universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

Mental Resources and Context in Mobile Interaction

verfasst von / submitted by
Svenja Schröder, MSc BSc

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of
Doktorin der technischen Wissenschaften (Dr. techn.)

Wien, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on student record sheet:	A-786 880
Dissertationsgebiet lt. Studienblatt / field of study as it appears on student record sheet:	Informatik
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Math. Dr. Peter Reichl

n

Contents

Contents **I**

Figures **IV**

Tables **VIII**

Acronyms **IX**

Acknowledgements **XIII**

Abstract **XV**

Kurzfassung **XVII**

1 Introduction **1**
- 1.1 Research Questions . . . 2
- 1.2 Contributions . . . 4
- 1.3 Published Work . . . 6
- 1.4 Overview . . . 6

2 Related Work **9**
- 2.1 Studying Context in the Field . . . 10
  - 2.1.1 Methods and Best Practices for Field Studies in HCI . . . 12
  - 2.1.2 Context . . . 14
- 2.2 Mental Conditions and Attention . . . 19
  - 2.2.1 Stress . . . 24
- 2.3 Mobile Interaction and Errors . . . 33
- 2.4 Conclusion . . . 36

3 Research Approach 39
3.1 Development of a Field Study Framework . . . 40
3.2 Measuring Context and Internal State . . . 40
3.3 Differentiating Types of Errors in Communication . . . 42
3.4 Model and Research Plan . . . 43

4 The CoConUT Framework 47
4.1 Introduction . . . 48
4.2 Related Software . . . 49
4.2.1 Funf Open Sensing Framework . . . 51
4.2.2 AWARE Framework . . . 52
4.2.3 Science Journal . . . 54
4.3 Requirements . . . 55
4.3.1 Online Survey . . . 55
4.3.2 Expert Interviews . . . 57
4.3.3 List of Requirements . . . 60
4.4 Concept . . . 62
4.5 CoConUT Sensing App . . . 65
4.5.1 Implementation . . . 67
4.5.2 Final Prototype . . . 67
4.5.3 External Sensors . . . 68
4.6 CoCoQuest App . . . 69
4.7 CoCoVis . . . 70
4.8 Other Components . . . 70
4.8.1 CoCoHat . . . 70
4.8.2 CoCoBand . . . 71
4.8.3 CoCoBoard . . . 71
4.8.4 CoCoBot . . . 72
4.9 Conclusion . . . 72

5 Laboratory Studies 79
5.1 Erroneous Mental Models . . . 80
5.1.1 Background . . . 82
5.1.2 Study Setup . . . 84
5.1.3 Technical Setup . . . 86
5.1.4 Pilot Study . . . 86
5.1.5 Results . . . 87
5.1.6 Discussion . . . 93

5.1.7 Conclusion . . . 95
5.2 Biophysical Measurements of Arousal . . . 96
5.2.1 Related Hardware . . . 97
5.2.2 Study Setup . . . 98
5.2.3 Technical Setup . . . 101
5.2.4 Results . . . 102
5.2.5 Discussion . . . 107
5.2.6 Conclusion . . . 108

6 Field Studies **111**
6.1 Exploring the Interplay of Context and Interaction in the Field . 111
6.1.1 Study Setup . . . 112
6.1.2 Results . . . 113
6.1.3 Discussion . . . 116
6.1.4 Conclusion . . . 121
6.2 Errors and Stress during Commute . . . 121
6.2.1 Introduction . . . 122
6.2.2 Related Work . . . 123
6.2.3 Study Setup . . . 123
6.2.4 Results . . . 126
6.2.5 Discussion . . . 132
6.2.6 Conclusion . . . 133

7 Discussion **135**
7.1 Supporting Mobile Field Studies . . . 135
7.2 Habituation of Mobile Device Usage . . . 137
7.3 Understanding Errors in Mobile Interaction . . . 139
7.4 Assessing Biophysical Signals in the Field . . . 141

8 Conclusions and Future Work **143**

A Appendix **149**
A.1 CoConUT Online Survey . . . 149
A.2 CoConUT Expert Interview Guideline . . . 154
A.3 Overview over different Biophysical Signals to Measure . . . 155
A.4 CoConUT Class Diagram . . . 157
A.5 Usability Evaluation of Signal . . . 158

