouped into several menu sections. Therefore, the driver can perform actions that are



(a) Context-sensitive context menu.

(b) Global context menu.

Figure 5.1.: Scribbles of the context menu concepts. The context-sensitive menu only lists options relevant for the current screen, the global context menu lists all available options.

not related to the current selected screen without changing the screen. On the other hand, the number of menu options and levels is increased which also increases the number of steps needed for specific actions. Both context menu variants offer advantages as well as disadvantages for specific use cases. The study should eliminate one variant in favor of the other.

5.3. Experiment 1 - User Study

To measure the performance of the CW method for the evaluation of an interaction concept in a driver display, a baseline is needed. As several publications (Andre et al., 1999; Landauer, 1995) propose, the results of the CW method are compared to the results of an empirical usability test. Therefore, the research question for this user study is how many usability problems the user study is able to reveal as well as their nature. This section introduces the applied method, the usability test procedure, and the results of the usability test.

5.3.1. Method

The experimental design of the comparative usability test of two different interaction concepts was within-subjects. The method consists of a combination of observation of users and the think aloud (TA) technique. The usability test was performed at the UI Studio of the Mercedes-Benz Technology Center in Sindelfingen.

Participants The participants were recruited via a department's mailing list. Therefore, and because of confidentiality issues, all participants were employees at Mercedes-Benz Research & Development from the department User Interaction & Software. In total 15 participants (including the pre-study participant) finished the experiment, eleven males and four females, with little to medium experience with the simulation mockup.

The age of the participants ranged from 21 to 57 years, mean age being 31.1 years (SD = 9.1). Three of the participants were 18 to 24 years old, eleven participants were 25 to 39 years old, and one participant was older than 55 years. All participants had a valid driver's license and drove between 2000 and 20,000 kilometers per year, mean kilometers per year being 13,866.7 (SD = 5527.4). The experiments were instructed and conducted in German and all participants understood and spoke German. The usability tests took place during the participant's work time.

Apparatus

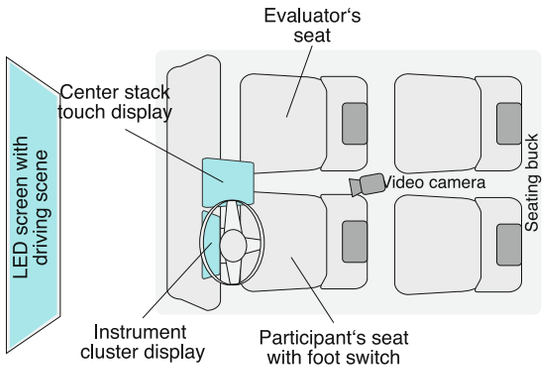
The lab, where the usability test was conducted, was equipped with a seating buck shown in Figure 5.2a. Besides seats for driver, front passenger and rear seat passengers, the seating buck contains a 12.3 inch (2.68 : 1 ratio) display for the instrument cluster and a 15.4 inch notebook covered to simulate a 12.9 inch (1.11 : 1 ratio) center stack touch display running a user interface simulation of the latest development state of the IVIS. The software simulation on the instrument cluster was rendered with Rightware Kanzi.[1] The simulation was not updated during the period where the tests were performed to ensure that every participant uses the same version of the simulation mockup. The seating buck was also equipped with non-functional pedals and a multifunctional steering wheel to control the instrument cluster simulation. The center stack simulation has not been investigated in this study, but was activated during the tests to enhance the experience of a real car cockpit. A 60 inches Sharp LC-60LE635E LED screen was used to display a non-interactive driving scene to simulate a dual task scenario. In the infinite loop video the brake lights of the leading vehicle were highlighted with visual markers. The participants were instructed to press a foot switch when these visual markers appear. Auditory signals indicated the success or failure of hitting the foot switch in time.

The context menu concepts were integrated in the instrument cluster simulation. To open and navigate inside the context menu, the participants used the touch control button on the multi-functional steering wheel. The button supported swipe gestures to toggle horizontal and vertical directions as well as pressing to confirm a selection. A separate hardware button allowed to go back in the menu tree or close the context menu. The user had to choose the correct menu item from the context menu in different tasks both during a simulated visual task and without a second task. The experiments were recorded on video including audio recording with a GoPro Hero 5 Black camera to be able to review sessions during analysis. Figure 5.2b shows a screenshot of a video recording.

Procedure

The sequence including the introduction, the exploratory phase, the secondary tasks, and the post-study questions was tested in a pretest with a single participant. Since the experiment design was not changed for the actual tests and the pre-study proceeded

[1] https://www.rightware.com/kanzi



(a) Setup of the seating buck and the visual driving simulation.



(b) A screenshot of a video recording showing the display positions and the steering wheel.

Figure 5.2.: Experimental setup of the seating buck with the position of the system displays and the visual driving simulation screen.

without issues, the data from the participant was analyzed together with the data of the actual test sessions.

The demographic data was collected in a pre-study interview. Besides age, gender, and job description, information about the participants main vehicle and the estimated mileage per year were collected. While the participant had to sit down on the driver seat, the experimenter took a seat on the front passenger seat of the seating buck. The participant was then given some time to explore the interface on their own. After a brief explanation of the basic interaction concept, the participants were instructed about the visual task. During a practice phase the participant should complete basic control tasks checking the understanding of the system and the dual task scenario.

After the introduction which took about 15 minutes, the participant was asked to complete several tasks arranged along a typical user journey using the first context menu variant. To minimize learning effects, the variants were counterbalanced. The user journey included three tasks without a secondary task (reset trip kilometers, check alerts, switch media source) and three tasks during a simulated drive (zoom in on navigational map, skip audio title, cancel route guidance). For each task, the participants performance was rated from *independent without errors*, *independent with searching/errors*, and *independent with help* to *much help needed*. Moreover, the participant was asked for a task rating — his subjective estimation regarding the number of operating steps and adequacy of menu levels on a scale consisting of *ok*, *acceptable*, *barely acceptable*, and *unacceptable*. Besides these measurements, the participant was encouraged to constantly express his thoughts about the interface via the TA technique. The experimenter noted down relevant statements from the participant to facilitate the review of the video material and reminded the participant to think aloud when necessary.

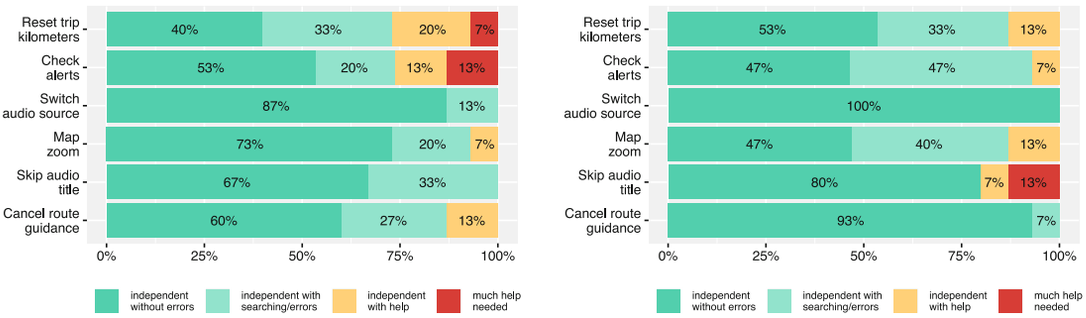
When the participant finished all tasks in the first context menu variant, the second menu variant was presented to the participant. To familiarize the participant with the changed context menu approach, another exploration phase offered the time to explore the interface. After completing all six tasks for each variant of the context menu, in a post-study interview, the participant was asked to answer summarizing questions regarding the comparison of both variants. Besides a list of pros and cons of the different variants, the participants should rate the variants on a ten point scale and mark their favorite variant.

Data Analysis

First, it was examined if the user behavior in terms of task performance was comparable under both variants. The task ratings were evaluated respectively, whether the result matches the observation from task performance. The context-sensitive approach can be considered as a baseline, as it is the traditional interaction concept in Mercedes-Benz cars. In a further step, the users' comments and TA video protocols were analyzed. The effectiveness of the secondary task — interacting with the context menu — was examined via the performance scale mentioned above. The experimenters subjective rating was verified by reviewing the video recording. For the task ratings, the participants' choices in the respective scale were analyzed.

Due to the relatively small sample and the ordinal scale of the data, as well as a violation of normal distribution of the paired differences calculated by the Shapiro-Wilk normality test (see Table B.1 in appendix B.1), the data was analyzed regarding positive correlation as proposed by Bortz and Schuster (2010, p. 125). The correlation analysis showed that most of the values for task performance as well as task ratings did not correlate significantly (see Table B.2 in appendix B.1). Therefore, the non-parametric alternative, the Wilcoxon signed rank test was used to compare the paired differences between task performance and task rating for the two context menu conditions.

Notes taken during the test from comments of the TA were used as a basis for the evaluation regarding usability problems of the interface. The list of problems was revised and complemented by reviewing the video recording and transcribing the usability problems using the categories of the Usability Problem Classifier (UPC). Thereby, all comments from the participant were collected in a list with some metadata, a description of the problem itself, and the classification according to the UPC. Table B.3 in appendix B.1 contains the complete list of usability problems. The list also contains multiple entries for a single problem, when the participant mentioned the problem multiple times, to emphasize the severity of the problem for the specific participant. To facilitate the analysis, the problems were assigned with a unique ID, which is repeated for repeated problem occurrence. From this list some statistical measures could be deduced, like the total number of usability problems, the total number of distinct usability problems,



(a) Context-sensitive menu condition.

(b) Global menu condition.

Figure 5.3.: Task performance distribution for both conditions context-sensitive and global context menu. Deviations to 100 % are due to rounding.

the total number of usability problems per participant, and the number of usability problems per classification category and between the context menu concepts.

5.3.2. Results

Task Performance

As Figure 5.3 illustrates, the task performance varied between the tasks as well as between the context menu variants. Investigating the context-sensitive menu variant in Figure 5.3a, the two tasks, switching the audio source (87 % *independent without errors*; 13 % *independent with searching/errors*) and skipping the audio title (67 % *independent without errors*; 33 % *independent with searching/errors*), cause only slight problems for the participants. However, during other tasks, setting the map zoom (7 %) and canceling route guidance (13 %), the participants sometimes needed help from the experimenter to solve the task. Here, for the map zoom task, 73 % had no problems at all and 20 % needed searching or made slight errors, while the task of canceling the route guidance lead to 60 % of participants with no problems and 27 % with slight errors or searching. Even more help is needed by the participants solving the tasks, resetting trip kilometers (20 % *independent with help*; 7 % *much help needed*) and checking alerts (13 % *independent with help*; 13 % *much help needed*). The task of resetting the trip kilometers was completed by 40 % of the participants without errors and 33 % had to search for the solution or made slight errors, while the task of checking alerts was completed without errors by 53 % of the participants and 20 % had to search or made minor errors. On average, 63.3 % of the participants completed the tasks without errors, 24.4% had to search or made slight errors, 8.9% needed help from the experimenter, and 3.3% needed much help to solve the tasks.

Looking at the results from the global menu variant in Figure 5.3b, the task of switching the audio source was completed by all participants without errors. Canceling the route

Task	V	z	p	r
Reset trip kilometers	16	-1.17	.24	.30
Check alerts	16	-0.75	.45	.19
Switch audio source	0	-0.94	.35	.24
Map zoom	33	-1.25	.21	.32
Skip audio title	18	-0.60	.55	.16
Cancel route guidance	0	-2.22	.03	.57

Note. V = sum of ranks assigned to the differences with positive sign; z = normally distributes z-score; r = Pearson's correlation coefficient as measure of effect size.

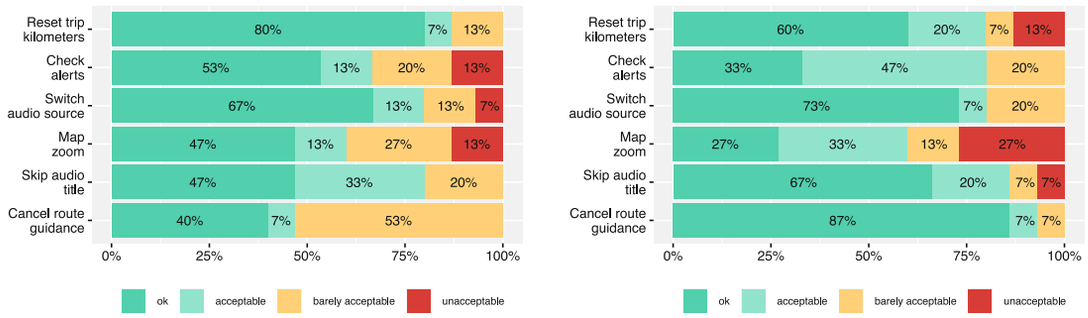
Table 5.1.: Differences in task performance between variants through Wilcoxon signed rank test.

guidance was completed by most of the participants (93 %) without errors and only a few participants (7 %) had to search for the solution or made minor errors. During the three tasks, reset trip kilometers (13 %), check alerts (7 %), and map zoom (13 %), some participants needed help from the experimenter, while the rest completed the tasks without errors (reset trip kilometers: 53 %; check alerts: 47 %; map zoom: 47 %) or with searching/errors (reset trip kilometers: 33 %; check alerts: 47 %; map zoom: 40 %). The only task for the global menu condition, where some participants (13 %) needed much help from the experimenter, was skipping the audio title. Another 80 % of participants completed the task without errors and 7 % needed slight help from the experimenter. On average, 70 % of the participants completed the tasks without errors, 21.1% had to search or made slight errors, 6.7% needed help from the experimenter, and 2.2% needed much help to solve the tasks.

Looking at Table 5.1, the differences between the task performance were only statistically significant for the task of canceling route guidance. As a result, the task performance for this specific task was significantly higher with the global context menu condition ($z = -2.22$, $p = .03$, $n = 15$). The effect size according to Cohen (1992) was $r = .57$ and therefore equals a large effect. The other task did not show significant differences in the task performance.

Task Rating

Beyond the task performance, the participants task ratings were analyzed regarding differences between the two context menu variants. As Figure 5.4 shows, the ratings between both conditions have slightly more differences than the task performance. While 80 % of the participants rated the task of resetting trip kilometers with the context-sensitive menu variant as *ok*, they same rating was applied only by 60 % of the participants to the global menu variant. By 7 % of the participants the task was rated as *acceptable* and 13 % of the participants rated the task with *barely acceptable* using the context-sensitive variant. Whereas, the global variant was rated by 20 % of the participants as *acceptable*, by 7 % as *barely acceptable*, and by 13 % as *unacceptable*. Using the context-specific menu



(a) Context-sensitive menu condition.

(b) Global menu condition.

Figure 5.4.: Task rating distribution for both conditions context-sensitive and global context menu. Deviations to 100 % are due to rounding.

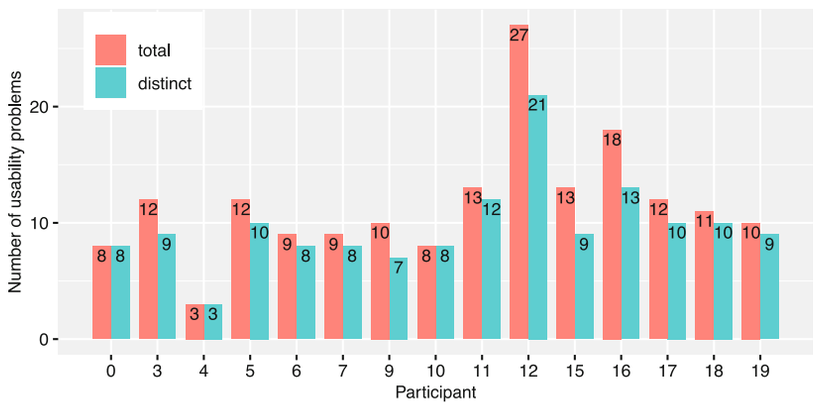
variant, 53 % rated the task of checking alerts as *ok*, 13 % as *acceptable*, 20 % as *barely acceptable*, and 13 % as *unacceptable*. The same task was rated by 33 % of the participants as *ok*, by 47 % as *acceptable*, and by 20 % as *barely acceptable* when using the global context menu variant. With 67 % of the participants rating as *ok*, 13 % each rating as *acceptable* and *barely acceptable*, and 7 % rating as *unacceptable*, the context-sensitive menu condition ranges slightly lower than the global menu condition (*ok*: 73 %, *acceptable*: 7 %, *barely acceptable*: 20 %) for the task of switching the audio source. The map zoom task was rated by 47 % of the participants as *ok*, by 13 % as *acceptable*, by 27 % as *barely acceptable*, and by 13 % as *unacceptable* for the context-specific condition. The same task using the global condition was rated by 27 % of the participants as *ok*, by 33 % as *acceptable*, by 13 % as *barely acceptable*, and by 27 % as *unacceptable*. While 47 % of the participants rated the task of skipping the audio title using the context-sensitive menu variant as *ok*, 33 % as *acceptable*, and 20 % as *barely acceptable*, the global menu variant was rated by 67 % of the participants as *ok*, by 20 % as *acceptable*, and by 7 % each as *barely acceptable* respectively *unacceptable*. For both conditions the cancel route guidance task was rated by 7 % of the participants as *acceptable*, while 40 % rated the task as *ok* for the context-sensitive condition and 87 % for the global condition. Respectively, 53 % of the participants rated the task as *barely acceptable* for the context-sensitive condition, and only 7 % for the global condition.

Equivalent to the task performance, the ratings were analyzed regarding differences between both conditions using the Wilcoxon signed rank test (see Table 5.2). Here again, the cancel route guidance task showed a significant difference. The rating using the global context menu variant was significantly higher compared to the context-sensitive menu condition ($z = -2.27$, $p = .02$, $n = 15$). With an effect size of $r = .59$ the difference showed a large effect according to Cohen (1992). As in the case of task performance, the differences in the task ratings did not show any further significant results.

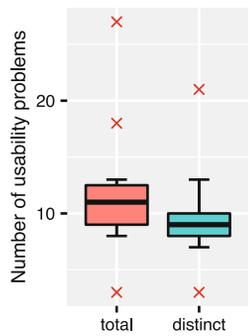
Task	V	z	p	r
Reset trip kilometers	21	-1.11	.26	.29
Check alerts	25	-0.21	.83	.05
Switch audio source	12	-0.26	.80	.07
Map zoom	52	-1.03	.30	.27
Skip audio title	8	-1.04	.30	.27
Cancel route guidance	5	-2.27	.02	.59

Note. V = sum of ranks assigned to the differences with positive sign; z = normally distributes z-score; r = Pearson's correlation coefficient as measure of effect size.

Table 5.2.: Differences between task ratings through Wilcoxon signed rank test.



(a) Histogram of the number of occurred usability problems per participant.



(b) Distribution of total and distinct number of usability problems.

Figure 5.5.: Number of usability problems.

Usability Problems

The experiments recorded a total number of 175 usability problems with a number of 63 distinct usability problems. On average each participant came across M_{total} = 11.7 in total and $M_{distinct}$ = 4.2 distinct usability problems. As Figure 5.5 shows, the number of total problems per participant is in most cases not significantly higher than the number of distinct usability problems. The participant with the least number of usability problems remarked three distinct usability problems, while the participant who came across the highest number of usability problems remarked 27 issue in total, with a distinct number of 21. The most frequently recorded usability problem was mentioned 12 times by 7 distinct participants, while the issue with the highest value regarding the number of distinct participants — as several usability problems were mentioned more than once by the same participant — was mentioned by 9 participants. In total 31 of the overall 63 distinct usability problems were mentioned more than once with an average of 2.1 mentions per distinct problem and participant. While the context-sensitive menu condition uncovered 29 distinct usability problems, for the global menu condition 34

issues were recorded. On average the participants came across 1.9 distinct issues using the context-sensitive menu respectively 2.3 distinct usability problems using the global context menu variant.

For the following distributions the total numbers of usability problems are taken into account, as the multiple occurrence of a single problem also gives an information about the severity. Looking at the categories of the UPC, 47.4 % of the usability problems were classified in the temporal category *Before* (n = 83), 38.3 % in the category *During* (n = 67), and 14.3 % in the category *After* (n = 25). The results described in the following paragraph are visualized in Figure 5.6 — deviations to 100 % are due to rounding. The detected issues in the temporal category *Before* split up almost evenly to the three subcategories *Determining next step* (38.6 %), *Determining how to do next step* (25.3 %), and *Not confident about next step* (36.1 %). From the task component classification *Before - Determining next step*, 43.8 % allot to the category *Cognitive attributes* of the object component classification, 56.3 % to the category *Physical attributes*. The distribution for the classification *Determining how to do next step* is 52.4 % for *Cognitive attributes* and 47.6 %, while usability problems that occurred in the category *Not confident about next step* were further classified by 46.7 % in the object component category *Cognitive attributes* and by 53.3 % in the category *Physical attributes*. The category *Cognitive attributes* deriving from the classification *Before - Determining next step* splits up into *Visual cues* (21.4 %) and *Content* (78.6 %), while the usability problems deriving from the category *Physical attributes* were all further classified into the category *Placement* The usability problems deriving from the category *Before - Determining how to do next step* were further classified into *Shape* (40 %) and *Placement* (60 %) in the category *Physical attributes* and into *Visual cues* (63.6 %) and *Content* (36.4 %) in the category *Cognitive attributes*. The remaining subcategory *Not confident about next step* is further classified into *Visual cues* (28.6 %) and *Content* (71.4 %) for the category *Cognitive attributes* and into *Size* (43.8 %) and *Placement* (56.3 %) for the category *Physical attributes*. The remaining subcategory *Shape* inside the category *Physical attributes* of the object component is not used for problems classified in the category *Before - Not confident about next step*. Looking at the task component category *During*, 88.1 % were classified as *Trouble performing next step* with 78 % in *Cognitive attributes* and 22 % in *Physical attributes*, while 11.9 % of the usability problems in the category *During* were classified as *Unexpected task automation* with a further classification into the categories *Cognitive attributes - Indirectness* (87.5 %) and *Physical attributes - Irritating* (12.5 %). The *Cognitive attributes* of the usability problems classified as *Trouble performing next step* split up into *Indirectness* (41.3 %) and *Manipulation concept* (58.7 %), while the *Physical attributes* split up into *Difficult to control* (69.2 %) and *Irritating* (30.8 %). The usability problems classified in the temporal category *After* were further classified into the categories *Uncertain of result* (48 %) and *Unexpected task automation* (52 %). The remaining category *Outcome did not match goal* was not used for

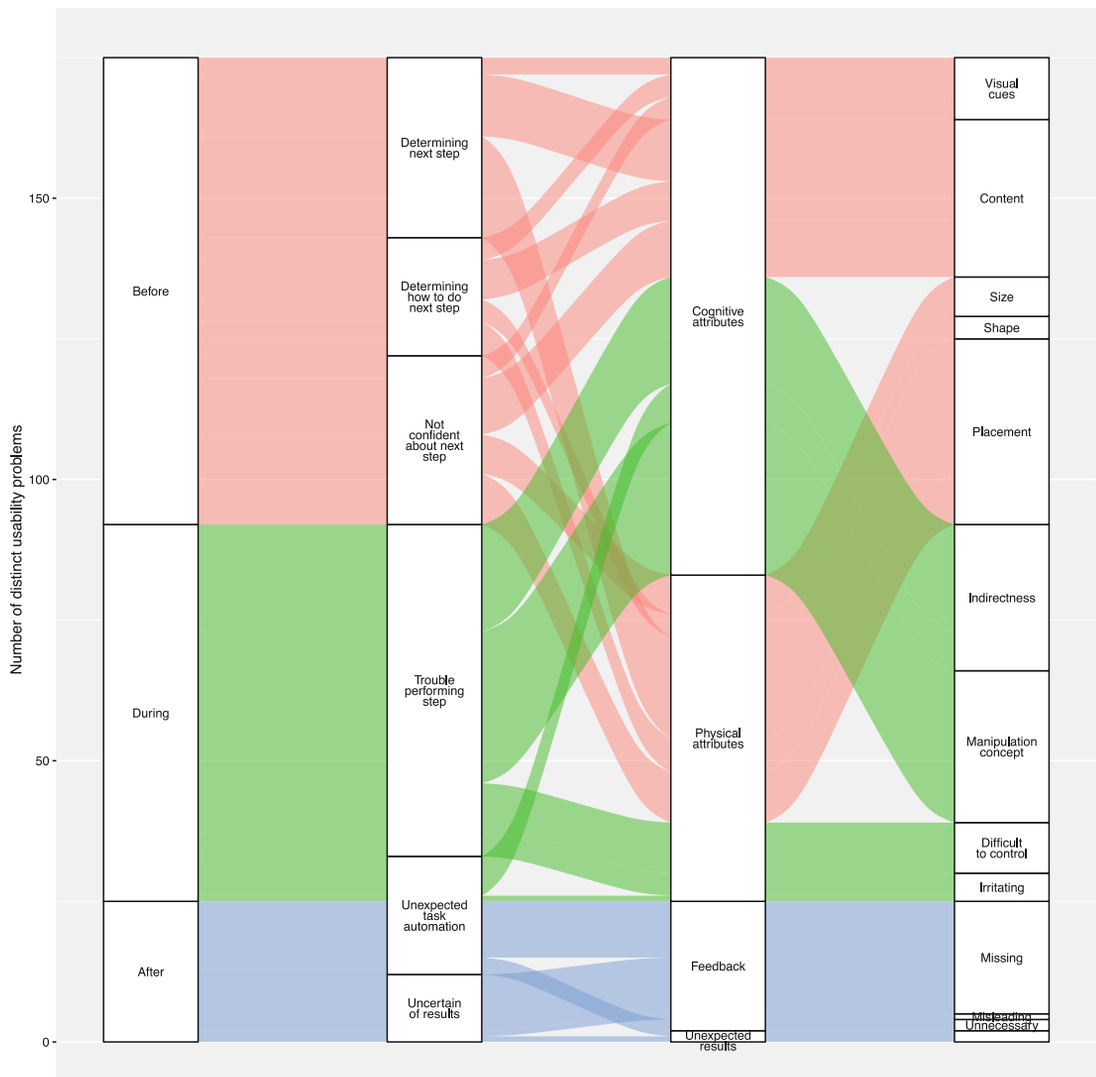


Figure 5.6.: The usability problem classification distribution for the user study using the UPC.

classification at all. Based on the category *Uncertain of result*, 91.7 % of the problems were further classified into the object component category *Feedback* which is separated into the categories *Missing* (90.9 %) and *Misleading* (9.1 %), while the remaining 8.3 % were classified into the category *Unexpected result* with no further separation. Those problems from the category *Unexpected task automation* in the temporal category *After*, were further classified by 92.3 % into the *Feedback* category with a proportion of 83.3 % for *Missing* and 16.7 % for *Unnecessary*, while the remaining 7.7 % fell into the category *Unexpected result*.

5.3.3. Discussion

As the user study results of the presented case study only serve as a benchmark for the results derived from the CW sessions, the discussion in this section is limited to information relevant for comparison of the two methods. Because the rating scale for the task performance was filled in by the experimenter it represents only a reasonably objective assessment of the actual performance. While the first two categories *independent without errors* and *independent with searching/errors* can be considered as adequate, the latter two *independent with help* and *much help needed* represent an insufficient result. Looking at the results from a bird's eye view, most of the participants had no severe problems using both interface variants. This is also supported by the results for task ratings, where most of the participants felt comfortable with using the interface, considering some exceptions. While most of the issues in the context-sensitive menu variant regarding task performance appeared during the first two use cases *Reset trip kilometer* and *Check alerts*, the global menu variant struggled with the task of skipping the audio title. Besides the first hurdle when using the system for the first time, the identification of the trip meter object is responsible for the setback in the task performance. The warning memory which should be found in the second task *Check alerts* is not supposed on the correspondent screen by all participants, therefore it took a while to find the correct way to solve the task. Other issues regarding the warning memory like the scroll direction and missing visual cues for interaction are not specific to the context-sensitive condition, but should be considered when looking at the task performance in general. Using the title skip function in the global context menu leads to problems for some users as the scroll direction for the next and previous title is not obvious from the design of the menu. Furthermore, users had to search in the menu tree to find the necessary option and did not get feedback from the system of a successful change of the audio title.

Looking at the task ratings, the category *ok* can be interpreted as "working without limitations", while the category *acceptable* represents a good rating with some limitations and the categories *barely acceptable* and *unacceptable* represent insufficient ratings During the evaluation of the context-sensitive menu variant, most of the critical ratings occurred

for the use cases *cancel route guidance*, *map zoom*, and *check alerts*. For the task of canceling the route guidance, the issue is twofold. On the one side, the participants had to change to a specific screen to select the option. On the other side, some participants were confused that there is no context menu at all in the screen from where the task was started. The problems recorded performing the *map zoom* task in the context-specific variant are not specific to the variant, while using the global menu condition, some of the problems address the menu depth in general as well as the step to select the appropriate menu item from the first level though the user already selected the appropriate main screen. Completing the *check alerts* task using the context-sensitive menu variant, the participants again reported the fact that they had to change to a specific screen to select the appropriate action.

The findings show that more than half of the detected usability problems occurred before the user executed a specific action. A similar finding is reported by Cuomo and Bowen (1992, 1994) as well as Andre (2000). As the selected participants were not trained in using the user interface, this result is not surprising. According to Andre (2000, p. 96), "for new users, the first translation [from an intention to an action description] for a task is generally the most difficult". Compared to the share of the two categories *Cognitive attributes* and *Physical attributes*, the number of usability problems classified in the object component category *Feedback* is relatively low. Furthermore, the distribution among the subcategories shows that most of the problems regarding feedback originated from missing user feedback, while only a few feedback messages are misleading or unnecessary. Most of the problems inside the category *Physical attributes* arose from issues regarding the *Placement* of the object of interaction, *Shape* and *Size* were hardly reported as the reason of usability issues. The object component category *Unexpected result* is even less represented than the other categories on the first level of the object component, which indicates that most of the participants were rarely surprised by the outcome of a specific action during a task.

5.4. Experiment 2 - Expert Review

The following sections present the expert-based approach of the comparison case study investigating a driver display context menu. The research question of the study results mainly from the subsequent comparison of both methods to identify the differences and similarities. These relate primarily to the nature of the usability problems and the resulting effectiveness of the expert-based method.

5.4.1. Method

The expert-based study presented in the following sections used the cognitive walkthrough (CW) method to identify usability problems in the interface variants. During the review the evaluators analyzed the interface variants presented in section 5.2 displayed on a prototype. The reviews took place in a conventional meeting room at the Mercedes-Benz Technology Center in Sindelfingen.

Experts The experts were selected according to the dimensions of Figure 4.1 in section 4.3. Four experts with different experience in domain, system and technique were selected for the experiment. One expert — a PhD student with a masters degree in media informatics — already had experience using the CW technique. Two interaction designers were recruited as domain experts, as both had been working in interface design in the automobile industry for more than three years. Another expert with focus on the system expertise was represented by an interaction designer who worked on a previous version of the interface. All four experts were employees of Mercedes-Benz in Sindelfingen.

Apparatus Unlike the user study which took place in a seating buck, the CW sessions were executed in a conventional meeting room using a prototype on a 2017 Apple iPad Pro (10.5 inch). The prototype contained an area for the instrument cluster display as well as the controls on the steering wheel below. It was created using the software Axure RP Pro 8[2] as well as Apache Cordova[3] to generate an iOS wrapper for the prototype. The prototype was operated with the integrated on-screen controls and supported the predetermined actions used for the given tasks. Moreover the prototype application contained a button to switch between the interface variants.

Additionally, the evaluators were given several material for the CW after the introduction and briefing. An overview of the target audience in the form of personae should help the evaluators to keep the actual users in mind. A list of the five questions asked including detailed descriptions and examples for each action which is already explained in section 5.1 was handed out to facilitate the review process. Furthermore, a printed figure of the UPC was used to give an overview of the classification of the usability problems.

Procedure After arriving and welcoming the expert at the meeting room, a short briefing introduced specific personae, the context of use, and the extension of the CW technique. The personae were created by a partner department and describe typical customers for two segments of Mercedes-Benz passenger cars. Furthermore, both interface variants were discussed regarding the similarities and differences.

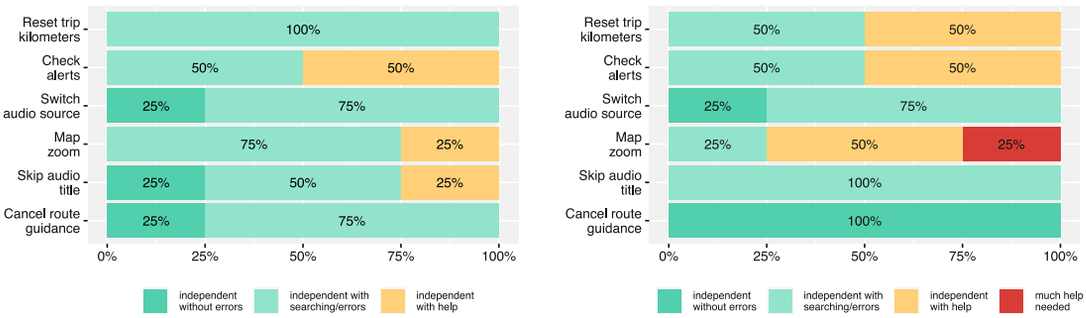
[2] https://www.axure.com/
[3] https://cordova.apache.org/

Throughout the session, the experimenter was present to guide through the CW and to support the expert in unclear situations as needed. Starting with the CW, the evaluator was presented with the six different use cases, already used for the user study — reset trip kilometers, check alerts, switch media source, zoom in on navigational map, skip audio title, and cancel route guidance. The order of the interface variants was counter-balanced between the experts, whereas each expert reviewed both variants. For each use case, the evaluator was asked to perform the necessary actions on the prototype and answer the five questions for each action. Arising issues where classified according to the UPC by the expert and documented by the experimenter. After each use case, the expert was asked how well the specified target users would be able to perform the task. In order to create comparable results to the user study, the same scale was used for the expert review. Therefore the estimated task performance was rated from *independent without errors*, *independent with searching/errors*, and *independent with help* to *much help needed*.

In a debriefing session with all four experts, the identified usability problems were discussed and consolidated in cases where similar problems were classified differently. Furthermore, the experts were invited to discuss several problems in the group and compare their results. The inspection itself took about two and a half hours including the briefing at the beginning per session with additional two hours for the debriefing session.

Data Analysis

The task performance rating collected from the user study was also estimated by the experts for each task. Therefore, the estimation raised by the experts can be compared to the findings from the user study. For each individual problem a unique ID was assigned together with metadata describing the problem. Table B.4 in appendix B.2 contains the complete list of usability problems. The statistical measures deduced from the list of usability problems include the total number of usability problems, the total number of distinct usability problems, the total number of usability problems per expert, and the number of usability problems per classification category and between the configuration interface variants. In order to compare the usability problem sets detected by the individual evaluators the measures average detection rate and average any-two agreement were used. While the detection rate is calculated by the number of problems detected by a single evaluator divided by the number of problems detected by all evaluators collectively, the any-two agreement is calculated by the number of overlapping problems two evaluators detected divided by the total number of problems they detected collectively.



(a) Context-sensitive menu condition.

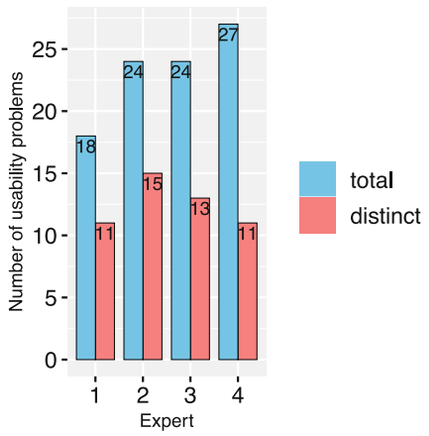
(b) Global menu condition.

Figure 5.7.: Estimation of the task performance for both context menu conditions context-sensitive and global.

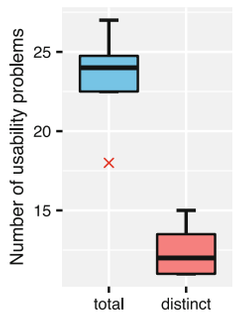
5.4.2. Results

Task Performance

While the task performance during the user study was collected through an estimation from the investigator, the same scale was applied to the task performance estimation from the experts during the CW. Figure 5.7 shows that the task performance estimation varied between both context menu conditions. While using the context-sensitive menu condition no expert assumed that the users will need much help using the menu, a single expert predicted a poor task performance for the *Map zoom* task using the global menu variant (25 % *independent with searching/errors*; 50 % *independent with help*; 25 % *much help needed*). On the other hand, for the context-sensitive menu condition performing the map zoom task, most of the experts predicted a task performance including searching and small mistakes (75 %) and a single expert rated with *independent with help*. In the context-sensitive condition, all four experts assumed that the users will have to search or make small mistakes along the way. The second task of checking the error memory is rated equally for both menu conditions, with a potential task performance split into *independent with searching/errors* (50 %) and *independent with help* (50 %), as well as the task of switching the audio source (25 % *independent without errors*; 75 % *independent with searching/errors*). The ratings for the task of skipping the audio title differ for the context-sensitive (25 % *independent without errors*; 50 % *independent with searching/errors*; 25 % *independent with help*) and the global menu condition, where all four experts predicted a task performance including searching and small mistakes. While the task of canceling the route guidance is assumed to raise no problems at all using the global menu condition, most of the experts predicted, that users might need searching or make small mistakes along the way (25 % *independent without errors*; 75 % *independent with searching/errors*). A statistical analysis using the Wilcoxon signed rank test as applied in the user study results of the task performance did not make much sense for the expert review data due to the small sample size of $N = 4$.



(a) Histogram of the number of occurred usability problems per expert.



(b) Distribution of total and distinct number of usability problems.

Figure 5.8.: Number of usability problems.

Usability Problems

The CW recorded a total number of 93 usability problems with a number of 29 distinct usability problems. Each expert detected on average M_{total} = 23.3 and $M_{distinct}$ = 7.3 usability problems. In most cases several usability problems were reported multiple times for different tasks and variants, therefore the number of total problems in Figure 5.8a lies in all cases above the number of distinct usability problems. However, the expert with the highest number of distinct usability problems (n = 15) did not report the highest number of total usability problems. The least number of distinct usability problems detected by two experts was 11. The most frequently recorded usability problem was detected 16 times and was found by all four experts. In total 12 usability problems were detected by more than a single evaluator. While the context-sensitive menu condition uncovered 7 distinct usability problems (n_{total} = 32), using the global menu variant the experts identified 22 distinct usability problems (n_{total} = 61). On average the experts detected 1.8 distinct usability problems using the context-sensitive context menu and respectively 5.5 usability problems using the global context menu.

For the following distributions among the categories of the UPC the total numbers of usability problems are taken into account, as the multiple occurrence of a single problem also gives an information about the severity. Looking at the list of usability problems, 53.8 % (n = 50) of the issues were classified in the task component category *Before*, 16.1 % (n = 15) fall into the category *During*, and 30.1 % (n = 28) fall into the category *After*. The results described in the following paragraph are visualized in Figure 5.9 — deviations to 100 % are due to rounding. The raised issues from the category *Before* split up into the three subcategories *Determining next step* (40 %), *Determining how to do next step* (30 %), and *Not confident about next step* (30 %). The ratio of problems from the classification

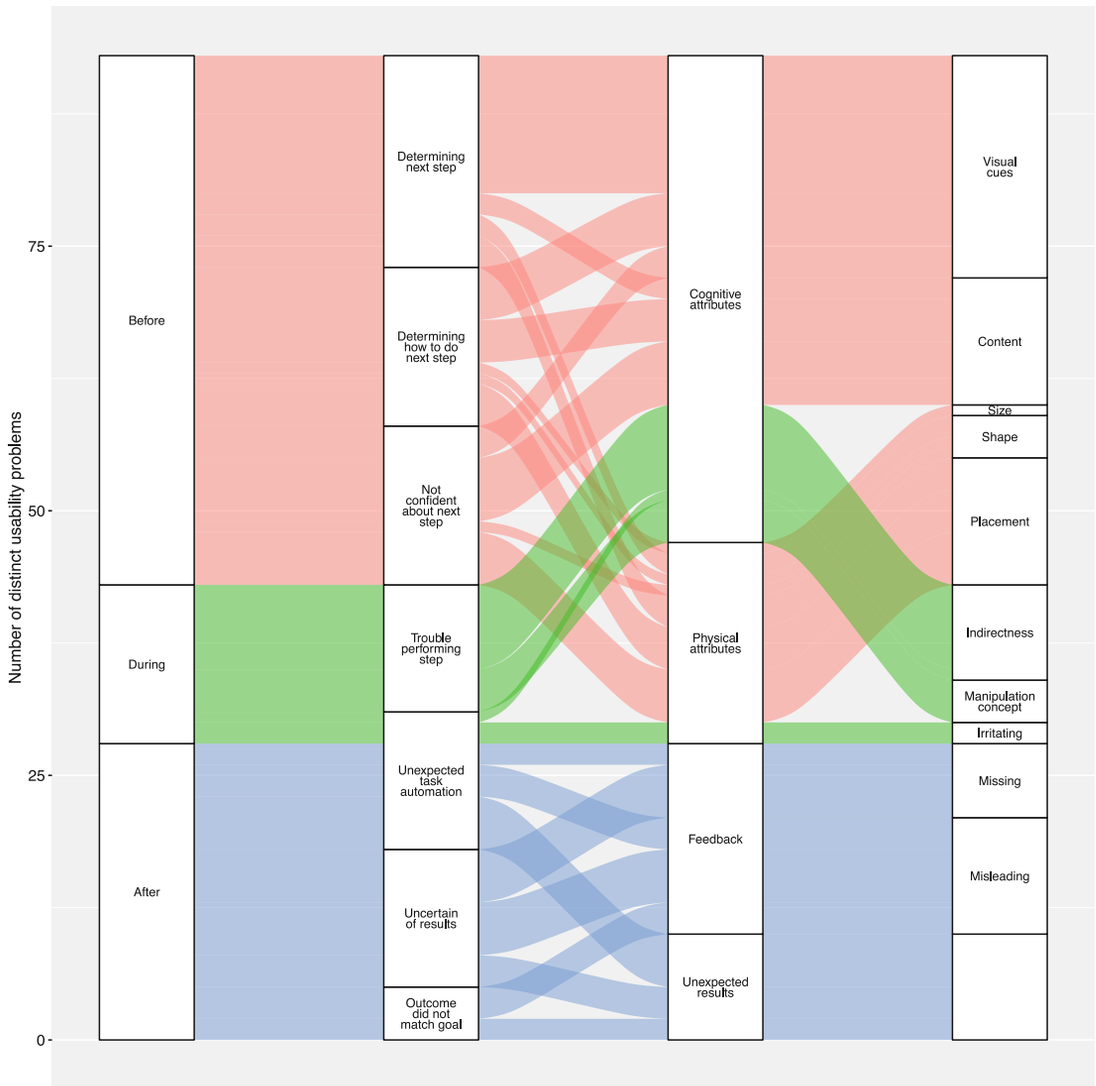


Figure 5.9.: The usability problem classification distribution for the CW using the UPC.