Bibliography **163**

Figures

2.1 Unified model of context in human-mobile computer interaction (CoU-HMCI) by Jumisko-Pyykkö and Vainio [JV10] 18
2.2 Relationship between attention, mental workload and situation awareness according to Vidulich and Tsang [VT12] 21
2.3 The 4-D Multiple Resource Model by Wickens [Wic08] 22
2.4 Workflow of attention by Sternberg and Sternberg [SS12] 23
2.5 "The Salience-Effort-Expectation-Value (SEEV) (Salience-Effort-Expectation-Value) attention model - extended with an effort awareness model" by Ferscha [Fer14] 24
2.6 Simplified model of stress related processes which could be related to health outcomes by Cacioppo et al. [CTB07] 25
2.7 Circumplex model of valence and arousal by Russell [Rus80] . . 26

3.1 The most important parts of the CoConUT framework: sensing app CoConUT, the study guide app CoCoQuest, the biofeedback wearable CoCoBand, the qualitative recording wearable CoCoHat and the visualization dashboard CoCoVis. 41
3.2 Different errors that can occur during chatting 42
3.3 Model and structure of the thesis at hand 44

4.1 The LiLiPUT prototype, which is a "wearable lab environment" for user tests in the form of a hat [Rei+07] 49
4.2 Screenshots of the app ContextPhone 50
4.3 The model of the app MyExperience (left) a screenshot of the app (right) 51
4.4 Input and output capabilities of the Funf framework (Source: Funf website) 52

4.5 Input and output capabilities of the AWARE framework (Source: AWARE website) 53

4.6 Screenshots of the AWARE framework in action. The left screen shows the starting screen with base information regarding the device. The screen in the middle lists all out-of-the-box available plugins. On the right screen, the feed shows the current measurements of activated sensor plugins. 53

4.7 Screenshots of Google's Science Journal. The left screenshot shows an open experiment. Science Journal structures the collected data into separate experiment folders, to which sensor measurements can be added. A list of potential sensors to collect is depicted in the next screenshot. The screenshot on the right shows a sensor measurement, in this case lux by the built-in light sensor. 54

4.8 Assessable context dimensions according to the Unified Model of Context in Human-Mobile Computer Interaction (CoU-HMCI) model by Jumisko-Pyykkö and Vainio [JV10]. The dimensions of the CoU-HMCI model are described on the outer boxes with their sensor and external representations in CoConUT in the inner boxes. Since most dimensions can also be assessed via qualitative means, a longitudinal box cuts the CoConUT representations. . 63

4.9 CoConUT screenshots, explained from the top left to the bottom right: 1) top menu of the app, 2) single user recording sessions within a study folder, 3) list of Bluetooth wearables nearby which can be connected, 4) settings menu where sensors can be activated and deactivated, 5) built-in visualization with first overview over gathered data of one sensor, 6) map visualization of gathered data, in this case data speed in km/h 75

4.10 Recording screen of CoConUT, together with a chest belt and a smartwatch 76

4.11 The CoCoQuest app with a loaded questionnaire. On the right a task description is displayed while on the left a question for rating is shown. 76

4.12 The CoCoQuest app in action during a mobile field study. . . . 77

4.13 The CoCoVis dashboard visualizing the data gathered in a study. 77

4.14 The CoCoHat with its different components: The core is a Raspberry Pi plus two accumulators. A PiCam films the user's interaction on the smartphone (see left). A USB microphone and a USB front camera record the surroundings (see right). 78
4.15 The CoCoBand wearable: optical heart rate sensor and galvanic skin measurement electrodes are applied on the fingers, while microboard and battery are attached to a sweat band [Leb17] . . 78

5.1 Verification of identity keys by scanning the each other's QR codes. On the left: a successful verification. On the right: warning because identity keys did not match. 85
5.2 Message delivery failure (1), notification about Bob's new identity (2) and new identity dialogue (3) 89
5.3 "Verify identity" option in the conversation settings (1 & 2). Key comparison page displaying Bob's key at the top and Alice's resp. the user's key at the bottom (3) 89
5.4 Setup of the experiment. The participant is instructed to sit still in front of the computer. Both hands are laying flat on the desk. On the left hand, the smartwatch (orange circle) and the CoCoBand (red circle) are applied. The chest strap is worn under the clothing (yellow dotted circle). On the screen, the Mental Arithmetic Tasks (MATs) are displayed (green box), and voice input is given over the headset. 100
5.5 Storyboard of the test procedure 100
5.6 The MATs tool poses an arithmetic task, evaluates it and indicates whether the solution is correct or wrong 102
5.7 BPM normalized in relation to relaxation phase 105
5.8 Boxplots of Root Mean Square of Differences Between Successive Heartbeat Intervals (RMSSD) across phases 105
5.9 RMSSD and errors per participant across phases 106
5.10 Relation of RMSSD to the percentage of correct answers for the different phases. The x-axis denotes the RMSSD value: The lower the value, the higher the participant is stressed. The y-axis denotes the percentage of correct answers. Each line represents a different phase of the MA tasks. It can be seen that the percentage of correct answers decreases when stress increases 107