Before - Determining next step into the object component categories *Cognitive attributes* and *Physical attributes* is 3:1, while the category of *Cognitive attributes* splits up into *Visual cues* by 86.7 % and *Content* by 13.3 % and the category of *Physical attributes* splits up into *Shape* by 40 % and *Placement* by 60 %. The usability problems inside the task component *Before - Determining how to do next step* split up into *Cognitive attributes* by 60 % with a further distribution into *Visual cues* (55.6 %) and *Content* (44.4 %), and *Physical attributes* by 40 % with a further distribution into *Size* (16.7 %), *Shape* (16.7 %), and *Placement* (66.7 %). The same distribution of *Cognitive attributes* (60 %) and *Physical attributes* (40 %) also applied to the usability problems in the task component category *Before - Not confident about next step*. One third of the issues in *Cognitive attributes* was further classified into *Visual cues*, the remaining two thirds were further classified into *Content*, and the problems from the object component category *Physical attributes* were distributed into *Shape* by one fifth and the rest into *Placement*. The task component category *During* split up into the categories *Trouble performing step* by 80 % and the category *Unexpected task automation* by 20 %. While all issues inside the classification *Trouble performing step* fell into the object component category *Cognitive attributes* with a distribution into *Indirectness* by 66.7 % and *Manipulation concept* by 33.3 %, the issues inside the classification *Unexpected task automation* split up into *Cognitive attributes - Indirectness* (33.3 %) and *Physical attributes - Irritating* (66.7 %). The task component category *After* split up into the three subcategories *Uncertain of results* by 46.4 %, *Outcome did not match goal* by 17.9 %, and *Unexpected task automation* 35.7 %. While the usability problems inside the category *Uncertain of results* split up into the categories *Feedback* (76.9 %) and *Unexpected results* (23.1 %), the issues inside the category *Feedback* are further classified evenly into the subcategories *Missing* and *Misleading*, and the usability problems deriving from the category *Unexpected results* are not further classified in the UPC by design. The usability problems inside the task component category *Outcome did not match goal* split up into the object component categories *Feedback - Misleading* by 60 % and *Unexpected results* by 40 %, while the usability problems inside the task component category *Unexpected task automation* split up evenly into the object component categories *Feedback* and *Unexpected results*. The classification of *Feedback* itself split up into the subcategories *Missing* by 40 % and *Misleading* by 60 %.

Having a look at the different usability problems reported by the four different evaluators, the average detection rate is 43.8 %. Considering the least possible detection rate of 25 % using four evaluators, with a range of the detection rate between 39.3 % and 50 %, the detection rate lies around the first third of the possible values between 25 % and 100 %. The average any-two agreement measure is lower with 29 % (range: 25 % to 33.3 %) compared to the detection rate, while the possible values range between 0 % and 100 %. Therefore, the average any-two agreement also lies on the border of the first third.

5.4.3. Discussion

The focus for the cognitive walkthrough was on the identification of usability problems in order to compare the results with those gathered during the user study. Regarding the predicted task performance, the tasks of resetting trip kilometers and zooming the navigation map was rated slightly better using the context-sensitive menu condition. Apart from a missing hint for a context menu at all which was reported for both menu conditions, the global context menu raised several problems addressing the number of possible options which are not related to the current screen. The different menu levels offer no cues for the user to predict whether a click results in another sublevel, the toggle of a setting, or a direct action. At first, the user has to navigate through several sublevels to perform the task which he has to move up again to actually see the result of his actions. The tasks to check alerts and switch the audio source were rated equally across the variants, while the tasks to skip the audio title and cancel the route guidance were rated slightly higher for the global menu condition. Regarding the alert memory, most of the experts reported that users might suppose the feature in another device like the center stack display as also reported by some experts for the task of switching the audio source. Furthermore, the audio screen of the system shows a label for the currently selected audio source which could be assumed as a button. Using the context-sensitive menu the experts recognized the inconvenience to change the current main screen as there is no option in the context menu. While the global context menu offers the task to be performed on every main screen through the context menu, again the number of menu levels was reported as an issue. The same applied for the task to cancel the route guidance; using the context-sensitive menu, the user has to change the current main screen, while the global menu condition offers the option on each screen. Furthermore, the swipe direction to skip the title in either the main audio screen or the global context menu was not self-explanatory. The applied rolling metaphor in the main screen could irritate users as they might assume to swipe in the opposite direction using the steering wheel controls.

The above findings are also supported by the analysis of the UPC categories. Most of the problems reported were due to *Visual cues* like the missing indicator for a context menu at all or missing cues for swipe direction as well as the missing predictability of menu levels. The second most commonly used categories were *Content* and *Placement* which reflects the issues with assuming several features in different screens or devices. Furthermore, *Feedback - Misleading* and *Unexpected results* were used as classification for several problems regarding the visual feedback or progress of actions and operations mostly using the global menu condition. From the temporal perspective, the experts classified most of the problems in the category *Before* which represents several of the problems users might have preparing their next operation in the overwhelming number

of sublevels of the global context menu. Nearly one third of the usability problems were classified in the temporal category *After* due to issues with missing or misleading feedback or unexpected results. The problems in the category *During* mostly address issues with indirect or potential misleading interaction behavior like the mentioned rolling metaphor for title skip or the manipulation of the map zoom.

Comparing the results to studies from the literature, Cuomo and Bowen (1992) report that most of the problems (75 %) found by a CW allot to the *Action Specification* stage of user activity which corresponds to the category *Before* of the UPC. However, Cuomo and Bowen (1992) report only a single usability problem in the stages *Perceive*, *Interpret*, and *Evaluate* which correspond to the category *After* of the UPC and two problems in the stage *Execute* that corresponds to the category *During* of the UPC.

5.5. Comparison

As already mentioned, the results of the user study are used as a baseline for comparison with the expert-based approach of the CW. The following section compares the results of both studies from different perspectives. The task performance ratings gathered from the user study are compared to the estimations collected during the CW. Furthermore the identified usability problems of both approaches are compared and several measures of effectiveness are applied to quantify the performance of the CW method compared to the user study.

5.5.1. Task Performance

While the task performance during the user study was collected using a rating scale filled out by the experimenter for each task, the experts during the CW were asked to give an estimation of the potential task performance for the target user group. Figure 5.10 shows the task performance from the CW as well as the user study for each task separated into the two tested context menu conditions. While for the task *reset trip kilometers* the performance during the user study was rated higher for the global condition, but lower during the CW, for the task *skip title*, the experts rated the task performance for the global condition higher than for the context-sensitive condition and the user study resulted in higher task performance for the context-sensitive variant. In two cases (*map zoom* and *cancel route guidance*), the assessment from the user study is consistent with the estimate from the CW, considering only which variant is rated higher. For the remaining tasks (*check alerts* and *switch audio source*), the experts estimated the same task performance for both variants, while the assessment from the user study shows higher task performance

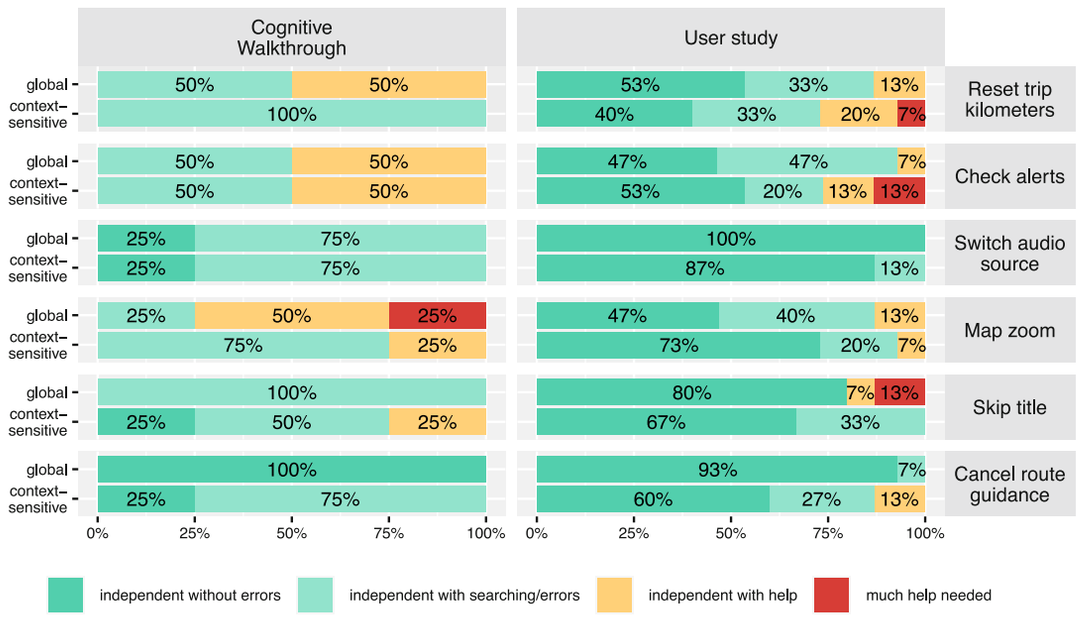


Figure 5.10.: Comparison of the task performance results for each task and variant.

values for the global context menu. Looking at Figure 5.10, the estimate from the CW is in general more conservative than the results from investigations with actual users. Hardly any task is rated with the highest category of the rating scale (*independent without errors*) with an exception for the task *cancel route guidance* using the global condition. Therefore, most of the experts assume that actual users are not able to perform most of the task without smaller interruptions. While the task performance during the user study is mostly rated with *independent without errors* and *independent with searching/errors*, the estimate from the CW is dominated by estimates from the categories *independent with searching/errors* and *independent with help*.

5.5.2. Usability Problems

The usability problem sets from both approaches were analyzed regarding similar problems and overlaps towards their classification in the UPC. Problems with different cause were grouped when possible and assigned with the same ID. For example, several issues occured because participants wanted to interact with the label displaying the current audio source in the audio screen. While this problem was reported most of the times due to the shape of the element, other participants stumbled upon the issue due to placement or missing visual cues. Analyzing the consolidated usability problem list, the CW raised 29 distinct usability problems, while the user study uncovered 63

distinct usability problems. Together both studies revealed 69 usability problems with an overlap of 23 (33.3 %) usability problems. Therefore, the list of usability problems contains 40 problems that were reported solely from the user study, while 6 issues were only reported by the experts during the CW.

Looking at these problems in detail, the issues that were only reported from the user study address several issues of first time usage. For example, several participants struggled to recognize the zoom slider during the task *map zoom* due to it's size. Other problems during this task were due to the nested menu structure of the zoom operation and the fact that horizontal swiping was not locked when zooming the map which lead to operating errors. Some major issues for the participants of the user study that were not addressed during the CW were due to the additional menu level using the global context menu which leads to even deeper menu structures. Moreover, the participants reported that using the global variant they had to select the main category even if they already switched to the appropriate main screen. On the other hand, some participants got irritated by the fact that some main screens do not offer contextual options at all using the context-sensitive variant. Some features were supposed in different screens or devices, while other functionalities were reported as unnecessary for the driver display and rather distracting while driving. Other problems regarding the global menu variant affected missing or misleading feedback for successful operations as well as the presentation of the current menu level.

On the other hand, the experts reported that due to the high number of sublevels in the global context menu variant the order as well as the underlying action is often not comprehensible. Another reported issue addressed the visual appearance of the menu to skip the current audio title in the global context menu which does not offer sufficient affordance and might look like a simple display. Moreover, the experts detected inconsistencies when operating the map zoom slider, since only for this task the context menu is temporarily hidden as well as the fact that already operated options when resetting the trip meter are still presented and not disabled like other menu items.

Among the usability problems reported by both approaches, several issues were mentioned multiple times. The most frequently reported problem by both methods describes the issue that the context menu is not closed automatically when an operation is completed ($n = 20$) or even occludes relevant information for the current task. This is followed by a group of issues ($n = 19$) where the corresponding information is assumed in another area of the user interface or even in another display like the center stack display. Further problems describe issues with insufficient feedback for successful operations ($n = 18$) or an undesirable change of the main screen to reach the requested context menu option using the context-sensitive menu condition ($n = 13$). Moreover, the direction of scroll or swipe gestures is noted as not marked appropriately, assumed the other way

Task component		Users (%)	Experts (%)
Before	Determining next step	18.3	21.5
	Determining how to do next step	12.0	16.1
	Not confident about next step	17.1	16.1
During	Trouble performing step	33.7	12.9
	Unexpected task automation	4.6	3.2
After	Uncertain of results	6.9	14.0
	Outcome did not match goal	0	5.4
	Unexpected task automation	7.4	10.8

Note. Deviations to 100 % are due to rounding.

Table 5.3.: Distribution of usability problems among UPC task component categories.

round as implemented, or along the wrong axis in the participants' opinion. Other problems address that the zoom slider during the map zoom task is displayed on the left edge of the display while the interaction was triggered from the context menu which appeared on the right edge of the display. Other issues address the number of menu levels in the global context menu variant which lead to higher eyes-off-road time and frustration due to the complex menu structure and indirect interactions. Furthermore, both approaches reported the lack of visual cues for the existence of a context menu at all as well as difficulties interpreting labels or the reason for grayed out menu options.

Looking at the individual categories of the UPC, both approaches raised the major part with around half of the problems in the temporal category *Before*, while the second largest group for the user study was the category *During* (38.3 %) and for the CW the category *After* (30.1 %). The distribution in Table 5.3 shows several differences within the subcategories of the task component. While the distribution in the category *Before* shows only small differences among the subcategories, the user study resulted in more than a twofold increase in usability problems classified into the subcategory *Trouble performing step* than the CW. On the other side, the results from the CW shows significantly higher shares for all three subcategories of the task component category *After*. This observation is also supported by the results of the performed Fisher's exact test with $p < .001$. With Cramer's V being $V = .44$ ($p < .001$), the effect size indicates a medium relation (Cohen, 1988) between the distribution among the subcategories of the task component and the applied UEM.

The same comparison was performed for the subcategories of the object component of the UPC. Here, Table 5.4 shows an almost fourfold increase in problems classified as issues due to *Visual cues* from the CW compared to user study. The share of problems in the category *Content* lies slightly higher in the problem set of the user study just

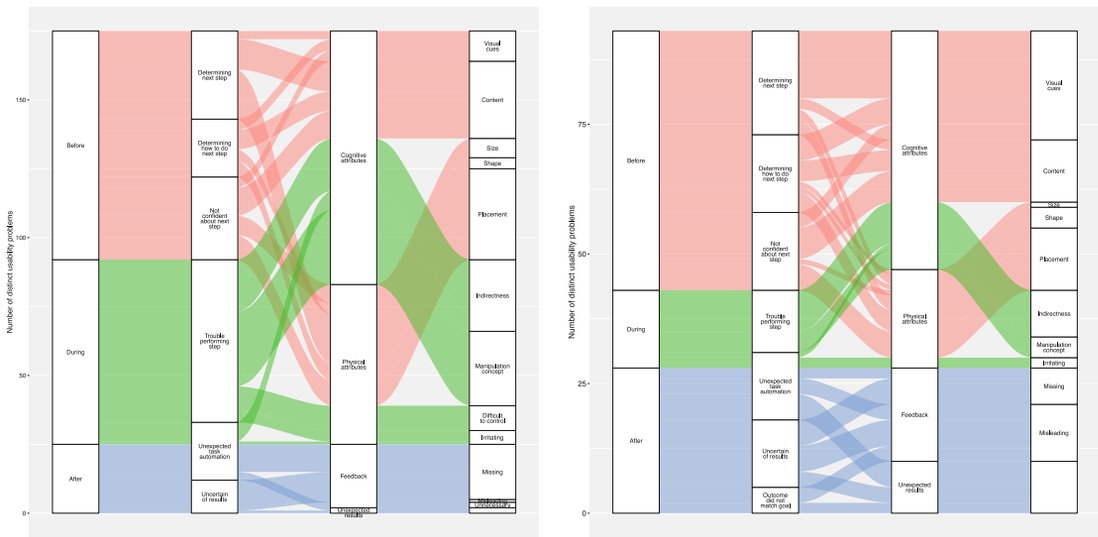
Task component	Object component		Users (%)	Experts (%)
Before	Cognitive attributes	Visual cues	6.3	24.7
		Content	16.0	12.9
	Physical attributes	Size	4.0	1.1
		Shape	2.3	4.3
		Placement	18.9	10.8
During	Cognitive attributes	Indirectness	14.9	9.7
		Manipulation concept	15.4	4.3
	Physical attributes	Not manipulable	0	0
		Difficult to control	5.1	0
		Irritating	2.9	2.2
After	Feedback	Missing	11.4	7.5
		Misleading	0.6	11.8
		Unnecessary	1.1	0
	Unexpected results		1.1	10.8

Note. Deviations to 100 % are due to rounding.

Table 5.4.: Distribution of usability problems among UPC object component categories.

like the physical attributes *Size* and *Placement*, while the share of problems inside the category *Shape* is slightly higher for the CW. For all subcategories of the task component *During*, the share of problems is higher for the user study, as it is for *Missing* and *Unnecessary* feedback. On the other hand, the CW identifies more problems due to *Misleading* feedback or *Unexpected results*. The differences in the distribution among the object component categories showed even more significant results. The Fisher's exact test with $p < .001$ and an effect size of Cramer's V being $V = .50$ ($p < .001$) indicate a large effect (Cohen, 1988) for the relation between the distribution among the categories and the applied UEM. Figure 5.11 shows a visual comparison of the distribution among the categories of the UPC between the two approaches.

With an overlap of one third of the usability problems between the methods user study and CW, the results showed that both approaches tend to identify problems of different nature. While the CW approach is able to identify more problems from the UPC category *After*, the user study is able to detect issues during the execution of single actions. Although the severity of the individual problems was not raised, both approaches are able to detect several major problems. Nonetheless, the CW approach missed to identify several severe problems regarding the navigation inside the complex menu structure of the global menu condition as well as some issues that emerged during the operation of specific functionalities like map zoom or the visibility of the current position in the



(a) Results of the user study.

(b) Results of the CW.

Figure 5.11.: Comparison of the distribution among the categories of the UPC. The readability of the graphics is neglected for a visual comparison.

menu structure. The comparison of the problem sets according to the classification in the UPC showed significant differences in the distribution between the two approaches for the task component as well as the object component.

5.5.3. Measuring Effectiveness

As already mentioned, the user study results are used as a gold standard to measure the performance of the results of the CW. The comparison uses several measures, with a focus on validity, thoroughness, reliability, and cost effectiveness.

Looking at the definition of validity in equations 4.3 and 4.4, with regard to the presented case study, the formula to calculate the validity of the CW results can be described as the following:

$$Validity = \frac{\mid E \cap U \mid}{\mid E \mid} \tag{5.1}$$

where E is defined as the problems detected during the expert review and U is defined as the problems detected through the user study. Therefore, $E \cap U$ describes the number of overlapping problems in both approaches. The validity for the underlying data is at a high level with 79.3 percent. Compared to similar studies applying the CW, this result shows a higher validity than reported by Cuomo and Bowen (1992) with 58 % for the CW. John and Marks (1997) reported a similar value of 73 % for the CW, while Desurvire

et al. (1992) on the other hand only reached a validity of 28 % for the CW compared with other techniques. According to Sears (1997) the validity measures the fit of a method for the purpose of identifying relevant usability problems. The results from the CW show that the technique is able to identify at least several, but not all of the relevant usability problems compared to a user test.

The measure of thoroughness describes the ability to find as many existing problems as possible. The following equation 5.2 shows the calculation depending on the number of problems detected during the user study (U) and the number of problems detected during the CW (E)

$$Thoroughness = \frac{\mid E \cap U \mid}{\mid U \mid} \tag{5.2}$$

This results in a thoroughness of 36.5 % which represents a higher value compared to similar studies reporting 17 % (Jeffries et al., 1991) and 15 % (John & Marks, 1997). Apart from this nearly twofold increase in thoroughness, the CW is not able to find a majority of the existing usability problems.

Another measure from the literature (Sears, 1997) is the reliability of a UEM defined by:

$$\begin{aligned} R_{Temp} &= 1 - \frac{\text{stdev}(\mid E \cap U \mid)}{\text{average}(\mid E \cap U \mid)} \\ Reliability &= \text{Maximum}(0, R_{Temp}) \end{aligned} \tag{5.3}$$

As this measure lacks a comprehensive examination of the usability problems itself and rather considers only the number of usability problems, the value should be interpreted with caution (Sears, 1997). The underlying data gives a reliability of 85.2 % which only indicates that all four experts identified almost the same number of relevant usability problems, without consideration of their nature.

Jeffries et al. (1991) propose a measure of cost effectiveness, the number of problems found per person-hour. As the original measure is applied using the severity ratings of the individual usability problems, the method is slightly adjusted in the presented case study. Considering that the number of real problems in the equations above is the number of problems that actually affect the user, the problems per person-hour measure can be calculated in two ways. Table 5.5 shows the results for the problems per person-hour measure differentiated between total problems based on the total number of problems the expert-based approach detected and real problems based on the problems detected that actually affect the users. The preparation time for the user study splits up into three hours for acquisition of participants as well as another one and a half hours for the performed pre-study for the participant and the experimenter, while the preparation of the interview manual is estimated with 20 person-hours. For the estimation of the execution time a basis of two person-hours for each session is taken to calculate the

	User study	Expert review	
Preparation time	26	9	
Execution time	30	30	
Analysis time	65	12	
Total time	**121**	**51**	
Number of problems	62	23	(29)
Problems per person-hour	**0.51**	**0.45**	**(0.57)**

Note. The times spent are listed in person-hours. The number of problems describes the number of real problems that affect the users with the values for the total number of problems in parentheses.

Table 5.5.: Benefit-cost ratios for user-based and expert-based approaches.

effort for the participants and the experimenter. The analysis of the user study made up the largest period of the user study, as the video protocols had to be analyzed for each participant to extract the usability problems. For the transcription of the videos including the classification three person-hours are estimated per participant, while the data analysis is estimated with another 20 hours. For the CW, one hour is estimated for acquisition of the experts and eight hours for preparation of study material and the review manual. The execution time is estimated with two and a half hours per expert as well as the experimenter. The following debriefing session is estimated with two hours per person attending and the analysis took around twelve hours, as the experts already classified all usability problems according to the UPC. Looking at similar studies, Doubleday et al. (1997) spent 125 hours for empirical user testing and 33.5 hours for a HE, while Jeffries et al. (1991) report 199 hours for user testing, 17 hours for guideline inspection, 20 hours for HE, and 27 hours for CW taking only the analysis into account. Looking at the resulting benefit-cost ratio, the user study outperforms the CW approach with 0.51 problems per person-hour. Taking only the problems into account that were also identified through the user study, the CW shows a benefit-cost ratio of 0.45 problems per person-hour, while considering the total number of usability problems leads to a ratio of 0.57 problems per person-hour which is slightly higher than the result from the user study.

5.5.4. Discussion

The presented case study addressed the relative effectiveness of the CW method for the usability evaluation of IVIS. The comparison of the results of the CW approach to those of a usability test showed an overlap as well as some differences. This indicates that the applied approaches have different focuses for evaluation of IVIS. While for some tasks, the observed and estimated task performance showed a similar trend, for

several tasks the estimation was too far off. The fact that all participants of the user study were employees at Mercedes-Benz Research & Development brings with it a certain bias towards technically skilled participants with experience in the operation of IVIS. On the other side, the issued personae illustrated a target user group from technically-skilled millennials to elderly and technically inexperienced users. Therefore, the task performance estimation from the experts tended to be more conservative than the observed task performance of the participants. As both the experts and the participants first had to get used to the system, there were also some initial difficulties in the evaluation of task performance for the first two tasks *reset trip kilometers* and *check alerts*.

The usability problem sets overlap in around a third of the total usability problems. As Cuomo and Bowen (1992) state, the CW method is successful in finding specific task-based usability problems, but mostly limited to the stage of action specification. This finding is also supported by the presented investigation, where the user study was able to detect issues with the operation of several interaction elements that were missed by the CW. Furthermore, the limitation of the CW due to it's focus on the "lower level interface issues" (Wharton et al., 1992, p. 388) could also be observed during the presented study, where the CW fails to interpret the investigated feature inside a bigger picture of the IVIS. In other comparisons of the CW approach with a HE by Desurvire et al. (1992), the CW is also observed to be limited with regard to the analyzed dimensions of an interface, while Jeffries et al. (1991) found that the usability problems detected by the CW were less general and less recurring. Several limitations to be considered for the interpretation of the CW results are also stated by John and Marks (1997), as three of the four experts were novices using the technique. Furthermore, the CW technique was "developed in the era where the paradigmatic human-computer interaction involved an office worker in front of a PC" (John & Marks, 1997, p. 197) and it was not intended to apply the technique for the specific context of use inside a vehicle. Sears (1997), who compared the CW technique with HE and a heuristic walkthrough approach — a combination of free-form and task-based evaluation supported by usability heuristics and "'thought-focusing' questions" (Sears, 1997, p. 219) — found that the CW method is able to identify more serious problems with less false positives compared to the HE but falls behind the heuristic walkthrough when it comes to thoroughness.

As already discussed by Andre (2000), the situation in a usability lab does not represent the real context in which the system would be used. Therefore, using the usability problem set from the usability test as a gold standard was rather acted from necessity than the decision as an ultimate criterion. Due to the nature of empirical studies, the occurrence of every single usability problem cannot be ensured and is influenced by the selection of participants and the design of the study. Because in practice the standard usability problem list does not exist — as usability testing would not be necessary if all problems are already known — the output of an empirical or analytical study cannot be

compared to such a list (Andre, 2000, pp. 40 ff.). Despite these limitations, the usability community places a high level of trust in lab-based usability testing as the mainstream method for formative usability evaluation. Most of the alternative approaches were even designed to reduce the cost of usability testing rather than increasing the quality of results. From another perspective the question should rather be how few usability problems are missed by a specific technique than how many a single method is able to find (Doubleday et al., 1997).

Not in all phases of product development the goal is the elimination of all existing usability issues. The choice of a usability evaluation method always requires the consideration of several factors, such as budget, timing, access to users, maturity level and many others. Therefore, the fact that each technique produces different kinds of results — elaborated through comparison studies between expert-based approaches and user testing (c.f. Desurvire et al., 1991; Doubleday et al., 1997; Karat et al., 1992) — should be taken into account. As already discovered by Desurvire et al. (1991), expert-based approaches are better suited to apply in earlier development stages, when the goal is to decide between different variants of an interface. While usability testing is suggested to be applied for baseline testing or key checkpoint tests during the development (Karat et al., 1992), a mixed methods approach adjusted to the specific needs of an individual project and research questions might achieve the best results.

The effectiveness of the CW measured through validity, thoroughness, and reliability shows somewhat similar results to already reported results in the literature. The validity of the results showed an even slightly higher value than most of the compared studies, while the results for thoroughness outperform the values reported in the literature. The reliability score has to be dealt with caution, as it considers only the number of relevant usability problems each expert detected rather than the overlap of common usability problems. Analyzing the benefit-cost ratio, both methods were on a similar level regarding the number of usability problems detected per person-hour. Taking not only the issues into account that were also reported by the usability test, the CW outperformed the usability test regarding the cost-benefit ratio only slightly. Therefore, the goal to reduce the cost of usability evaluation through the CW technique is not fulfilled.

6. Case Study: A Guideline Review of a Driver Display Screen Configuration Application

Parts of the following chapter have already been published in Lamm and Wolff (2021). The second case study investigates the application of the guideline review technique to an in-vehicle information system (IVIS). Therefore, the results from a guideline review are compared to the results from a user study. In order to compare both methods regarding the efficiency in detecting usability problems, several measures are applied.

6.1. Background

The following section introduces the guideline review method and its similarities and differences to the heuristic evaluation (HE). Furthermore, potential guidelines from the literature are presented in an overview section.

6.1.1. Guideline Review Method

As already mentioned in section 2.4.3, the guideline review is similar to the HE with the difference that the list of guidelines, which are used instead of heuristics, are often a much larger and more specific collection of items. Similar to the procedure of a HE, several evaluators review the interface and compare it against the specified guidelines.

After the specification of the interface or part of the interface, the researcher puts together a collection of guidelines against which the interface should be evaluated. These can be derived from different sources like styleguides, design documents, and organizational or governmental standards. A review of relevant guidelines for the presented case study is described in the following section 6.1.2. The selection of experts has already been discussed in section 4.3. This process should consider to invite experts from different domains, to get as many different perspectives as possible. During the evaluation, the

reviewers note the usability problems with their severity or classification and which guideline is violated.

6.1.2. Literature Review for Existing Guidelines

As the guideline review method requires a collection of guidelines to evaluate an interface, a literature review of existing guidelines gives an overview of the state of the art. Besides general HMI guidelines, the following list contains specific guidelines for the interaction inside vehicles as well as guidelines addressing accessibility, situation awareness, and persuasion (the number of items are given in square brackets):

- Nielsen's Heuristics (Nielsen, 1993, 1995a) — [10]
- Shneiderman's 8 Golden Rules of Interface Design (Shneiderman et al., 2018, pp. 95 ff.) — [8]
- A Guide to Carrying Out Usability Reviews (Turner, 2011) — [45]
- Ergonomic Criteria for the Evaluation of Human-Computer Interfaces (Bastien & Scapin, 1993) — [83]
- Interaction principles (International Organization for Standardization, 2019a) — [27]
- Principles on Driver Interaction with Advanced In-Vehicle Information and Communication Systems (Alliance of Automobile Manufacturers, 2006) — [25]
- Guideline for In-Vehicle Display Systems (Japan Automobile Manufacturers Association, 2004) — [23]
- The SAE Handbook (Society of Automotive Engineers, 2001)
- Designing Future Automotive In-Vehicle Devices: Issues and Guidelines (Bhise, 2002)
- Principles on the Design of Human-Machine Interface (Commission of the European Communities, 2008) — [29]
- Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices (National Highway Traffic Safety Administration, 2013) — [30]
- Human Factor Guidelines for the Design of Safe In-Car Traffic Information Services (Kroon et al., 2016) — [26]
- Design Guidelines for Web Readability (Miniukovich et al., 2017) — [12]
- Web Content Accessibility Guidelines (Caldwell et al., 2008) — [12]
- Self-Report Motivational Model (de Vicente & Pain, 2002) — [9]
- Criteria for the Assessment of Technological Persuasion (Némery et al., 2011) — [8]
- Situation Awareness Theory (Endsley, 1995) — [8]
- Simplified Situation Awareness Guidelines for Intelligent Transport Systems (M. L. Matthews et al., 2001) — [8]

- Human Factors in Engineering and Design (Sanders & McCormick, 1993)
- Apple Human Interface Guidelines (Apple Computer, 1995)
- Material Design Guidelines (Google Inc., 2019)

The detailed guidelines including their descriptions are listed in appendix C.1. As the platform specific guidelines from Apple and Google are very detailed and often do not apply to IVIS, they are not considered during the review. Furthermore, the guidelines from Society of Automotive Engineers (2001), Bhise (2002), and Sanders and McCormick (1993) were not researchable and therefore are not included in the list.

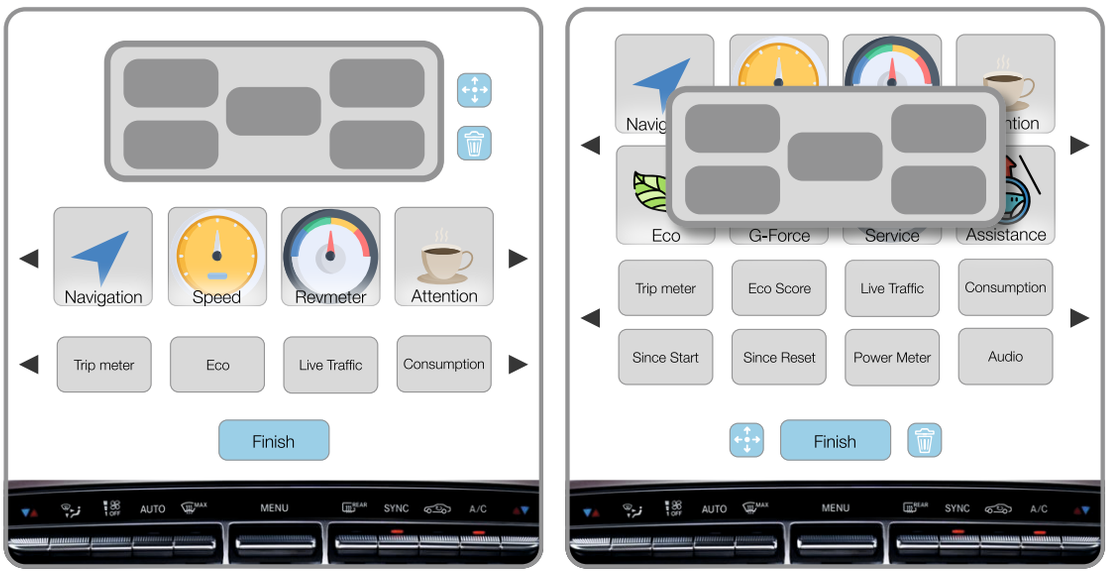
While the Nielsen heuristics as well as the golden rules of interface design by Shneiderman contain relatively abstract instructions for designing interfaces, the expert review template designed by Turner (2011) holds 45 recommendations for website usability regarding different categories features and functionality, starting page, navigation, search, control and feedback, forms, errors, content and text, help, and performance. While the heuristics from Nielsen and Shneiderman are both already introduced in section 2.4.3, the interaction principles contain the following principles: suitability for the user's tasks, self-descriptiveness, conformity with user expectations, learnability, controllability, error robustness, and user engagement (International Organization for Standardization, 2019a). Furthermore, the ISO/FDIS 9241-110:2019 contains a similar checklist, which was not available at the time the study was conducted. Since the automotive-specific guidelines (Alliance of Automobile Manufacturers, 2006; Commission of the European Communities, 2008; Japan Automobile Manufacturers Association, 2004; Kroon et al., 2016; National Highway Traffic Safety Administration, 2013) overlap in many parts, the different sets are compared in Table C.1 in appendix C.1.9. The different categories addressed in most of the guidelines are installation principles, information presentation, interaction with displays and controls, system behavior, and information about the system. The guidelines for web readability by Miniukovich et al. (2017) contain 12 guidelines derived from workshops with design and dyslexia experts and address several aspects of readability. Among them are recommendations to structure sentences in a simple and direct style as well as avoiding complex language and jargon. Other guidelines address formatting of the text with a minimum font size, the avoidance of italics or large blocks of underlined text, as well as text and background color and a plain sans serif font style. The guidelines regarding accessibility by Caldwell et al. (2008) focus mainly on the presentation of content like text, images, or time-based media as well as on offering help for understanding the content and the page structure of a website. The collection of guidelines by de Vicente and Pain (2002) presents rules to evaluate students' motivational state and the guidelines by Némery et al. (2011) address persuasion of interfaces. While the recommendations by Endsley (1995) introduce detailed information about situation awareness using the example of military aircraft, the guidelines by M. L.

Matthews et al. (2001) are based on the situation awareness theory but contain rather concise rules to increase situation awareness in driving situations.

6.2. Object of Investigation - A Customizable Driver Display

As with the first case study in chapter 5, the scope and maturity of the concept influences the selection of the object of investigation. In the guideline review, the configuration application could be examined in its entirety. Also in this case, a practical interest and the availability of the experts play a significant role.

In the presented case study a customizable driver display is evaluated through a user study and a guideline review. The interface concept allows the driver to configure the driver display screen through an interface on a touch display in the center stack. The scribbles in Figure 6.1 show both interface variants. While the fixed variant in Figure 6.1a uses a combined preview and dropzone that is fixed in the upper part of the screen, the dynamic variant in Figure 6.1b hides the overlaying preview and dropzone when the user is not interacting with the screen. This variant was introduced to avoid restrictions on the use of smaller screens with a landscape screen ratio. Therefore the dynamic variant offers more space for the content element lists, which are separated into small and big items. The fixed variant presents these content elements as single-row list using pagination via the arrow buttons, while the available space in the dynamic variant allows to use double-space lists. The different content elements can be arranged in the preview and dropzone via drag and drop or direct touch. This is also visualized in a short tutorial video which is started when the user opens the configuration mode. The preview contains a grid that visualizes the available space. The left and right side of the customizable driver display screen offer space for two small or one big content elements, while in the center area a small or a big content element can be assigned. The buttons placed to the right of the fixed preview in the fixed variant are used to activate two interaction modes; the switch mode to switch content elements between containers and the deletion mode to delete individual items or the whole screen. These buttons are placed besides the finish button in the lower screen area for the dynamic variant. While for the dynamic variant these buttons are necessary to open the preview, the switch and deletion actions can be performed by simple swipe gestures in the preview using the fixed variant. The climate controls on the bottom are included to illustrate the permanent area of the center stack display.



(a) Variant with fixed preview.

(b) Variant with dynamic preview.

Figure 6.1.: Scribbles of the individualization screen variants on the center stack display.

6.3. Experiment 1 - User Study

Since the goal of the case study is to compare the expert-based method with traditional usability testing, the following chapter introduces the user study part. The primary research question of this study is therefore the number of identified usability problems. The following sections introduce the applied method, the usability test procedure, and the results of the usability test.

6.3.1. Method

A within-subjects usability test was used to compare two different interaction concepts. For the experiment a combination of observation of users and the think aloud (TA) technique was applied. The usability test was performed in a stationary 2015 Mercedes-Benz C-Class Sedan at the Mercedes-Benz Technology Center in Sindelfingen.

Participants

The participants for the experiment were recruited via several department's mailing lists to ensure confidentiality. Therefore, all participants were employees at Mercedes-Benz from different departments including Quality Management, Marketing & Sales, and Research & Development. In total 18 participants (including the pre-study participant) finished the experiment — eleven males and seven females, with little to medium experience with the simulation mockup.



Figure 6.2.: Experimental setup of the test vehicle with the position of the displays.

The age of the participants ranged between 22 and 53 years with an average of 30 years ($SD = 7.7$). Four of the participants were between 18 and 24 years old, eleven participants were between 25 and 39 years old, and one participant was between 40 and 54 years old. All participants had a valid driver's license. The experiments were instructed and conducted in German and all participants understood and spoke German. The usability tests took place during the participants' work time.

Apparatus The test vehicle had a modified display line-up including a 12.3 inch (2.68 : 1 ratio) screen for the driver display and a 15.4 inch notebook covered to simulate a 12.9 inch (1.11 : 1 ratio) touch screen in the center stack. The positioning of both displays is shown in Figure 6.2. The software simulation on the driver display was rendered with Rightware Kanzi[1]; the simulation on the center stack display used a custom simulation software based on Qt.[2]

The configuration screen, the participants had to operate was triggered via the driver display simulation controlled by a touch control button on the multi-functional steering wheel. The button supported swipe gestures to toggle horizontal and vertical directions as well as pressing to confirm a selection. A separate hardware button allowed to go back in the menu tree or close context menus. The center stack display, where the configuration screen was integrated, was controlled via touch gestures. Most of the time, the participant had to control the center stack display with the exception of starting the configuration process via the steering wheel. The experimenters encouraged the participant to continuously express his thoughts and verbalize his goals and expectations which were recorded in a written protocol.

[1] https://www.rightware.com/kanzi
[2] https://www.qt.io

Procedure

The experiments consisted of an interview with an introduction to the general operation of the vehicle's controls and functions, an exploration phase, the testing of operating tasks, and a final survey. Each experiment lasted an hour in total. The experiment design was tested in a pretest with a single participant. As there were no changes in the design between the pretest and the actual experiment, the data from the pretest is included in the analyzed dataset.

The participant was welcomed at the specified meeting point and brought to the test vehicle. After sitting down at the driver seat, the participant was instructed on the purpose of the study and the experiment procedure. Each experiment began with the pre-study interview for collecting demographic data, information on the experience in general IVIS interaction, and the participant's attitude towards technology and individualization. Information about the age, gender, handedness, and job description of the participant was collected. The attitude towards technology was raised through a self-assessment by the participant into one of the following categories, for the sake of simplicity:

- I always want to be one of the first to try out new technologies. In doing so, I also accept so-called "teething troubles".
- I tend to be mainstream and adopt new technologies as soon as the "teething troubles" of the devices have disappeared.
- I accept new technologies only when I have to. Dealing with technology is a waste of time for me.

Finally, the participant was asked to answer two questions regarding individualization on his personal devices using a seven-point Likert scale:

- Do you use customization/individualization features on your devices (e.g. smartphone, tablet, PC)? *(Very strongly disagree — Strongly disagree — Disagree — Neutral — Agree — Strongly agree — Very strongly agree)*
- How often do you use customization/individualization features on your devices? *(once after purchase — once a month or less — two to three times a month — about once a week — several times a week — every day)*

After an instruction on the general interaction concept — excluding the individualization feature — the participant was given some time on his own to explore the IVIS. The focus for the exploration lay on the driver display, in order for the participant to get to know the different features of the system, which were customizable during the experiment. Following the introduction, the first variant was presented to the participant. The order of the variants was counter-balanced between the subjects. After familiarizing with the individualization screen, the participant had to perform several tasks in order to depict a usage scenario for the core functionality of the system. The first task of copying a given

configuration represented the situation of configuring the system with a specific result in mind from scratch, while in the second task the participant was asked to overwrite the configuration with another given configuration. The tasks of switching content elements between containers, overwriting and deleting specific content elements and finalizing the configuration represent typical use cases for a customizable system. For each task, the participant's performance was rated from *independent without errors*, *independent with searching/errors*, and *independent with help* to *much help needed* by the investigator. Besides these measurements, the participant was encouraged to constantly express his thoughts about the interface through TA. The investigator noted down relevant statements from the participant and reminded the participant to think out loud when necessary. After the last task, the participant was asked to complete the System Usability Scale (SUS) questionnaire for the first variant. When the participant finished all tasks using the first variant, he was presented the second variant, given again some time for exploration. After completing the same tasks using the changed variant, the participant was again asked to complete the SUS for the second variant. In a post-study interview, the participant was asked about his favorite variant as well as an overall rating for each variant on a scale from one to ten.