6.1 Violin plot of correlation distribution between different dimensions of context (speed, light, Bluetooth devices, sound level) and interaction 116

6.2 3D plot of speed (x-axis), interaction (y-axis) and nearby Bluetooth devices (z-axis) 117

6.3 Boxplot of different typing behaviors (none, slow and fast) . . . 118

6.4 Mean and SD of touch interaction (Global Positioning System (GPS) accuracy $\leq$ 10m). Labels on the means indicate the number of data points in this range. 118

6.5 Route the participants had to take. It consisted of walking on secure sideways, standing on two stations and waiting for the tramways, and taking two tramways. 125

6.6 The polygons on the map used for sorting the data points into categories. The orange and green polygons are the parts of the routes to be taken with tramways, while the yellow and red forms are the two stations the participants had to wait on. Finally, the purple polygon was the part of the route to be walked by the participants. 128

6.7 Boxplots of sensor measurements across the different categories 129

6.8 RMSSD values and their underlying measurement durations across categories. 130

6.9 Mean and sd of screen touch interactions per second over walking speed for the *Walking* phase. The numbers of touches for each speed are specified in the graph. Only GPS values with a sufficient accuracy of less than 10 meters were taken into account 131

6.10 In this figure the relation between RMSSD and error ratio in % over the whole span of the experiment is shown. 132

A.1 Class diagram of the CoConUT sensing app 157

Tables

2.1 General-purpose contextual parameters 17
2.2 Summarizing all three error types and their distinctions after Reason [Rea90] 37

4.1 List of methods used in field studies 56
4.2 Requirements for the CoConUT framework. 61
4.3 Requirements and the way CoConUT meets them 74

5.1 False verification strategies ($n = 12$) 90
5.2 Assumptions about the attack 91
5.3 Mental model of the app (questions were answered on a Likert scale from 1 to 5 where 1 is completely disagree and 5 is completely agree) 92
5.4 Possible mitigation strategies as expressed by the participants . 93
5.5 Heart rate values across phases 104

6.1 Table with summary about assessed sensor values of the CoConUT app. 114
6.2 Table with summary about assessed sensor values of the CoConUT app. 127
6.3 Mean, sd, median, min and max of error rates in percent for different parts of the route 131

Acronyms

ANS Autonomic Nervous System. 25
AR Augmented Reality. 27, 58
BCG Ballistocardiography. 29
BLE Bluetooth Low Energy. 68, 103, 124
BMDW Bundesministerium für Digitalisierung und Wirtschaftsstandort. XIII
BMVIT Bundesministerium für Verkehr, Innovation und Technologie. XIII
BPM Beats per Minute. 29, 68, 71, 101–104, 107, 127, 129, 130
COMET Competence Centers for Excellent Technologies. XIII
COSY Cooperative Systems Research Group at the University of Vienna. XIII, 85, 86, 97, 124, 125
CoU-HMCI Unified Model of Context in Human-Mobile Computer Interaction. V, 16, 62, 63, 65
CPT Cold Pressor Test. 99
CSV Comma-Separated Values. 59
CVT Cardiac Vagal Tone. 31
ECG Electrocardiogram. 20, 29, 31, 133, 155
EDA Electrodermal Activity. 28
EEG Electroencephalography. 20, 28, 29, 156
EMG Electromyography. 156
ESM Experience Sampling Method. 52, 57, 59
FFG Österreichische Forschungsförderungsgesellschaft. XIII
FFT Fast Fourier Transformation. 30
GDPR General Data Protection Regulation. 125
GEMS Generic Error-Modeling System. 35
GOMS Goals, Operators, Methods, and Selections rules. 36
GPS Global Positioning System. VII, 65, 67, 114, 115, 118, 119, 124, 129, 131
GSR Galvanic Skin Response. 28, 71, 97, 99, 101, 107, 108, 155