Data Analysis

The task effectiveness for the different variants was examined through an evaluation of the task performance scale mentioned above. With 18 participants, the sample is relatively small and not normally distributed (see Table C.2 in appendix C.2). As proposed by Bortz and Schuster (2010, p. 125), a correlation analysis was performed in order to check the prerequisites for the parametrized t-test. The correlation analysis showed that most of the values did not correlate significantly (see Table C.2 in appendix C.2). Therefore, the non-parametric alternative, the Wilcoxon signed rank test was used to compare the paired differences between task performance and task rating for the two conditions.

In a further step, the users' comments and the TA protocols were analyzed manually. The notes taken during the test were transcribed and categorized as a basis of the usability problems list. The usability problems used the classification of the Usability Problem Classifier (UPC) to ensure comparability with the results from the expert review technique. Thereby, all comments from the participant were collected in a list with some metadata, a description of the problem itself, and the classification according to the UPC. Table C.3 in appendix C.2 contains the complete list of usability problems. The list also contains multiple entries for a single problem, when the participant mentioned the problem multiple times, to emphasize the severity of the problem for the specific participant. To facilitate the analysis, the problems were assigned with a unique ID, which is repeated for repeated problem occurrence. From this list some statistical measures could be deduced, like the total number of usability problems, the total number of distinct usability problems, the number of usability problems per participant, and the

number of usability problems per classification category and between the configuration interface variants.

The SUS which was used as post-study questionnaire for each condition was analyzed regarding the overall score for each of the two conditions. In order to calculate possibly significant differences between the SUS scores, a Wilcoxon signed rank test is performed. As Brooke (1996) describes, although the score ranges between 0 and 100, it represents a sum of the individual items rather than a percentage value. Furthermore, he points out that looking at the scores of individual items does not deliver detailed results on the domain of usability problems.

6.3.2. Results

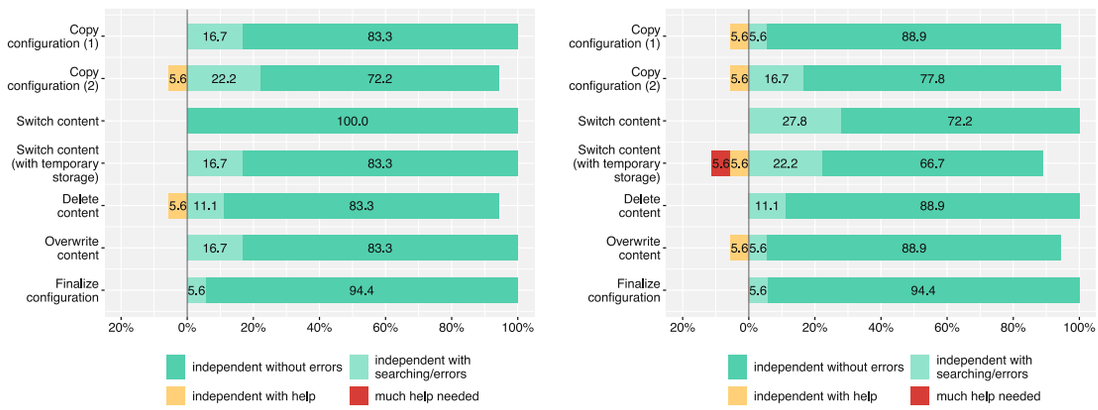
Task Performance

As Figure 6.3 illustrates, the task performance did not differ that much compared between both preview conditions. All seven tasks were performed *independent without errors* by most of the participants for both conditions. While the first two tasks of copying a given configuration were solved without errors by slightly more participants using the dynamic preview condition, the switch content tasks were both performed without errors by more participants using the fixed preview variant. Looking at the overall differences in the category *independent without errors*, the highest values were reported during the two tasks of switching content between container elements (Switch content: 100 % fixed, 72.2 % dynamic; Switch content (with temporary storage): 83.3 % fixed, 66.7 % dynamic). Only one participant needed much help solving the task of switching content using the temporary storage in the dynamic preview condition. Summing up, the participants needed help only in two cases using the fixed preview condition, while the dynamic preview variant lead to five situations where help was needed by the participants.

The results from the applied Wilcoxon signed rank test in Table 6.1 show that most of the tasks did not differ significantly between the two test conditions regarding the task performance. The only significant difference with $z = -2.09$ and a significance value of $p = .03$ is reported for the task of switching content. With an effect size of $r = .49$, the result shows a large effect, according to Cohen (1992). While all participants solved the task without errors using the fixed preview variant, more than a quarter of the participants needed searching or made errors using the dynamic preview condition.

Rating & Usability

The ratings for both variants in Figure 6.4a show different distributions for both variants. The median for the dynamic preview variant is 6, while it lies at 8 for the fixed variant. While for the dynamic preview condition 50 % of the participants rate between 5 and 7, the ratings of 50 % of the participants lie between 8 and 9. The box plot for the fixed preview condition is comparatively short which indicates a higher level of agreement



(a) Fixed preview condition.

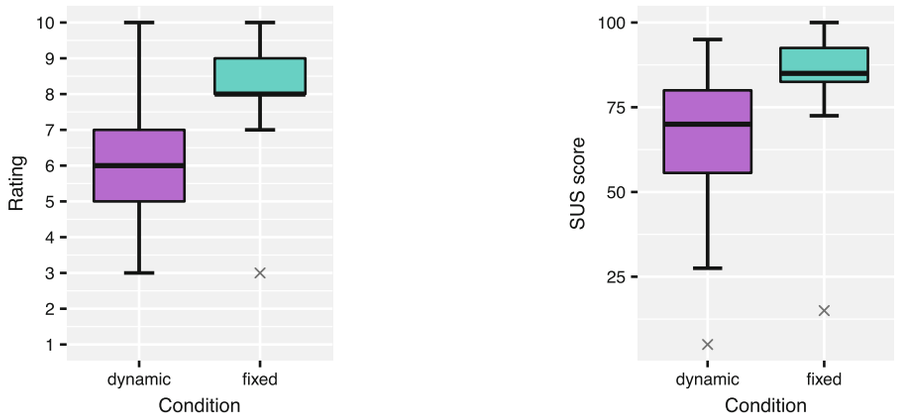
(b) Dynamic preview condition.

Figure 6.3.: Task performance distribution for both conditions fixed and dynamic. Deviations to 100 % are due to rounding.

Task	V	z	p	r
Copy configuration (1)	1.5	0	1	0
Copy configuration (2)	12	−0.22	.82	.05
Switch content	0	−2.09	.03	.49
Switch content (with temporary storage)	2.5	−1.62	.10	.38
Delete content	4.5	−0.54	.59	.13
Overwrite content	7.5	0	1	0
Finalize configuration	1.5	0	1	0

Note. V = sum of ranks assigned to the differences with positive sign; z = normally distributed z-score; r = Pearson's correlation coefficient as measure of effect size.

Table 6.1.: Results of a Wilcoxon signed rank test comparing the task performance for the two interface variants. The only significant difference occurred for the task *Switch content* with $p = .03$.



(a) Distribution of overall user ratings (1-10).

(b) Distribution of SUS scores (0-100).

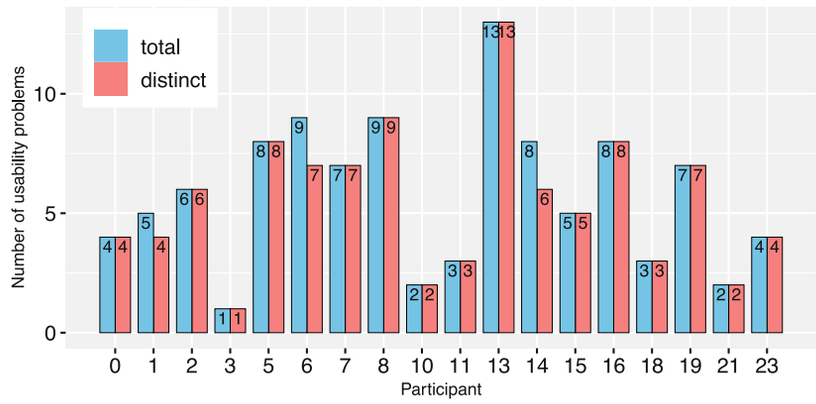
Figure 6.4.: Distribution of ratings and SUS scores between both preview conditions.

for this preview variant. The performed Wilcoxon signed rank test ($z = -3.14$, $p = .002$) shows a significant difference between the ratings with a large effect size of $r = .74$, according to Cohen (1992). Therefore, the fixed preview variant is rated significantly higher than the dynamic preview variant.

This finding is also supported by the distribution of the SUS scores for both variants. The box plot for the fixed variant in Figure 6.4b is also comparatively short. Both conditions report an outlier in the category *worst imaginable* according to Bangor et al. (2009), which is actually from the same participant. While the lower whisker for the dynamic condition goes down to 27.5 inside the category *poor*, the lower whisker for the fixed condition stops at 72.5 and therefore on the lower end of the *good* category. With a median of 85 and an upper quartile of 92.5 as well as a lower quartile of 82.5 using the fixed condition, 50 % of the participants report a SUS score on the border between the categories *good* and *excellent*. Looking at the dynamic preview condition, 50 % of the participants report a SUS score between 55.6 and 80 ($Q_1 = 55.625$, $Q_2 = 70$, $Q_3 = 80$). Therefore, these results are divided between the categories *ok* and *good* of the range of Bangor et al. (2009). The Wilcoxon signed rank test also shows a significant difference for the SUS scores between both conditions with $z = -3.56$ and a significance value of $p < .001$. With $r = .84$ the effect size is even higher than those for the ratings of both variants.

Usability Problems

The experiments recorded a total number of 104 usability problems with in total 54 distinct usability problems. The mean average for each participant was $M_{total} = 5.78$ for the number of total usability problems and $M_{distinct} = 3$ for the number of distinct usability problems. As shown in Figure 6.5a, the number of total problems per participant is in most cases not significantly higher than the number of distinct usability problems.



(a) Histogram of the number of occurred usability problems per participant.

(b) Distribution of total and distinct number of usability problems.

Figure 6.5.: Number of usability problems from the user study.

The maximum number of distinct usability problems a single participant encountered was 13. The participant with the least number of distinct usability problems reported only a single problem. The most frequently reported usability problem was mentioned 8 times by also 8 different participants. Overall, 17 of the usability problems were reported more than once by different participants. While the number of distinct usability problems for both preview conditions was 14, the participants identified a total number of 36 usability problems using the dynamic preview variant and 22 usability problems using the fixed preview variant. Moreover, the participants identified 46 usability problems respectively 26 distinct usability problems independent from the condition. These usability problems could not be traced back to issues depending on the condition, but rather represent general usability problems of the IVIS.

For the distribution among the categories of the UPC, in the following the total numbers of usability problems are taken into account, as the multiple occurrence of a single problem also gives an information about the severity. Taking a look at the classification of the reported usability problems, 60.6 % of the issues were classified in the temporal category *Before* ($n = 63$), 27.9 % in the category *During* ($n = 29$), and 11.5 % in the category *After* ($n = 12$). The results described in the following paragraph are visualized in Figure 6.6; deviations to 100 % are due to rounding. The share of usability problems in the temporal category *Before* splits up by 9.5 % into the category *Determining next step* ($n = 6$), 38.1 % into the category *Determining how to do next step* ($n = 24$), and 52.4 % into the category *Not confident about next step* ($n = 33$). One third of the problems identified as *Determining next step* was further categorized in the object category *Physical attributes* with all problems in the subcategory *Placement*, while two thirds fall into the category *Cognitive attributes* with a distribution of one quarter into subcategory *Visual cues* and three quarters into subcategory *Content*. In the category *Determining how to do next step* 41.7 % of the usability

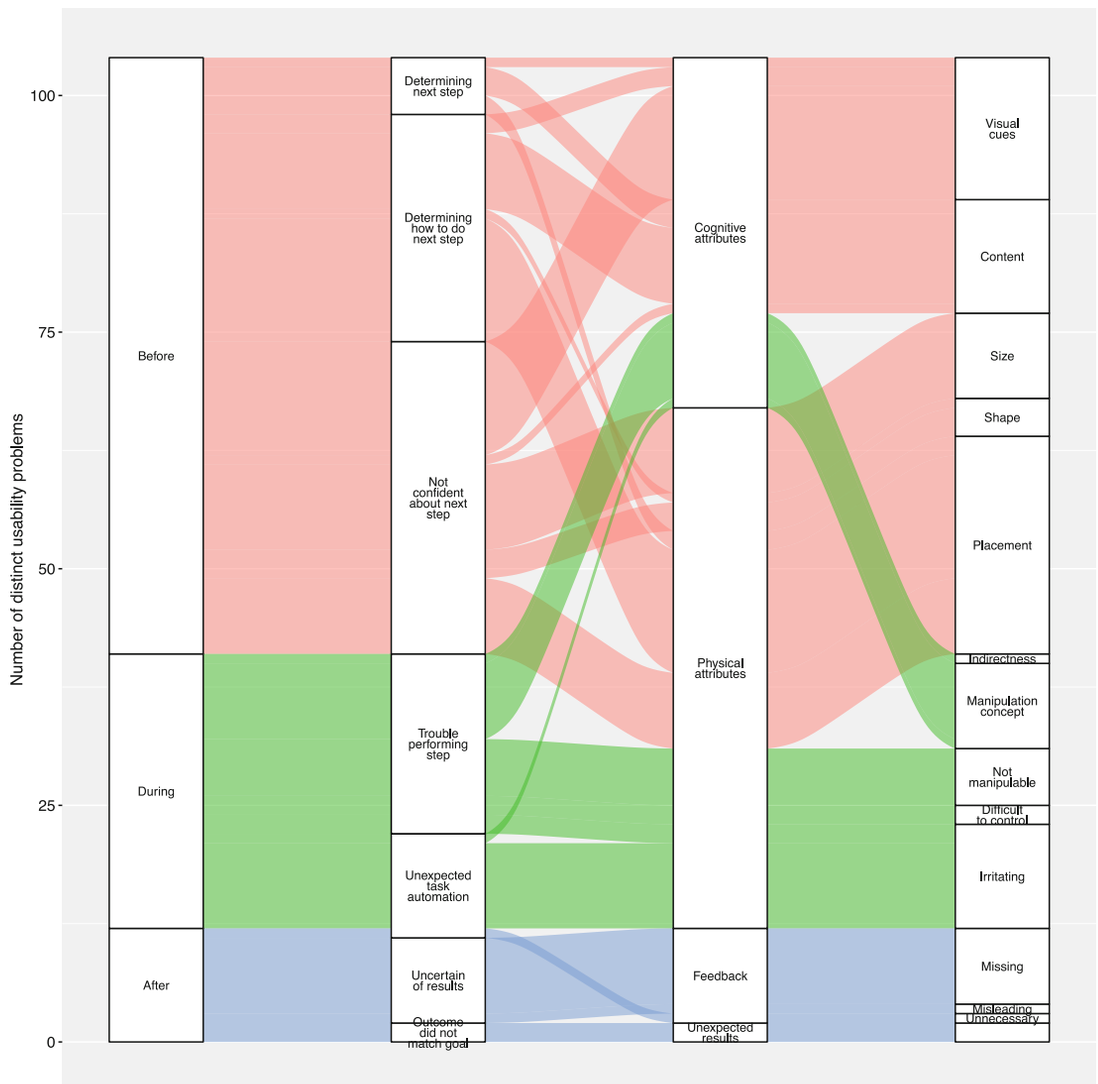


Figure 6.6.: The usability problem classification distribution using the UPC.

problems allot to the category *Cognitive attributes*, 58.3 % to the category *Physical attributes*. The usability problems in the category *Cognitive attributes*, deriving from the classification *Before - Determining how to do next step*, split into the subcategories *Visual cues* (20 %) and *Content* (80 %), while the problems in the category *Physical attributes* are further classified into the subcategory *Shape* by 7.1 % and into the subcategory *Placement* by 92.9 %. The category *Not confident about next step* is separated into the object component categories *Cognitive attributes* (39.4 %) and *Physical attributes* (60.6 %), while 92.3 % of the issues in the category *Cognitive attributes* were raised due to *Visual cues* with the remaining 7.7 % of the issues are falling into the category *Content*. A share of 45 % of the usability problems classified in the category *Physical attributes*, deriving from the classification *Before - Not confident about next step*, allot to the subcategory *Size*, 15 % are further classified into the subcategory *Shape*, and 40 % fall into the subcategory *Placement*. The usability problems from the temporal category *During* were further classified with 65.5 % into the category *Trouble performing step* and with 34.5 % into the category *Unexpected task automation*. The issues from the category *Trouble performing step* are further categorized into the category *Cognitive attributes* by 47.4 % which is separated into *Indirectness* (11.1 %) and *Manipulation concept* (88.9 %), and into the category *Physical attributes* by 52.6 % which is separated into *Not manipulable* (60 %), *Difficult to control* (20 %), and *Irritating* (20 %). While 10 % of the usability problems classified as *During - Unexpected task automation* split up into *Cognitive attributes* and *Physical attributes* by 90 %, all problems inside the category *Cognitive attributes* are classified due to the *Manipulation concept* and all issues from the category *Physical attributes* fall into the subcategory *Irritating*. The usability problems inside the temporal category *After* are separated into the categories *Uncertain of results* (75 %), *Outcome did not match goal* (16.7 %), and *Unexpected task automation* (8.3 %). While all of the problems inside the category *Outcome did not match goal* fall into the subcategory *Unexpected result* and all of the problems inside the category *Unexpected task automation* fall into the subcategories *Feedback - Unnecessary*, the problems from the category *Uncertain of results* are separated inside the category *Feedback* into *Missing* (88.9 %) and *Misleading* (11.1 %).

6.3.3. Discussion

For the presented case study the results from the user study, especially the list of usability problems, is used as a baseline for comparison with the expert-based guideline review. The rating scale to record the participants task performance was filled in by the experimenter and therefore only represents a subjective measurement of the actual task performance. Because time measurements interfere with TA, the focus in this study was to collect qualitative data on the existing usability problems in two different interface variants. Regardless, the results for task performance show that both interface

variants can be used by the participants without serious problems. The task performance categories *independent without errors* and *independent with searching/errors* outweigh for both preview conditions. The only statistically significant difference between both conditions was reported for the task of switching content elements between different containers. While using the fixed preview variant all participants were able to solve the task without issues, more than a quarter of the participants needed some time for searching or made errors using the dynamic preview variant. When looking at the variant ratings, the majority of the participants preferred the fixed preview variant. This finding which is also statistically significant is also supported by the examination of the SUS scores for the different conditions. This also showed a statistically significant positive effect towards the fixed preview condition.

The greatest share of usability problems accounts to the categories of *Placement* which indicates a non-ideal placement of different features like the dynamic preview occluding content elements or the buttons to enter the different configuration modes. With the second largest category of usability problems — *Visual cues*, the participants reported missing cues for interaction possibilities, for example whether to use drag and drop or direct touch. Furthermore, some content elements are not self-explanatory and their highlighting in different states of the configuration stays unclear. An issue caused due to limits in the simulated interface with the detection of swipe gestures to scroll in lists was most often reported for the category *Manipulation concept*. Moreover, several participants missed the functionality to delete with a long press or undo the last configuration operations. The problems in the category *Size* most frequently describe issues with the size of the content elements itself, as several participants would have preferred specific content elements for small containers rather than large containers. Moreover, several participants perceived the number of configurable content elements as too limited. The issues in the category *Feedback* almost entirely allot to the missing permanent preview as feedback for configuration operations in the dynamic condition, while the issues reported for the category *Unexpected results* describe the inconsistent functionality of saving the current configuration when pressing the back button on the steering wheel, as participants rather expected to abort and reset the configuration.

Similar to the results from the first case study in chapter 5 more than half of the usability problems occurred in the *Before* category of the UPC which is not surprising for untrained users. Cuomo and Bowen (1992, 1994) as well as Andre (2000) report similar findings in their studies. The users first need to interpret their aim in the form of a concrete sequence of actions in order to perform a task, which is supposed as the most difficult part of the task (Andre, 2000). This is also reflected in the results from the presented user study, as most of the usability problems occurred in the temporal category *Before*. While most of the problems in the temporal category *After* are due to missing feedback, most of the problems that occurred during an action could be resolved by improving the

way of manipulating some objects and features in the interface to a more natural form of interaction.

6.4. Experiment 2 - Expert Review

The following sections present the expert-based approach of the comparing case study investigating a driver display screen configuration application. The research question of the study results mainly from the subsequent comparison of both methods to identify the differences and similarities. These relate primarily to the nature of the usability problems and the resulting effectiveness of the expert-based method.

6.4.1. Method

The expert-based study within the described case study used the guideline review technique to identify usability problems of the interface. During the review, the experts evaluated both interface variants presented in section 6.2 in a seating buck at the UI Studio of the Mercedes-Benz Technology Center in Sindelfingen.

Experts The experts were selected according to the dimensions of experts in section 4.3 modified from Nielsen (1993). Therefore, four experts with different focuses on domain, system, and technique were selected for the experiment. The expert in the dimension system was represented by an interaction designer who worked on a previous version of the interface — the technique expert was represented by a PhD student with a masters degree in media informatics and experience in performing heuristic evaluations. Two domain expert were represented by interaction designers which were unfamiliar with the interface, but had experience in designing interfaces for in-vehicle information systems. All four experts were employees of Mercedes-Benz in Sindelfingen.

Apparatus For the expert review, the lab already described in section 5.3.1 was used. The seating buck was also equipped with simulation displays for instrument cluster and center stack display. The configuration screens were integrated in the center stack display simulation — supporting touch gestures.

In order to review the interface according to guidelines for IVIS, several sources of guidelines that potentially relate to the domain were reviewed (see section 6.1.2) and consolidated. While most of the guidelines address specific environments like accessibility of websites (cf. Caldwell et al., 2008) or motivation of students (cf. de Vicente & Pain, 2002), these guidelines could be neglected for the presented guideline review. Other rather specific guidelines like the situation awareness guidelines by M. L. Matthews et al.

(2001) are integrated in the form of abstract guidelines into most of the automotive HMI guidelines (cf. Alliance of Automobile Manufacturers, 2006; Commission of the European Communities, 2008; Japan Automobile Manufacturers Association, 2004; Kroon et al., 2016; National Highway Traffic Safety Administration, 2013), while the ergonomic criteria by Bastien and Scapin (1993) are mostly covered by recent usability guidelines like the website usability guidelines by Turner (2011).

Because the guidelines by Turner (2011) already cover a broad range of categories, these guidelines were taken as a basis which was adapted to the specific context of use. As the categories search, forms, and help do not apply to automotive user interfaces and describe specific elements for websites, the categories have been removed from the basis. The categories feedback and errors have been consolidated together and adapted to the context of use by removing irrelevant items addressing form elements. Furthermore, several items addressing specific elements of a website like hyperlinks, traditional form elements as well as hardware issues and mouse and keyboard interaction have been removed. The list of guidelines has also been extended by items from the collected guidelines on automotive HMI (cf. Alliance of Automobile Manufacturers, 2006; Commission of the European Communities, 2008; Japan Automobile Manufacturers Association, 2004; Kroon et al., 2016; National Highway Traffic Safety Administration, 2013). Some of these guidelines address hardware-specific issues like position or the luminosity and contrast of displays which do not fall within the responsibility of usability evaluation methods and were therefore not taken into account. Whereas other items regarding driver distraction through interaction are integrated into the guideline review template in Table C.4 in appendix C.3. Since the review was carried out exclusively with German native speakers, the guidelines where formulated in German and extended to include an additional description or examples.

The list of guidelines was sent via email to six independent HMI experts from different departments at Mercedes-Benz Research & Development with the request of weighting the guidelines according to their importance for usability between low (1) and high (5). Figure 6.7 shows the distribution of the experts weightings for each of the 27 guidelines including the corresponding mean weight. Several guidelines (numbers 1, 3, 10, 13, 14, 22, 26, and 27) show relatively low scattering with average weights between 4 and 5 compared to items with a high disagreement (numbers 9, 12, and 18). While the former items seem to belong to a common understanding of HMI guidelines, the experts have disparate views on the latter items addressing an easy access to the application, an indication for the current location, and the recovering from errors. The items with the highest importance according to the experts were related to an easy accessible and consistent navigation (10) as well as performance (26) and system errors (27) with a weight of 5 by all six experts. Furthermore, guidelines addressing the fulfillment of user goals and objectives (1) and the influence on the driving task (22) showed high importance

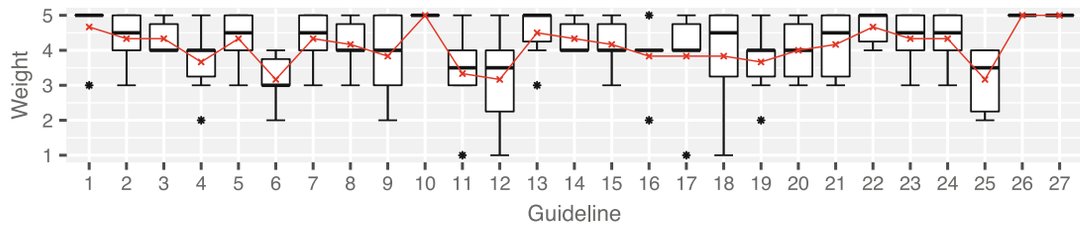


Figure 6.7.: The distribution of guidelines weights returned from the six HMI experts. The red line indicates the mean weight.

with a weight of 4.7, followed by guidelines on supporting the user's workflow (2) and often used tasks (3), the appearance of call to actions (5), orientation on the main screen (7), reversing of user actions (14) as well as driving related guidelines on continuous gazing at the screen (23) and interruptible actions (24) with a weight of 4.3.

In order to create a comparable metric from these weighted guidelines, the Guideline Compliance Scale (GCS) is introduced for the guideline review. Similar to the SUS which is used for the user study, the evaluators are asked to rate the 27 different guidelines regarding their compliance in the presented interface. The following formula calculates a value between 0 and 100 from the given ratings:

$$GCS = \frac{\sum_{k=1}^{n} w_k * r_k - \sum_{k=1}^{n} w_k * r_{min}}{\sum_{k=1}^{n} w_k * r_{max} - \sum_{k=1}^{n} w_k * r_{min}} \tag{6.1}$$

where n is the number of guidelines, w_k is the weight of a guideline, r_k is the rating of a guideline, r_{min} is the minimum possible rating, and r_{max} is the maximum possible rating.

Procedure

After welcoming the expert at the UI Studio, a briefing on specific personae, the context of use, and common tasks was performed. The personae were created by a partner department and describe typical customers for two segments of Mercedes-Benz passenger cars. The context of use explains the basic concept of the instrument cluster display for the expert in order to understand the specific use case for the customizable screen. Furthermore, several prerequisites like the constraint to use the system while driving and the limited screen area due to technical restrictions and static display of information were discussed. The briefing also introduced typical tasks the users would perform using the system (configuration, switch content elements, delete content, finish configuration) as well as similarities and differences of the two variants being observed (c.f. section 6.2). Finally, the selected guidelines which were used for the review as well as the documentation of the usability problems with the UPC were presented to the expert.

During the review, the expert was asked to inspect the interface according to the guidelines and assign identified usability problems to the guideline that was violated

processing the introduced tasks one by one. Furthermore, the usability problems had to be classified in the categories of the UPC. The two different variants of the system were counterbalanced, whereas each expert reviewed both variants. After the investigation of each variant, the expert was asked to rate the overall compliance with each of the selected guidelines on a 5-point Likert scale.

In a debriefing session with all four experts, the identified usability problems were reviewed together. The focus here was to refine and consolidate the list of usability problems, especially in cases where more than one expert described similar problems with slightly different classification in the UPC or different guidelines assigned to the problem. Furthermore, the experts were invited to discuss several problems in the group and review the individual inspection sessions. The inspection itself took about one and a half hours per expert with an additional hour for the debriefing session.

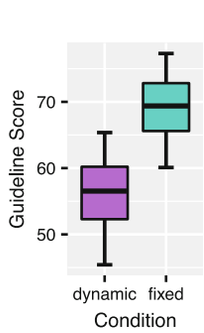
Data Analysis

Unlike the user study, the expert review did not collect task performance or task rating. The refined list of usability problems from the debriefing session was prepared in the same way as the list of usability problems from the user study. For each individual problem a unique ID was assigned together with metadata describing the usability problem. Table C.5 in appendix C.3 contains the complete list of usability problems. The statistical measures deduced from the list of usability problems include the total number of usability problems, the total number of distinct usability problems, the number of usability problems per expert, and the number of usability problems per classification category and between the configuration interface variants. In order to compare the usability problem sets detected by the individual evaluators the measures average detection rate and average any-two agreement were used. While the detection rate is calculated by the number of problems detected by a single evaluator divided by the number of problems detected by all evaluators collectively, the any-two agreement is calculated by the number of overlapping problems two evaluators detected divided by the total number of problems they detected collectively. As counterpart to the SUS which was used in the user study, the GCS of each interface variant is used. In order to analyze the evaluator agreement for the Guideline Compliance Scale (GCS), the intraclass correlation coefficient (ICC) for two-way models is used.

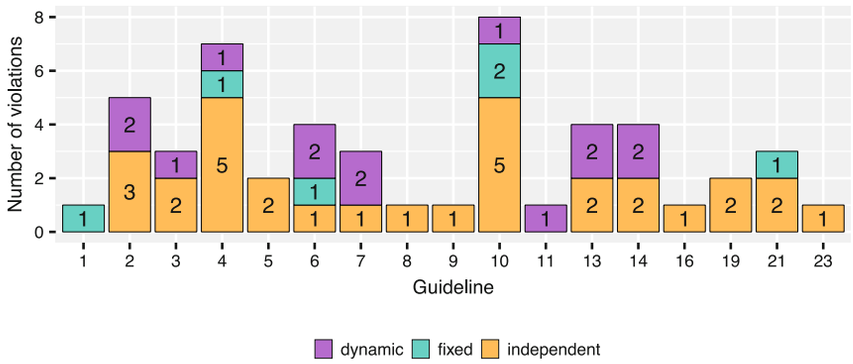
6.4.2. Results

Guideline Compliance

The boxplot in Figure 6.8a shows the distribution of the GCS among the four experts. With an interquartile range of 7.9 for the dynamic condition and 7.2 for the fixed condition, the values do not show a remarkable difference between the variants. The median for the fixed condition lies with 69.4 only 12.8 points higher than the median for the dynamic condition with 56.5. Applying the same categories from the adjective rating scale of



(a) Distribution of GCS.



(b) Histogram of guideline violations per variant.

Figure 6.8.: The GCS and the number of guideline violations per variant.

	Single measures				Average measures			
Condition	$ICC(C,1)$	95% CI	F	p	$ICC(C,k)$	95% CI	F	p
fixed	.235	[.061, .460]	2.271	.003	.552	[.205, .773]	2.271	.003
dynamic	.276	[.097, .498]	2.614	.001	.603	[.301, .799]	2.614	.001

Note. $ICC(C,1)$ = ICC (two-way mixed, single measures, consistency); $ICC(C,k)$ = ICC (two-way mixed, average measures, consistency); F = F-test value

Table 6.2.: Intraclass correlation between the four experts for GCS.

the SUS (Bangor et al., 2009), the median for the fixed condition lies slightly below the border to the category *good*. The lowest score for the fixed condition lies with 60.1 in the center of the category *ok*. The upper whisker of the boxplot for the fixed condition does not reach the category *excellent* with a highest score of 77.3. Whereas the boxplot of the dynamic condition does not even reach the category *good* with the highest score of 65.4. With the lowest score of 45.4, the lower whisker of the dynamic condition extends into the category *poor*, while the box lies in the category *ok* with a lower quartile of 52.3 and an upper quartile of 60.2. The performed Wilcoxon signed rank test shows no significant difference for the GCS between both conditions with $z = -1.53$ and a significance value of $p = 0.125$. In order to analyze the inter-rater agreement for the GCS, the intraclass correlation coefficient (ICC) is calculated. Table 6.2 shows the results separated into the single and average measures methods. The correlation for both variants shows only a poor connection between the different raters according to Koo and Li (2016), looking at the single measures of the four experts. On the other side, looking at the average scores of the four experts, the correlation coefficients for both variants lie in the moderate range.

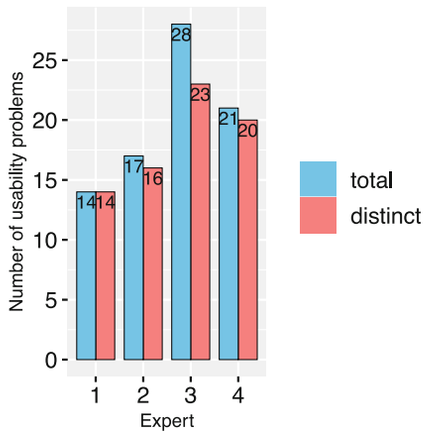
Both correlation coefficients are statistically significant at the 1% level, which is also supported by the relatively large 95% confidence interval sizes.

Looking at Figure 6.8b, the most frequently violated guideline with a total of 8 violations reads as follows: *The navigation within the application is easy to find, intuitive, and consistent* (No. 10). The original set of guidelines in German language is listed in Table C.4 in appendix C.3. Other frequently violated guidelines were:*Users are adequately supported according to their level of expertise* (No. 4, $n = 7$), *Features and functionality support users desired workflows* (No. 2, $n = 5$), *The main screen provides a clear snapshot and overview of the content, features, and functionality available* (No. 6, $n = 4$), *Prompt and appropriate feedback is given* (No. 13, $n = 4$), *Users can easily undo, go back and change or cancel actions* (No. 14, $n = 4$), and *Text and content is legible and scanable, with good typography and visual contrast* (No. 21, $n = 4$). As most of the reported usability problems are variants-independent, most of the guideline violations were not assigned to a specific variant. Only two problems constitute a guideline violation that were unique for one of the two variants. While in the fixed condition, one usability problem violated the guideline *Features and functionality meet common user goals and objectives* (No. 1), another guideline *The navigation has sufficient flexibility to allow users to navigate by their desired means* (No. 11) was only violated once using the dynamic condition. For the dynamic preview variant 14 violations were reported, while the fixed preview variant yielded only 6 violations and the greatest share of violations was variants-independent.

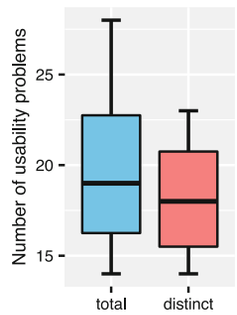
Nine of the 27 guidelines were not violated by any of the interface variants. The category *Performance* including the two guidelines *Application performance doesn't inhibit the user experience* and *Errors and reliability doesn't inhibit the user experience* is not violated by the interface variants at all. Furthermore, the category *Driver Distraction* includes the guidelines *The driver is able to assimilate the presented information with a few glances that do not affect driving*, *The driver is able to interrupt a system input at any time*, and *The system aids the driver when resuming a task after an interruption* that are not reported as violations. The remaining satisfied guidelines are *The current location is clearly indicated*, *Errors messages are concise, written in easy to understand language and describe what's occurred and what action is necessary*, *Users are able to easily recover from errors*, and *Terms, language, and tone are consistent*.

Usability Problems

The expert reviews recorded a total number of 80 usability problems with in total 51 distinct usability problems. The mean average for each expert was $M_{total} = 20$ for the number of total usability problems and $M_{distinct} = 12.75$ for the number of distinct usability problems. Figure 6.9a shows a histogram of the number of usability problems identified by the experts involved in the reviews, while Figure 6.9b shows the distribution of the total and distinct number of usability problems. Both figures show that the number of total usability problems is not significantly higher than the number of distinct



(a) Histogram of the number of identified usability problems per expert.



(b) Distribution of total and distinct number of identified usability problems.

Figure 6.9.: Number of usability problems from the expert review.

usability problems for most of the experts. The maximum number of distinct usability problems identified by a single expert was 23. The most frequently identified usability problem was identified by three experts. Overall, 16 of the usability problems were identified more than once by different experts. Looking at the problem sets of the different evaluators, the average detection rate is with 35.8 % relatively low, taking into account the minimum detection rate of a single evaluator of 27.5 %. Considering the number of four experts, the least possible detection rate is 25 % when each problem is only identified by a single expert. The average any-two agreement measure — ranging from 0 percent to 100 percent — adds up to 15.4 % (range: 2.8 % to 26.5 %.

The experts identified 14 distinct usability problems (n_{total} = 22) specific to the dynamic preview condition, while the number of distinct as well as total usability problems specific for the fixed preview condition was 6. Furthermore, the experts identified a number of 31 distinct usability problems (n_{total} = 52) which were independent from the condition.

The following section describes the distribution of the usability problems among the categories of the UPC. Since several usability problems that were reported by more than one expert were classified into different categories of the UPC, these usability problems occur multiple times in the classification. Looking at the first three categories of the task component in Figure 6.10, 38.75 % of the usability problems are classified in the category *Before*, another 38.75 % in the category *During*, and 22.5 % in the category *After*. Most of the problems classified in the task component category *Before* are further classified as issues *Determining how to do next step* (64.5 %, n = 20). The remaining issues were classified as *Not confident about next step* (19.4 %, n = 6) or *Determining next step* (16.1 %,

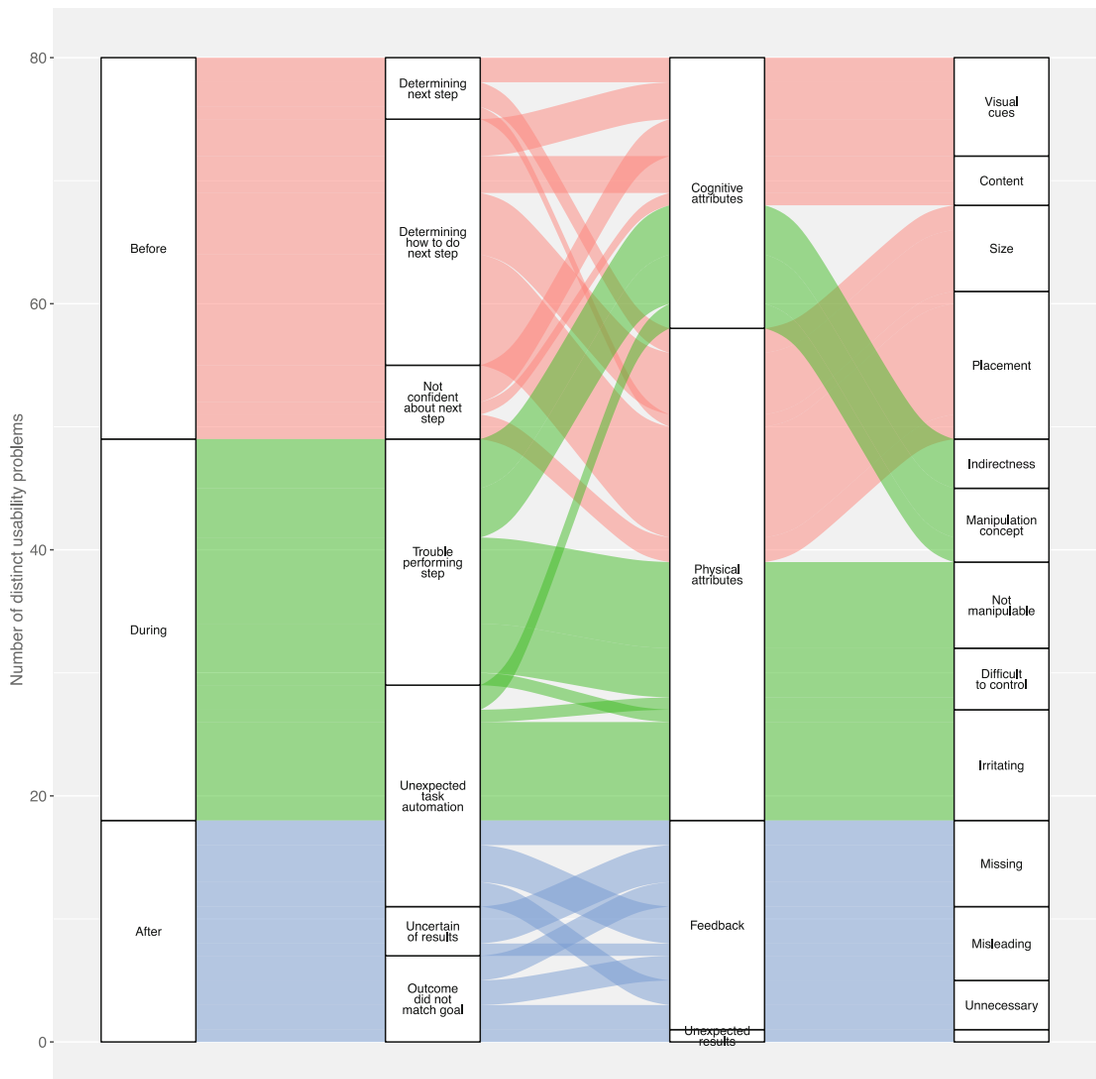


Figure 6.10.: The usability problem classification distribution using the UPC.

$n = 5$). Taking a look at the object component, the usability problems classified as *Before - Determining how to do next step* were further classified into *Cognitive attributes* by 30 %, which split up evenly into issues with *Content* and *Visual cues*, and into *Physical attributes* by 70 %, which split up into issues with *Placement* (64.3 %) and *Size* (35.7 %). The five usability problems classified as *Before - Determining next step* were classified further into the object component categories *Cognitive attributes - Visual cues* by 40 % and *Physical attributes* by 60 %, which split up into the subcategories *Placement* (33.3 %) and *Size* (66.7 %). Whereas the problems classified as *Before - Not confident about next step* split up into the categories *Cognitive attributes* by two thirds that are further classified as *Visual cues* (75 %) and *Content* (25 %), and the categories *Physical attributes - Placement* by one third.

The problems in the task component category *During* split up into the categories *Trouble performing step* by 64.5 % ($n = 20$) and *Unexpected task automation* by 35.5 % ($n = 11$). The problems in the category *Unexpected task automation* split into *Cognitive attributes* by 18.2 % — where all problems were further classified as *Manipulation concept* — and *Physical attributes* by 81.8 % — where 11.1 % were further classified as *Difficult to control* and 88.9 % into the category *Irritating*. While the 40 % of the problems classified as *Trouble performing step - Cognitive attributes* split up evenly into *Indirectness* and *Manipulation concept*, the 60 % of the problems classified as *Trouble performing step - Physical attributes* are divided into the categories *Not manipulable* by 58.3 %, *Difficult to control* by 33.3 %, and *Irritating* by 8.3 %.

The problems classified in the task component category *After* were further separated into the categories *Uncertain of results* by 22.2 % ($n = 4$) as well as *Outcome did not match goal* and *Unexpected task automation* by 38.9 % ($n = 7$) each. While the problems in the categories *Uncertain of results* and *Unexpected task automation* are all further classified as issues regarding *Feedback*, the problems in the category *Outcome did not match goal* split up into the categories *Feedback* by 85.7 % and *Unexpected results* by 14.3 %. The category *Feedback* is further separated evenly into *Missing*, *Misleading*, and *Unnecessary* for the category *Outcome did not match goal*. Whereas for the classification *Unexpected task automation - Feedback* the usability problems are further classified into the categories *Missing* by 28.6 %, *Misleading* by 42.9 %, and *Unnecessary* by 28.6 % (deviation to 100 % is due to rounding). Furthermore, the usability problems classified as *Uncertain of results - Feedback* were issues regarding *Missing* feedback by 75 % and *Misleading* feedback by 25 %.

6.4.3. Discussion

As in the user study that serves as a baseline for comparison, the focus for the guideline review was to identify usability problems for the two different interface variants. Furthermore, the focus during the expert review was on identifying violations of specific HMI guidelines which lead to usability problems. Regarding the guideline compliance, the experts reported more than three times as much violations for the dynamic preview variant as for the fixed preview variant. This is also supported by the distribution of the GCS which lies between the categories *ok* and *good* for the fixed condition, while the values for the dynamic condition range between *poor* and *ok*. A similar result is reported from the user study, where the difference even shows a statistically significant positive effect. Due to the sample size of four experts, a statistical test is not meaningful for the expert review. Looking at the relation of the GCS rated by the four evaluators using the ICC, the results indicate that the experts identify different usability problems, based on their experience, and therefore rate different for the GCS. As the evaluators were selected to cover the variety of different dimensions of experts (c.f. 4.3) on purpose, the result is not surprising. Rather, the different perspectives of different experts are welcome to increase the detection rate.

The most frequently violated guidelines are from the categories *Navigation*, *Features & Functionality*, *Main Screen*, and *Control, Feedback & Errors*. On the one hand the navigation within the application is often criticized as not being easy to find, intuitive, or consistent. The issues leading to a violation of the guideline are related to the affordance of the mode buttons as well as an automatic mode change when interacting with elements in the preview. Another problem reported relating to the guideline is the recognizability of the buttons to scroll in the content lists. On the other hand the experts report issues with supporting the users in their workflows as well as according to their level of expertise. The buttons to change the configuration mode, to scroll in content lists, and the missing support for swipe gestures in content lists were also reported for the guideline *Features and functionality support users desired workflows* as well as issues with the temporary storage. Issues with supporting the users according to their level of expertise primarily address missing support through highlights for possible interactions as well as irritating highlights for user errors. The main screen does not provide a clear snapshot and overview of the content, features, and functionality which is mostly reported either due to the dynamic preview occluding content elements or due to the missing preview in the dynamic condition. Furthermore, there were issues with feedback which were reported due to an inappropriate delay for the disappearance of the preview in the dynamic condition as well as problems with missing visual cues as hints for possible interactions.

Regarding the usability problems, most experts did not tend to report similar problems more than once. Most of the usability problems were classified inside the object component category *Placement*. These problems describe suboptimal placement of control elements to scroll in content lists and the buttons to enter a specific mode on the one hand, but also issues with the position of the dynamic preview. The category *Placement* is followed by the categories *Irritating*, *Visual cues*, and *Missing* feedback. Only few problems address the object component categories *Unexpected results*. The issues classified as *Visual cues* mainly describe problems with hints to support configuration like highlighting of possible element containers in the preview as well as the short tutorial video running after the configuration has been started. Issues regarding the dynamic preview, the temporary storage, and grayed out list elements are classified as *Irritating*. The usability problems in the category *Manipulation concept* describe issues with the temporary storage feature, with drag and drop interaction, and the support of swipe gestures in content lists. From the temporal perspective, most of the problems were classified in the task component category *Before*, followed by the categories *During* and *After*. As already mentioned for the user study, these results are similar to other studies (Andre, 2000; Cuomo & Bowen, 1992, 1994).

6.5. Comparison

The following section compares the results from both studies regarding the metrics SUS and GCS as well as the usability problem sets of both studies. The distributions among the categories of the UPC are compared for both studies in order to unveil differences in the focus of both approaches. Furthermore, both approaches are compared regarding effectiveness applying the measures of validity, thoroughness, reliability, and a benefit-cost ratio.

6.5.1. System Usability Scale and Guideline Conformity Score

While the SUS represents a subjective score assessing the usability by users, the introduced GCS served as a counterpart for the guideline review. Figure 6.11 shows a direct comparison for the two measures separated by the two conditions, the fixed and dynamic preview variants. Looking at the ratings for the two conditions fixed and dynamic the SUS showed bigger differences for mean (M_{fixed} = 82.9, $M_{dynamic}$ = 64.2), median (Mdn_{fixed} = 85, $Mdn_{dynamic}$ = 70), and standard deviation (SD_{fixed} = 18.5, $SD_{dynamic}$ = 22.96). While the SUS as well as the GCS for the fixed variant were on average slightly higher than for the dynamic variant (ΔM = 29.2%, ΔMdn = 21.4%), the standard deviation was higher for the dynamic variant (ΔSD = 6.1%). Whereas

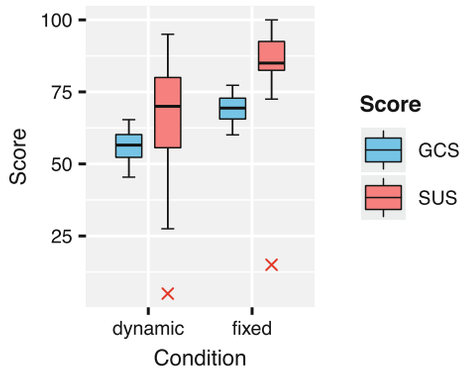


Figure 6.11.: Comparing the results of the SUS rated by the user study with the GCS observed by the guideline review.

the average SUS for both conditions lied above the average GCS (ΔM_{fixed} = 20.1%, $\Delta M_{dynamic}$ = 14.7%), the standard deviation for the SUS was substantially higher for both conditions compared to the standard deviation of the GCS (ΔSD_{fixed} = 156.6%, $\Delta SD_{dynamic}$ = 175.6%). Therefore, the scores from the user study disseminate significantly more than those from the expert review.

Unlike Brooke (1996, p. 193) stating that "scores for individual items are not meaningful on their own" for the SUS, the individual guidelines of the constructed GCS deliver meaningful insights into possible issues for a human-machine interface. Although the comparison of the SUS and GCS showed a similar trend using both measures, it stays unclear whether the comparison of the scores is reasonable. Several comparison studies would be necessary to analyze the relation between both scores.

6.5.2. Usability Problems

In order to compare the problem sets from both studies, they were analyzed regarding similar problems. An additional step served to group several similar usability problems with different origins. For example, issues regarding the list scroll buttons where reported due to their size as well as their placement, which makes them hard to recognize. These problems were assigned with the same ID as they describe problems operating the list scroll. Evaluating the results of the consolidation of both individual problem sets, the expert review uncovered 45 usability problems, while the user study reported 54 issues. Together the two approaches gathered 78 usability problems in total with an overlap of 21 (26.9 %) usability problems. Therefore the list of usability problems contains 24 issues reported solely by the evaluators of the expert review, and 33 problems that were only detected by the user study.

Looking at the problems in detail, most of the problems that were only reported from the user study deal with issues regarding the content of the application, initial difficulties, and problems with placement of different interactive elements. The subjects of the user study pleaded for more flexibility regarding the size of the content elements as well as an improvement of recognition of individual content elements. Another category of problems that were often detected only by the user study were problems regarding the initial usage. Despite the short tutorial animation at the startup, some users had problems recognizing how to perform the configuration. While these users had problems after entering the configuration screen, the evaluators of the expert review assumed problems accessing the configuration via the steering wheel controls. Several problems that were only identified by the expert review address issues with the affordance of interactive elements like the buttons to activate the switch and delete mode. Moreover, the experts identified problems concerning the temporary storage which did not occur during the user test sessions. For example, deselecting or removing items from the temporary storage can therefore be considered as edge cases.

On the other hand, the usability problems detected by both approaches mostly address the plausibility of the dynamic preview. The behavior where the preview is placed is not comprehensible, the preview itself occludes content elements, and the fade out of the preview appears tardy. Established interaction techniques like swipe gestures to scroll in lists were expected but not supported, while the supported drag and drop interaction sometimes lead to problems in individual cases. Other issues reported during both studies include content-related problems like the order of content elements or their self-descriptiveness.

Looking at the individual categories of the UPC, both approaches reported the most problems in the temporal categories *Before* and *During* of the task component. The distribution within the task component in detail is presented in Table 6.3. Inside the category *Before* most of the problems reported by the experts were issues regarding the user to determine how to do the next step, while the user study itself yielded a higher share of problems where the user was not confident about the next step. Both studies uncovered several usability problems due to trouble performing the step in the temporal category *During* of the UPC, whereas the evaluators revealed slightly more issues regarding unexpected task automation during the expert review. Looking at the category *After*, the expert review reported slightly more problems in the categories *Outcome did not match goal* and *Unexpected task automation*. Therefore, the frequencies of problems in the different categories between the user study and the expert review show a significant relation to the applied UEM. This is also supported by the results of Fisher's exact test with $p = .029$. With an effect size calculated through Cramer's V of $V = .34$ ($p = .024$) there is a medium relation (Cohen, 1988) between the number of problems in the different categories and the applied UEM.

Task component		Users (%)	Experts (%)
Before	Determining next step	7.4	6.2
	Determining how to do next step	24.1	25.0
	Not confident about next step	29.6	7.5
During	Trouble performing step	24.1	25.0
	Unexpected task automation	5.6	13.8
After	Uncertain of results	3.7	5.0
	Outcome did not match goal	3.7	8.8
	Unexpected task automation	1.9	8.8

Note. Deviations to 100 % are due to rounding.

Table 6.3.: Distribution of usability problems among UPC task component categories.

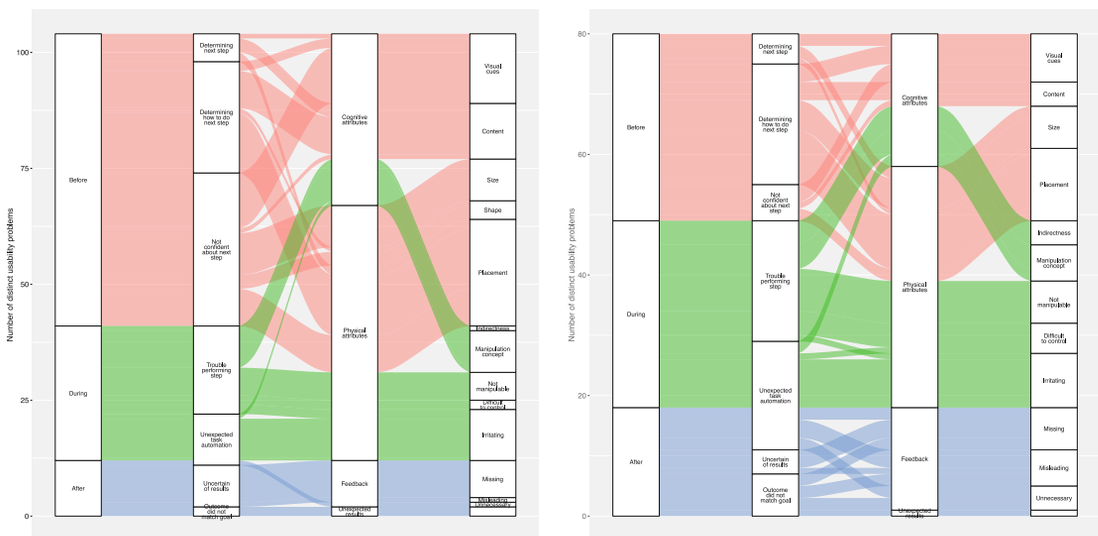
The same comparison could be performed for the object component of the UPC. Table 6.4 shows that the user study reported slightly more problems for both categories *Visual cues* and *Content* of the main category *Cognitive attributes* in the temporal classification *Before*. A similar observation is apparent for the category *Physical attributes*. Regarding the category *Cognitive attributes* of the temporal classification *During*, the evaluators reported slightly more problems due to *Indirectness* during the expert review than the users, whereas the users revealed more problems deriving from the *Manipulation concept*. In the category *Physical attributes* the experts revealed slightly more issues for all three subcategories which also applies for the *Feedback* category. Unlike the task component, the relationship between the applied UEM and the number of problems in the different categories of the object component was not statistically significant applying the Fisher's exact test ($p = .36$). Figure 6.12 shows a visual comparison of the distribution among the categories of the UPC between the two approaches.

With an overlap of slightly more than a quarter of the overall problems, the results showed that the user study discovered several other kinds of problems compared to the expert review. Although the severity of the individual problems was not raised, both approaches detected most of the obvious usability issues regarding the dynamic preview as well as established interaction techniques like swipe or drag and drop gestures that are not supported. The results of the comparison of the classification according to the UPC showed that the choice of one approach over the other has an influence on the temporal character of the problems. While both approaches identify several problems due to issues determining how to do next step before a specific action as well as issues with trouble performing step during an action, the user-based approach also shows a peak for issues due to the user not being confident about next step. However, the expert review approach classified several usability problems into diverse categories

Task component	Object component		Users (%)	Experts (%)
Before	Cognitive attributes	Visual cues	18.5	10.0
		Content	9.3	5.0
	Physical attributes	Size	7.4	8.8
		Shape	3.7	0
		Placement	22.2	15.0
During	Cognitive attributes	Indirectness	1.9	5.0
		Manipulation concept	9.3	7.5
	Physical attributes	Not manipulable	7.4	8.8
		Difficult to control	3.7	6.2
		Irritating	7.4	11.2
After	Feedback	Missing	1.9	8.8
		Misleading	1.9	7.5
		Unnecessary	1.9	5.0
	Unexpected results		3.7	1.2

Note. Deviations to 100 % are due to rounding.

Table 6.4.: Distribution of usability problems among UPC object component categories.



(a) Results of the user study.

(b) Results of the CW.

Figure 6.12.: Comparison of the distribution among the categories of the UPC. The readability of the graphics is neglected for a visual comparison.

of the UPC. Looking at the object component of the UPC, the distribution among the different categories showed no significant differences between both approaches.

6.5.3. Measuring Effectiveness

Looking in detail at the problems identified by the expert review — as the user study is used as the gold standard in the comparison — several measures can be calculated. As described in section 4.5, the focus in this case study lies on the quantifiable and reproducible measures validity, thoroughness, reliability, and a benefit-cost ratio.

Recapitulating the definition of the validity measure from equations 4.3 and 4.4 with regard to the underlying data of the presented case study, the formula to calculate the validity of the expert review results can be described as the following:

$$Validity = \frac{\mid E \cap U \mid}{\mid E \mid} \tag{6.2}$$

where E is defined as the problems detected during the expert review and U is defined as the problems detected through the user study. Therefore, $E \cap U$ describes the number of overlapping problems in both approaches. The validity for the underlying data is at a medium level with 47 %. Compared to a comparative study performed by Desurvire et al. (1992), the data shows similar results when compared with the HE technique ($Validity_{HE}$ = 44%). Another study by John and Marks (1997) on the other hand reported a validity of 17 % for HE. Following the definition mentioned by Andre (2000, p. 48) — "validity is a measure of how well a method does what it is intended to do" — the results show that the guideline review method is able to identify at least several problems that are *real* usability problems. But, the technique also reveals usability problems where there are no issues for the users.

While high validity means that a UEM only finds problems that are real problems, high thoroughness means that a UEM is able to find as many existing problems as possible. For the presented case study, equation 6.3 calculates the thoroughness for the expert review results.

$$Thoroughness = \frac{\mid E \cap U \mid}{\mid U \mid} \tag{6.3}$$

With a thoroughness of 39 %, the result from the presented case study lies somewhere in the lower third of the spread reported from studies by Jeffries et al. (1991) with 50 %, John and Marks (1997) with 31 %, and Virzi et al. (1993) with 81 %, compared with the HE technique. Therefore, the guideline review technique is not able to find most of the usability problems that exist in the interface, but it is able to find the most severe issues (c.f. section 6.5.2).

	User study	Expert review	
Preparation time	20	11	
Execution time	34	17	
Analysis time	60	12	
Total time	114	40	
Number of problems	54	21	(45)
Problems per person-hour	0.47	0.53	**(**1.13**)**

Note. The times spent are listed in person-hours. The number of problems describes the number of real problems that affect the users with the values for the total number of problems in parentheses.

Table 6.5.: Benefit-cost ratios for user-based and expert-based approaches.

Another measure of effectiveness is the reliability that measures the independence of the individual performing the usability evaluation. Equation 6.4 adapted from Sears (1997) shows the calculation of reliability in two steps.

$$\begin{aligned} R_{Temp} &= 1 - \frac{\text{stdev}(\mid E \cap U \mid)}{\text{average}(\mid E \cap U \mid)} \\ \textit{Reliability} &= \text{Maximum}(0, R_{Temp}) \end{aligned} \tag{6.4}$$

As also noted by Sears (1997), the above reliability measure does not consider the usability problems itself rather than the number of problems detected by the different methods. Therefore, the relatively strong reliability of 83 % stands against the poor values of the ICC observed during the comparison of the guideline compliance for the individual evaluators (c.f. section 6.4.2). This shows that the reliability regarding the number of usability problems that are real problems in the interface is relatively high, but the individual evaluators tend to detect different issues when looking at the problem descriptions.

As a measure of cost effectiveness Jeffries et al. (1991) use the measure of problems found per person-hour. As the authors in the article used the severity ratings of the individual usability problems instead of the number of problems itself, the measure has to be adapted to fit the presented case study. Considering that the number of real problems in the equations above is the number of problems that actually affect the user, the problems per person-hour measure can be calculated in two ways. Table 6.5 shows the results for the problems per person-hour measure differentiated between total problems based in the total number of problems the expert-based approach detected and real problems based on the problems detected that actually affect the users. The preparation time for the user study is roughly estimated with three hours for acquisition of participants, another one and a half hours for the performed pre-study for the participant as well as the experimenter, and the preparation of the interview manual with 14 person-hours. The execution time is estimated with one hour per session for each of the 17 participants

as well as the experimenter, while the analysis is estimated with two person-hours per session for the transcription of the thinking aloud protocol and classification of usability problems as well as 24 hours for data analysis. For the expert review, one hour is estimated for acquisition of the experts and ten hours for preparation of the review manual. The execution time considers one and a half hours for each expert as well as the experimenter and an additional debriefing session with all experts with one person-hour per participant. As the experts already classified the usability problems according to the UPC as well as the violated guidelines, the data analysis for the expert review is estimated with 12 person-hours. Looking at other studies the time spent for the different approaches shows similar values to those reported here. Doubleday et al. (1997) spent 125 hours for empirical user testing and 33.5 hours for a HE, while Jeffries et al. (1991) report 199 hours for user testing, 17 hours for guideline inspection, 20 hours for HE, and 27 hours for CW taking only the analysis into account. While the benefit-cost ratio for the expert review is only slightly higher than that for the user study when considering the number of real problems, the ratio for the total number of usability problems identified by the experts is more than twice as high as that of the user study.

6.5.4. Discussion

The comparison of the two raised scores — the SUS from the user study and the GCS from the guideline review — shows a similar trend for both of the interface variants. Even though, there is further research needed in order to analyze the composition of the GCS. Several additional comparison studies between the SUS and the constructed GCS would be needed to make a meaningful conclusion on the relationship of both scores.

The fact that the problem sets from both studies overlap in only a quarter of the cases shows that the two approaches are suited to identify different kinds of problems. As already discovered by Desurvire et al. (1991), the applied expert-based method — several heuristic evaluations — tends to find usability problems that do not occur during user testing. Similar results are reported by Karat et al. (1992), where about a third of the significant problems were common across all methods. In their study comparing empirical testing and two walkthrough approaches, the expert-based methods miss several severe problems that occurred during user testing. Contrary results are reported by Doubleday et al. (1997) who compared user testing with heuristic evaluation. The expert-based approach identified more problems than the user-based approach. But when looking at the overlap of both problem sets as in the presented case study to ensure that issues actually affect the users, the overlap is 27 %. For the presented case study, the guideline review method is able to identify the most severe usability issues in both interface variants, but also reported several problems that did not occur during the

user testing used as a baseline. Karat et al. (1992) emphasize the influence of evaluator expertise for the result of the expert-based approaches. Unlike the results from Jeffries et al. (1991) who found a larger number of severe problems through HE than through usability testing, the severity in the presented case study was not raised explicitly. However, the results regarding the comparison of the UPC classifications show that user testing tends to find more problems related to the preparation of actions, while experts detect slightly more problems that occur during or after specific user actions. As an assumption this is due to the nature of experts who are less prone to uncertainty because of their expertise. Despite the provision of personae and a detailed briefing, the experts are not able to fully empathize with the potential users.

Another factor that influences the interpretation of the presented results is the determination of *real* problems. The applied approach to only define the problems identified through user testing as *real* problems contains some weaknesses. The nature of empirical studies cannot ensure the occurrence of every single usability issue during a usability study. This is influenced by several factors like the definition of tasks the users has to perform, the selection of target users, and the setup of the study itself. Another approach to determine realness of usability problems is described by Andre (2000, pp. 40 ff.) as standard usability problem list. In practice, there is usually no such list of usability problems with which the output of any UEM can be compared. Moreover, the application of a UEM would not be necessary at all if all possible usability problems in an interface would be familiar. Andre (2000, pp. 40 ff.) also argue that the results from lab-based usability testing "does not meet the definition for an ultimate criterion", because the situation in the lab does not represent the real context of use. Nonetheless, lab-based usability testing has a high level of confidence in the usability community as the mainstream UEM for formative evaluation. Furthermore, they argue that most alternative approaches arose from the need to find methods that are able to reduce the costs rather than increasing a higher quality of results. Another issue when comparing UEMs is the approach to measure the number of problems an individual UEM is able to detect. From a theoretical point of view "the important, and much harder, measure, is how *few* the technique has *missed*" (Doubleday et al., 1997, p. 104).

As already stated by Desurvire et al. (1991, p. 59), the expert-based methods are better suited when "competing interface alternatives are being considered" in order to narrow the variants for a following user study. Furthermore, Karat et al. (1992) suggest applying empirical usability testing for "baseline and other key checkpoint tests in the development cycle". In other phases during the project it may not be essential to identify all of the significant problems, but rather ease the choice between different interface variants. Other studies comparing expert-based approaches with user testing (c.f. Desurvire et al., 1991; Doubleday et al., 1997; Karat et al., 1992) support the fact that the "actual results produced by each technique are quite different in kind" (Doubleday et al., 1997, p. 108).

This leads to the conclusion that expert-based approaches are better suited to apply in earlier development phases or in order to eliminate interface variants. User-based approaches on the other hand can exploit its strengths as baseline testing or when it is necessary to identify most of the significant problems in an interface that actually affect real users. Therefore, the best results will be achieved by selecting a custom-tailored mixed methods approach for the specific research question of a project.

Looking at the different measures to assess the effectiveness of the guideline review approach for the automotive context of use, the values for the presented study show similar but also contrary results to studies from the literature of the usability community. The validity of the guideline review results ranks on a medium level, while thoroughness may be improved by consulting more experts. The reliability calculated from the number of real usability problems of each method shows a strong effect, which is contrary to the correlation calculated between the GCSs of the individual evaluators. The benefit-cost ratio analysis shows that the guideline review is not able to identify all of the problems that occurred during user testing, but the ratio of problems per person-hour is slightly higher for the expert-based approach. This result strengthens the assumption to increase the number of evaluators in order to increase the performance of the guideline review technique. On the other hand, in a real-world application, problems would already be fixed that seem irrelevant to real users. This results in unnecessary costs that can be avoided by using real users. Furthermore, design changes always bear the risk of causing new usability problems.

A comprehensive comparison to established methods like the HE or a traditional usability walkthrough was not performed in this case study, but could deliver valuable insights. Due to restrictions regarding confidentiality the recruitment of experts for several comparison studies is only possible to a limited extent. As the studies were performed in an industrial context, the availability of HMI experts was rather limited. In order to strengthen the findings regarding the relationship between the SUS and the GCS further research is needed.

7. Summary & Outlook

This dissertation is dedicated to the investigation of expert-based usability evaluation methods for the assessment of in-vehicle information systems. The application of the developed exploratory literature review approach in chapter 3 has shown that throughout the literature on human-computer interaction for in-vehicle information systems the research on expert-based usability evaluation methods is rather limited. After a profound discussion on several definitions of usability and the methods to evaluate usability, the conditions and implications of the specific context of use in the dual-task environment of a vehicle has been demonstrated. The already mentioned literature review exposed a lack of research on expert-based usability evaluation methods for in-vehicle information systems. Therefore, the implications of the context of use are discussed regarding the assumption or adaptation of existing expert-based usability evaluation methods for the application in the automotive domain. The two case studies demonstrated that the selected techniques cognitive walkthrough and guideline review are able to detect several problems that were also reported during separate usability tests, but lack acceptable values regarding the thoroughness. In summary, the expert-based usability evaluation methods turned out to suit well as cost-effective alternatives in earlier phases of the development cycle in order to eliminate interface variants or when it is not necessary to identify all potential usability problems. Nonetheless, expert-based testing is not a replacement to testing with actual users, a mixed methods approach adjusted to the specific characteristics of a project will achieve the best results — considering the following:

> "Expert-based tests are often used in conjunction with user-based tests, but the expert-based tests always come first." (Lazar et al., 2017, p. 268)

7.1. Main Contributions and Discussion

The following sections summarize the main contributions of this doctoral thesis to the body of knowledge in usability evaluation of in-vehicle information systems. This section is therefore divided roughly into the individual parts of the dissertation.

7.1.1. Usability Engineering in an Automotive Context

The chapter *Usability Engineering* gives a in-depth overview of the concept of usability and different definitions of usability throughout the literature. Brian Shackel, the grandfather of human-computer interaction (HCI) (Dillon, 2009, p. 367) defined usability — the interaction between the user, the task, and the environment — through the factors effectiveness, learnability, flexibility, and attitude also known as LEAF. Jakob Nielsen also used the factor of learnability as on of his five components of usability between efficiency, memorability, errors, and satisfaction. Furthermore, Jakob Nielsen introduced the concept of discount usability with the background that some usability engineering activities are still better than none at all because of high cost. The techniques user observation with think aloud, scenarios, and heuristic evaluation should provide the researcher with simple and cost-effective methods to assess the usability of a product. Donald Norman's focus on the shift of attention of usability from the product to the user played an important role for todays understanding of user-centered design. Moreover, Norman has shaped the need for uniform principles of interaction, suggesting the principles of affordances, signifiers, constraints, mappings, feedback, and conceptual models. Another concept of universal usability introduced by Ben Shneiderman is based on challenges to meet the needs of users with different background and characteristics as well as the variety of different technology. These challenges are also addressed by the eight golden rules of interface design.

Besides an overview of usability in different international standards, chapter 2 gives an overview of the user-centered design process and its individual stages. Since this dissertation focuses on the phase of evaluation in the process, different usability evaluation methods are presented and discussed. Apart from different survey tools the techniques interview and focus group as well as traditional usability testing belong to the group of methods that directly involve target users. The section also introduces several measures of human performance like task performance, specific tracking of user behavior, or techniques to measure mental workload. Some of these measurements serve as a basis for model-based approaches like the goals, operators, methods, and selection rules (GOMS) or the keystroke-level model (KLM) that use more or less abstract representations to predict human performance. Another aspect of usability testing is the automation of evaluation through different stages of automation from automatic capturing to automatic analysis and automatic critic. In section 2.4.3 several expert-based usability evaluation techniques are presented that do not involve users. Traditional methods from this category are the heuristic evaluation and the cognitive walkthrough as well as the guideline review and the consistency inspection. Furthermore, a more recent approach based on human thinking is represented by the metaphors of human thinking technique.

During the rather general considerations of different aspects of usability, the context of use is a constant part of several individual definitions. Since the context of use inside a vehicle is characterized by the aspects of a dual task environment, this special context of use is elaborated in section 2.5. Not only the shift of attention in this dual task environment which is among other things also influenced by environmental conditions, the range of users, or the frequency of use, but also the key concepts of multimodal interaction play an important role when designing interfaces for the use inside a vehicle. The design space developed by Kern and Schmidt (2009) addresses this trade-off between different tasks while driving and multimodality. By analyzing the different input and output modalities critical use cases as well as new potentials for user interfaces can be disclosed. Regarding critical use cases several techniques for the assessment of driver distraction are discussed in section 2.5.3. Besides different guidelines throughout the industry (c.f. Alliance of Automobile Manufacturers, 2006; Bhise, 2002; Commission of the European Communities, 2008; Kroon et al., 2016; National Highway Traffic Safety Administration, 2013; Society of Automotive Engineers, 2001), the section introduces common techniques to measure driver distraction from survey tools like the NASA Task Load Index, the Subjective Workload Assessment Technique, or the Driving Activity Load Index, to user testing with the occlusion technique, peripheral detection tasks, electrocardiogram and electroencephalogram, or the standardized lane change task.

As the chapter *Usability Engineering* shows, different definitions of usability overlap in several key concepts and influence todays understanding of usability and it's role in the user-centered design process. The craftsmanship of interface design is supported by a number of techniques to assess the usability of interfaces in different stages of the development and a wide research community is working on the development of new and advanced techniques for a changing environment. The concept of the context of use strongly influences this craftsmanship, which is why the specific context of use in the vehicle plays a decisive role in the design of in-vehicle information systems. Special techniques and guidelines were developed to overcome the challenges of this dual task environment shaped by aspects of multimodality and driver distraction to create an environment of safe driving.

7.1.2. Literature Review of the Current State of the Art

The literature review in the chapter *Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Information Systems* aimed to get an overview of the research landscape in HCI for in-vehicle information systems. The exploratory approach using the graph database system *Neo4j* enabled the application of network analytics to investigate relationships in the research community and offered various possibilities for visualization

of information from the network. In addition to an overview of which methods are currently being most commonly applied in the field, the chapter provides an insight into existing communities of individual authors and institutions. The structured approach using a schematic representation of the relationships between authors, papers, and venues where the papers got published is not specific to research on HCI and could therefore be applied to several other research areas. While the applied community detection algorithms are able to identify community structures in research landscapes, the centrality measures show relative importance of individuals for different research communities and the network.

Regarding the applied evaluation techniques, most of the articles examined report the application of methods with direct involvement of users like performance measurements, observation, and questionnaires. This result is indeed gratifying when considering that user-centered design thrives on involving the user as early as possible, but also highlights a lack of research addressing other techniques including expert-based or model-based approaches. As several model-based approaches have already been investigated for the specific context of use inside a vehicle (c.f. section 4.1), this dissertation focused on the application of expert-based approaches for usability evaluation of in-vehicle information systems, as this group of methods has already been elaborated successfully for other contexts than a vehicle (c.f. section 4.2). Moreover, the result that the most commonly investigated interface types during the literature review were traditional interfaces like touch displays located in the center stack of a car or dashboard displays is hardly surprising.

Besides the already mentioned rather obvious evaluations, network analytics offers more in-depth metrics to identify key roles in the community. While most authors in the network are affiliated with institutions in Germany, the most productive institutions, when looking at the number of publications, are located in the UK, Austria and the USA. Looking at the PageRank metric, Gary Burnett from the University of Nottingham as well as Manfred Tscheligi from the University of Salzburg take key roles in their corresponding research community. The community around Gary Burnett contains several individual authors acting as bridges between smaller sub-communities. The average community contains six authors with around two publications in the investigated time period of 2015 to 2017.

The literature review is not only able to give results on a snapshot of the research landscape of HCI for in-vehicle information systems, but also highlights underrepresented research areas. One of them — the application of expert-based approaches for the assessment of usability for interfaces inside the context of use of a vehicle — was discussed in detail in this dissertation.

7.1.3. Expert Reviews in Automotive

The chapter *Expert Reviews for Automotive User Interfaces* delimited the scope of the case studies in the following chapters 5 and 6. As expert-based methods were not the only underrepresented category of techniques, an introductory section discussed among others the usage of model-based approaches in the automotive context. While some model-based techniques are already established in the field through the Society of Automotive Engineers recommended practice J2365 or the extended keystroke-level model (KLM) approach by Pettitt et al. (2007) to predict the visual load during secondary tasks, the practice of automated evaluation is still in its infancy. The automatic logging of usage data from in-vehicle information systems faces technical challenges like data transfer as well as legal restrictions regarding data protection.

Expert-based methods and inspection techniques have already been successfully tested in other contexts of use than a vehicle (c.f. section 4.2). Therefore, expert-based approaches offer cost-effective and intuitive as well as quickly and early applicable alternatives to traditional usability testing. The dimensions of users expertise by Nielsen (1993) can be easily adapted into dimensions of experts expertise using axes for technique, domain, and system, in order to support the selection of experts for a expert-based evaluation study. Because most of the expert review methods profit from selecting experts from different backgrounds or perspectives, the measurement of agreement plays a subordinate role. In further step several definitions of the term usability problem were discussed. Due to its comprehensible theoretic foundation and its structured separation in task and object component, the Usability Problem Classifier aims to support the expert during a review to focus on the problem's effect on a user rather than the feature itself.

Since the introduction of expert-based usability evaluation approaches, the measurement of their effectiveness is of paramount interest. The section *Measuring the Effectiveness of Usability Evaluation Methods* presents several measures from the literature including validity, thoroughness, reliability, and cost effectiveness to compare inspection techniques to traditional usability testing. These measures as well as the specifications on the selection of experts using the adapted dimensions of expert expertise and the documentation of usability problems using the Usability Problem Classifier were applied in the case studies in chapters 5 and 6 to compare the expert-based techniques to traditional user testing.

7.1.4. Case Study: Cognitive Walkthrough

The case study in the chapter *Case Study: A Cognitive Walkthrough of a Driver Display Context Menu* compared the cognitive walkthrough technique with the results of a

usability test regarding the relative effectiveness. The cognitive walkthrough technique was extended by a question regarding the operation while driving as well as an estimation of task performance by the experts. The comparison was performed for the collected data of task performance as well as the usability problems detected by each approach. While for some of the carried out user tasks both methods showed a similar trend, several tasks resulted in different assessments of task performance. It should be noted that the use of solely employees at Mercedes-Benz Research & Development favors a bias towards technically-skilled participant of the usability test.

The lists of usability problems from both approaches overlapped in around a third of the total usability problems. The cognitive walkthrough reported task-based usability issues with a focus on the stage of action specification and failed to interpret individual features in the overall context of an in-vehicle information system. Most of these observations concerning the cognitive walkthrough technique are consistent with the literature (c.f. Cuomo & Bowen, 1992; Desurvire et al., 1992; Jeffries et al., 1991; John & Marks, 1997; Sears, 1997; Wharton et al., 1992) Due to the original development of the cognitive walkthrough to evaluate tasks performed by office workers in front of a computer, the technique is only suitable for the application in a dual task environment like a vehicle to a limited extent. Regardless of this and the extension of the cognitive walkthrough for the case study, the usability test did not find significantly more problems associated with distraction from the driving task. Looking at the measures of effectiveness, the cognitive walkthrough could produce valid results, showing even higher values than most of the compared studies, which also applied for the measure of thoroughness. The cognitive walkthrough produced a high reliability, but the value should be interpreted with caution as it does only consider the number of problems each expert detected. When it comes to cost-effectiveness, the results from the cognitive walkthrough did not significantly outperform the usability test value.

In summary, the cognitive walkthrough is able to detect several problems of the in-vehicle information system, but shows limited performance with regard to general and recurring problems as well as issues with the specific operation of individual interaction elements. The values for validity and thoroughness were satisfying compared to studies from the literature, while the benefit-cost ratio for the cognitive walkthrough did not outperform the usability test. Since one characteristic of expert-based usability evaluation methods is the reduction of cost while at the same time reasonably reducing the number of usability problems found, the method is not recommended. The cognitive walkthrough did not deliver adequate results without reducing the costs compared to a traditional usability test.

7.1.5. Case Study: Guideline Review

The chapter *Case Study: A Guideline Review of a Driver Display Screen Configuration Application* presented the second case study comparing the guideline review technique to the results of a usability test regarding relative effectiveness. The guideline review was applied using a consolidated list of guidelines from human-machine interaction (HMI) research, specific guidelines for the interaction in vehicles by governmental or non-profit organizations as well as from other research areas like website usability, accessibility, situation awareness, and persuasion. Therefore, a review of several guidelines was performed whereby each guideline was examined with regard to its relevance for the investigated interface. The resulting list of 27 guidelines got reviewed individually by six experts in order to weight them according to their importance which also represents the basis of the introduced metric — the Guideline Compliance Scale (GCS). The GCS closes the gap that most expert-based inspection techniques lack comparable metrics for effective communication of results to the stakeholders, or to compare different versions of a product in an iterative development. Several survey tools exist for this use case in traditional user testing (c.f. Chin et al., 1988; Kirakowski & Corbett, 1993), while the System Usability Scale (SUS) (c.f. Brooke, 1996) with only ten items provides comparable simplicity and is more likely to be applied in industrial usability evaluation. Therefore, the GCS metric from the guideline review was compared to SUS scores from the usability tests which showed a common trend between the two tested interface variants.

The lists of usability problems from both approaches overlapped in around a quarter of the total usability problems. The guideline review approach tended to find several interface issues that did not occur during usability testing with users. Although the severity of the individual usability was not reported for both studies, from a subjective perspective, the guideline review is able to identify most of the severe problems. Nonetheless, the traditional usability test unveils several problems regarding the planning of actions that were missed by the guideline review approach. An assumption is that this is due to the nature of experts who are less prone to uncertainty because of their expertise. Another factor that needs to be considered for the interpretation is the determination of *real* problems. The results of the case study were discussed on the basis that only problems that also occurred during user testing were considered as *real* problems. However, the guideline review approach was able to detect several usability problems that did not appear in the list of issues from the usability test, but therefore should not be simply dropped. There were no differences between the methods used to detect problems associated with driver distraction or related to the dual task environment. Therefore, the guideline review method appeared to be well suited to examine in-vehicle information systems.

Looking at the measures of effectiveness for the guideline review, the validity turned out on a medium level, while the value for thoroughness was reasonably similar to studies describing the heuristic evaluation technique (c.f. Jeffries et al., 1991; John & Marks, 1997). As in the first case study, the measure of reliability lacked to consider the actual substance of the individual problems and therefore contradicted the results for evaluator agreement. When it comes to cost-effectiveness, the benefit-cost ratio of the guideline review outperformed the usability study slightly when only considering the problems detected also by the user study, but surpassed it more than twofold considering all detected usability problems.

Summarizing the results, the guideline review was able to detect several usability problems of the in-vehicle information system, but also unveiled a list of potential false positives or phantom problems. Compared to studies from the literature the values for validity and thoroughness were acceptable. From a pragmatic perspective, the more important measure of effectiveness is not the number of problems that were detected by the guideline review rather than the number of problems it misses (Doubleday et al., 1997), and the guideline review has proven to find most of the severe problems. With the Guideline Compliance Scale (GCS) a new metric was introduced closing a gap in expert-based usability evaluation, although it needs further investigation regarding the relation to other metrics of usability. As a conclusion, the guideline review is well suited to be applied in earlier stages of development in order to eliminate interface variants or when it is not necessary to detect every single problem of an interface, as already stated in the literature (c.f. Desurvire et al., 1991; Doubleday et al., 1997). Furthermore, the benefit-cost ratio clearly speaks for the guideline review.

7.2. Outlook and Future Work

While the previous sections presented a summary and discussion of the main contributions of this dissertation, the following sections should highlight several opportunities for future work. Since the two case studies in this dissertation could only cover a part of the expert-based evaluation methods available, there remains material for consecutive studies. Moreover, this dissertation identified a need for a structured selection aid of usability evaluation methods for the specific context of automotive user interfaces.

7.2.1. Expanding the Scope of Literature Review

The exploratory literature review in chapter 3 identified trends in the research fields of HCI for in-vehicle information systems. While the exemplary results are only able

to convey a snapshot of the results, the provided web interface coming with the *Neo4j* framework allows detailed investigation of the network and the relationships between nodes. In order to develop a deeper understanding of the research landscape, besides the investigation of additional publications from different sources, an observation of the publication behavior over a longer period of time can be of interest.

Moreover, the schema could be further extended by adding the different departments of listed institutions, authors are affiliated with, or capturing keywords for the classified papers. There are a lot of sources not included in the presented literature review, which could be classified and added to the dataset. Especially different sources of human-computer interaction research, for example the journals *Proceedings of the Human Factors and Ergonomics Society*, *Ergonomics* or *Human Factors* as well as further conference proceedings like the *International Conference on Advances in Computer-Human Interfaces (ACHI)* or the *Conference on Universal Usability (CUU)* should be considered.

Another opportunity for future work could be the automated classification through machine learning and text mining techniques. The dataset created in the presented literature review could serve as training data, whereupon several studies could be added automatically. In such a scenario the automated analysis of change in a specific domain over time would be possible, as the review in this paper only captures a snapshot of the current research situation.

Since the graph schema is very generic, the method can also be applied to other research areas than HCI for in-vehicle information systems. The entities of type *Interface* may then be adapted to the specific needs. The same holds for an analysis aiming at research methods.

7.2.2. Consecutive Studies

As already mentioned in the discussion of the second case study (c.f. section 6.5.4), a comparison of the developed guideline review technique to the already established usability walkthrough would be of interest. While the guideline review is able to produce reasonable results by reducing the costs, a comparison to traditional methods could provide a further classification.

Other approaches from the presented methods in section 2.4.3 like metaphors of human thinking might need adjustments to fit the specific context of use of in-vehicle information systems. Since Frøkjær and Hornbæk (2008) state that the metaphors of human thinking technique is able to detect more serious usability problems than a heuristic evaluation, an investigation for the application on in-vehicle information systems could be useful. Techniques like the consistency inspection could deliver

meaningful results for automobile manufacturers across their entire model range. Since most of the components of an in-vehicle information system are not developed by a single team, but in cross-departmental cooperation, the results could contribute to an overall consistency of the model range or between different devices in the vehicle.

Besides the comparison of the presented results with further methods from the category of expert-based evaluation techniques, a comparison with other method categories can be of interest. Therefore, a comparison with already established model-based approaches like keystroke-level model (KLM) or goals, operators, methods, and selection rules (GOMS) which had been discussed in section 4.1 could provide meaningful insights.