HCI Human-Computer Interaction. 3, 19, 26–28, 32, 33, 36, 49, 55, 80, 87, 112, 113, 145, 155, 156
HF High Frequency. 30, 31
HPA Hypothalamic-Pituitary-Adrenal Axis. 24
HR Heart Rate. 29, 32, 66, 101, 127, 128
HRV Heart Rate Variability. 29–33, 41, 66, 68, 102, 103, 107, 109, 127, 128, 142
IAPS International Affective Picture System. 32
ICT Information and Communications Technology. 1, 15
IoT Internet of Things. 49
ISO International Organization for Standardization. 15
JSON JavaScript Object Notation. 65, 67, 69
LF Low Frequency. 30, 31
LF Very Low Frequency. 30
LNRMSSD Natural Logarithm of the Root Mean Square of Differences Between Successive Heartbeat Intervals. 31, 32, 66, 142
MATs Mental Arithmetic Tasks. VI, 43, 79, 99–104, 108
MITM Man-in-the-Middle. 81, 82, 84–86, 88–94, 145
Mobile HCI Mobile Human-Computer Interaction. 11, 12, 19, 42, 59
NASA-TLX NASA Task Load Index. 58, 64
OTR Off-the-Record Messaging. 82
PGP Pretty Good Privacy. 2, 81, 82
pNN50 Percentage of Successive Normalized Sinus (N-N) Intervals that Differ from Each Other by More Than 50 Milliseconds. 30, 31, 66
PNS Parasympathetic Nervous System. 26, 30
PSS-10 Perceived Stress Scale with 10 items. 64
PTSD Post-Traumatic Stress Disorder. 32
QoE Quality of Experience. 16, 27, 29
RCF Resource Competition Framework. 9
RF Respiratory Frequency. 33
RMSSD Root Mean Square of Differences Between Successive Heartbeat Intervals. VI, VII, 30–32, 66, 103–107, 130–133, 138, 142
SAM Self-Assessment Manikin. 32, 58
SDK Software Development Kit. 67, 98
SDNN Standard Deviation of the Inter-Beat Interval of Normal Sinus Beats. 30, 66
SEEV Salience-Effort-Expectation-Value. IV, 23, 24
SFG Steirische Wirtschaftsförderung. XIII
SMS Short Message Service. 83, 115

SNS Sympathetic Nervous System. 26
SSL Secure Sockets Layer. 86
TSST Trier Social Stress Test. 98
TÜV Technischer Überwachungsverein. 141
ULF Ultra Low Frequency. 30
UX User Experience. 13, 14, 16, 112
VR Virtual Reality. 27, 29
VRVis VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH. XIII
WLAN Wireless LAN. 86
XMPP Extensible Messaging and Presence Protocol. 112

Acknowledgements

First of all, I would like to thank my supervisor Peter Reichl for all the feedback, encouragement and endless proofreading. Next, I would like to thank Sebastian Möller for his support, the critical discussions and the commitment despite the upcoming sabbatical. Furthermore, Michael Sedlmair deserves my particular gratitude for patiently listening to all my concerns and being a great friend. Thank you also to Torsten Möller, Renate Motschnig, Simone Kriglstein and Manfred Bijak for the advice and being there when I had questions. Thank you to Albert Rafetseder and Valentin for proofreading and helping me shape the thesis in its final stages. Everybody at the Cooperative Systems Research Group at the University of Vienna (COSY) deserves a big thank you for the support and being patient even when I was stressed out the most. Last but not least, I would like to thank my family and my friends for their endless support during the last years, and everybody else who I might have forgotten.

Part of this research has been conducted in the framework of the cooperation between the University of Vienna and the VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH (VRVis). VRVis is funded by Bundesministerium für Verkehr, Innovation und Technologie (BMVIT), Bundesministerium für Digitalisierung und Wirtschaftsstandort (BMDW), Styria, Steirische Wirtschaftsförderung (SFG) and Vienna Business Agency in the scope of Competence Centers for Excellent Technologies (COMET) (854174) which is managed by the Österreichische Forschungsförderungsgesellschaft (FFG). The support resulting from this cooperation is gratefully acknowledged.

This work is furthermore supported with a netidee scholarship by the Internet Foundation Austria.