7.2.3. Toward a Toolbox for Usability Evaluation of In-Vehicle Information Systems

The scope of this dissertation has been narrowed down to two concrete expert-based usability evaluation methods in the form of comparative case studies. However, the range of available methods is much wider, which can cause difficulties in their selection. When extending the scope of expert-based usability evaluation methods for the evaluation of in-vehicle information system to a more wider context of general usability evaluation methods for the evaluation of in-vehicle information systems, a structured selection aid could support professionals in the industry. As mentioned in the introduction of this dissertation (c.f. section 1.1), the field of HMI for in-vehicle information systems is still in its infancy. Therefore, professionals with a more technical background, as it is common in the automobile industry, would benefit from a structured toolbox for usability evaluation.

References

100 Jahre Tachometer. Tempomesser kam nur langsam auf Touren. (2002). Retrieved April 19, 2018, from http://www.spiegel.de/auto/werkstatt/100-jahre-tachometer-tempomesser-kam-nur-langsam-auf-touren-a-222464.html

Aston Martin Lagonda. (2018). Retrieved April 25, 2018, from https://de.wikipedia.org/wiki/Aston_Martin_Lagonda

Ford Model T. (2018). Retrieved April 20, 2018, from https://en.wikipedia.org/wiki/Ford_Model_T

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, *5*(4), 594–607.

Agah, A. (2000). Human interactions with intelligent systems: Research taxonomy. *Computers & Electrical Engineering*, *27*(1), 71–107.

Aggarwal, C. C. (2011). An Introduction to Social Network Data Analytics. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 1–15). Springer US.

Alexander von Humboldt-Stiftung. (2009). *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen* (Second). Alexander von Humboldt-Stiftung.

Alliance of Automobile Manufacturers. (2006). *Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems* (tech. rep.).

Andre, T. S., Belz, S. M., McCrearys, F. A., & Hartson, H. R. (2000). Testing a Framework for Reliable Classification of Usability Problems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(37), 573–576.

Andre, T. S., Hartson, H. R., Belz, S. M., & McCreary, F. A. (2001). The user action framework: A reliable foundation for usability engineering support tools. *International Journal of Human-Computer Studies*, *54*(1), 107–136.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060.

Andre, T. S., Williges, R. C., & Hartson, H. R. (1999). The Effectiveness of Usability Evaluation Methods: Determining the Appropriate Criteria. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(20), 1090–1094.

Andre, T. S. (2000). *Determining the Effectiveness of the Usability Problem Inspector: A Theory-Based Model and Tool for Finding Usability Problems* (Doctoral dissertation). Virginia Polytechnic Institute and State University. https://pdfs.semanticscholar.org/89a8/b40b7a06932509dc89c50d6529bf045965e4.pdf

Apple Computer. (1992). *Macintosh Human Interface Guidelines*. Addison-Wesley Publishing Company.

Apple Computer. (1995). *Macintosh Human Interface Guidelines*. Addison-Wesley Publishing Company.

Arnaboldi, V., Dunbar, R. I. M., Passarella, A., & Conti, M. (2016). Analysis of Co-authorship Ego Networks. In A. Wierzbicki, U. Brandes, F. Schweitzer, & D. Pedreschi (Eds.), *Proceedings of the 12th International Conference and School on Network Science* (pp. 82–96). Springer.

Atterer, R., Wnuk, M., & Schmidt, A. (2006). Knowing the user's every move. *Proceedings of the 15th International Conference on World Wide Web - WWW '06*, 203.

Bach, K. M., Jæger, M. G., Skov, M. B., & Thomassen, N. G. (2008). You can touch, but you can't look: Interacting with in-vehicle systems. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, 1139.

Baldwin, C. L. (2002). Designing in-vehicle technologies for older drivers: Application of sensory-cognitive interaction theory. *Theoretical Issues in Ergonomics Science*, *3*(4), 307–329.

Balbo, S. (1995). Automatic evaluation of user interface usability: Dream or reality. *Proceedings of the Queensland Computer-Human Interaction Symposium*.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, *4*(3), 114–123.

Barreto, A., Gao, Y., & Adjouadi, M. (2008). Pupil Diameter Measurements: Untapped Potential to Enhance Computer Interaction for Eye Tracker Users? *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility - Assets '08*, 269.

Barnes, R. (1968). *Motion and Time Study* (Sixth). Wiley.

Bastien, J. M. C., & Scapin, D. L. (1993). *Ergonomic Criteria for the Evaluation of Human-Computer Interfaces* (tech. rep.).

Bastien, J. M. C., & Scapin, D. L. (1995). Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction*, *7*(2), 105–121.

Bederson, B. B., & Shneiderman, B. (Eds.). (2003). *The craft of information visualization: readings and reflections*. Morgan Kaufmann. http://www.loc.gov/catdir/description/els051/2002116252.html

Benz & Co. (1886). *Fahrzeug mit Gasmotorenbetrieb*. Kaiserliches Patentamt.

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps*. University of Wisconsin Press.

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies*, *55*(4), 533–552.

Bevan, N. (1995). Usability is Quality of Use. *Advances in Human Factors/Ergonomics* (pp. 349–354).

Beyer, H., & Holtzblatt, K. (1998). *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann.

Bhise, V. (2002). Designing Future Automotive In-Vehicle Devices: Issues and Guidelines. *Annual Meeting of the Transportation Research Board*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 8.

Böhm, V., & Wolff, C. (2013). Usability Engineering-Methoden im interkulturellen Kontext. In S. Boll, S. Maaß, & R. Malaka (Eds.), *Mensch & Computer 2013 - Workshopband* (pp. 457–462). Oldenburg Verlag. https://dl.gi.de/handle/20.500.12116/7673

Böhm, P., Wolff, C., & Schneidermeier, T. (2014). Heuristiken für information appliances. In A. Butz, M. Koch, & J. Schlichter (Eds.), *Mensch & computer 2014 - tagungsband* (pp. 275–284). De Gruyter Oldenbourg.

Böhm, V., & Wolff, C. (2014). A Review of Empirical Intercultural Usability Studies. In M. Aaron (Ed.), *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. DUXU 2014. Lecture Notes in Computer Science, vol 8517* (pp. 14–24). Springer International Publishing.

Böhm, V., Langer, A., & Wolff, C. (2018). Validierung von Web-Usability-Heuristiken für eine ältere Zielgruppe. In R. Dachselt & G. Weber (Eds.), *Mensch und computer 2018 - workshopband*. Gesellschaft für Informatik e.V.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Auflage). Springer
OCLC: 845714518.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107–117.

Brookhuis, K. A., de Vries, G., & de Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention*, *23*(4), 309–316.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis.

Burmester, M., Hassenzahl, M., & Koller, F. (2002). Usability ist nicht alles - Wege zu attraktiven Produkten (Beyond Usability - Appeal of interactive Products). *i-com*, *1*(1/2002), 32–40.

Burns, P. C., Parkes, A. M., Burton, S., Smith, R. K., & Burch, D. (2002). *How dangerous is driving with a mobile phone? benchmarking the impairment to alcohol* (tech. rep.). TRL Limited.

Burmester, M., Graf, R., Hellbrück, J., & Meroth, A. (2008). Usability — Der Mensch im Fahrzeug. *Infotainmentsysteme im Kraftfahrzeug* (pp. 321–355). Vieweg.

Burnett, G. E. (2000). Usable Vehicle Navigation Systems : Are We There Yet? *Vehicle Electronic Systems 2000 European Conference and Exhibition*, 3.1.1–3.1.11.

Burghardt, M. (2014). *Engineering Annotation Usability. Toward Usability Patterns for Linguistic Annotation Tools.* (Doctoral dissertation). Universität Regensburg.

Butz, A., & Krüger, A. (2017). *Mensch-Maschine-Interaktion*. Walter de Gruyter GmbH.

Byrne, M. D. (2003). Cognitive Architecture. In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook* (pp. 1–57). Erlbaum.

Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008). Web Content Accessibility Guidelines (WCAG) 2.0. Retrieved November 28, 2018, from https://www.w3.org/TR/2008/REC-WCAG20-20081211/

Card, S., Moran, T., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410.

Card, S., Newell, A., & Moran, T. (1983). *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc.

Carroll, J. M., & Rosson, M. B. (1992). Getting around the task-artifact cycle: How to make claims and design by scenario. *ACM Transactions on Information Systems*, *10*(2), 181–212.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision To Think*.

Chin, J. P., Diehl, V. A., & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '88*, 213–218.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed., reprint). Psychology Press
OCLC: 642919193.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159.

Commission of the European Communities. (2008). *Commission Recommendation of 26 May 2008 on Safe and Efficient In-Vehicle Information and Communication Systems* (tech. rep.). http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008H0653&from=EN

Coutaz, J., & Balbo, S. (1994). Evaluation des interfaces utilisateur: Taxonomie et recommandations. *IHM'94, Lille*, 211–218. http://iihm.imag.fr/publication/CB94a/

Cox III, E. P. (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research*, *17*(4), 407.

Cuomo, D. L., & Bowen, C. D. (1992). Stages of User Activity Model as a Basis for User-System Interface Evaluations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*(16), 1254–1258.

Cuomo, D. L., & Bowen, C. D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers*, *6*(1), 86–108.

Curzon, P., Blandford, A., Butterworth, R., & Bhogal, R. (2002). Interaction design issues for car navigation systems. *People and Computers XVI - Memorable Yet Invisible: Proceedings of HCI 2002*, 4.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies. *Proceedings of the 1st International Conference on Intelligent User Interfaces - IUI '93*, *6*, 193–200.

Damiani, S., Deregibus, E., & Andreone, L. (2009). Driver-vehicle interfaces and interaction: Where are they going? *European Transport Research Review*, *1*(2), 87–96.

Davis, J., & Rebelsky, S. A. (2007). Food-First Computer Science: Starting the First Course Right with PB&J. *ACM SIGCSE Bulletin*, *39*(1), 372.

Davidson, T. L. (1975). When ... if ever ... are Focused Groups a Valid Research Tool? In E. M. Mazze (Ed.), *1975 Combined Proceedings, Series No. 37: Marketing in Turbulent Times and Marketing: The Challenges and the Opportunities*. American Marketing Association.

de Vicente, A., & Pain, H. (2002). Informing the Detection of the Students' Motivational State: An Empirical Study. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems* (pp. 933–943). Springer Berlin Heidelberg.

De Lange, C., & Glänzel, W. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part I. Development of a new model using a series expansion approach. *Scientometrics*, *40*(3), 593–604.

De S. Price, D. (1981). Multiple Authorship. *Science*, *212*(4498), 986–986.

Desurvire, H., Lawrence, D., & Atwood, M. (1991). Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *ACM SIGCHI Bulletin*, *23*(4), 58–59.

Desurvire, H., Kondziela, J., & Atwood, M. E. (1992). What is gained and lost when using methods other than empirical testing. *Posters and Short Talks of the 1992 SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 125.

Design Council UK. (2005). The "double diamond"design process model. Retrieved August 7, 2018, from https://www.designcouncil.org.uk/news-opinion/design-process-what-double-diamond

Desurvire, H. (1994). Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 173–202). John Wiley & Sons, Inc. Retrieved May 8, 2019, from http://dl.acm.org/citation.cfm?id=189200.189217

Devi, M. S., & Bajaj, P. R. (2008). Driver Fatigue Detection Based on Eye Tracking. *2008 First International Conference on Emerging Trends in Engineering and Technology*, 649–652.

Di Caro, L., Cataldi, M., & Schifanella, C. (2012). The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*, *93*(3), 583–607.

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research*, *8*(2), 181–192.

Diestel, R. (2017). *Graph Theory* (Vol. 173). Springer Berlin Heidelberg.

Dillon, A. (2009). Inventing HCI: The grandfather of the field. *Interacting with Computers*, *21*(5-6), 367–369.

Dingus, T. A., McGehee, D., Hulse, M., Jahns, S., Manakkal, N., Mollenbauer, M., & Fleischman, R. (1995). *Travtek Evaluation Task C3 - Camera Car Study* (tech. rep.). Performance and Safety Sciences, Inc.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. *Proceedings of the Conference on Designing Interactive Systems Processes, Practices, Methods, and Techniques - DIS '97*, 101–110.

Duchowski, A. T. (2017). *Eye Tracking Methodology*. Springer International Publishing.

Dumas, J. S., & Loring, B. A. (2008). *Moderating Usability Tests: Principles and Practices for Interacting*. Morgan Kaufmann.

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In D. Lalanne & J. Kohlas (Eds.), *Human Machine Interaction: Research Results of the MMI Program* (pp. 3–26). Springer Berlin Heidelberg.

Dumas, J. S., & Redish, J. C. (1999). *A Practical Guide to Usability Testing*. Intellect Books.

Durham, T. (1998). Science of the appliance. Retrieved February 27, 2020, from https://www.timeshighereducation.com/features/science-of-the-appliance/109788.article

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Edge, D. (1979). Quantitative Measures of Communication in Science: A Critical Review. *History of Science*, *17*(2), 102–134.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 97–120.

Fang, X., Xu, S., Brzezinski, J., & Chan, S. S. (2006). A Study of the Feasibility and Effectiveness of Dual-Modal Information Presentations. *International Journal of Human–Computer Interaction*, *20*(1), 3–17.

Faraday, P. (2000). Visually Critiquing Web Pages. *Multimedia '99*, 155–166.

Fastrez, P., & Haué, J.-B. (2008). Designing and evaluating driver support systems with the user in mind. *International Journal of Human-Computer Studies*, *66*(3), 125–131.

Faulkner, L. (2003). Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383.

Fern, E. F. (1983). Focus Groups: A Review of Some Contradictory Evidence, Implications, and Suggestions for Future Research. *Advances in Consumer Research*, *10*, 121–126. http://acrwebsite.org/volumes/6093/volumes/v10/NA-10

Feuerstack, S., Wortelen, B., Kettwich, C., & Schieben, A. (2016). Theater-system Technique and Model-based Attention Prediction for the Early Automotive HMI Design Evaluation. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - Automotive'UI 16*, 19–22.

Finstad, K. (2010a). Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales. *Journal of Usability Studies*, *5*(3), 104–110.

Finstad, K. (2010b). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5), 323–327.

Frank, T. L., Noy, Y. I., & Klachan, C. (2002). Occlusion Paradigm As a Tool To Assess Visual Distraction From in-Vehicle Telematics. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, 1863–1867.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Frøkjær, E., & Hornbæk, K. (2002). Metaphors of Human Thinking in HCI : Habit, Stream of Thought, Awareness, Utterance, and Knowing. *Proceedings of HF2002/OzCHI2002*.

Frøkjær, E., & Hornbæk, K. (2005). Cooperative usability testing. *CHI '05 Extended Abstracts on Human Factors in Computing Systems - CHI '05*, *2*, 1383.

Frøkjær, E., & Hornbæk, K. (2008). Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction*, *14*(4), 1–33.

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, *37*(3), 461–472.

Geiser, G. (1985). Man machine interaction in vehicles. *ATZ*, *87*(74-77), 56.

Glänzel, W., & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics*, *40*(3), 605–626.

Google Inc. (2019). Material Design Guidelines. Retrieved September 9, 2019, from https://material.io/design/

Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing Letters With a Simulated Listening Typewriter. *Communications of the ACM*, *26*(4), 295–308.

Gray, W. D., & Salzman, M. C. (1998). Damaged Merchandise? a Review of Experiments That Compare Usability Evaluation Methods. *Human–Computer Interaction*, *13*(3), 203–261.

Greenberg, J., Tijerina, L., Curry, R., Artz, B., Cathey, L., Kochhar, D., Kozak, K., Blommer, M., & Grant, P. (2003). Driver Distraction: Evaluation with Event Detection Paradigm. *Transportation Research Record: Journal of the Transportation Research Board*, *1843*(1), 1–9.

Green, P. A., Hoekstra, E., & Williams, M. (1993). *On-the-Road Tests of Driver Interfaces. Examination of a Route Guidance System and a Car Phone* (tech. rep.). University of Michigan.

Green, P. A. (2012). Motor Vehicle-Driver Interfaces. In J. A. Jacko (Ed.), *The Human-Computer Interaction Handbook* (Third Edit, pp. 749–771). CRC Press.

Green, P. A. (1999). Estimating Compliance with the 15-Second Rule for Driver-Interface Usability and Safety. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*, 987–991.

Grudin, J. (2009). Brian Shackel's contribution to the written history of Human–Computer Interaction. *Interacting with Computers*, *21*(5-6), 370–374.

Grudin, J. (1989). The case against user interface consistency. *Communications of the ACM*, *32*(10), 1164–1173.

Haigney, D., Taylor, R., & Westerman, S. (2000). Concurrent mobile (cellular) phone use and driving performance: Task demand characteristics and compensatory processes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*(3), 113–121.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *15*(1), 145–181.

Harvey, C., & Stanton, N. A. (2013). *Usability Evaluation for In-Vehicle Systems*. CRC Press.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* (pp. 139–183). North-Holland.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

Harvey, C. (2011). *Modelling and evaluating drivers' interactions with in-vehicle information systems (IVIS)* (Doctoral dissertation). University of Southampton.

Harley, A. (2018). UX Expert Reviews. Retrieved March 20, 2019, from https://www.nngroup.com/articles/ux-expert-reviews/

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '00*, *2*, 201–208.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). B.G. Teubner.

Hassenzahl, M., Burmester, M., & Koller, F. (2008). Der User Experience auf der Spur: Zum Einsatz von www.attrakdiff.de. In H. Brau, S. Diefenbach, M. Hassenzahl, F. Koller, M. Peissner, & K. Röse (Eds.), *Usability Professionals 2008* (pp. 78–82). Fraunhofer Verlag. http://attrakdiff.de/sience.html

Hassenzahl, M. (2001). The Effect of Perceived Hedonic Quality on Product Appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481–499.

Haslegrave, C. M. (1993). Visual aspects in vehicle design. *Automotive ergonomics* (pp. 79–98). Taylor & Francis. http://worldcat.org/isbn/0748400052

Healey, C. G., Booth, K. S., & Enns, J. T. (1996). High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *3*(2), 107–135.

Hertzum, M., & Jacobsen, N. E. (2001). The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *13*(4), 421–443.

Hertzum, M., & Jacobsen, N. E. (1999). The Evaluator Effect during First-Time Use of the Cognitive Walkthrough Technique. *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*, 1063–1067. http://dl.acm.org/citation.cfm?id=647943.743299

Herriotts, P. (2005). Identification of vehicle design requirements for older drivers. *Applied Ergonomics*, *36*(3), 255–262.

Hix, D., & Hartson, H. R. (1993). *Developing User Interfaces: Ensuring Usability Through Product and Process*. John Wiley & Sons.

Hjälmdahl, M., & Várhelyi, A. (2004). Speed regulation by in-car active accelerator pedal. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(2), 77–94.

Holtzblatt, K., & Beyer, H. (2017). *Contextual Design: Design for Life* (Second). Morgan Kaufmann.

Holcomb, R., & Tharp, A. (1989). An amalgamated model of software usability. *[1989] Proceedings of the Thirteenth Annual International Computer Software & Applications Conference*, 559–566.

Holcomb, R., & Tharp, A. L. (1991). What users say about software usability. *International Journal of Human-Computer Interaction*, *3*(1), 49–78.

Hopcroft, J., & Tarjan, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. *Communications of the ACM*, *16*(6), 372–378.

Hulse, M., Dingus, T., Mollenhauer, M., Liu, Y., Jahns, S., Brown, T., & McKinney, B. (1998). *Development of human factors guidelines for advanced traveler information*

systems and commercial vehicle operations : Identification of the strengths and weaknesses of alternative information display formats (tech. rep. FHWA-RD-96-142). Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/14092

Hurwitz, J. B., & Wheatley, D. J. (2002). Using Driver Performance Measures to Estimate Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(22), 1804–1808.

Hurtado, S., & Chiasson, S. (2016). An Eye-tracking Evaluation of Driver Distraction and Unfamiliar Road Signs. In P. A. Green, S. Boll, G. Burnett, J. Gabbard, & S. Osswald (Eds.), *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 153–160). ACM Press.

Hwang, W., & Salvendy, G. (2010). Number of People Required for Usability Evaluation: The 10±2 Rule. *Communications of the ACM*, *53*(5), 130.

International Organization for Standardization. (2001). *Software engineering - Product quality - Part 1: Quality model* (International Standard ISO 9126-1).

International Organization for Standardization. (2002). *Ergonomics of human-system interaction - Usability methods supporting human-centered design* (International Standard ISO 16982).

International Organization for Standardization. (2008). *Ergonomics of human-system interaction - Part 110: Dialogue principles* (International Standard ISO 9241-110).

International Organization for Standardization. (2011a). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems* (International Standard ISO 9241-210).

International Organization for Standardization. (2011b). *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models* (International Standard ISO 25010).

International Organization for Standardization. (2017). *Road vehicles - Ergnonomic aspects of transport information and control systems - Specifications and test procedures for in-vehicle visual presentation* (International Standard ISO 15008).

International Organization for Standardization. (2018). *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts* (International Standard ISO 9241-11). https://www.iso.org/standard/63500.html

International Organization for Standardization. (2019a). *Ergonomics of human-system interaction - Part 110: Interaction principles* (International Standard ISO/FDIS 9241-110).

International Organization for Standardization. (2019b). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems* (International Standard ISO 9241-210).

International Organization for Standardization. (1991). *Information technology - Software product evaluation - Quality characteristics and guidelines for their use* (International Standard ISO 9126).

International Organization for Standardization. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability* (International Standard ISO 9241-11).

Ioannidis, J. P. A. (2008). Measuring Co-Authorship and Networking-Adjusted Scientific Impact (E. Joly, Ed.). *PLoS ONE*, *3*(7), e2778.

Islinger, T., Köhler, T., & Ludwig, B. (2011). Driver Distraction Analysis based on FFT of steering wheel angle. *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interac-Tive Vehicular Applications*, 21–22.

Islinger, T., Köhler, T., & Wolff, C. (2011a). A Functional Driver Analyzing Concept. *Advances in Human-Computer Interaction*, *2011*, 1–4.

Islinger, T., Köhler, T., & Wolff, C. (2011b). Human modeling in a driver analyzing context: Challenge and benefit. *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11*, 99–102.

Itoh, M., Mizuno, Y., Mori, S., & Shin Yamamoto. (2005). Drivers Status Monitor. *21st International Conference on Data Engineering Workshops (ICDEW'05)*, 1201–1201.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, *33*(4), 470–516.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(19), 1336–1340.

Jaccard, P. (1902). Lois de Distribution Florale dans la Zone Alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *38*(144), 69–130.

Japan Automobile Manufacturers Association. (2004). *Guideline for In-vehicle Display Systems - Version 3.0* (tech. rep.).

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology - CHI '91*, 119–124.

Jenness, J. W., Lattanzio, R. J., O'Toole, M., & Taylor, N. (2002). Voice-Activated Dialing or Eating a Cheeseburger: Which is More Distracting during Simulated Driving? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(4), 592–596.

Jiang, X., Atkins, M. S., Tien, G., Bednarik, R., & Zheng, B. (2014). Pupil Responses during Discrete Goal-directed Movements. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 2075–2084.

John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, *16*(4-5), 188–202.

Jordan, P. W. (1998). *An Introduction to Usability*. Taylor & Francis.

Kahn, M. J., & Prail, A. (1994). Formal usability inspections. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 141–171). John Wiley & Sons.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 397–404.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18.

Kayl, K. (2000). The Networked Car: Where the Rubber Meets the Road. Retrieved January 16, 2018, from http://www.yorku.ca/dzwick/networked%20car%20open%20standards%20and%20eGas.doc

Keenan, S. L., Hartson, H. R., Kafura, D. G., & Schulman, R. S. (1999). The Usability Problem Taxonomy: A Framework for Classification and Analysis. *Empirical Software Engineering*, *4*(1), 71–104.

Keenan, S. L. (1996). *Product usability and process improvement based on usability problem classification* (Doctoral dissertation). Virginia Polytechnic Institute and State University. Retrieved May 8, 2019, from https://vtechworks.lib.vt.edu/handle/10919/39100

Kern, D., & Schmidt, A. (2009). Design space for driver-based automotive user interfaces. *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '09*, 3–10.

Kern, D. (2012). *Supporting the Development Process of Multimodal and Natural Automotive User Interfaces* (Doctoral dissertation). Universität Duisburg-Essen.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212.

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). *The Impact of Driver Inattention On Near-Crash/Crash Risk. An Analysis Using the 100-Car Naturalistic Driving Study Data* (tech. rep.). Virginia Polytechnic Institute and State University.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications - ETRA '08*, 69–72.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Kroon, E. C. M., Martens, M. H., Brookhuis, K. A., Hagenzieker, M. P., Alferdinck, J. W. A. M., Harms, I. M., & Hof, T. (2016). *Human Factor Guidelines for the Design of Safe In-Car Traffic Information Services* (tech. rep.). DITCM.

Kuhn, K. (2000). Problems and Benefits of Requirements Gathering With Focus Groups: A Case Study. *International Journal of Human-Computer Interaction*, *12*(3-4), 309–325.

Lamm, L., & Wolff, C. (2019). Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Systems. *11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19)*.

Lamm, L., & Wolff, C. (2021). GCS: A Quick and Dirty Guideline Compliance Scale. *Journal of Usability Studies*, *16*(3), 179–202.

Lamm, L. (2019). Exploratory Analysis of the Research Literature on Evaluation of In-Vehicle Systems.

Lansdown, T. C., Brook-Carter, N., & Kersloot, T. (2002). Primary Task Disruption from Multiple In-Vehicle Systems. *Journal of Intelligent Transportation Systems*, *7*(2), 151–168.

Landauer, T. K. (1995). *The Trouble with Computers: Usefulness, Usability, and Productivity*. MIT Press.

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. *Mensch & Computer 2006 - Mensch Und Computer Im Strukturwandel*, 125–134.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, *16*(4-5), 246–266.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods In Human-Computer Interaction* (Second). Morgan Kaufmann.

Lee, D., Oh, S., Heo, S., & Hahn, M. (2008). Drowsy Driving Detection Based on the Driver's Head Movement using Infrared Sensors. *2008 Second International Symposium on Universal Communication*, 231–236.

Lewis, J. R., & Sauro, J. (2009). The Factor Structure of the System Usability Scale. In M. Kurosu (Ed.), *Human Centered Design* (pp. 94–103). Springer Berlin Heidelberg.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When There's No Time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 2099.

Lewis, C. H., Polson, P. G., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90*, 235–242.

Lewis, C. H., Polson, P. G., & Rieman, J. (1991). *Cognitive walkthrough forms and instructions* (tech. rep. ICS 91-14).

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.

Ling, C., & Salvendy, G. (2005). Extension of heuristic evaluation method: A review and reappraisal. *Ergonomia*, *27*(3).

Lindgaard, G., & Chattratichart, J. (2007). Usability Testing: What Have We Overlooked? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 1415.

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4-5), 267–277.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, *41*(6), 1462–1480.

Lohmüller, V., Schmaderer, D., & Wolff, C. (2018). Heuristiken für Second-Screen-Anwendungen. In R. Dachselt & G. Weber (Eds.), *Mensch und computer 2018 - tagungsband*. Gesellschaft für Informatik e.V.

Lohmüller, V., Schmaderer, D., & Wolff, C. (2019). A Heuristic Checklist for Second Screen Applications. *i-com*, *18*(1), 55–65.

Lu, H., & Feng, Y. (2009). A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships. *Scientometrics*, *81*(2), 499–511.

Ludvigsen, K. E. (1997). A Century of Automobile Comfort and Convenience. *The Automobile: A Century of Progress* (pp. 99–120). SAE International.

Mack, R. L., & Montaniz, F. (1994). Observing, predicting, and analyzing usability problems. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 295–339). John Wiley & Sons.

Mack, R. L., & Nielsen, J. (1994). Executive summary. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 1–23). John Wiley & Sons.

Mackinlay, J. D. (1986). *Automatic design of graphical presentations* (Doctoral dissertation). Stanford University. Stanford, CA, USA.

MacEachren, A. M. (1995). *How Maps Work: Representation, Visualization, and Design*. The Guilford Press.

Mahajan, R., & Shneiderman, B. (1997). Visual and textual consistency checking tools for graphical user interfaces. *IEEE Transactions on Software Engineering*, *23*(11), 722–735.

Manseer, M., & Riener, A. (2014). Evaluation of Driver Stress while Transiting Road Tunnels. In L. N. Boyle, G. E. Burnett, P. Fröhlich, S. T. Iqbal, E. Miller, & Y. Wu (Eds.), *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1–6). ACM Press.

Manakhov, P., & Ivanov, V. D. (2016). Defining Usability Problems. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, (December), 3144–3151.

Manes, D., Green, P. A., & Hunter, D. (1998). *Prediction of destination entry and retrieval times using keystroke-level models* (tech. rep.). The University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/1188

Marchiori, M., & Latora, V. (2000). Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, *285*(3-4), 539–546.

Marcus, A. (2004). The Next Revolution: Vehicle User Interfaces. *interactions*, *11*(1), 40–47.

Matthews, M. L., Bryant, D. J., Webb, R. D. G., & Harbluk, J. L. (2001). Model for Situation Awareness and Driving: Application to Analysis and Research for Intelligent Transportation Systems. *Transportation Research Record: Journal of the Transportation Research Board*, *1779*(1), 26–32.

Matthews, R., Legg, S., & Charlton, S. (2003). The effect of cell phone type on drivers subjective workload during concurrent driving and conversing. *Accident Analysis & Prevention*, *35*(4), 451–457.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. I. (2015). The Psychometrics of Mental Workload : Multiple Measures Are Sensitive but Divergent. *Human Factors*, *57*(1), 125–143.

Mattes, S. (2003). *The Lane Change Task as a Tool For Driver Distraction Evaluation* (tech. rep.).

McSweeney, R. (1992). Sumi–a psychometric approach to software evaluation. *Unpublished MA (Qual) thesis in Applied Psychology, University College Cork, Ireland*.

Meier, E.-M., Böhm, P., & Wolff, C. (2017). Comparing Heuristic Walkthrough and User Studies in Evaluating Digital Appliances. In M. Gäde, V. Trkulja, & V. Petras (Eds.), *Everything Changes, Everything Stays the Same? understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)* (pp. 146–157). Verlag Werner Hülsbusch.

Meixner, G., & Müller, C. (Eds.). (2017). *Automotive User Interfaces*. Springer International Publishing.

Mezzanzanica, M., Mercorio, F., Cesarini, M., Moscato, V., & Picariello, A. (2018). GraphDBLP: A system for analysing networks of computer scientists through graph databases. *Multimedia Tools and Applications*, 1–32.

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, *1*(1), 61–67.

Minge, M., Thüring, M., Wagner, I., & Kuhr, C. V. (2017). The meCUE Questionnaire: A Modular Tool for Measuring User Experience. In M. Soares, C. Falcão, & T. Z. Ahram (Eds.), *Advances in Ergonomics Modeling, Usability & Special Populations* (pp. 115–128). Springer International Publishing.

Miniukovich, A., De Angeli, A., Sulpizio, S., & Venuti, P. (2017). Design Guidelines for Web Readability. *Proceedings of the 2017 Conference on Designing Interactive Systems - DIS '17*, 285–296.

Miniukovich, A., Scaltritti, M., Sulpizio, S., & De Angeli, A. (2019). Guideline-Based Evaluation of Web Readability. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12.

Mitsopoulos-Rubens, E., Trotter, M. J., & Lenne, M. G. (2011). Usability evaluation as part of iterative design of an in-vehicle information system. *IET Intelligent Transport Systems*, *5*(2), 112–119.

Mohs, C., Hurtienne, J., Scholz, D., & Rotting, M. (2006). Intuitivität: Definierbar, beeinflussbar, überprüfbar! *VDI Berichte* (pp. 215–224).

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*, *33*(3), 338–348.

Munzner, T. (2008). Process and Pitfalls in Writing Information Visualization Research Papers. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization* (pp. 134–153). Springer Berlin Heidelberg.

National Highway Traffic Safety Administration. (2013). Visual-Manual NHTSA Driver Distraction Guidelines for In- Vehicle Electronic Devices. *Federal Register*, *78*(81), 24818–24890.

Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A., & Hußlein, S. (2007). Intuitive Use of User Interfaces: Defining a Vague Concept. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics* (pp. 128–136). Springer Berlin Heidelberg.

Naumann, A., & Hurtienne, J. (2010). Benchmarks for intuitive interaction with mobile devices. *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '10*, 401.

Némery, A., Brangier, E., & Kopp, S. (2011). First Validation of Persuasive Criteria for Designing and Evaluating the Social Influence of User Interfaces: Justification of a Guideline. In A. Marcus (Ed.), *Design, User Experience, and Usability. Theory, Methods, Tools and Practice* (pp. 616–624). Springer Berlin Heidelberg.

Neo4j Inc. (2018a). Neo4j Graph Algorithms. Retrieved December 9, 2018, from http://neo4j-contrib.github.io/neo4j-graph-algorithms

Neo4j Inc. (2018b). Neo4j's Graph Query Language: An Introduction to Cypher. Retrieved December 7, 2018, from https://neo4j.com/developer/cypher-query-language/

Neo4j Inc. (2018c). What is a Graph Database? Retrieved December 7, 2018, from https://neo4j.com/developer/graph-database/

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*, 404–409.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, *70*(5), 056131.

Newman, M. E. J. (2018). *Networks* (Second). Oxford University Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In J. C. Chew & J. Whiteside (Eds.), *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90* (pp. 249–256). ACM Press.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 206–213.

Nielsen, J. (2000). Novice vs. Expert Users. https://www.nngroup.com/articles/novice-vs-expert-users/

Nielsen, J. (2012). How Many Test Users in a Usability Study. Retrieved August 14, 2018, from https://www.nngroup.com/articles/how-many-test-users/

Nielsen, J. (1989a). The Matters that really matter for hypertext usability. *Proceedings of the Second Annual ACM Conference on Hypertext - HYPERTEXT '89*, 239–248.

Nielsen, J. (1989b). Usability Engineering at Discount. *Designing and Using Human-Computer Interfaces and Knowledge Based Systems: Proceedings of the 3rd International Conference on Human-Computer Interaction*, 394–401. https://dl.acm.org/citation.cfm?id=92499

Nielsen, J. (1990a). Big Paybacks from 'discount' Usability Engineering. *IEEE Software*, *7*(3), 107–108.

Nielsen, J. (1990b). Paper versus Computer Implementations as Mockup Scenarios for Heuristic Evaluation. *INTERACT '90 Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction*, 315–320. https://dl.acm.org/citation.cfm?id=725312

Nielsen, J. (1990c). Traditional dialogue design applied to modern user interfaces. *Communications of the ACM*, *33*(10), 109–118.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 373–380.

Nielsen, J. (1993). *Usability Engineering*. Academic Press.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Conference Companion on Human Factors in Computing Systems - CHI '94*, 210.

Nielsen, J. (1994b). Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier. In R. G. Bias & D. J. Mayhew (Eds.), *Cost-justifying Usability* (pp. 245–272). Academic Press. http://dl.acm.org/citation.cfm?id=186524.186639

Nielsen, J. (1995a). 10 Usability Heuristics for User Interface Design. Retrieved November 28, 2018, from https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. (1995b). Severity Ratings for Usability Problems. Retrieved November 26, 2018, from https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard Coefficient for Keywords Similarity. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *Vol 1*, 380–384.

Noel, E., Nonnecke, B., & Trick, L. (2005). A Comprehensive Learnability Evaluation Method for In-Car Navigation Devices. *SAE Technical Papers*.

Norman, D. A., & Draper, S. W. (1986). *User centered system design; new perspectives on human-computer interaction*. L. Erlbaum Associates Inc.

Norman, D. A. (2013). *The Design of Everyday Things*. Basic Books.

Norman, D. A. (1983). Design principles for human-computer interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '83*, 1–10.

Norman, D. A. (1988). *The psychology of everyday things.* Basic Books.

Nunnally, J. C. (1978). *Psychometric theory* (2. ed.). McGraw-Hill.

Nuutila, E., & Soisalon-Soininen, E. (1994). On finding the strongly connected components in a directed graph. *Information Processing Letters*, *49*(1), 9–14.

Östlund, J., Nilsson, L., Carsten, O., Merat, N., Jamson, S., Janssen, W., Mouta, S., Carvalhais, J., Santos, J., Anttila, V., Sandberg, H., Waard, D., Brookhuis, K. A., Johansson, E., Engstrom, J., Victor, T. A., Harbluk, J. L., & Brouwer, R. (2004). *Deliverable 2 - HMI and Safety-Related Driver Performance* (tech. rep.). Human Machine Interface And the Safety of Traffic in Europe.

Ozok, A. A., & Salvendy, G. (2001). How consistent is your web design? *Behaviour & Information Technology*, *20*(6), 433–447.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web* (tech. rep.). Stanford InfoLab. Retrieved May 25, 2018, from http://ilpubs.stanford.edu:8090/422

Parush, A., Nadir, R., & Shtub, A. (1998). Evaluating the Layout of Graphical User Interface Screens: Validation of a Numerical Computerized Model. *International Journal of Human-Computer Interaction*, *10*(4), 343–360.

Patten, C. J., Kircher, A., Östlund, J., & Nilsson, L. (2004). Using mobile telephones: Cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, *36*(3), 341–350.

Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems*, *2*(4), 315.

Pettitt, M., Burnett, G., & Stevens, A. (2006). Extending the Keystroke Level Model (KLM) to assess the visual demand of in-vehicle information systems (IVIS). *Proceedings of the 13th ITS World Congress*.

Pettitt, M., Burnett, G. E., & Stevens, A. (2007). An extended keystroke level model (KLM) for predicting the visual demand of in-vehicle information systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 1515–1524.

Pettitt, M. (2008). *Visual demand evaluation methods for in-vehicle interfaces* (Doctoral dissertation). University of Nottingham.

Polson, P. G., & Lewis, C. H. (1990). Theory-Based Design for Easily Learned Interfaces. *Human–Computer Interaction*, *5*(2-3), 191–220.

Purucker, C., Naujoks, F., Prill, A., Krause, T., & Neukum, A. (2014). Vorhersage von Blickabwendungszeiten mit Keystroke-Level-Modeling. In A. Butz, M. Koch, & J. Schlichter (Eds.), *Mensch & Computer 2014 - Workshopband* (pp. 239–248). De Gruyter Oldenbourg.

Purucker, C., Naujoks, F., Prill, A., & Neukum, A. (2017). Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling. *Applied Ergonomics*, *58*, 543–554.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3).

Reed, M. P., & Green, P. A. (1999). Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task. *Ergonomics*, *42*(8), 1015–1037.

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Advances in Psychology* (pp. 185–218). North-Holland.

Reid, G. B., Potter, S. S., & Bressler, J. R. (1989). *Subjective Workload Assessment Technique (SWAT): A User's Guide (U)* (tech. rep.). Armstrong Aerospace Medical Research Laboratory. http://ar.iiarjournals.org/lookup/doi/10.21873/anticanres.12641

Rieman, J., Davies, S., Hair, D. C., Esemplare, M., Polson, P., & Lewis, C. (1991). An automated cognitive walkthrough. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology - CHI '91*, 427–428.

Roche, A., Lespinet-Najib, V., & Andre, J.-M. (2014). Use of usability evaluation methods in France: The reality in professional practices. *2014 3rd International Conference on User Science and Engineering (i-USEr)*, 180–185.

Rowley, D. E., & Rhoades, D. G. (1992). The cognitive jogthrough: A fast-paced user interface evaluation procedure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 389–395.

Rubin, J., & Chisnell, D. (2008). *Handbook of Usability Testing* (2nd ed). Wiley Publishing.

Rygula, A. (2009). Driving Style Identification Method Based on Speed Graph Analysis. *2009 International Conference on Biometrics and Kansei Engineering*, 76–79.

Sanders, M. S., & McCormick, E. J. (1993). *Human Factors in Engineering and Design*. McGraw-Hill.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, 1609.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the User Experience. Practical Statistics for User Research* (2nd Edition). Morgan Kaufmann.

Schank, T., & Wagner, D. (2005). Finding, Counting and Listing All Triangles in Large Graphs, an Experimental Study. In S. E. Nikoletseas (Ed.), *Experimental and Efficient Algorithms* (pp. 606–609). Springer Berlin Heidelberg.

Schneegaß, S., Pfleging, B., Kern, D., & Schmidt, A. (2011). Support for modeling interaction with automotive user interfaces. *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11*, 71.

Schmargendorf, M., Schuller, H.-M., Böhm, P., Isemann, D., & Wolff, C. (2018). Autonomous driving and the elderly: Perceived risks and benefits. In R. Dachselt & G. Weber (Eds.), *Mensch und computer 2018 - workshopband*. Gesellschaft für Informatik e.V.

Scholtz, J., Laskowski, S., & Downey, L. (1998). Developing Usability Tools and Techniques for Designing and Testing Web Sites. *In Proceedings of the 4th Conference on Human Factors & the Web*.

Schuh, C. (2009). Publikationsverhalten im Überblick – eine Zusammenfassung der einzelnen Diskussions- beiträge. *Diskussionsbeiträge der Alexander von Humboldt-Stiftung. Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen* (Second, pp. 6–13). Alexander von Humboldt-Stiftung.

Schmettow, M. (2012). Sample Size in Usability Studies. *Communications of the ACM*, *55*(4), 64.

Scott, J. (2017). *Social Network Analysis: A Handbook* (Fourth). Sage Publications.

Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction*, *9*(3), 213–234.

Shackel, B. (1959). Ergonomics for a Computer. *Design*, *120*, 36–39.

Shackel, B. (1986). Ergonomics in design for usability. *Proceedings of the Second Conference of the British Computer Society, Human Computer Interaction Specialist Group on People and Computers: Designing for Usability*, 44–64.

Shapiro, D. W. (1994). The Contributions of Authors to Multiauthored Biomedical Research Papers. *JAMA: The Journal of the American Medical Association*, *271*(6), 438.

Shackel, B. (1997). Human-computer Interaction—Whence and whither? *Journal of the American Society for Information Science*, *48*(11), 970–986.

Shirey, R. W. (1969). *Implementation and Analysis of Efficient Graph Planarity Testing Algorithms* (Doctoral dissertation). The University of Wisconsin - Madison.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., & Elmqvist, N. (2018). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (sixth ed.). Pearson.

Shneiderman, B. (2000). Universal usability. *Communications of the ACM*, *43*(5), 84–91.

Sivak, M. (1996). The Information That Drivers Use: Is it Indeed 90% Visual? *Perception*, *25*(9), 1081–1089.

Smith, D. C., Irby, C., Kimball, R., Verplank, B., & Harslem, E. (1982). Designing the Star User Interface. *Byte*, *7*(4), 242–282.

Snyder, B. (2015). Apple exec: 'The car is the ultimate mobile device'. Retrieved March 21, 2018, from http://fortune.com/2015/05/27/apple-cars-tesla/

Society of Automotive Engineers. (2001). *The SAE Handbook*.

Society of Automotive Engineers. (2016). *Calculation and Measurement of the Time to Complete In-Vehicle Navigation and Route Guidance Tasks (SAE Recommended Practice J2365)* (tech. rep.). Society of Automotive Engineers. Warrendale, PA. http://standards.sae.org/j2365_201607/

Solomon, J. (2009). Programmers, Professors, and Parasites: Credit and Co-Authorship in Computer Science. *Science and Engineering Ethics*, *15*(4), 467–489.

Spool, J., & Schroeder, W. (2001). Testing Web Sites: Five Users Is Nowhere Near Enough. *CHI '01 Extended Abstracts on Human Factors in Computing Systems - CHI '01*, 285.

Srinivasan, R., & Jovanis, P. P. (1997). Effect of in-vehicle route guidance on driver workload and choice of vehicle speed: Findings from a driving simulator experiment. In I. A. Noy (Ed.), *Ergonomics and safety of intelligent driver interfaces* (pp. 97–114). Lawrence Erlbaum Associates, Inc.

Stanton, N. A., & Baber, C. (1992). Usability and EC Directive 90/270. *Displays*, *13*(3), 151–160.

Stevens, A., Quimby, A., Board, A., Kersloot, T., & Burns, P. (2002). Design guidelines for safety of in-vehicle information systems. *Department of Transport, Local Government and the Regions*, 1–55.

Stevens, A., Byrgave, S., Brook-Carter, N., & Luke, T. (2004). *Occlusion as a technique for measuring In-Vehicle Information System (IVIS) visual distraction: A research literature review* (tech. rep.). Transport Research Laboratory. http://www.trl.co.uk/store/report_detail.asp?srid=5443%5Cnhttp://trid.trb.org/view.aspx?id=741077

Stevens, A., & Cynk, S. (2011). Checklist for the assessment of in-vehicle information systems (Transport Research Laboratory, Ed.). (MIS005).

Steinfeld, A., Manes, D., Green, P. A., & Hunter, D. (1996). *Destination Entry and Retrieval with the Ali-Scout Navigation System* (tech. rep.). The University of Michigan.

Streff, F., & Spradlin, H. (2000). *Driver Distraction, Aggression, and Fatigue: A Synthesis of the Literature and Guidelines for Michigan Planning* (tech. rep.). University of Michigan. Ann Arbor, MI.

Strayer, D. L., & Johnston, W. A. (2001). Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychological Science*, *12*(6), 462–466.

Strayer, D. L., & Drew, F. A. (2004). Profiles in Driver Distraction: Effects of Cell Phone Conversations on Younger and Older Drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 640–649.

Tarjan, R. (1972). Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing*, *1*(2), 146–160.

Theng, Y. L., & Marsden, G. (1998). Authoring tools: Towards continuous usability testing of web documents.

The Interaction Design Foundation. (2017). How to Conduct a Cognitive Walkthrough. Retrieved July 19, 2019, from https://www.interaction-design.org/literature/article/how-to-conduct-a-cognitive-walkthrough

Thimbleby, H. (1997). Gentler: A tool for systematic web authoring. *International Journal of Human-Computer Studies*, *47*(1), 139–168.

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264.

Tijerina, L., Parmer, E., & Goodman, M. J. (1998). Driver Workload Assessment of Route Guidance System. *Proceedings of the 5th ITS World Congress*, 1–8.

Tokuda, S., Palmer, E., Merkle, E., & Chaparro, A. (2009). Using Saccadic Intrusions to Quantify Mental Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(12), 809–813.

Tokuda, S., Obinata, G., Palmer, E., & Chaparro, A. (2011). Estimation of Mental Workload Using Saccadic Eye Movements in a Free-Viewing Task. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 4523–4529.

Tönnis, M., Broy, V., & Klinker, G. (2006). A Survey of Challenges Related to the Design of 3D User Interfaces for Car Drivers. *Proceedings of IEEE Symposium on 3D User Interfaces - 3DUI 2006*, 127–134. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1647518

Tory, M., & Möller, T. (2005). Evaluating Visualizations: Do Expert Reviews Work? *IEEE Computer Graphics and Applications*, *25*(5), 8–11.

Toutkoushian, R. K., Porter, S. R., Danielson, C., & Hollis, P. R. (2003). Using Publications Counts to Measure an Institution's Research Productivity. *Research in Higher Education*, *44*(2), 121–148.

Treffner, P. J., & Barrett, R. (2004). Hands-free mobile phone speech while driving degrades coordination and control. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(4-5), 229–246.

Treat, J. R. (1980). A Study of Pre-Crash Factors Involved in Traffic Accidents. *HSRI Research Review*, *10/11*, 1–35.

Trösterer, S., Meschtscherjakov, A., Wilfinger, D., & Tscheligi, M. (2014). Eye Tracking in the Car: Challenges in a Dual-Task Scenario on a Test Track. In L. N. Boyle, G. E. Burnett, P. Fröhlich, S. T. Iqbal, E. Miller, & Y. Wu (Eds.), *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1–6). ACM Press.

Turner, N. (2011). A guide to carrying out usability reviews. Retrieved September 9, 2019, from http://www.uxforthemasses.com/usability-reviews/

Ullrich, D., & Diefenbach, S. (2010). INTUI. Exploring the Facets of Intuitive Interaction. In J. Ziegler & A. Schmidt (Eds.), *Mensch & Computer 2010* (pp. 251–260). OLDENBOURG WISSENSCHAFTSVERLAG.

van Rens, L. S. (1997). *Usability Problem Classifier* (Unpublished Master's Thesis). Virginia Polytechnic Institute and State University. Blacksburg, VA.

Van Raan, a. F. J. (1998). The influence of international collaboration on the impact of research results. *Scientometrics*, *42*(3), 423–428.

Vhaduri, S., Ali, A., Sharmin, M., Hovsepian, K., & Kumar, S. (2014). Estimating Drivers' Stress from GPS Traces. In L. N. Boyle, G. E. Burnett, P. Fröhlich, S. T. Iqbal, E. Miller, & Y. Wu (Eds.), *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1–8). ACM Press.

Virzi, R. A., Sorce, J. F., & Herbert, L. B. (1993). A Comparison of Three Usability Evaluation Methods: Heuristic, Think-Aloud, and Performance Testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *37*(4), 309–313.

Virzi, R. A. (1992). Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 457–468.

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '92*, 381–388.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 105–140). John Wiley & Sons.

Wharton, C. (1992). *Cognitive walkthroughs: Instructions, forms, and examples* (Technical Report CU-ICS-92-17). Institute of Cognitive Science University of Colorado. Boulder.

Wickens, C. D., Helleberg, J., Goh, J., Xu, X., & Horrey, W. J. (2001). *Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning* (tech. rep.). University of Illinois at Urbana-Champaign.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wierwille, W. W., & Casali, J. G. (1983). A Validated Rating Scale for Global Mental Workload Measurement Applications. *Proceedings of the Human Factors Society Annual Meeting*, *27*, 129–133.

Wierwille, W. W. (1993). Visual and manual demands of in car controls and displays. *Automotive ergonomics* (pp. 299–320). Taylor & Francis. http://worldcat.org/isbn/0748400052

Wilson, C. E., & Coyne, K. P. (Eds.). (2001). The Whiteboard: Tracking Usability Issues: To Bug or Not to Bug? *interactions*, *8*(3), 15–19.

Winner, H., Hakuli, S., Lotz, F., & Singer, C. (Eds.). (2016). *Handbook of Driver Assistance Systems*. Springer International Publishing. http://link.springer.com/10.1007/978-3-319-09840-1

Wixon, D. (2003). Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. *interactions*, *10*(4), 28–34.

Wu, Q. (2009). An overview of driving distraction measure methods. *2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design*, 2391–2394.

Young, K. L., Regan, M. A., & Hammer, M. (2007). Driver distraction: A review of the literature. In I. J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter, & J. D. Irwin (Eds.), *Distracted driving.* (pp. 379–405). Monash University Accident Research Centre.

Young, K. L., Regan, M. A., & Lee, J. D. (2009). Measuring the Effects if Driver Distraction: Direct Driving Performance Methods and Measures. In M. A. Regan, J. D. Lee, & K. L. Young (Eds.), *Driver Distraction. Theory, Effects, and Mitigation* (pp. 85–105). CRC Press.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools* (Dissertation (PhD)). Delft University of Technology. http://repository.tudelft.nl/view/ir/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f/

Appendices

A. Exploratory Literature Review

Table A.1.: List of communities detected by the *Louvain* algorithm with the corresponding number of authors, publications, institutions, and ratio of publications per author.

Community ID	Authors	Publications	Publications per author	Institutions
3	45	27	0.6	11
9	37	16	0.432	7
15	27	7	0.259	7
6	20	9	0.45	4
42	20	7	0.35	6
8	16	6	0.375	4
5	14	5	0.357	5
4	13	6	0.462	3
64	12	4	0.333	3
1	11	8	0.727	2
7	11	6	0.545	4
23	11	5	0.455	4
58	11	6	0.545	3
69	11	3	0.273	2
31	10	4	0.4	2
101	10	1	0.1	2
0	9	5	0.556	4
11	9	2	0.222	3
14	9	1	0.111	2
81	9	2	0.222	4
45	8	1	0.125	8
89	8	2	0.25	6
95	8	2	0.25	5
99	8	1	0.125	3
30	7	2	0.286	2
47	7	2	0.286	5
52	7	2	0.286	2
76	7	1	0.143	3
92	7	1	0.143	2
2	6	1	0.167	1
20	6	1	0.167	2

Continued on next page

Table A.1 - *Continued from previous page*

Community ID	Authors	Publications	Publications per author	Institutions
26	6	1	0.167	2
32	6	2	0.333	2
85	6	1	0.167	1
87	6	1	0.167	2
88	6	1	0.167	1
96	6	1	0.167	2
16	5	2	0.4	2
22	5	2	0.4	3
28	5	1	0.2	3
29	5	1	0.2	1
41	5	1	0.2	2
43	5	1	0.2	1
48	5	1	0.2	2
57	5	1	0.2	2
59	5	1	0.2	2
61	5	1	0.2	2
70	5	1	0.2	2
71	5	1	0.2	3
75	5	1	0.2	4
78	5	1	0.2	3
79	5	1	0.2	2
91	5	1	0.2	2
10	4	1	0.25	1
12	4	2	0.5	2
17	4	1	0.25	3
24	4	1	0.25	1
27	4	1	0.25	1
38	4	1	0.25	2
39	4	1	0.25	1
40	4	1	0.25	1
44	4	1	0.25	1
49	4	1	0.25	1
53	4	1	0.25	1
62	4	1	0.25	3
83	4	1	0.25	1
84	4	1	0.25	2
86	4	1	0.25	2
100	4	1	0.25	1
102	4	1	0.25	3
103	4	1	0.25	2
18	3	1	0.333	2
33	3	2	0.667	1
35	3	1	0.333	1

Continued on next page

Table A.1 - *Continued from previous page*

Community ID	Authors	Publications	Publications per author	Institutions
36	3	1	0.333	2
46	3	1	0.333	1
51	3	1	0.333	2
54	3	1	0.333	2
55	3	1	0.333	1
63	3	1	0.333	1
65	3	1	0.333	1
67	3	1	0.333	1
68	3	1	0.333	2
72	3	1	0.333	1
74	3	1	0.333	2
94	3	1	0.333	1
13	2	1	0.5	1
19	2	1	0.5	1
21	2	1	0.5	1
25	2	1	0.5	2
34	2	1	0.5	1
37	2	1	0.5	1
50	2	1	0.5	2
56	2	1	0.5	2
66	2	1	0.5	1
73	2	1	0.5	2
77	2	1	0.5	3
80	2	1	0.5	2
82	2	1	0.5	1
93	2	1	0.5	1
97	2	1	0.5	1
98	2	1	0.5	2
60	1	1	1.0	1
90	1	1	1.0	1

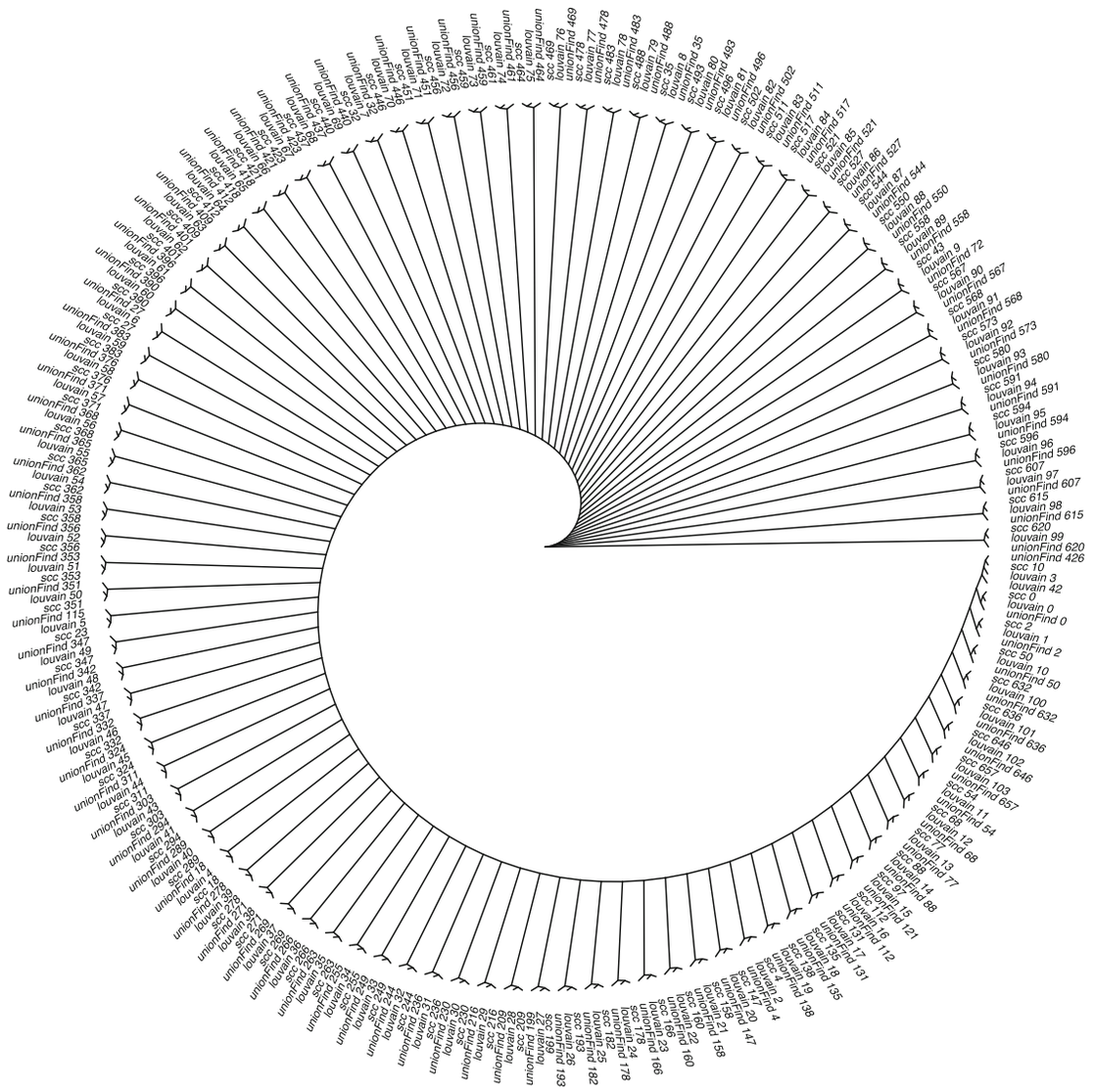


Figure A.1.: Hierarchical cluster analysis of community detection algorithms results.

B. Case Study: Cognitive Walkthrough

B.1. User Study

Table B.1.: Results of Shapiro-Wilk test for normality for task performance and task rating between context menu conditions.

	Task performance		Task rating	
Task	W	p	W	p
Reset trip kilometers	.87	.03	**.89**	.07
Check alerts	.87	.04	**.92**	.16
Switch audio source	.41	<.001	**.89**	.06
Map zoom	**.89**	.07	**.91**	.13
Skip audio title	.78	.002	.80	.003
Cancel route guidance	.71	<.001	.74	<.001

Note. Values that can be assumed to be normally distributed are in boldface. W = test statistic of Shapiro-Wilk test.

Table B.2.: Results of correlation analysis for task performance and task rating between context menu conditions.

	Task performance		Task rating	
Task	r_s	p	r_s	p
Reset trip kilometers	.19	.49	.25	.38
Check alerts	.19	.51	.06	.82
Switch audio source			-.06	.83
Map zoom	-.08	.77	.47	.08
Skip audio title	.02	.93	**.62**	.01
Cancel route guidance	.46	.08	-.29	.29

Note. Significant positive correlations are in boldface. For empty cells, the correlation coefficient could not be calculated due to invariant values.

Table B.3.: List of identified usability problems through user testing; classified according to the UPC.

#	Participant	Variant	Task	ID	Description	Task Component			Object Component	
1	0	Context-sensitive	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
2	3	Global	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
3	5	Context-sensitive	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
4	10	Context-sensitive	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
5	12	Context-sensitive	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
6	15	Global	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept
7	16	Context-sensitive	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble	performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
8	18	Global	Check alerts	1	The scroll direction is not identifiable. The user is not able to see which direction he has to swipe.	During	Trouble performing step	Cognitive attributes	Manipulation concept
9	5	Global	Skip audio title	1	The user is not able to identify in which direction he has to swipe to get the next/previous title.	Before	Determining how to do next step	Cognitive attributes	Visual cues
10	11	Context-sensitive	Skip audio title	1	The user is not able to identify in which direction he has to swipe to get the next/previous title.	Before	Determining how to do next step	Cognitive attributes	Visual cues
11	0	Context-sensitive	Switch audio source	2	The user tries to move a cursor to the labels indicating current selected audio source.	Before	Determining how to do next step	Physical attributes	Shape
12	12	Context-sensitive	Switch audio source	2	The user tries to move a cursor to the labels indicating current selected audio source.	Before	Determining how to do next step	Physical attributes	Shape
13	18	Global	Switch audio source	2	The user tries to move a cursor to the labels indicating current selected audio source.	Before	Determining how to do next step	Physical attributes	Shape
14	0	Context-sensitive	Map zoom	3	No direct interaction for map zoom available. Menu item not intuitive. The user is searching for a way to directly manipulate the map zoom.	During	Trouble performing step	Cognitive attributes	Indirectness
15	7	Global	Map zoom	3	No direct interaction for map zoom available. Menu item not intuitive. The user is searching for a way to directly manipulate the map zoom.	During	Trouble performing step	Cognitive attributes	Indirectness

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
16	10	Context-sensitive	Map zoom	3	No direct interaction for map zoom available. Menu item not intuitive. The user is searching for a way to directly manipulate the map zoom.	During	Trouble performing step	Cognitive attributes	Indirectness
17	0	Context-sensitive	Map zoom	4	Zoom slider difficult to recognize.	Before	Not confident about next step	Physical attributes	Size
18	10	Context-sensitive	Map zoom	4	Zoom slider difficult to recognize.	Before	Not confident about next step	Physical attributes	Size
19	12	Context-sensitive	Map zoom	4	Zoom slider difficult to recognize.	Before	Not confident about next step	Physical attributes	Size
20	15	Global	Map zoom	4	Zoom slider difficult to recognize.	Before	Not confident about next step	Physical attributes	Size
21	19	Context-sensitive	Map zoom	4	Zoom slider difficult to recognize.	Before	Not confident about next step	Physical attributes	Size
22	0	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
23	5	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
24	5	Global	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
25	12	Context-sensitive	Map zoom	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
26	6	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
27	11	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
28	12	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
29	19	Context-sensitive	Skip audio title	5	Direction of swipe gesture reversed in the users' opinion.	During	Trouble performing step	Cognitive attributes	Manipulation concept
30	0	Global	Reset trip kilometers	6	The user does not understand the word used for trip meter.	Before	Not confident about next step	Cognitive attributes	Content
31	5	Global	Reset trip kilometers	6	The user does not understand the word used for trip meter.	Before	Not confident about next step	Cognitive attributes	Content
32	9	Global	Reset trip kilometers	6	The user does not understand the word used for trip meter.	Before	Not confident about next step	Cognitive attributes	Content
33	0	Global	Check alerts	7	List of other menu options is unnecessary to the user.	Before	Not confident about next step	Cognitive attributes	Content
34	0	Global	Check alerts	8	Categories of context menu options are inconsistent. Some derive from thematic screens others not.	Before	Not confident about next step	Cognitive attributes	Content
35	12	Global	Switch audio source	8	Categories of context menu options are inconsistent. Some derive from thematic screens others not.	Before	Not confident about next step	Cognitive attributes	Content
36	3	Global	Reset trip kilometers	9	The user is confused that the cursor automatically preselects the "Yes" option.	During	Unexpected task automation	Physical attributes	Irritating
37	3	Global	Switch audio source	10	Insufficient feedback of successful operation when switching audio source.	After	Uncertain of results	Feedback	Missing
38	12	Context-sensitive	Switch audio source	10	Insufficient feedback of successful operation when switching audio source.	After	Uncertain of results	Feedback	Missing
39	15	Context-sensitive	Switch audio source	10	Insufficient feedback of successful operation when switching audio source.	After	Uncertain of results	Feedback	Missing

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
40	16	Global	Switch audio source	10	Insufficient feedback of successful operation when switching audio source.	After	Uncertain of results	Feedback	Missing
41	19	Global	Switch audio source	10	Insufficient feedback of successful operation when switching audio source.	After	Uncertain of results	Feedback	Missing
42	9	Global	Cancel route guidance	10	Missing feedback for canceled route guidance.	After	Uncertain of results	Feedback	Missing
43	16	Global	Cancel route guidance	10	Insufficient feedback of successful operation when canceling route guidance.	After	Uncertain of results	Feedback	Missing
44	3	Global	Switch audio source	11	List ending not visualized in context menu. Scrollbar is overlooked.	Before	Not confident about next step	Physical attributes	Size
45	11	Global	Check alerts	11	List ending not visualized in context menu. Scrollbar is overlooked.	Before	Not confident about next step	Physical attributes	Size
46	3	Global	Switch audio source	12	The user stays unclear what grayed out list items should imply.	Before	Not confident about next step	Cognitive attributes	Visual cues
47	6	Global	Check alerts	12	The user stays unclear what grayed out list items should imply.	Before	Not confident about next step	Cognitive attributes	Visual cues
48	3	Global	Map zoom	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
49	3	Global	Cancel route guidance	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
50	5	Global	Cancel route guidance	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
51	12	Context-sensitive	Cancel route guidance	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
52	12	Global	Reset trip kilometers	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
53	12	Global	Switch audio source	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
54	4	Global	Skip audio title	13	The context menu does not close automatically.	After	Unexpected task automation	Feedback	Missing
55	3	Context-sensitive	Reset trip kilometers	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
56	3	Context-sensitive	Switch audio source	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
57	5	Context-sensitive	Cancel route guidance	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
58	7	Context-sensitive	Cancel route guidance	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
59	9	Context-sensitive	Reset trip kilometers	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
60	11	Context-sensitive	Cancel route guidance	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
61	15	Context-sensitive	Reset trip kilometers	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
62	15	Context-sensitive	Check alerts	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
63	15	Context-sensitive	Switch audio source	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
64	16	Context-sensitive	Cancel route guidance	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
65	3	Context-sensitive	Skip audio title	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
66	7	Context-sensitive	Skip audio title	14	The user needs to change the screen to select specific options.	During	Trouble performing step	Cognitive attributes	Indirectness
67	3	Context-sensitive	Check alerts	15	The error memory is per se context-independent.	Before	Determining next step	Physical attributes	Placement
68	17	Context-sensitive	Check alerts	15	The error memory is not associated with the respective screen.	Before	Determining next step	Physical attributes	Placement
69	19	Context-sensitive	Check alerts	15	The user supposes the error memory between the instruments.	Before	Determining next step	Physical attributes	Placement
70	3	Context-sensitive	Cancel route guidance	16	The user has to remember the screen order to navigate to the desired screen quickly.	During	Trouble performing step	Cognitive attributes	Indirectness
71	4	Global	Check alerts	17	The user supposes the error memory directly on the screen, rather than in the context menu.	Before	Determining next step	Cognitive attributes	Content
72	7	Global	Check alerts	17	The user supposes the error memory directly on the screen, rather than in the context menu.	Before	Determining next step	Cognitive attributes	Content
73	10	Context-sensitive	Check alerts	17	The user supposes the error memory directly on the screen, rather than in the context menu.	Before	Determining next step	Cognitive attributes	Content

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
74	19	Context-sensitive	Check alerts	17	The user supposes the error memory directly on the screen, rather than in the context menu.	Before	Determining next step	Cognitive attributes	Content
75	4	Context-sensitive	Cancel route guidance	18	The user is confused that there is no context menu at all in some screens.	Before	Determining next step	Physical attributes	Placement
76	5	Context-sensitive	Cancel route guidance	18	The user is confused that there is no context menu at all in some screens.	Before	Determining next step	Physical attributes	Placement
77	11	Context-sensitive	Cancel route guidance	18	The user is confused that there is no context menu at all in some screens.	Before	Determining next step	Physical attributes	Placement
78	15	Context-sensitive	Reset trip kilometers	18	The user is confused that there is no context menu at all in some screens.	Before	Determining next step	Physical attributes	Placement
79	18	Context-sensitive	Cancel route guidance	18	The user is confused that there is no context menu at all in some screens.	Before	Determining next step	Physical attributes	Placement
80	5	Context-sensitive	Reset trip kilometers	19	The user is not able to identify the object as trip meter.	Before	Determining how to do next step	Cognitive attributes	Content
81	9	Global	Reset trip kilometers	19	The user is not able to identify the object as trip meter.	Before	Determining how to do next step	Cognitive attributes	Content
82	12	Context-sensitive	Reset trip kilometers	19	The user is not able to identify the object as trip meter.	Before	Determining how to do next step	Cognitive attributes	Content
83	5	Context-sensitive	Check alerts	20	The scroll direction should be vertical rather than horizontal.	During	Trouble performing step	Cognitive attributes	Manipulation concept
84	15	Global	Check alerts	20	The scroll direction should be vertical rather than horizontal.	During	Trouble performing step	Cognitive attributes	Manipulation concept
85	16	Context-sensitive	Check alerts	20	The scroll direction should be vertical rather than horizontal.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component			Object Component	
86	18	Global	Check alerts	20	The scroll direction should be vertical rather than horizontal.	During	Trouble step	performing	Cognitive attributes	Manipulation concept
87	15	Global	Skip audio title	20	The scroll direction was supposed to be horizontal rather than vertical.	During	Trouble step	performing	Cognitive attributes	Manipulation concept
88	16	Global	Skip audio title	20	The scroll direction was supposed to be horizontal rather than vertical.	During	Trouble step	performing	Cognitive attributes	Manipulation concept
89	5	Context-sensitive	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble step	performing	Physical attributes	Difficult to control
90	10	Global	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble step	performing	Physical attributes	Difficult to control
91	12	Context-sensitive	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble step	performing	Physical attributes	Difficult to control
92	12	Global	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble step	performing	Physical attributes	Difficult to control
93	15	Context-sensitive	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble step	performing	Physical attributes	Difficult to control

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
94	17	Context-sensitive	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble performing step	Physical attributes	Difficult to control
95	19	Context-sensitive	Map zoom	21	The horizontal swipe is not locked (as it is in other context menu options). The user accidentally switches between screens.	During	Trouble performing step	Physical attributes	Difficult to control
96	5	Global	Map zoom	22	The first level of categories in the global context menu overstrains the user. The user has to make a choice on the first level which requires an extra cognitive process.	Before	Determining next step	Cognitive attributes	Content
97	7	Global	Check alerts	22	The first level of categories in the global context menu overstrains the user. The user has to make a choice on the first level which requires an extra cognitive process.	Before	Determining next step	Cognitive attributes	Content
98	10	Global	Reset trip kilometers	22	The first level of categories in the global context menu overstrains the user. The user has to make a choice on the first level which requires an extra cognitive process.	Before	Determining next step	Cognitive attributes	Content

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
99	12	Global	Check alerts	22	The first level of categories in the global context menu overstrains the user. The user has to make a choice on the first level which requires an extra cognitive process.	Before	Determining next step	Cognitive attributes	Content
100	12	Global	Switch audio source	22	The first level of categories in the global context menu overstrains the user. The user has to make a choice on the first level which requires an extra cognitive process.	Before	Determining next step	Cognitive attributes	Content
101	18	Global	Skip audio title	23	The menu options would be easier to find if the list would use icons.	Before	Determining next step	Cognitive attributes	Visual cues
102	6	Context-sensitive	Map zoom	24	The user associates the context menu with resetting values.	Before	Determining next step	Cognitive attributes	Content
103	6	Context-sensitive	Skip audio title	25	Functionality is not needed in driver display.	Before	Not confident about next step	Physical attributes	Placement
104	6	Context-sensitive	Cancel route guidance	25	Functionality is not needed in driver display.	Before	Not confident about next step	Physical attributes	Placement
105	11	Global	Switch audio source	25	Functionality is not needed in driver display.	Before	Not confident about next step	Physical attributes	Placement
106	17	Global	Skip audio title	25	Functionality is not needed in driver display.	Before	Not confident about next step	Physical attributes	Placement
107	17	Global	Cancel route guidance	25	Functionality is not needed in driver display.	Before	Not confident about next step	Physical attributes	Placement

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
108	6	Global	Reset trip kilometers	26	The user is overwhelmed by the choice of menu items in different levels.	Before	Determining how to do next step	Cognitive attributes	Content
109	12	Global	Switch audio source	26	The user is overwhelmed by the choice of menu items in different levels.	Before	Determining next step	Cognitive attributes	Content
110	17	Global	Skip audio title	26	The user is overwhelmed by the choice of menu items in different levels.	Before	Determining how to do next step	Cognitive attributes	Content
111	18	Global	Reset trip kilometers	26	The user is overwhelmed by the choice of menu items in different levels.	Before	Determining how to do next step	Cognitive attributes	Content
112	6	Global	Reset trip kilometers	27	Swiping in submenus is annoying.	During	Trouble performing step	Physical attributes	Difficult to control
113	11	Global	Map zoom	27	Swiping in submenus is annoying.	During	Trouble performing step	Physical attributes	Difficult to control
114	6	Global	Map zoom	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness
115	11	Global	Switch audio source	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness
116	11	Global	Map zoom	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness
117	12	Global	Switch audio source	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness
118	17	Global	Switch audio source	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
119	17	Global	Map zoom	28	The user has to pass first menu level, even if he already selected the related screen.	During	Unexpected task automation	Cognitive attributes	Indirectness
120	6	Global	Skip audio title	29	The option inside the context menu is too complicated to reach.	Before	Determining next step	Physical attributes	Placement
121	9	Global	Map zoom	29	The menu item to change the map zoom is too deep in the menu tree.	Before	Determining how to do next step	Physical attributes	Placement
122	18	Global	Reset trip kilometers	29	The menu item to reset the trip kilometers is too deep in the menu tree.	Before	Determining how to do next step	Physical attributes	Placement
123	18	Global	Map zoom	29	The menu item to change the map zoom is too deep in the menu tree.	Before	Determining how to do next step	Physical attributes	Placement
124	9	Context-sensitive	Switch audio source	29	The interaction for switching the audio source takes too many steps.	During	Trouble performing step	Cognitive attributes	Indirectness
125	9	Global	Skip audio title	29	The interaction for changing the audio title takes too many steps.	During	Trouble performing step	Cognitive attributes	Indirectness
126	7	Global	Reset trip kilometers	30	Insufficient feedback of successful operation when resetting trip kilometers. The success message looks like a menu item.	After	Uncertain of results	Feedback	Missing
127	7	Global	Switch audio source	31	The possibility to switch the media source from every screen is confusing.	Before	Determining how to do next step	Physical attributes	Placement
128	7	Global	Skip audio title	32	Insufficient feedback of successful operation when changing the audio title. The user tries to confirm the operation by clicking.	After	Unexpected task automation	Feedback	Missing

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
129	11	Global	Skip audio title	32	Insufficient feedback of successful operation when changing the audio title. The user tries to confirm the operation by clicking.	After	Unexpected task automation	Feedback	Missing
130	7	Context-sensitive	Skip audio title	33	Missing hint for direct interaction. The user is looking for a way to skip the audio title.	Before	Determining how to do next step	Cognitive attributes	Visual cues
131	17	Context-sensitive	Map zoom	34	Ordering of menu options is not optimal. The option to change the map zoom could be one line higher.	Before	Determining how to do next step	Physical attributes	Placement
132	18	Context-sensitive	Map zoom	34	Prioritization of menu options should change according to the context. If route guidance is active the zoom option should be before last destinations and favorites in the list.	During	Trouble performing step	Cognitive attributes	Indirectness
133	9	Global	Map zoom	35	The interaction to change the map zoom is quite complex. The user would not use the function while driving.	During	Trouble performing step	Cognitive attributes	Manipulation concept
134	9	Context-sensitive	Map zoom	35	The interaction to change the map zoom is quite complex. The user would not use the function while driving.	During	Trouble performing step	Cognitive attributes	Manipulation concept
135	18	Global	Map zoom	35	The interaction to change the map zoom is quite complex. The user would not use the function while driving.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
136	19	Global	Switch audio source	36	Menu moves one level up after selecting an audio source which is inconsistent with menus on the infotainment display.	After	Unexpected task automation	Unexpected results	
137	19	Global	Switch audio source	37	The currently selected audio source should be highlighted in the list.	Before	Not confident about next step	Cognitive attributes	Visual cues
138	9	Context-sensitive	Check alerts	38	The user takes a long time to find the error memory location.	Before	Determining next step	Physical attributes	Placement
139	17	Context-sensitive	Check alerts	38	The user takes a long time to find the error memory location.	Before	Determining next step	Physical attributes	Placement
140	17	Context-sensitive	Reset trip kilometers	39	On the first look it is labeled "Tageswegstrecke - Nein/Ja" which is not very declarative.	Before	Not confident about next step	Cognitive attributes	Content
141	10	Context-sensitive	Reset trip kilometers	40	The user supposes the trip meter in another screen element.	Before	Determining next step	Physical attributes	Placement
142	11	Context-sensitive	Reset trip kilometers	40	Trip meter is supposed to be on another screen	Before	Determining how to do next step	Physical attributes	Placement
143	18	Global	Reset trip kilometers	40	The user tries to find information about trip meter on trip screen, because he can reset it there.	After	Uncertain of results	Unexpected results	
144	10	Global	Check alerts	41	The user does not identify the related first level menu item as category for error memory.	Before	Determining how to do next step	Cognitive attributes	Content
145	11	Global	Reset trip kilometers	42	The screen does not indicate that pressing the OK-Button is possible.	Before	Determining how to do next step	Cognitive attributes	Visual cues
146	11	Global	Switch audio source	43	Not all first level menu items are visible at once. Scrolling is needed to get to the last menu item.	During	Trouble performing step	Physical attributes	Irritating

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
147	17	Context-sensitive	Reset trip kilometers	44	First line of text and text inside the box should be switched to set focus on current action that should be performed.	Before	Not confident about next step	Physical attributes	Placement
148	12	Context-sensitive	Reset trip kilometers	45	The user is irritated by the wrap-around list of main screens.	During	Trouble performing step	Physical attributes	Irritating
149	12	Context-sensitive	Check alerts	46	The user supposes an indicator whether the error memory has items that should be considered.	Before	Determining next step	Cognitive attributes	Visual cues
150	12	Context-sensitive	Check alerts	47	List endings without wrap-around are inconsistent regarding the interaction in other screens.	During	Trouble performing step	Physical attributes	Irritating
151	12	Context-sensitive	Switch audio source	48	The screen appears to have an active cursor on the current playing audio track. The user is irritated because he has to click the current track to select another audio source.	Before	Determining next step	Cognitive attributes	Visual cues
152	12	Context-sensitive	Map zoom	49	The user is unable to find the related option in the context menu.	Before	Determining next step	Physical attributes	Placement
153	17	Global	Check alerts	50	The user is confused that the category contains only two options. It would be better to show both options on the first level.	During	Unexpected task automation	Cognitive attributes	Indirectness
154	12	Global	Check alerts	51	Wording in breadcrumbs and menu items is not consistent.	Before	Not confident about next step	Cognitive attributes	Content
155	12	Global	Check alerts	52	The user wants an option to quickly close the context menu when navigating in lower levels.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
156	12	Global	Skip audio title	52	The user wants an option to quickly close the context menu when navigating in lower levels.	During	Trouble performing step	Cognitive attributes	Manipulation concept
157	12	Global	Skip audio title	53	The functionality in the context menu does not show previous and next audio title as the main menu does.	Before	Not confident about next step	Cognitive attributes	Content
158	16	Global	Skip audio title	53	The preview of next audio title is missing in context menu.	Before	Not confident about next step	Cognitive attributes	Content
159	15	Global	Reset trip kilometers	54	The buttons for canceling or submitting the operation do not differ enough from on-screen text.	Before	Determining how to do next step	Physical attributes	Shape
160	15	Global	Map zoom	55	Zoom slider on the left side is irritating because interaction was triggered on the right side.	Before	Determining next step	Physical attributes	Placement
161	15	Context-sensitive	Map zoom	55	Zoom slider on the left side is irritating because interaction was triggered on the right side.	Before	Determining next step	Physical attributes	Placement
162	16	Context-sensitive	Map zoom	55	Zoom slider on the left side is irritating because interaction was triggered on the right side.	Before	Determining next step	Physical attributes	Placement
163	19	Context-sensitive	Map zoom	55	Zoom slider on the left side is irritating because interaction was triggered on the right side.	Before	Determining next step	Physical attributes	Placement
164	19	Global	Map zoom	55	Zoom slider on the left side is irritating because interaction was triggered on the right side.	Before	Determining next step	Physical attributes	Placement
165	16	Global	Cancel route guidance	56	The participant is irritated by preselecting a menu item that is not specific to the current screen.	During	Trouble performing step	Physical attributes	Irritating

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
166	16	Context-sensitive	Switch audio source	57	Scrollbar right to the content list is unnecessary.	After	Unexpected task automation	Feedback	Unnecessary
167	16	Context-sensitive	Map zoom	57	Scrollbar right to the content list is unnecessary.	After	Unexpected task automation	Feedback	Unnecessary
168	16	Context-sensitive	Map zoom	58	Grayed out list elements should rather be removed from the list, as they are not accessible.	Before	Not confident about next step	Cognitive attributes	Visual cues
169	16	Context-sensitive	Skip audio title	59	Differentiation between preview and actual title skip is not possible for the participant.	After	Uncertain of results	Feedback	Misleading
170	16	Context-sensitive	Skip audio title	60	There is no progress bar for audio playback.	After	Uncertain of results	Feedback	Missing
171	16	Global	Reset trip kilometers	61	The breadcrumbs in the top line are irritating rather than supportive.	Before	Not confident about next step	Physical attributes	Placement
172	16	Global	Check alerts	61	The breadcrumbs in the top line are irritating rather than supportive.	Before	Not confident about next step	Physical attributes	Placement
173	16	Global	Switch audio source	61	The breadcrumbs in the top line are irritating rather than supportive.	Before	Not confident about next step	Physical attributes	Placement
174	16	Global	Reset trip kilometers	62	The system should return to the next level up after successful interaction rather than remaining in the current menu level.	After	Unexpected task automation	Feedback	Missing

Continued on next page

Table B.3 - *Continued from previous page*

#	Participant	Variant	Task	ID	Description	Task Component		Object Component	
175	12	Global	Reset trip kilometers	63	Insufficient feedback of successful operation when resetting trip kilometers. The user is irritated that the reset value is not presented to him.	After	Uncertain of results	Feedback	Missing

Note. The issues were numbered according to their occurrence during the tests. Gaps in the participant numbering occur because they were numbered according to the available time slots in the schedule. The ID column contains a numerical identifier for unique usability problems.

Table B.4.: List of identified usability problems through user testing; classified according to the UPC.

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
1	8	1	specific	Check alerts	Service is not identified as a context	Before	Not confident about next step	Physical attributes	Shape
2	8	1	specific	Switch audio source	Media source assumed to be in HU	Before	Determining how to do next step	Physical attributes	Placement
3	13	1	specific	Map zoom	Closing context menu on zoom is irritating	After	Outcome did not match goal	Unexpected results	
4	8	1	specific	Skip title	Skip title assumed in different device (HU/Steering wheel)	Before	Determining how to do next step	Cognitive attributes	Visual cues
5	6	1	specific	Cancel route guidance	Progress when canceling route guidance not recognized	After	Uncertain of results	Feedback	Missing
6	1	1	global	Reset trip kilometers	Global options may irritate	During	Unexpected task automation	Physical attributes	Irritating
7	2	1	global	Reset trip kilometers	Number of options for current task are overwhelming	Before	Not confident about next step	Cognitive attributes	Content
8	7	1	global	Reset trip kilometers	Trip meter not recognizable in option list	After	Uncertain of results	Feedback	Misleading

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
9	2	1	global	Reset trip kilometers	Number of menu levels is overwhelming	After	Outcome did not match goal	Feedback	Misleading
10	2	1	global	Reset trip kilometers	Number of options for current task are overwhelming	Before	Not confident about next step	Cognitive attributes	Content
11	3	1	global	Reset trip kilometers	Context menu not closing after action	After	Unexpected task automation	Unexpected results	
12	11	1	global	Reset trip kilometers	No progress visible when resetting trip meter because of overlaying context menu	After	Uncertain of results	Feedback	Misleading
13	10	1	global	Check alerts	Label for service options not identifiable	Before	Determining next step	Cognitive attributes	Content
14	6	1	global	Switch audio source	No progress visible when switching audio source	After	Uncertain of results	Feedback	Misleading
15	12	1	global	Map zoom	No direct option to zoom map	During	Trouble performing step	Cognitive attributes	Indirectness
16	8	1	global	Map zoom	Zoom option not assumed in context menu	Before	Determining how to do next step	Physical attributes	Placement
17	13	1	global	Map zoom	Closing context menu on zoom is irritating	After	Outcome did not match goal	Unexpected results	

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
18	25	1	global	Skip title	Option to switch audio title not identifiable in sub menu	Before	Determining how to do next step	Physical attributes	Shape
19	4	2	global	Reset trip kilometers	No hint for context menu	Before	Determining next step	Cognitive attributes	Visual cues
20	1	2	global	Reset trip kilometers	Context menu might not be recognizable as global function.	Before	Determining next step	Cognitive attributes	Visual cues
21	5	2	global	Reset trip kilometers	Options that had not been selected are still visible/prominent.	Before	Not confident about next step	Cognitive attributes	Content
22	11	2	global	Reset trip kilometers	Context menu occludes relevant information.	Before	Not confident about next step	Physical attributes	Placement
23	3	2	global	Reset trip kilometers	Context menu is not closing when action is finished.	After	Uncertain of results	Unexpected results	
24	3	2	global	Reset trip kilometers	Context menu is not closing when action is finished.	After	Unexpected task automation	Feedback	Missing
25	8	2	global	Check alerts	Error storage is not assumed in current display.	Before	Determining next step	Cognitive attributes	Content
26	8	2	global	Switch audio source	Media settings not assumed to be in current screen.	Before	Determining how to do next step	Physical attributes	Placement

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
27	6	2	global	Switch audio source	Action when selecting audio source switch is not self-explanatory.	During	Trouble performing step	Cognitive attributes	Indirectness
28	3	2	global	Switch audio source	Context menu is not closing when action is finished.	After	Uncertain of results	Unexpected results	
29	29	2	global	Map zoom	Context menu is opening in the right side, control elements are on the left side.	During	Trouble performing step	Cognitive attributes	Manipulation concept
30	16	2	global	Map zoom	Context menu is on the right, zoom slider on the left.	Before	Not confident about next step	Physical attributes	Placement
31	19	2	global	Skip title	Current title could be not recognizable as song list	After	Uncertain of results	Feedback	Misleading
32	20	2	global	Skip title	Skip title interaction should be horizontal	During	Trouble performing step	Cognitive attributes	Manipulation concept
33	21	2	global	Skip title	Next title is not visible	Before	Not confident about next step	Cognitive attributes	Content
34	3	2	global	Cancel route guidance	Context menu is not closing when action is finished.	After	Uncertain of results	Unexpected results	
35	4	2	specific	Reset trip kilometers	No hint for context menu	Before	Determining next step	Cognitive attributes	Visual cues

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
36	14	2	specific	Switch audio source	Current audio source does not seem clickable.	Before	Not confident about next step	Cognitive attributes	Visual cues
37	15	2	specific	Switch audio source	Highlight for cursor is on current title.	Before	Determining how to do next step	Cognitive attributes	Visual cues
38	16	2	specific	Map zoom	Context menu is on the right, zoom slider on the left.	Before	Not confident about next step	Physical attributes	Placement
39	22	2	specific	Skip title	Screen does not appear to have an option for skipping titles.	Before	Determining how to do next step	Physical attributes	Size
40	16	2	specific	Skip title	Metaphor for barrel roll is irritating. Direction is wrong.	During	Trouble performing step	Cognitive attributes	Manipulation concept
41	16	2	specific	Skip title	Direction to skip title is the other way round.	During	Trouble performing step	Cognitive attributes	Manipulation concept
42	3	2	specific	Cancel route guidance	Context menu is not closing when action is finished.	After	Unexpected task automation	Feedback	Missing
43	2	3	global	Reset trip kilometers	Number of menu levels is overwhelming.	Before	Determining how to do next step	Cognitive attributes	Content
44	9	3	global	Reset trip kilometers	Confirm operation is on the same level as the selection of resetting object.	After	Uncertain of results	Feedback	Misleading

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
45	3	3	global	Reset trip kilometers	Context menu not closing after action.	After	Unexpected task automation	Unexpected results	
46	6	3	global	Reset trip kilometers	No visual feedback for operations success	After	Uncertain of results	Feedback	Missing
47	8	3	global	Check alerts	Information is assumed in center stack display.	Before	Determining next step	Cognitive attributes	Visual cues
48	10	3	global	Check alerts	Menu option "Service" has no affordance regarding alerts.	Before	Determining how to do next step	Cognitive attributes	Content
49	8	3	global	Switch audio source	Information is assumed in center stack display.	Before	Determining next step	Cognitive attributes	Visual cues
50	3	3	global	Switch audio source	Context menu not closing after action.	After	Unexpected task automation	Unexpected results	
51	6	3	global	Switch audio source	No visual feedback for operations success	After	Uncertain of results	Feedback	Missing
52	18	3	global	Map zoom	The user might not recognize, that changing to another screen is necessary.	Before	Determining next step	Physical attributes	Placement

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
53	17	3	global	Map zoom	The user might not understand why list options are grayed out.	Before	Determining how to do next step	Cognitive attributes	Visual cues
54	16	3	global	Map zoom	Context menu is on the right, zoom slider on the left.	Before	Not confident about next step	Physical attributes	Placement
55	3	3	global	Map zoom	Context menu not closing after action.	After	Unexpected task automation	Unexpected results	
56	23	3	global	Skip title	Direction for title skip is not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
57	24	3	global	Skip title	The number of menu levels leads to several interactions on back to entirely close the context menu	During	Unexpected task automation	Cognitive attributes	Indirectness
58	6	3	global	Cancel route guidance	Visual feedback after cancelling route guidance might not be sufficient.	After	Uncertain of results	Feedback	Missing
59	8	3	specific	Check alerts	Information is assumed in center stack display.	Before	Determining next step	Cognitive attributes	Visual cues

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
60	10	3	specific	Check alerts	Service screen has no affordance due to several differnet objects on one screen.	Before	Determining how to do next step	Cognitive attributes	Content
61	14	3	specific	Switch audio source	Label for current audio source looks like a button.	Before	Determining next step	Physical attributes	Shape
62	15	3	specific	Switch audio source	Seek bar looks like a cursor highlight, OK is not associated with audio sources.	Before	Determining next step	Cognitive attributes	Visual cues
63	16	3	specific	Map zoom	Context menu is on the right, zoom slider on the left.	Before	Not confident about next step	Physical attributes	Placement
64	3	3	specific	Map zoom	Context menu not closing after action.	After	Unexpected task automation	Unexpected results	
65	23	3	specific	Skip title	Direction for title skip is not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
66	6	3	specific	Cancel route guidance	Visual feedback after cancelling route guidance might not be sufficient.	After	Uncertain of results	Feedback	Missing
67	4	4	specific	Reset trip kilometers	No indication for context menu.	Before	Determining next step	Cognitive attributes	Visual cues

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
68	8	4	specific	Check alerts	Looking for error messages in different device.	Before	Determining next step	Cognitive attributes	Visual cues
69	6	4	specific	Check alerts	Progress not recognizable.	During	Trouble performing step	Cognitive attributes	Indirectness
70	28	4	specific	Check alerts	Menu items are not predictable (Arrow for next level, Toggle for setting, Nothing for direct action).	Before	Determining next step	Physical attributes	Shape
71	8	4	specific	Switch audio source	Looking for media settings in different device.	Before	Determining next step	Physical attributes	Placement
72	14	4	specific	Switch audio source	User might try to select label items on the right.	Before	Determining how to do next step	Physical attributes	Placement
73	12	4	specific	Map zoom	Direct interaction for map zoom necessary	Before	Determining how to do next step	Cognitive attributes	Visual cues
74	16	4	specific	Map zoom	Zoom slider on the left while context menu on the right.	After	Outcome did not match goal	Feedback	Misleading
75	8	4	specific	Skip title	Looking for title skip in different device.	Before	Determining next step	Cognitive attributes	Visual cues
76	8	4	specific	Cancel route guidance	Looking for route guidance settings in different device.	Before	Determining next step	Physical attributes	Placement

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
77	3	4	specific	Cancel route guidance	Context menu should close after cancelling route guidance	After	Unexpected task automation	Feedback	Misleading
78	4	4	global	Reset trip kilometers	No indication for context menu.	Before	Determining next step	Cognitive attributes	Visual cues
79	27	4	global	Reset trip kilometers	Current most relevant list entry should be on top.	During	Unexpected task automation	Physical attributes	Irritating
80	3	4	global	Reset trip kilometers	Context menu not closing when finishing action.	After	Unexpected task automation	Feedback	Misleading
81	8	4	global	Check alerts	Error memory not supposed in menu "Service".	Before	Determining how to do next step	Cognitive attributes	Content
82	8	4	global	Check alerts	Looking for error messages in different device.	Before	Determining next step	Cognitive attributes	Visual cues
83	6	4	global	Check alerts	Progress not recognizable.	During	Trouble performing step	Cognitive attributes	Indirectness
84	26	4	global	Switch audio source	Too many menu levels for operation.	During	Trouble performing step	Cognitive attributes	Indirectness
85	6	4	global	Switch audio source	Progress not recognizable.	During	Trouble performing step	Cognitive attributes	Indirectness
86	3	4	global	Switch audio source	Context menu not closing when finishing action.	After	Unexpected task automation	Feedback	Misleading

Continued on next page

Table B.4 - *Continued from previous page*

#	ID	Expert	Variant	Task	Description	Task Component		Object Component	
87	12	4	global	Map zoom	Direct interaction for map zoom necessary	Before	Determining how to do next step	Cognitive attributes	Visual cues
88	16	4	global	Map zoom	Zoom slider on the left while context menu on the right.	After	Outcome did not match goal	Feedback	Misleading
89	8	4	global	Skip title	Looking for title skip in different device.	Before	Determining next step	Cognitive attributes	Visual cues
90	26	4	global	Skip title	Too many menu levels for operation.	During	Trouble performing step	Cognitive attributes	Indirectness
91	2	4	global	Skip title	Number of menu levels is overwhelming and distracting	Before	Not confident about next step	Cognitive attributes	Content
92	6	4	global	Cancel route guidance	Progress not recognizable.	During	Trouble performing step	Cognitive attributes	Indirectness
93	2	4	global	Cancel route guidance	Number of menu levels is overwhelming and distracting	Before	Not confident about next step	Cognitive attributes	Content

Note. The issues were numbered according to their occurrence during the review. The ID column contains a numerical identifier for unique usability problems.

C. Case Study: Guideline Review

C.1. Literature Review for Existing Guidelines

C.1.1. General HMI Guidelines

Ten Usability Heuristics for User Interface Design (Nielsen, 1993, 1995a):

- **Visibility of system status**
  The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
- **Match between system and the real world**
  The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
- **User control and freedom**
  Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
- **Consistency and standards**
  Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
- **Error prevention**
  Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
- **Recognition rather than recall**
  Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
- **Flexibility and efficiency of use**
  Accelerators — unseen by the novice user — may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
- **Aesthetic and minimalist design**
  Dialogues should not contain information which is irrelevant or rarely needed. Every

extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

- **Help users recognize, diagnose, and recover from errors**
  Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- **Help and documentation**
  Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Shneiderman's Eight Golden Rules of Interface Design (Shneiderman et al., 2018, pp. 95 ff.):

- **Strive for consistency.**
  Consistent sequences of actions should be required in similar situations; identical terminology should be used in prompts, menus, and help screens; and consistent commands should be employed throughout.
- **Enable frequent users to use shortcuts.**
  As the frequency of use increases, so do the user's desires to reduce the number of interactions and to increase the pace of interaction. Abbreviations, function keys, hidden commands, and macro facilities are very helpful to an expert user.
- **Offer informative feedback.**
  For every operator action, there should be some system feedback. For frequent and minor actions, the response can be modest, while for infrequent and major actions, the response should be more substantial.
- **Design dialog to yield closure.**
  Sequences of actions should be organized into groups with a beginning, middle, and end. The informative feedback at the completion of a group of actions gives the operators the satisfaction of accomplishment, a sense of relief, the signal to drop contingency plans and options from their minds, and an indication that the way is clear to prepare for the next group of actions.
- **Offer simple error handling.**
  As much as possible, design the system so the user cannot make a serious error. If an error is made, the system should be able to detect the error and offer simple, comprehensible mechanisms for handling the error.
- **Permit easy reversal of actions.**
  This feature relieves anxiety, since the user knows that errors can be undone; it thus encourages exploration of unfamiliar options. The units of reversibility may be a single action, a data entry, or a complete group of actions.
- **Support internal locus of control.**
  Experienced operators strongly desire the sense that they are in charge of the system and that the system responds to their actions. Design the system to make users the initiators of actions rather than the responders.
- **Reduce short-term memory load.**
  The limitation of human information processing in short-term memory requires that displays be kept simple, multiple page displays be consolidated, window-motion frequency

be reduced, and sufficient training time be allotted for codes, mnemonics, and sequences of actions.

C.1.2. Website Usability Guidelines

A guide to carrying out usability reviews (Turner, 2011):

- Features & Functionality
  - **Features and functionality meet common user goals and objectives**
    Key and common user goals and objectives (e.g. carry out some transaction, find some information, carry out some research etc. . .) should have been identified and addressed. Ideally the site or application should allow users to meet all of their key goals and objectives.
  - **Features and functionality support users desired workflows**
    The site or application should support or at least be compatible with the way that users wish to work. For example, users might want to be able to carry out bulk transactions or be able to save and return to their work.
  - **Frequently-used tasks are readily available (e.g. easily accessible from the homepage) and well supported**
    For example short cuts and a login to retrieve details might be provided to speed up the completion of frequently carried out tasks.
  - **Users are adequately supported according to their level of expertise**
    For example, novice users are given help and instructions and features are progressively disclosed (e.g. advanced features not being shown by default).
  - **Calls to action (e.g. register, add to basket, submit) are clear, well labelled and appear clickable**
    Possible actions should always be clear and the primary call to action (i.e. the most common or desirable user action) should stand out on the page or screen.
- Homepage / Starting Page
  - **The Homepage / starting page provides a clear snapshot and overview of the content, features and functionality available**
    For example, an introduction and overview of the site is provided together with section snapshots and example content.
  - **The homepage / starting page is effective in orienting and directing users to their desired information and tasks**
    Users should be able to work out where they need to go to complete a given task (e.g. carry out some research, complete a transaction).
  - **The homepage / starting page layout is clear and uncluttered with sufficient 'white space'**
    Users should be able to quickly scan the homepage and make sense of both the content available and of how the site is structured.
- Navigation

  - **Users can easily access the site or application**
    For example, the URL is predictable and is returned by search engines. If a user attempts to find the site via a search engine, it should ideally be returned on the first page of search results for likely queries.
  - **The navigational scheme is easy to find, intuitive and consistent**
    Users should be able to very easily locate and use the navigational scheme (e.g. left hand menu, top menu, tabbed menu), and it should not be significantly different across the site or application (unless a decision has been made to specifically differentiate a given section or area).
  - **The navigation has sufficient flexibility to allow users to navigate by their desired means**
    For example a user might want to be able to search for an item or browse by size, name or type. Although not all user preferences can or indeed should be addressed, the most useful and common navigational means should be supported.
  - **The site or application structure is clear, easily understood and addresses common user goals**
    For example, gathering information, submitting data, carrying out research. Users should be able to work out where they need to go to carry out common user goals and be able to quickly gain an understanding of how the site or application is structured.
  - **Links are clear, descriptive and well labelled**
    Links should be clearly 'clickable' (e.g. underlined or colourised) and it should be clear to users where any given link goes to. Non-descriptive links such as 'click here' should be avoided and any links going to an external website or opening a new window should be identified as such.
  - **Browser standard functions (e.g. 'back', 'forward', 'bookmark') are supported**
    Users should be able to bookmark a page (or be presented with a URL to use) and go back and forth without breaking the site or losing any information they have entered.
  - **The current location is clearly indicated (e.g. breadcrumb, highlighted menu item)**
    Users should always know where they are in the site or application.
  - **Users can easily get back to the homepage or a relevant start point**
    For example, a homepage link might be part of the breadcrumb or a home link might be available as part of the header.
  - **A clear and well structure site map or index is provided (where necessary)**
    The sitemap might be part of the header or footer and should ideally be available from every page on the site.

- Search

  - **A consistent, easy to find and easy to use search function is available throughout**
    The search function (where required) should be directly available from most pages on the site or application and should be consistently positioned (e.g. top left, top right or top centre).
  - **The search interface is appropriate to meet user goals**
    For example users are able to filter search results, an advanced search is available (if

necessary) and common search conventions such as quotation marks (") and natural language searches are handled.
  - **The search facility deals well with common searches, misspellings and abbreviations**
    Ideally synonyms (e.g. 'coat' should also match 'jacket') should mean that logical and appropriate search results are returned for common user queries. Popular search results (e.g. top matches) should also be identified for common queries.
  - **Search results are relevant, comprehensive, precise, and well displayed**
    It should be easy for users to see what has been returned, to work out why something has been returned and to determine how many results there are.

- Control & Feedback
  - **Prompt and appropriate feedback is given**
    For example, a confirmation message is shown following a successful transaction, input errors are promptly highlighted and it's made clear to users when a page has been updated.
  - **Users can easily undo, go back and change, or cancel actions**
    If an action can not be undo then users should at least be given the chance to confirm an action before committing (e.g. before placing an order). For example, users can return to a step and change their options or dynamically change a value without having to start again. Where an action can't be undone (e.g. a deletion), this should be made clear to users.
  - **Users can easily give feedback**
    For example, via email or an online feedback / contact us form. There should be an indication of how long users can expect to wait for a response if a query has been made.

- Forms
  - **Complex forms and processes are broken up into readily understood steps and sections**
    For example, a checkout process might be broken up in to 'address', 'delivery options', 'payment' and 'confirmation'. Where a process is used a progress indicator is present with clear numbers or named stages.
  - **A minimal amount of information is requested and where necessary justification is given for asking for information**
    For example a site might outline that a telephone number is required in case there is an issue with a transaction. Users shouldn't be asked for extraneous information and where possible information should be auto populated (e.g. postcode lookup, code lookup) to keep input to a minimum.
  - **Required and optional form fields are clearly indicated (e.g. using text or '*')**
    Where most fields are required the optional fields should be identified and when most fields are optional the required fields should be identified.
  - **Appropriate input fields are used and required formats are indicated**
    Appropriate input fields might include calendar for date selection, drop downs for selection and radio button for small selections. Text might be used to indicate the required format or an example might be provided. Field lengths should correspond

to the expected input so for example an email input field should be long, where as an initials input field should be very short.
  - **Help and instructions (e.g. examples, information required) are provided where necessary**
    Where input is non trivial or is likely to require some explanation this should be provided. Where a-lot of explanation is necessary a link to a page outlining what is required should be provided.

- Errors
  - **Errors are clear, easily identified and appear in appropriate locations**
    Errors should be immediately apparent to users and ideally be located close to the offending input or function (e.g. adjacent to an input entry field). Inputs causing an error should be highlighted, together with an explanation for the error.
  - **Error messages are concise, written in easy to understand language and describe what's occurred and what action is necessary**
    Errors should avoid using very technical terms or jargon and should be written from the user's perspective.
  - **Common user errors have been taken into consideration and where possible prevented**
    Common user errors might be missing fields, invalid formats and invalid selections. For example, fields might limit input to particular a format (e.g. numbers only) or only become available once certain criteria have been met. JavaScript might also be utilised to provide immediate feedback for common formatting errors or errors caused by missing fields.
  - **Users are able to easily recover (i.e. not have to start again) from errors**
    For example, users might be able to re-edit and resubmit a form or enter a different value.

- Content & Text
  - **Content available (e.g. text, images, video, audio) is appropriate and sufficiently relevant, and detailed to meet user goals**
    Content should also be appropriately formatted, so for example videos and audio should be directly playable (i.e. shouldn't need to be downloaded to be played) and images should be of a sufficient quality.
  - **Links to other useful and relevant content (e.g. related pages, external websites or documents) are available and shown in context**
    For example there might be links from an article to related articles, related content or related external websites.
  - **Language, terminology and tone used is appropriate and readily understood by the target audience**
    Jargon should be kept to a minimum and plain language should be used where ever possible.
  - **Terms, language and tone used are consistent (e.g. the same term is used throughout)**
    Capitalisation (e.g. 'Main title'; 'Main Title'; 'MAIN TITLE') and grammar should

be consistent, together with the use of formal or informal terms (e.g. could not vs couldn't; what's vs what is etc...).
  - **Text and content is legible and scanable, with good typography and visual contrast**
    Users should be able to quickly scan headers and body text, in order to get an overview of what's available.

- Help
  - **Online help is provided and is suitable for the user base**
    Help should be written in easy to understand language and only uses recognised terms. Users should be able to easily find and access help and where appropriate contextual help should be available, such as help for a specific page, feature or process.
  - **Online help is concise, easy to read and written in easy to understand language**
    Help should cover the essentials without providing excessive detail and shouldn't use jargon or technical terminology that isn't likely to be understood by users.
  - **Accessing online help does not impede users**
    Users should be able to resume work where they left off after accessing help. Ideally help should be available directly on a page or using a new window. If help is provided in the form of a document, it should be formatted for the web (e.g. PDF, rather than a Word document).
  - **Users can easily get further help (e.g. telephone or email address)**
    If a telephone help number is provided the hours of operation should be shown. If an email address or online form is provided, an indication should be given of how long a response is likely to take (e.g. within the next 24 hrs).

- Performance
  - **Site or application performance doesn't inhibit the user experience (e.g. slow page downloads, long delays)**
    Web page downloads shouldn't take longer than 5 seconds and on page interactions (e.g. using an application or AJAX functionality) shouldn't take any longer than 1 second to respond. Interactions taking longer than 1 second to respond should provide suitable feedback to show that something is taking place (e.g. an hour glass or swirling graphic).
  - **Errors and reliability issues don't inhibit the user experience**
    Sites and applications should be free of bugs and shouldn't have any broken links.
  - **Possible user configurations (e.g. browsers, resolutions, computer specs) are supported**
    Websites should be usable at a 800x600 screen resolution and should work with the most common browsers (IE, Firefox, Opera, Chrome etc. . .). Applications should be usable with common computer specifications (operation system, memory, available disk space) and screen resolutions (e.g. 800x600, 1025x768).

C.1.3. Ergonomic Criteria

Ergonomic Criteria for the Evaluation of Human-Computer Interfaces (Bastien & Scapin, 1993):

- Guidance
  - Prompting
    - For data entry, provide the user with the required formats and acceptable values; e.g., include in a field label additional cueing of data format (e.g., Date (mm/dd/yy): _ _ / _ _ / _ _).
    - Display measurement units for data entry.
    - Indicate all status information (e.g., modes, values, etc.).
    - For each data field, display an associated label.
    - Provide cues on the acceptable length of entries.
    - Provide a title for each window.
    - Provide on-line help and guidance.
  - Grouping/Distinction
    - by Location
      - Organise items in hierarchical lists.
      - Organise the options of a menu dialogue as a function of the objects to which they apply.
      - When several options are presented, their organisation must be logical, i.e., the organisation must represent a significant or relevant functional organisation (alphabetic order, functional, frequency of use, etc.).
    - by Format
      - Provide clear visual distinction of areas having different functions (command zone, message zone, etc.).
      - Provide clear visual distinction of data fields and their labels.
  - Immediate Feedback
    - All users' entries should be displayed except for secure entries. Even in this case, every keyed entry should produce a perceptible feedback (e.g., symbols such as stars).
    - Following user interruption of data processing, display an advisory message assuring the user that the system has returned to its previous status.
    - When computer processing is lengthy, information concerning the state of the processing should be provided to the user.
  - Legibility
    - Titles should be centred.
    - Labels should be displayed in upper case letters.
    - Cursors should be distinguished from other displayed items.
    - When space for text display is limited, display a few long lines of text rather than many short lines of text.
    - Display continuous text in wide columns, containing at least 50 characters per line.
    - Right justification should be employed if it can be achieved by variable spacing, maintaining constant proportional differences in spacing between and within words, and consistent spacing between words in a line.
    - In display of textual material, keep words intact, with minimal breaking by hyphenation between lines.

- Workload
  - Brevity
    - Concision
      - For numerical data, entry of leading zeros should not be necessary.
      - If codes are longer than 4 or 5 characters, use mnemonics or abbreviations.
      - Allow users short data entries.
      - When a measurement unit is associated with a particular data field, include that unit as part of the field label rather than requiring a user to enter it.
    - Minimal Actions
      - Minimise the number of steps required to make a selection in a menu.
      - Do not require data entry by the user when the data can be derived by the computer.
      - Avoid users' entries of commands that include punctuation.
      - For data entry, display currently defined default values in their appropriate data fields.
      - For long, multipage displays, it should be possible to request a particular page directly, without having to go through all intermediary pages.
  - Information Density
    - Provide only necessary and immediately usable data for any transaction; do not overload displays with extraneous data.
    - Data should not require unit translations.
    - Query language should use the minimum of quantifiers in query formulation.
    - Do not require users to remember data accurately from one display frame to another.
    - Provide automatic computation of derived data, so that a user does not have to calculate and enter any number that can be derived from data already accessible to the computer.
- Explicit Control
  - Explicit User Action
    - Always require a user to take an explicit ENTER action to initiate processing of entered data; do not initiate processing as a side effect (e.g., updating a file) of some other action (e.g., printing a file).
    - If menu selection is accomplished by pointing, provide for dual activation, in which the first action (positions a cursor) designates the selected option, followed by a separate second action that makes an explicit control entry.
    - Users' command entries should be completed with an ENTER action following editing facilities.
  - User Control
    - Allow users to pace their data entry, rather than having the pace being controlled by computer processing or by external events.
    - The cursor should not be automatically moved without users' control (except for stable and well known procedures, such as in form-filling).
    - Users should have the control over the screen pages.

        * Allow users to interrupt or cancel a current transaction or process.
        * Provide a CANCEL option which will have the effect of erasing any changes just made by the user and restoring the current display to its previous version.

- Adaptability
    - Flexibility
        * When user requirements are uncertain, provide users with some means to control display configuration.
        * When interface designers cannot predict which default values will be helpful, permit users to define, change or remove default values for data entry.
        * When some displays are unnecessary, the users should be able to remove them temporarily.
        * Provide some means for users to change the data entry sequence to respect their preferred sequence.
        * When text formats cannot be predicted in advance, allow users to specify and store for future use the formats that might be needed.
        * Users should be able to assign names to data fields they have created.
    - User Experience
        * Allow experienced users to by-pass a series of menu selections and make an equivalent command entry or keyboard shortcuts directly.
        * Allow experienced users to key in a series of commands at one time, and inexperienced users only step by step keying.
        * Dialogue types must be designed to match the needs of different users.
        * Different dialogue types should be provided as a function of the experience of the various users groups (e.g., provide prompting as an optional guidance feature that can be selected by novice users but can be omitted by experienced users.
        * When techniques adopted for user guidance may slow an experienced user, provide alternative paths or modes permitting a user to by-pass standard guidance procedures.
        * Following the output of an error message, permit users to request a more detailed explanation of the error that is adapted to their level of knowledge.

- Error Management
    - Error Protection
        * When a user requests LOG-OFF and if any pending transaction will not be completed, or if data will be lost, display an advisory message requesting user confirmation.
        * Protect field labels from accidental change by users.
        * Fields designed for information display should be protected: users should not be allowed to change the information contained in these fields.
        * Ensure that user interface software will deal appropriately with all possible user errors, including accidental inputs.
    - Quality of Error Messages

    - If the user selects an invalid function key, no system action should result except a message indicating the functions appropriate for that transaction step.
    - For error messages, adopt task-oriented wording.
    - Make the error messages as specific as possible.
    - Make error messages brief but informative.
    - Adopt neutral wording for error messages; do not imply blame on the part of the user, or personalise the computer, or attempt to make a message humorous.
  - Error Correction
    - Users should be allowed to edit an extended command during its composition before taking an explicit action to ENTER the command.
    - Following error detection, require the user to re-enter only that portion of a data/command entry which is not correct.
    - If a data entry transaction has been completed and errors detected, permit users to make corrections directly and immediately.

- Consistency
  - Window titles should always be located in the same place.
  - Use similar screen formats.
  - Use similar procedures to access menu options.
  - Consistent phrasing and punctuation should be used in all prompts.
  - Prompts for data or command entry should be displayed in a standard location.
  - Data entry fields should always be the same.

- Significance of Codes
  - Titles should be distinct and meaningful.
  - Make abbreviation rules explicit.
  - Codes should be meaningful and familiar rather than arbitrary (e.g., M for Male, and F for Female rather than 1 and 2).

- Compatibility
  - When data entry involves transcription from source documents, ensure that form-filling displays match those documents.
  - Dialogues should reflect data structures or organisations which are perceived by users as being natural.
  - Calendar formats should follow users' customs (American vs European calendar).
  - Labels, prompts, and user guidance messages should be familiar to the users and task-oriented.
  - Units of measurement should be familiar to the user.
  - Displays of textual data, messages, or instructions, should follow design conventions for printed text.

C.1.4. Readability

Design Guidelines for Web Readability (Miniukovich et al., 2017):

- Use short, simple sentences in a direct style.
- Avoid complex language and jargon.
- Consider using short paragraphs.
- Put the main point of sentence or paragraph into the beginning of the sentence or paragraph.
- Use section headings to organize the content.
- Limit the amount of content on a page to avoid scrolling.
- Avoid using italics in the main body of the text.
- Avoid underlining large blocks of text as it makes reading harder.
- Use a minimum of text size 12pt or 14pt.
- Avoid formatting texts in large-width columns.
- Use an off-white color for your background, like light gray or tan; use dark gray for text instead of pure black. Use a plain, evenly spaced sans serif font such as Arial and Comic Sans.

C.1.5. Accessibility

Web Content Accessibility Guidelines (Caldwell et al., 2008):

- Perceivable
  - Provide text alternatives for an non-text content so that it can be changed into other forms people need, such as large print, braille, speech, symbols or simpler language.
  - Provide alternatives for time-based media.
  - Create content that can be presented in different ways (for example simpler layout) without losing information or structure.
  - Make it easier for users to see and hear content including separating foreground from background.
- Operable
  - Make all functionality available from a keyboard.
  - Provide users enough time to read and use content.
  - Do not design content in a way that is known to cause seizures.
  - Provide ways to help users navigate, find content, and determine where they are.
- Understandable
  - Make text content readable and understandable.
  - Make Web pages appear and operate in predictable ways.
  - Help users avoid and correct mistakes.
- Robust
  - Maximize compatibility with current and future user agents, including assistive technologies.

C.1.6. Motivation

Self-Report Motivational Model (de Vicente & Pain, 2002):

- Permanent Traits
  - **Control**
    Refers to the degree of control that the student likes having over the learning situation (i.e. does he like to select which exercises to do, in which order, etc. rather than let the instructor take these decisions?).
  - **Challenge**
    Refers to the degree that the student enjoys having challenging situations during the instruction (i.e. does he like to try difficult exercises that represent a challenge for him?).
  - **Independence**
    Refers to the degree that the student prefers to work independently, without asking others for help (i.e. does he prefer to work on his own, even if he finds some difficulties, and try to solve them by himself rather than asking for collaboration or help from others?).
  - **Fantasy**
    Refers to the degree that the student appreciates environments that evoke mental images of physical or social situations not actually present (i.e. does he like the learning materials being embedded in an imaginary context?).
- Transient States
  - **Confidence**
    Refers to the student's belief in being able to perform the task at hand correctly.
  - **Sensory Interest**
    Refers to the amount of curiosity aroused through the interface presentation (i.e. appeal of graphics, sounds, etc.).
  - **Cognitive Interest**
    Refers to curiosity aroused through the cognitive or epistemic characteristics of the task (i.e. regardless of the presentation issues, does the student find the task at hand cognitively appealing?).
  - **Effort**
    Refers to the degree that the student is exerting himself in order to perform the learning activities.
  - **Satisfaction**
    Refers to the overall feeling of goal accomplishment (i.e. does the student think that the instruction is satisfying and that it is getting him closer to his goals?).

C.1.7. Persuasion

Criteria for the Assessment of Technological Persuasion (Némery et al., 2011):

- Static Criteria
  - **Credibility of Interaction**
    Giving enough information to the user enables him to identify the source of informa-

tion to be reliable, expert and trustworthy.
Example: *Presenting updated information and the date of the update.*
- **Guarantee of Privacy**
Do not persuade the user to do something that publicly exposes his private life and which he would not consent to do.
Example: *The system should preserve wherever possible the right of a user to remain anonymous to the larger user community as well as the providers of the system.*
- **Personalization**
Present information adapted to the user or to the user group.
Example: *Remembering the state from a user's last interactions.*
- **Attractiveness**
Presentation of elements in a way that is engaging and visually appealing.
Example: *The choice of colors as a reinforcement of the message.*

- Dynamic Criteria
  - **Solicitation**
  To solicit user in a light way, to catch his attention with an engaging effect.
  Example: *Display many invitations.*
  - **Priming**
  Trigger user interaction by creating a point of entry, stirring interest.
  Example: *Click to immediately view relevant and interesting content for free.*
  - **Commitment**
  Reinforce repeat behavior and frequency, as well as individual engagement.
  Example: *Improve the frequency of the final behavior or attitude expected.*
  - **Ascendancy and Possibility of Addiction**
  Show engaging scenario completion, follow up its influence and control its evolution over time.
  Example: *The individual accepts information that he would not have accepted voluntarily.*

C.1.8. Situation Awareness

Toward a Theory of Situation Awareness in Dynamic Systems (Endsley, 1995):

- As attention and working memory are limited, the degree to which displays provide information that is processed and integrated in terms of Level 2 and 3 SA requirements will positively affect SA. For Instance, directly portraying the amount of time and distance available on the fuel remaining in an aircraft would be preferable to requiring the pilot to calculate this information based on lower-level data (e.g., fuel, speed, altitude, etc.).
- The degree to which information is presented in terms of the operator's major goals will positively affect SA. Many systems provide information that is technology oriented – based on physical system parameters and measurements (e.g., oil pressure or temperature). To improve SA, this information needs to be SA oriented. That is, it should be organized so that the information needed for a particular goal is colocated and directly answers the major decisions associated with the goal. For example, for the goal of weapons employment, factors such as opening/closing velocity, weapon selected and firing envelope, probability

of kill, target selected, and time to employment would be relevant elements that should be presented in an integrated form for this goal.

- Considering that mental models and schemata are hypothesized to be key tools for achieving the higher levels of SA in complex systems, the critical cues used for activating these mechanisms need to be determined and made salient in the interface design. In particular those cues that will indicate the presence of prototypical situations will be of prime importance. Kaplan and Simon (1990) found decision making is facilitated if the critical attributes are perceptually salient.
- Designs need to take into consideration both top- down and bottom-up processing. In this light, environmental cues with highly salient features will tend to capture attention away from current goal- directed processing. Salient design features, such as those indicated by Treisman and Paterson (1984), should be reserved for critical cues that indicate the need for activating other goals and should be avoided for noncritical events.
- A major problem for SA occurs when attention is directed to a subset of information and other important elements are not attended to, either intentionally or unintentionally (Endsley and Bolstad, 1993). It is hypothesized that designs that restrict access to SA elements (via information filtering, for instance) will contribute to this problem. A preferred design will provide global SA-an over- view of the situation across operator goals-at all times, while providing the operator with detailed information related to his or her immediate goals, as required. Global SA is hypothesized to be important for determining current goals and for enabling projection of future events.
- Although filtering out information on relevant SA elements is hypothesized to be detrimental, the problem of information overload in many systems must still be considered. The filtering of extraneous information (not related to SA needs) and reduction of data (by processing and integrating low-level data to arrive at SA requirements) should be beneficial to SA.
- One of the most difficult and taxing parts of SA is the projection of future states of the system. This is hypothesized to require a fairly well developed mental model. System-generated support for projecting future events and states of the system should directly benefit Level 3 SA, particularly for less-experienced opera tors.
- The ability to share attention between multiple tasks and sources of information will be very important in any complex system. System designs that support parallel processing of information should directly benefit SA. For example, the addition of voice synthesis or three-dimensional audio cues to the visually overloaded cockpit is predicted to be beneficial on this basis.

Model for Situation Awareness and Driving: Application to Analysis and Research for Intelligent Transportation Systems (M. L. Matthews et al., 2001):

- For warning and alerting systems, maximize the rapid detection of new information.
- For information systems, minimize the impact on the driver's existing processes of scanning and information acquisition.
- Design information content in a format that allows it to be rapidly integrated with other information that is held in SA.
- Ensure that the format of the information is consistent with the driver's relevant mental model.

- For multiple in-vehicle systems, ensure that visual and auditory warnings and alerts are unambiguous and consistent with the relevant driver's mental model.
- In the absence of a specific mental model, avoid in designs information density that exceeds the capacity of working memory.
- Minimize the requirement for the driver to develop new mental models to process the information to avoid imposing a cognitive load associated with the requirement to juggle mental models.
- Consider any training requirements that may be necessary to facilitate the rapid integration of the new technology into the driver's existing mental models.

C.1.9. Automotive HMI Guidelines

Table C.1.: Consolidated automotive HMI guidelines from Alliance of Automobile Manufacturers (2006) as AAM, Japan Automobile Manufacturers Association (2004) as JAMA, Commission of the European Communities (2008) as ESoP, National Highway Traffic Safety Administration (2013) as NHTSA, and Kroon et al. (2016) as DITCM.

No.	Source	Guideline
Installation Principles		
1	AAM	The system should be located and fitted in accordance with relevant regulations, standards, and the vehicle components manufacturers' instructions for installing the systems in vehicles.
1	JAMA	In case of retrofit display systems, vehicle manufacturers shall take measures to ensure that such display systems be installed at proper positions inside their vehicles.
1	ESoP	The system should be located and securely fitted in accordance with relevant regulations, standards and manufacturers' instructions for installing the system in vehicles.
1	DITCM	The display should always be fixed to the car with a holder, preferably in 10 to 20 cm reach of the hand.
2	AAM	No part of the system should obstruct the driver's field of view as defined by applicable regulations.
2	JAMA	A display system shall not obstruct any part of the driver's visual field that is necessary for driving.
2	ESoP	No part of the system should obstruct the driver's view of the road scene.
2	NHTSA	No part of the physical device should, when mounted in the manner intended by the manufacturer, obstruct a driver's view.
3	AAM	No part of the physical system should obstruct any vehicle controls or displays required for the driving task.
3	JAMA	A display system shall not interfere with the operation of the steering device nor limit the visibility of the various meters from the driver.
3	ESoP	The system should not obstruct vehicle controls and displays required for the primary driving task.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
3	NHTSA	No part of the physical device should, when mounted in the manner intended by the manufacturer, obstruct a driver's view of any vehicle controls or displays required for the driving task.
4	NHTSA	The mounting location for a device should not be in a location that is difficult to see and/or reach (as appropriate) while driving.
5	NHTSA	The 2D Maximum Downward Angle is equal to 30.00 degrees for a vehicle with the height above the ground of the nominal driver eye point less than or equal to 1700 millimeters above the ground.
6	NHTSA	The 2D Maximum Downward Angle is given by the following equation for nominal driver eye point heights greater than 1700 millimeters above the ground: $\theta_{2DMax} = 0.01303h_{Eye} + 15.07$; θ_{2DMax} is the 2D Maximum Downward Angle. h_{Eye} is the height above the ground of the nominal driver eye point.
7	NHTSA	The 3D Maximum Downward Angle is equal to 28.16 degrees for a vehicle with the height above the ground of the nominal driver eye point less than or equal to 1146.2 millimeters above the ground.
8	NHTSA	The 3D Maximum Downward Angle is given by the following equation for nominal driver eye point heights greater than 1146.2 millimeters above the ground: $\theta_{3DMax} = 57.2958tan^{-1}[0.829722tan(0.263021 + 0.000227416h_{Eye})]$; θ_{3DMax} is the 3D Maximum Downward Angle. h_{Eye} is the height above the ground of the nominal driver eye point.
9	NHTSA	Visual displays that present information highly relevant to the driving task and/or visually-intensive information should have downward viewing angles that are as close as practicable to a driver's forward line of sight. Visual displays that prevent information less relevant to the driving task should have lower priority, when it comes to locating them to minimize their downward viewing angles, than displays that present information highly relevant to the driving task.
10	AAM	Visual displays that carry information relevant to the driving task and visually-intensive information should be positioned as close as practicable to the driver's forward line of sight.
10	JAMA	The operating section of a display system shall not be located at a position that causes the driver to be substantially displaced from driving posture when operating the display system.
10	ESoP	Visual displays should be positioned as close as practicable to the driver's normal line of sight.
10	NHTSA	Visual displays that present information relevant to the driving task and/or visually-intensive information should be laterally positioned as close as practicable to a driver's forward line of sight.
11	AAM	Visual displays should be designed and installed to reduce or minimize glare and reflections.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
11	JAMA	The luminous intensity, contrast, colors and other display conditions of a display system shall be such that the driver is not dazzled by the display at night.
11	ESoP	Visual displays should be designed and installed to avoid glare and reflections.
Information Presentation		
1	AAM	Systems with visual displays should be designed such that the driver can complete the desired task with sequential glances that are brief enough not to adversely affect driving.
1	JAMA	The visual information to be displayed shall be sufficiently small in volume to enable the driver to comprehend it in a short time or shall be presented in portions for the driver to scan them in two or more steps.
1	ESoP	Visually displayed information presented at any one time by the system should be designed in such a way that the driver is able to assimilate the relevant information with a few glances which are brief enough not to adversely affect driving.
1	DITCM	Information should not lead to glances that exceed 2 seconds eyes off the road.
2	AAM	Where appropriate, internationally agreed upon standards or recognized industry practice relating to legibility, icons, symbols, words, acronyms, or abbreviations should be used. Where no standards exist, relevant design guidelines or empirical data should be used.
2	JAMA	It is desirable that a display system be designed, where possible, to comply with an internationally agreed standard respecting readability, audibility, icons, symbols, letters, abbreviations, and other factors relating to the manner of information display.
2	ESoP	Internationally and/or nationally agreed standards relating to legibility, audibility, icons, symbols, words, acronyms and/or abbreviations should be used.
2	DITCM	The traffic information service should use the formal national signs and signals of the local country (no modifications).
3	AAM	Available information relevant to the driving task should be timely and accurate under routine driving conditions.
3	ESoP	Information relevant to the driving task should be accurate and provided in a timely manner.
3	DITCM	The content of the information should be relevant and in line with the traffic scenario at that moment in time to be valid.
4	JAMA	A display system shall not present the kind of information that impairs the safety and smooth flow of road traffic.
5	DITCM	Information should make sense in the situation, not conflicting with perceived feasibility.
6	ESoP	Information with higher safety relevance should be given higher priority.
6	DITCM	Safety related warnings have priority over non-safety related information

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
7	DITCM	Safety related warnings should always be combined with an auditory attention cue.
8	DITCM	Information that is related to the maneuver or control level of the driving task has priority over information related to the navigation level of the driving task.
9	DITCM	Information which is always of (high) relevance to the driving task can be displayed best continuously at a fixed position on the screen.
10	DITCM	Information that requires behavioral change has priority over information that does not.
11	DITCM	Emotional content should be avoided.
12	DITCM	The display does not present more than 4 separate types of information units simultaneously in relation to an event, next to the continuously shown navigation information.
13	DITCM	A 'neutral' auditory sound should be used when warning for hazardous situations rather than emotion-laden sounds.
14	DITCM	Information presented should not be interpretable in multiple ways.
15	DITCM	In-car information is in accordance with local road side information, discrepancies between different information resources should not occur.
16	DITCM	Dynamic information provision such as a sudden speed limit change or closing of a traffic lane should be accompanied by an argument.
17	AAM	The system should not produce uncontrollable sound levels liable to mask warnings from within the vehicle or outside or to cause distraction or irritation.
17	JAMA	A display system shall not be capable of generating an uncontrollable volume of sound that may cancel out alarms sounded from inside or outside of the vehicle.
17	ESoP	System-generated sounds, with sound levels that cannot be controlled by the driver, should not mask audible warnings from within the vehicle or the outside.
17	NHTSA	Devices should not produce uncontrollable sound levels liable to mask warnings from within the vehicle or outside or to cause distraction or irritation.
Interaction with Displays/Controls		
1	DITCM	The information service should not require any manual control input from the driver while driving.
2-3	AAM	The system should allow the driver to leave at least one hand on the steering control.
2-3	JAMA	The operation of a display system shall not cause the driver to remove both hands simultaneously from the steering wheel.
2-3	ESoP	The driver should always be able to keep at least one hand on the steering wheel while interacting with the system.
2	NHTSA	A driver's reach to the device's controls should allow one hand to remain on the steering control at all times.
3	NHTSA	When manual device controls are placed in locations other than on the steering control, no more than one hand should be required for manual input to the device at any given time during driving.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
4	NHTSA	When device controls are located on the steering wheel and both hands are on the steering wheel, no device task should require simultaneous manual inputs from both hands.
5	NHTSA	Reach of the whole hand through steering wheel openings should not be required for operation of any device controls.
6	JAMA	The operation of a display system shall not result in a marked obstruction of forward field visibility.
7	AAM	Speech-based communication systems should include provision for hands-free speaking and listening. Starting, ending, or interrupting a dialog, however, may be done manually. A hands-free provision should not require preparation by the driver that violates any other principle while the vehicle is in motion.
8-10	AAM	The system should not require uninterruptible sequences of manual/visual interactions. The driver should be able to resume an operator-interrupted sequence of manual/visual interactions with the system at the point of interruption or at another logical point in the sequence.
8	JAMA	Information to be presented by a display system shall not cause the driver to gaze at the screen continuously.
9	JAMA	The content of visual information to be displayed while the vehicle is in motion shall relate exclusively to driving, but shall not necessitate the driver gazing at it continuously.
10	JAMA	Preferably, the visual information to be displayed is sufficiently small in volume or is presented in portions so that the display system can be operated in separate steps.
10	ESoP	The driver should be able to resume an interrupted sequence of interface with the system at the point of interruption or at another logical point.
10	NHTSA	A visual display of previously-entered data or current device state should be provided to remind a driver of where the task was left off.
11	ESoP	The system should not require long and uninterruptible sequences of manual-visual interface. If the sequence is short, it may be uninterruptible.
11	NHTSA	No device-initiated loss of partial driver input (either data or command inputs) should occur automatically.
12	NHTSA	If feasible, necessary, and appropriate, the device should offer to aid a driver in finding the point to resume the input sequence or in determining the next action to be taken. Possible aids include, but are not limited to: • A visually displayed indication of where a driver left off • A visually displayed indication of input required to complete the task • An indication to aid a driver in finding where to resume the task
13	AAM	In general (but with specific exceptions) the driver should be able to control the pace of interaction with the system. The system should not require the driver to make time-critical responses when providing input to the system.
13	JAMA	When the driver is to input data into a display system, the display system shall not demand immediate responses from the driver.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
13	ESoP	The driver should be able to control the pace of interface with the system. In particular the system should not require the driver to make time-critical responses when providing inputs to the system.
14	ESoP	System controls should be designed in such a way that they can be operated without adverse impact on the primary driving controls.
15	JAMA	Preferably, a display system is so designed that its display of information can be discontinued by the driver.
15	DITCM	Upon request of the driver, it should always be possible to turn off the application, to adjust the brightness of the screen, and tune the volume. Furthermore, operating buttons should require minimal visual guidance.
16	JAMA	Preferably, when its display of information is discontinued, a display system is capable of resuming the display from the point of discontinuation or a point enabling the understanding of the displayed information as a whole.
17	JAMA	A display system shall be equipped with a means of controlling auditory information, but not including alarms, for the driver who may find auditory information distracting or irritating.
17	ESoP	The driver should have control of the loudness of auditory information where there is likelihood of distraction.
18	AAM	The system's response (e.g. feedback, confirmation) following driver input should be timely and clearly perceptible.
18	JAMA	Information, such as the reporting of system state and operation that is displayed in response to the data inputted by the driver shall be quickly and easily comprehensible.
18	ESoP	The system's response (e.g. feedback, confirmation) following driver input should be timely and clearly perceptible.
18	NHTSA	A device's response (e.g. feedback, confirmation) following driver input should be timely and clearly perceptible. The maximum device response time to a device input should not exceed 0.25 second. If device response time exceeds 0.25 second, a clearly perceptible indication should be given indicating that the device is responding.
19	DITCM	Information should be presented preferably about 36 seconds before the point of action or 200m before the first road sign.
20	DITCM	Information should be presented minimally 9 seconds before the point of action.
21	AAM	Systems providing non-safety-related dynamic (i.e. moving spatially) visual information should be capable of a means by which that information is not provided to the driver.
21	ESoP	Systems providing non-safety-related dynamic visual information should be capable of being switched to a mode where that information is not provided to the driver.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
21	NHTSA	Devices providing dynamic (i.e. moving) non-safety-related visual information should provide a means by which that information is not seen by a driver. A device visually presenting dynamic non-safety-related information should make the information not seen by a driver through at least one of the following mechanisms: • Dimming the display information • Turning off or blanking the displayed information • Changing the state of the display so that the dynamic, non-safety-related information cannot be seen by a driver while driving • Positioning or moving the display so that the dynamic, non-safety-related information cannot be seen while driving
22	NHTSA	Drivers may initiate commands that erase driver inputs.
23	NHTSA	Devices may revert automatically to a previous state without the necessity of further driver input after a device defined time-out period, provided: • It is a low priority device state (one that does not affect safety-related functions or way finding) • The state being left can be reached again with low driver effort. In this context, low driver effort is defined as either a single driver input or not more than four presses of one button
System Behavior		
1	AAM	Visual information not related to driving that is likely to distract the driver significantly (e.g., video and continuously moving images and automatically-scrolling text) should be disabled while the vehicle is in motion or should be only presented in such a way that the driver cannot see it while the vehicle is in motion.
1	JAMA	A display system's functions that are not presumed to be used by the driver during driving operation shall be inoperative by the driver while the vehicle is in motion.
1	ESoP	While the vehicle is in motion, visual information not related to driving that is likely to distract the driver significantly should be automatically disabled, or presented in such a way that the driver cannot see it.
1	NHTSA	Device functions and tasks not intended to be used by a driver while driving should always be inaccessible for performance by the driver while driving.
2	NHTSA	Manual text entry by the driver for the purpose of text-based messaging, other communication, or internet browsing should always be inaccessible for performance by the driver while driving.
3	NHTSA	Displaying dynamic or static visual photographic or graphical images not related to driving including, but not limited to: • Video-based entertainment in view of the driver • Video-based communications including video phone calls and other forms of video communication

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
4	NHTSA	Displaying non-video graphical or photographic images should always be inaccessible for performance by the driver while driving.
5	NHTSA	The display of scrolling text that is moving at a pace not controlled by the driver should always be inaccessible for performance by the driver while driving.
6	NHTSA	The visual presentation of the following types of non-driving-related task textual information should always be inaccessible for performance by the driver while driving: books, periodical publications, web page content, social media content, text-based advertising and marketing, text-based messages and correspondence.
7	AAM	System functions not intended to be used by the driver while driving should be made inaccessible for the purpose of driver interaction while the vehicle is in motion.
7	ESoP	System functions not intended to be used by the driver while driving should be made impossible to interact with while the vehicle is in motion, or, as a less preferred option, clear warnings should be provided against the unintended use.
7	NHTSA	Any secondary task that draws a driver's attention from the primary driving task to the point where safety is reduced should be locked out unless either: • The vehicle's engine is not running • The vehicle's transmission is in "Park" (automatic transmission vehicles) or the vehicle's transmission is in "Neutral" and the parking brake is on (manual transmission vehicles)
8	AAM	The system should clearly distinguish between those aspects of the system, which are intended for use by the driver while driving, and those aspects (e.g. specific functions, menus, etc) that are not intended to be used while driving.
8	ESoP	The behavior of the system should not adversely interfere with displays or controls required for the primary driving task and for road safety.
8	NHTSA	Devices should clearly distinguish between those aspects of a device which are intended for use by a driver while driving, and those aspects (e.g., specific functions, menus, etc.) that are not intended to be used while driving.
9	AAM	Information about current status, and any detected malfunction, within the system that is likely to have an adverse impact on safety should be presented to the driver.
9	ESoP	Information should be presented to the driver about current status and any malfunction within the system that is likely to have an impact on safety.
9	NHTSA	Information about current status, and any detected malfunction, within the device that is likely to have an adverse impact on safety should be presented to the driver.
10	DITCM	The occurrence of false alarms and misses should be minimized, to ensure reliable information.
11	DITCM	Information can be presented best when the workload of the primary task is low (tedious for some, to a long, time), e.g. driving on a quiet road with low traffic density and activity for a long time.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
12	DITCM	In complex situations, depending on the complexity of the infrastructure, the traffic density and the speed that is being driven, information provided to the driver should be minimized; less urgent messages should be postponed.
13	DITCM	An advice should not lead to higher speeds, and particularly avoid large speed differences between different drivers (maximum 20 km/h differences in operating speed).
14	DITCM	City centers, school areas, and other safety critical areas should be avoided (if it is not the final destination).
Information About the System		
1	AAM	The system should have adequate instructions for the driver covering proper use and safety-relevant aspects of installation and maintenance.
1	JAMA	The importance of safe driving shall be appealed to the users of display systems through pamphlets and operation manuals and by fully educating the vehicle dealers and the sellers of retrofit display systems.
1	ESoP	The system should have adequate instructions for the driver covering use and relevant aspects of installation and maintenance.
2	AAM	Safety instructions should be correct and simple.
2	JAMA	Safety information to be presented to the users of display systems shall be accurate, simple and clear.
2	ESoP	System instructions should be correct and simple.
3	AAM	System instructions should be in a language or form designed to be understood by drivers in accordance with mandated or accepted regional practice.
3	JAMA	The operation manuals of display systems shall be so written and graphically designed as to be easily comprehensible by display system users.
3	ESoP	System instructions should be in languages or forms designed to be understood by the intended group of drivers.
3	DITCM	Text and sound are preferably displayed in the driver's preferred language.
4	AAM	The instructions should distinguish clearly between those aspects of the system that are intended for use by the driver while driving, and those aspects (e.g. specific functions, menus, etc) that are not intended to be used while driving.
4	ESoP	The instructions should clearly state which functions of the system are intended to be used by the driver while driving and those which are not.
5	ESoP	Product information should be designed to accurately convey the system functionality.
6	AAM	Product information should make it clear if special skills are required to use the system or if the product is unsuitable for particular users.
6	ESoP	Product information should make it clear if special skills are required to use the system as intended by the manufacturer or if the product is unsuitable for particular users.

Continued on next page

Table C.1 - *Continued from previous page*

No.	Source	Guideline
7	AAM	Representations of system use (e.g. descriptions, photographs, and sketches) provided to the customer with the system should neither create unrealistic expectations on the part of potential users, nor encourage unsafe or illegal use.
7	ESoP	Representations of system use (e.g. descriptions, photographs and sketches) should neither create unrealistic expectations on the part of potential users nor encourage unsafe use.

C.2. User Study

Table C.2.: Results of Shapiro-Wilk test for normality and correlation analysis for task performance between preview mode conditions.

	Normality		Correlation	
Task	W	p	r_s	p
Copy configuration (1)	.48	<.001	**.79**	<.001
Copy configuration (2)	.74	<.001	.28	.27
Switch content	.57	<.001		
Switch content (with temporary storage)	.75	<.001	.36	.14
Delete content	.56	<.001	.29	.24
Overwrite content	.70	<.001	-.16	.53
Finalize configuration	.48	<.001	-.06	.82

Note. W = test statistic of Shapiro-Wilk test. Significant positive correlations are in boldface. For empty cells, the correlation coefficient could not be calculated due to invariant values.

Table C.3.: List of identified usability problems through user testing; classified according to the UPC.

#	Participant	Variant	ID	Description	Task Component		Object Component	
1	0	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating
2	0	dynamic	2	The preview overlay occludes content elements.	Before	Determining how to do next step	Physical attributes	Placement
3	0	dynamic	3	The fade out duration of the preview is too long.	During	Trouble performing step	Physical attributes	Not manipulable
4	0	-	4	Drag & Drop operation is cumbersome.	During	Trouble performing step	Physical attributes	Difficult to control
5	1	-	5	The order of the content elements is not comprehensible.	Before	Determining how to do next step	Physical attributes	Placement
6	1	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
7	2	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
8	6	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
9	1	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
10	2	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating
11	6	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
12	8	-	29	A on-screen help function is missing.	Before	Determining next step	Cognitive attributes	Content
13	2	-	10	Swiping in content lists is not possible.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
14	2	fixed	11	Grayed out content elements are not manipulable.	During	Trouble performing step	Physical attributes	Not manipulable
15	2	dynamic	12	Switching content elements via mode is not self-explanatory.	Before	Determining how to do next step	Physical attributes	Placement
16	3	fixed	13	The configuration takes too much time.	During	Trouble performing step	Cognitive attributes	Indirectness
17	5	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating
18	5	-	10	Swiping in content lists is not possible.	During	Trouble performing step	Cognitive attributes	Manipulation concept
19	13	-	39	Some content elements are available in two sizes which is confusing to differentiate.	Before	Not confident about next step	Cognitive attributes	Content
20	5	fixed	15	The buttons to select a specific mode are unnecessary and confusing.	Before	Not confident about next step	Physical attributes	Placement
21	5	-	16	The grid for small and large content elements is not comprehensible for the user.	During	Trouble performing step	Cognitive attributes	Manipulation concept
22	5	-	17	The temporary storage is confusing at first sight.	During	Unexpected task automation	Cognitive attributes	Manipulation concept
23	5	dynamic	18	The content elements could be smaller in shape.	Before	Not confident about next step	Physical attributes	Size
24	13	dynamic	44	The screen is perceived overcrowded.	Before	Determining next step	Cognitive attributes	Content
25	14	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
26	15	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content
27	6	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
28	6	-	10	Swiping in content lists is not possible.	During	Trouble performing step	Cognitive attributes	Manipulation concept
29	16	dynamic	44	The screen is perceived overcrowded.	Before	Determining next step	Cognitive attributes	Content
30	6	-	21	Starting the configuration was assumed from the central display.	Before	Determining how to do next step	Physical attributes	Placement
31	6	-	22	Some content elements would be preferred in different size.	Before	Not confident about next step	Physical attributes	Size
32	6	-	22	Some content elements would be preferred in different size.	Before	Not confident about next step	Physical attributes	Size
33	6	-	23	The user is irritated by the highlighting when selecting empty space.	During	Unexpected task automation	Physical attributes	Irritating
34	7	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating
35	7	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
36	7	-	22	Some content elements would be preferred in different size.	Before	Not confident about next step	Physical attributes	Size
37	7	-	24	Buttons should be grouped in a specified location.	Before	Determining how to do next step	Physical attributes	Placement
38	7	-	25	The preview does not completely match the real world display.	Before	Not confident about next step	Physical attributes	Shape
39	7	fixed	26	The user is not able to differentiate between empty and already configured containers in the preview.	During	Trouble performing step	Physical attributes	Irritating
40	7	fixed	27	The button to finish the configuration should be placed elsewhere.	Before	Determining how to do next step	Physical attributes	Placement
41	19	-	6	The list of content elements is too limited.	Before	Determining how to do next step	Cognitive attributes	Content

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
42	19	fixed	52	Scrolling through the content elements lists reduces the overview.	Before	Determining how to do next step	Cognitive attributes	Content
43	1	fixed	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
44	8	-	30	Using the back button on the steering wheel control does not discard changes.	After	Outcome did not match goal	Unexpected results	
45	8	-	31	The content elements labels are used inconsistently	After	Uncertain of results	Feedback	Misleading
46	8	-	32	The difference between the content elements sizes stays unclear.	Before	Not confident about next step	Physical attributes	Size
47	1	dynamic	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
48	8	-	34	Content elements in the temporary storage are hard to recognize.	Before	Not confident about next step	Physical attributes	Size
49	8	dynamic	35	The preview is not manipulable regarding its position.	During	Trouble performing step	Physical attributes	Not manipulable
50	10	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
51	10	fixed	15	The buttons to select a specific mode are unnecessary and confusing.	Before	Not confident about next step	Physical attributes	Placement
52	11	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
53	11	-	23	The user is irritated by the highlighting when selecting empty space.	During	Unexpected task automation	Physical attributes	Irritating
54	11	dynamic	36	The preview is occluded by the users hand when dragging down.	Before	Determining how to do next step	Physical attributes	Placement
55	13	dynamic	3	The fade out duration of the preview is too long.	During	Trouble performing step	Physical attributes	Not manipulable

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
56	2	fixed	9	Graying out of already configured content elements stays unclear.	Before	Determining how to do next step	Cognitive attributes	Visual cues
57	13	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
58	13	-	18	The content elements could be smaller in shape.	Before	Not confident about next step	Physical attributes	Size
59	13	-	37	The startup animation is unnecessary.	After	Unexpected task automation	Feedback	Unnecessary
60	5	dynamic	14	The user is not able to identify how the configuration is performed.	Before	Determining how to do next step	Cognitive attributes	Visual cues
61	5	dynamic	19	The content elements could use icons to be better recognizable.	Before	Not confident about next step	Cognitive attributes	Visual cues
62	6	-	20	The highlight color is not distinguishable.	Before	Not confident about next step	Cognitive attributes	Visual cues
63	13	fixed	41	The button to discard the entire configuration is to prominent.	Before	Not confident about next step	Physical attributes	Placement
64	13	fixed	42	The order of touch events when interacting via direct touch is inverted.	During	Trouble performing step	Physical attributes	Irritating
65	8	-	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
66	8	-	28	The user is not able to identify how the configuration is performed.	Before	Determining next step	Cognitive attributes	Visual cues
67	13	dynamic	45	The icon for switch mode is not clear.	Before	Determining how to do next step	Physical attributes	Shape
68	14	-	5	The order of the content elements is not comprehensible.	Before	Determining how to do next step	Physical attributes	Placement
69	8	-	33	The highlighting of interaction elements stays unclear.	Before	Not confident about next step	Cognitive attributes	Visual cues

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
70	13	dynamic	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
71	14	-	22	Some content elements would be preferred in different size.	Before	Not confident about next step	Physical attributes	Size
72	14	fixed	15	The buttons to select a specific mode are unnecessary and confusing.	Before	Not confident about next step	Physical attributes	Placement
73	13	-	38	There is no hint that configuration is possible via drag&drop and direct touch.	Before	Not confident about next step	Cognitive attributes	Visual cues
74	14	fixed	25	The preview does not completely match the real world display.	Before	Not confident about next step	Physical attributes	Shape
75	14	-	5	The order of the content elements is not comprehensible.	Before	Determining how to do next step	Physical attributes	Placement
76	15	-	22	Some content elements would be preferred in different size.	Before	Not confident about next step	Physical attributes	Size
77	13	fixed	40	The user is not able to differentiate between available and already configured content elements.	Before	Not confident about next step	Cognitive attributes	Visual cues
78	15	fixed	23	The user is irritated by the highlighting when selecting empty space.	During	Unexpected task automation	Physical attributes	Irritating
79	15	fixed	10	Swiping in content lists is not possible.	During	Trouble performing step	Cognitive attributes	Manipulation concept
80	15	dynamic	46	The position of the temporary storage is not optimal.	Before	Not confident about next step	Physical attributes	Placement
81	16	fixed	15	The buttons to select a specific mode are unnecessary and confusing.	Before	Not confident about next step	Physical attributes	Placement
82	16	fixed	47	Content elements should be deleted using long press.	During	Trouble performing step	Cognitive attributes	Manipulation concept

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
83	13	fixed	43	The user is missing a possibility to see the content containers for small and large content elements at once.	Before	Not confident about next step	Cognitive attributes	Visual cues
84	16	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
85	16	dynamic	2	The preview overlay occludes content elements.	Before	Determining how to do next step	Physical attributes	Placement
86	16	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating
87	16	dynamic	3	The fade out duration of the preview is too long.	During	Trouble performing step	Physical attributes	Not manipulable
88	16	dynamic	48	There is no possibility to show the preview without selecting content elements.	Before	Determining next step	Physical attributes	Placement
89	18	fixed	49	The user needs to switch his hands to operate the configuration.	During	Trouble performing step	Physical attributes	Difficult to control
90	18	dynamic	8	A permanent preview of the configured screen is missing.	After	Uncertain of results	Feedback	Missing
91	18	-	50	Using the back button of the center stack display does not discard changes.	After	Outcome did not match goal	Unexpected results	
92	19	-	51	The user misses a possibility to configure static screen elements.	During	Trouble performing step	Physical attributes	Not manipulable
93	14	-	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
94	14	-	7	Some content elements are not self-explanatory.	Before	Not confident about next step	Cognitive attributes	Visual cues
95	19	-	5	The order of the content elements is not comprehensible.	Before	Determining how to do next step	Physical attributes	Placement
96	19	-	53	The discard all function is not available via long press on delete button.	Before	Determining how to do next step	Physical attributes	Placement

Continued on next page

Table C.3 - *Continued from previous page*

#	Participant	Variant	ID	Description	Task Component		Object Component	
97	19	fixed	25	The preview does not completely match the real world display.	Before	Not confident about next step	Physical attributes	Shape
98	19	dynamic	2	The preview overlay occludes content elements.	Before	Determining how to do next step	Physical attributes	Placement
99	21	dynamic	48	There is no possibility to show the preview without selecting content elements.	Before	Determining next step	Physical attributes	Placement
100	21	dynamic	47	Content elements should be deleted using long press.	During	Trouble performing step	Cognitive attributes	Manipulation concept
101	23	fixed	15	The buttons to select a specific mode are unnecessary and confusing.	Before	Not confident about next step	Physical attributes	Placement
102	23	dynamic	46	The position of the temporary storage is not optimal.	Before	Not confident about next step	Physical attributes	Placement
103	23	-	54	There is no possibility to undo the last operation.	During	Trouble performing step	Cognitive attributes	Manipulation concept
104	23	dynamic	1	The behavior of the preview overlay is not comprehensible.	During	Unexpected task automation	Physical attributes	Irritating

Note. The issues were numbered according to their occurrence during the tests. Gaps in the participant numbering occur because they were numbered according to the available time slots in the schedule. The *ID* column contains a numerical identifier for unique usability problems.

C.3. Guideline Review

Table C.4.: Guidelines included in the review template in German language with a description or examples.

No.	Guideline	Description/Examples
Features & Funktionalität		
1	Features und Funktionalität bedienen die Ziele des Nutzers.	Hauptziele und Vorgaben der Nutzer sollten identifiziert und angesprochen werden. Idealerweise sollte die Applikation alle Ziele aller Nutzer abdecken.
2	Features und Funktionalität unterstützen die gewünschten Workflows der Nutzer.	Die Applikation sollte die Vorgehensweise seiner Nutzer unterstützen oder zumindest mit der Arbeitsweise kompatibel sein.
3	Häufig benötigte Tätigkeiten sind bequem verfügbar und werden unterstützt.	Zum Beispiel können Shortcuts für häufig genutzte Funktionen angeboten werden, um die Bearbeitung typischer Aufgaben zu beschleunigen.
4	Nutzer werden entsprechend ihrem Kenntnisstand unterstützt.	Z.B. bekommen unerfahrene Nutzer Hilfe oder Anweisungen angeboten und werden Scrittweise an die Bedienung herangeführt oder werden spezielle Funktionen für Experten (z.B. Shortcuts) angeboten.
5	Handlungsaufforderungen sind ersichtlich, angemessen beschriftet und erscheinen klickbar.	Mögliche Aktionen sollten klar ersichtlich sein. Die primäre Handlungsaufforderung (die häufigste oder meistgenutzte Nutzeraktion) sollte klar hervortreten.
Main Screen		
6	Der Hauptscreen bietet einen klaren Auszug, sowie einen Überblick über den Inhalt, die Features und deren Funktionalität.	Z.B. wird eine Einführung und Übersicht über die Applikation angeboten mit ggf. beispielhaften Inhalten.
7	Der Hauptscreen bietet eine effektive Orientierung für den Nutzer.	Es sollte dem Nutzer möglich sein zu Erkennen wie sich eine gestellte Aufgabe mit der Applikation lösen lässt.
8	Das Layout des Hauptscreens ist klar strukturiert und wirkt nicht überladen.	Ein Nutzer sollte möglichst schnell beim Überfliegen des Hauptscreens sowohl den Inhalt als auch die Struktur erfassen können.
Navigation		
9	Die Applikation ist einfach und logisch erreichbar.	Ggf. sollte der entsprechende Menüpunkt deutlich beschriftet und entsprechend seiner Wichtigkeit untergebracht sein.
10	Die Navigation innerhalb der Applikation ist einfach zu finden, intuitiv und konsistent.	Nutzer sollten in der Lage sein die Navigation innerhalb der Applikation zu lokalisieren und ggf. die Menüstruktur zu erfassen. Die Struktur sollte sich innerhalb der Applikation nicht wesentlich verändern oder anders dargestellt werden.

Continued on next page

Table C.4 - *Continued from previous page*

No.	Guideline	Description/Examples
11	Die Navigation innerhalb der Applikation ist flexibel und erlaubt dem Nutzer mit seiner gewünschten Möglichkeit zu navigieren.	Z.B. möchten Nutzer einerseits die Möglichkeit haben per Touch zu interagieren oder eine Remote-Bedienung durchführen. Darüber hinaus suchen manche Nutzer lieber gezielt nach dem gewünschten Inhalt, andere stöbern durch Listen von Inhalten. Hierbei müssen nicht alle Nutzerpräferenzen abgedeckt werden, aber die gängisten.
12	Die aktuelle Position in der Applikation ist eindeutig angegeben.	Der Nutzer sollte zu jedem Zeitpunkt wissen wo er sich gerade befindet.
Kontrolle, Fehler & Feedback		
13	Die Applikation reagiert mit umgehendem und angemessenem Feedback.	Erfolgreiche Operationen werden entsprechend und angemessen bestätigt, Eingabefehler werden hervorgehoben und eine Änderung am Zustand der Applikation ist klar ersichtlich.
14	Nutzerinteraktionen können auf einfache Weise rückgängig gemacht, geändert oder abgebrochen werden.	Sollte eine Benutzeraktion nicht auf einfache Weise rückgängig gemacht werden können, sollte zumindest eine Möglichkeit zur Bestätigung angeboten werden, bevor das Ergebnis verbindlich ist. Eingaben sollte auf einfache Weise geändert werden können, falls keine Möglichkeit angeboten wird Eingaben rückgängig zu machen.
15	Gängige Bedienfehler werden berücksichtigt und wo möglich vermieden.	Gängige Bedienfehler umfassen fehlende Eingaben, ungültige Eingaben oder eine ungültige Auswahl. Diese Bedienfehler sollten im Konzept berücksichtigt sein oder bestenfalls vermieden werden.
16	Fehlerhinweise sind klar, können einfach identifiziert werden und erscheinen an angemessenen Positionen.	Fehler sollten dem Nutzer unmittelbar aufgezeigt werden und dabei idealerweise so nah wie möglich an der entsprechenden Fehlerposition angezeigt werden.
17	Fehlermeldungen sind präzise, in verständlicher Sprache verfasst und beschreiben was passiert ist und was als nächstes zu tun ist.	Fehlermeldungen sollten technischen Jargon vermeiden und aus der Perspektive des Nutzers verfasst sein. Bei Bedarf sollten zusätzlich zu einem Highlight entsprechende Informationen und Handlungsaufforderungen zum Fehler angezeigt werden.
18	Nutzer sollten sich auf einfache Weise von Fehlern "erholen" können.	Z.B. sollte es vermieden werden, dass Nutzer nach einem Fehler ihre Aufgabe von Anfang an neu beginnen müssen. Eine erneute Bearbeitung der Eingabe oder eine erneuts Speichern mit anderen Werten sollte ermöglicht werden.
Inhalt		

Continued on next page

Table C.4 - *Continued from previous page*

No.	Guideline	Description/Examples
19	Die verwendete Sprache, Begrifflichkeit und der Umgangston sind angemessen und einfach verständlich für die Zielnutzer.	Fachsprache und Jargon sollten minimiert werden. Klare und verständliche Sprache sollte an allen möglichen Stellen verwendet werden.
20	Begriffe, Sprache und Umgangston werden konsistent verwendet.	Z.B. Kapitalisierung, Grammatik sollte konsistent verwendet werden. Entsprechend sollten auch formelle oder informelle Begrifflichkeiten immer gleich verwendet werden (z.B. Du/Sie, geht's/geht es, etc.).
21	Inhalt und Text sind lesbar, können durch überfliegen erfasst werden und weisen gute Typografie und visuellen Kontrast auf.	Nutzer sollten in der Lage sein die Hauptüberschriften und Labels im Screen durch überfliegen zu erfassen um einen Überblick über den Screen zu bekommen.
Driver Distraction		
22	Der Fahrer kann die angezeigte Information mit wenigen kurzen Blicken erfassen, ohne die Fahraufgabe negativ zu beeinflussen.	Die Struktur des Systems sollte es dem Nutzer ermöglichen sich schnell einen Überblick über die angebotene Funktion zu verschaffen, ohne dabei die Hauptaufgabe im Fahrzeug zu beeinflussen.
23	Die angezeigte Information führt nicht dazu, dass der Fahrer ununterbrochen auf den Bildschirm blickt.	Die Information kann z.B. durch das System in einzelnen kleinen Paketen angeboten werden, wodurch die Bedienung in einzelnen kleinen Schritten, statt einem großen/langen Schritt erfolgt.
24	Die Eingabe in das System kann zu jeder Zeit unterbrochen werden.	Z.B. werden keine zeitkritischen Eingaben vom Nutzer gefordert. Der Nutzer kann sich während der Interaktion mit dem System uneingeschränkt seiner Hauptaufgabe als Fahrer (Führen des Fahrzeugs) widmen.
25	Bei einer Unterbrechung der Eingabe in das System werden Hinweise zur Fortsetzung der Aufgabe gegeben.	Z.B. können visuelle Hinweise gegeben werden, mit welchem Objekt als letztes interagiert wurde oder welche Schritte noch erforderlich sind, um die Aufgabe abzuschließen.
Performance		
26	Die Performance des Systems beeinträchtigt die User Experience nicht.	Interaktionen, die länger als eine Sekunde dauern sollte mit entsprechendem Feedback versehen werden (z.B. Ladeanimation).
27	Systemfehler und Probleme mit der Zuverlässigkeit haben keinen negativen Einfluss auf die User Experience.	Die Interaktion sollte ohne Systemfehler möglich sein. Fehlerhafte Darstellungen oder Bugs sollten vermieden werden.

Table C.5.: List of identified usability problems through user testing; classified according to the UPC.

#	Expert	ID	Variant	Description	Guideline	Task Component		Object Component	
1	1	1	-	Graying out of content elements is not comprehensible for the user. Why are they not available?	16	During	Trouble performing step	Physical attributes	Not manipulable
2	1	2	dynamic	Dynamic preview is overwhelming the user.	3	During	Trouble performing step	Physical attributes	Difficult to control
3	1	3	-	Swipe in scroll lists not available.	2	During	Trouble performing step	Cognitive attributes	Indirectness
4	2	3	-	Swipe in scroll lists not available.	3	During	Trouble performing step	Physical attributes	Difficult to control
5	3	3	-	Swipe in content lists is not working	2	During	Trouble performing step	Cognitive attributes	Manipulation concept
6	4	3	-	Swipe in content lists not available.	11	During	Trouble performing step	Physical attributes	Difficult to control
7	1	4	-	Button to scroll in lists too far away from the actual content list.	2	Before	Determining how to do next step	Physical attributes	Placement
8	1	5	dynamic	Preview occluded by users hand.	2	Before	Determining how to do next step	Physical attributes	Placement
9	1	6	-	Mode-Buttons do not appear to activate a specific mode.	10	Before	Determining how to do next step	Cognitive attributes	Visual cues
10	4	4	-	List scroll arrow not recognizable.	10	Before	Determining next step	Physical attributes	Size
11	1	7	-	The Switch-Mode button is not self-explanatory.	19	Before	Determining next step	Cognitive attributes	Visual cues
12	1	8	-	Content switch when moving items inside the preview is not indicated/visualized.	13	After	Unexpected task automation	Feedback	Misleading

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task	Component	Object	Component
13	1	9	-	Elements in temporary storage are very small.	21	Before	Determining how to do next step	Physical attributes	Size
14	1	10	-	User is not able to identify what happens to items in the temporary storage.	4	After	Uncertain of results	Feedback	Missing
15	1	11	-	Content elements order is not comprehensible.	4	Before	Determining how to do next step	Physical attributes	Placement
16	1	12	-	Differentiation in list between content elements with the same icon but different size is not comprehensible.	7	Before	Determining how to do next step	Cognitive attributes	Content
17	2	12	-	Differentiation in list between content elements with the same icon but different size is not comprehensible.	7	Before	Determining how to do next step	Cognitive attributes	Content
18	4	4	-	List scroll arrow not recognizable.	10	Before	Determining next step	Physical attributes	Placement
19	1	13	fixed	The user might be irritated how to leave the automatically entered switch mode.	10	After	Unexpected task automation	Feedback	Missing
20	1	14	-	Content icons are not self-explanatory.	19	Before	Determining how to do next step	Cognitive attributes	Content
21	3	15	dynamic	Too many possible options for dynamic preview. Users might be irritated.	7	During	Unexpected task automation	Physical attributes	Irritating
22	4	6	-	Buttons do not appear to activate mode.	2	Before	Not confident about next step	Cognitive attributes	Visual cues
23	2	16	dynamic	Interaction is not self-explanatory when preview is missing.	6	During	Trouble performing step	Cognitive attributes	Indirectness
24	3	16	dynamic	Permanent preview is missing.	3	Before	Not confident about next step	Cognitive attributes	Content
25	4	12	-	Differentiation between content elements of different size but same icon not possible.	19	Before	Determining next step	Cognitive attributes	Visual cues

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task Component		Object Component	
26	2	17	dynamic	User might be unclear whether a configuration was successful because of missing preview.	13	After	Uncertain of results	Feedback	Missing
27	3	17	dynamic	Permanent preview is missing.	6	After	Unexpected task automation	Feedback	Misleading
28	3	18	dynamic	Dynamic preview occludes content elements. Possibilities for configuration are limited.	11	During	Trouble performing step	Physical attributes	Not manipulable
29	4	15	dynamic	Preview position not self-explanatory.	14	Before	Not confident about next step	Physical attributes	Placement
30	2	19	dynamic	It might be confusing that preview disappears after placing an element in a container.	6	After	Unexpected task automation	Feedback	Missing
31	3	20	dynamic	Tutorial shows different grid highlights than in configuration mode.	4	Before	Not confident about next step	Cognitive attributes	Visual cues
32	3	21	dynamic	Dynamic preview shows on top when selecting mode.	7	During	Unexpected task automation	Physical attributes	Irritating
33	3	22	dynamic	Preview fade out animation when selecting empty space takes too long.	13	After	Outcome did not match goal	Feedback	Misleading
34	4	16	dynamic	Fixed preview is missing.	7	Before	Determining how to do next step	Physical attributes	Placement
35	3	23	dynamic	Users cannot deselect items in temporary storage.	14	During	Trouble performing step	Physical attributes	Not manipulable
36	3	23	dynamic	Users cannot deselect items in temporary storage.	1	During	Trouble performing step	Physical attributes	Not manipulable
37	3	24	dynamic	Users cannot remove items from temporary storage.	14	During	Trouble performing step	Physical attributes	Not manipulable
38	3	24	dynamic	Users cannot remove items from temporary storage.	1	During	Trouble performing step	Physical attributes	Not manipulable

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task Component		Object Component	
39	3	25	dynamic	Unexpected deletion of temporary storage and closing of preview when selecting empty space.	2	During	Unexpected task automation	Physical attributes	Difficult to control
40	4	18	dynamic	Dynamic preview occludes content elements.	6	Before	Determining how to do next step	Physical attributes	Placement
41	4	22	dynamic	User has to wait for preview to disappear.	13	After	Outcome did not match goal	Feedback	Misleading
42	2	27	fixed	Delete button suggests a drop zone to delete elements.	1	During	Trouble performing step	Cognitive attributes	Manipulation concept
43	2	28	fixed	Delete icon is too small for a drop zone.	21	Before	Determining how to do next step	Physical attributes	Size
44	3	29	fixed	Preview enters switch mode when aborting delete operation.	10	After	Unexpected task automation	Feedback	Misleading
45	3	30	fixed	Users might suppose limited possibilities when whole page is already configured.	4	Before	Determining how to do next step	Physical attributes	Placement
46	4	25	-	Items in temporary storage disappear when interacting with content list.	2	During	Unexpected task automation	Cognitive attributes	Manipulation concept
47	2	32	-	Drag and drop interaction may be confusing for older users.	3	During	Trouble performing step	Cognitive attributes	Manipulation concept
48	2	33	-	The highlight of containers in the preview when selecting an element is confusing.	5	Before	Determining how to do next step	Cognitive attributes	Visual cues
49	4	26	dynamic	Hand occludes content of the screen.	10	Before	Determining how to do next step	Physical attributes	Placement
50	3	34	-	Highlighting when selecting empty space is irritating.	4	After	Outcome did not match goal	Feedback	Unnecessary
51	4	31	fixed	Available space is not highlighted in preview	6	Before	Not confident about next step	Cognitive attributes	Visual cues
52	4	33	-	Grid highlighting switch for big and small irritates.	6	After	Uncertain of results	Feedback	Misleading

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task Component		Object Component	
53	2	36	-	Mode buttons are unnecessary.	2	During	Trouble performing step	Cognitive attributes	Indirectness
54	3	36	-	Switch mode is unnecessary for experienced users.	4	During	Unexpected task automation	Physical attributes	Irritating
55	3	37	-	When starting to switch an element, the system changes to switch mode.	10	After	Unexpected task automation	Feedback	Unnecessary
56	3	37	-	Selecting single switch mode after short press on content	15	During	Unexpected task automation	Physical attributes	Irritating
57	3	38	-	Missing delete mode button when screen is empty might irritate.	4	Before	Not confident about next step	Physical attributes	Placement
58	4	34	-	Highlight when selecting empty space is irritating.	13	After	Unexpected task automation	Feedback	Unnecessary
59	2	41	-	Switch icon in temporary storage item looks like a plus sign instead of the switch icon.	21	Before	Determining how to do next step	Physical attributes	Size
60	2	41	-	Elements in temporary storage are very small.	21	Before	Determining how to do next step	Physical attributes	Size
61	3	41	-	Items in temporary storage are too small and not recognizable.	21	Before	Determining how to do next step	Physical attributes	Size
62	4	35	-	Selected content element in preview not highlighted during drag in switch mode.	13	During	Unexpected task automation	Physical attributes	Irritating
63	3	42	-	Temporary storage might overflow.	8	During	Trouble performing step	Physical attributes	Irritating
64	3	43	-	It might be easier for user to select content elements from content list than from temporary storage.	10	During	Trouble performing step	Physical attributes	Difficult to control
65	4	39	-	Delete mode not directly available when selecting grayed out item.	10	During	Trouble performing step	Physical attributes	Not manipulable

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task Component		Object Component	
66	2	44	-	Back on steering wheel should reset configuration. Inconsistent usage of button.	10	After	Outcome did not match goal	Unexpected results	
67	4	41	-	The items in temporary storage are not readable.	21	Before	Determining next step	Physical attributes	Size
68	2	45	-	Access to screen via steering wheel is confusing.	9	Before	Determining how to do next step	Physical attributes	Placement
69	2	46	-	Tutorial should run until user interacts with the screen.	6	Before	Determining how to do next step	Cognitive attributes	Visual cues
70	4	43	-	Temporary storage has high potential for irritation.	2	During	Unexpected task automation	Cognitive attributes	Manipulation concept
71	3	48	-	Selecting a grayed out content element should not be possible.	5	During	Unexpected task automation	Physical attributes	Irritating
72	3	48	-	Selecting a grayed out content element should not be possible.	12	During	Unexpected task automation	Physical attributes	Irritating
73	3	49	-	No possibility to reset when discarding all items. Dialog might help to avoid errors.	14	After	Outcome did not match goal	Feedback	Missing
74	3	49	-	No possibility to reset when discarding all items. Dialog might help to avoid errors.	10	After	Outcome did not match goal	Feedback	Missing
75	4	44	-	Back button on steering wheel should reset screen.	13	After	Uncertain of results	Feedback	Missing
76	2	4	-	Button to scroll in lists too far away from the actual content list.	21	Before	Determining how to do next step	Physical attributes	Placement
77	4	47	-	Tutorial animation with possibility to hide for next time.	4	After	Outcome did not match goal	Feedback	Unnecessary
78	4	50	-	The user should get a dialog when switching big content element with two small elements.	14	During	Unexpected task automation	Physical attributes	Irritating

Continued on next page

Table C.5 - *Continued from previous page*

#	Expert	ID	Variant	Description	Guideline	Task Component			Object Component	
79	3	51	-	The application is extremely distracting while driving.	23	During	Trouble step	performing	Cognitive attributes	Indirectness
80	2	40	-	Redundant ways of moving content elements to a container may confuse users.	3	During	Trouble step	performing	Cognitive attributes	Manipulation concept

Note. The issues were numbered according to their occurrence during the reviews. The *ID* column contains a numerical identifier for unique usability problems. The *Guideline* column contains a reference to the violated guideline in Table C.4.