Abstract

Mobile interaction describes the interaction of human users with mobile devices outside of stationary settings like a desktop workstation. Since the introduction of modern smartphones a decade ago, the interaction with mobile devices has changed: users more and more communicate while being on the go. Hence, *revisiting the fundamentals of technology-supported communication becomes necessary*. When interacting with a device in the field, errors increasingly happen, since the user has to monitor their contextual surroundings, which binds mental resources. This thesis explores the interplay between the user's mental state (which results from limited mental resources), the contexts in the field, and mobile interaction. A particular focus is put on the occurrence of errors, which are examined on three levels: knowledge-based mistakes, rule-based mistakes, and skill-based slips [Rea90]. These errors can disrupt mobile communication.

To explore the interplay as mentioned above, and the occurrence of errors, the prevalent method of this thesis are field studies. Since existing frameworks do not suffice to support mobile field studies as required by this thesis, the Open Source CoConUT field study framework is developed. Overall, in this thesis, two field studies and two laboratory studies are conducted. The laboratory studies address aspects that can not be tested in the field.

Knowledge-based mistakes during mobile communication are assessed in a laboratory study since an extensive qualitative analysis is required. The occurrence of rule-based mistakes under stress is also tested in the laboratory since the probands had to remain seated. Finally, two consecutive field studies investigate the interplay of context, internal state, and interaction in commute-like situations. A second study places its focus on skill-based slips.

Findings reveal that at present users do not stop for typing on their smartphones any more during walking outdoors, despite physical activity and increased stress leading to a higher typing error slips. Stress also has a negative influence on the occurrence of rule-based errors. Additionally, erroneous mental models can lead

to disrupted mobile communication due to incomplete or false knowledge. Overall, the CoConUT framework proves to be reliable to support mobile field studies, and an overview of the gathered sensor data is given. Consumer wearables like chest belts are considered to be a robust and affordable solution for measuring the user's arousal as an indicator of their internal state.

To conclude, the thesis provides *essential contributions to understand mobile interaction in the field after the widespread adoption of mobile devices and puts a particular focus on the occurrence of different types of errors.*

Kurzfassung

Mobile Interaktion beschreibt die Interaktion zwischen menschlichen Nutzer_innen und mobilen Geräten außerhalb eines stationären Szenarios wie beispielsweise eines klassischen Computerarbeitsplatzes. Seit der Einführung moderner Smartphones vor rund zehn Jahren hat sich die Interaktion mit mobilen Geräten maßgeblich geändert: Nutzer_innen kommunizieren mehr und mehr unterwegs. Folglich ist es *notwendig die Grundlagen technologisch gestützter Kommunikation neu zu beleuchten.* Während der Interaktion im Feld geschehen vermehrt Fehler, da der_die Nutzer_in die Umgebung beobachten muss, was mentale Ressourcen erfordert. Diese Dissertation untersucht das Zusammenspiel von mentalem Status des_der Nutzer_in (welcher sich aus den limitierten mentalen Ressourcen ergibt), Kontext im Feld und mobiler Interaktion. Ein besonderer Fokus wird auf das Auftreten von Fehlern gelegt, welche auf drei Ebenen untersucht werden: wissensbasierte Fehler, regelbasierte Fehler und handwerkliche Flüchtigkeitsfehlern [Rea90]. All diese Fehlerarten können hierbei mobile Kommunikation stören.

Um dieses Zusammenspiel und das Auftreten von Fehlern zu erforschen, greift diese Dissertation auf Feldstudien als vorherrschende Methode zurück. Da existierende Software-Frameworks zur Durchführung der für diese Arbeit erforderlichen Feldstudien nicht ausreichen, wird das Open Source Feldstudien-Framework CoConUT entwickelt. Insgesamt werden im Rahmen dieser Arbeit zwei Feldstudien und zwei Laborstudien durchgeführt. Die Laborstudien adressieren dabei Aspekte, die nicht im Feld realisierbar sind.

Wissensbasierte Fehler, die während mobiler Kommunikation auftreten können, werden in einer Laborstudie untersucht, da dies eine umfangreiche qualitative Analyse erfordert. Das Auftreten von regelbasierten Fehlern unter Stress wird ebenfalls im Labor getestet, da die Proband_innen sich in einer sitzenden Position befinden müssen. Anschließend werden zwei Feldstudien durchgeführt, welche

das Zusammenspiel von Kontext, internem Status und Interaktion in nahtransportähnlichen Situationen untersucht.

Die *Ergebnisse* zeigen, dass Nutzer_innen unterwegs nicht mehr anhalten um auf ihren Smartphones zu tippen, obwohl sowohl körperliche Aktivität als auch ein erhöhtes Stresslevel zu mehr Tippfehlern führen. Des weiteren hat Stress einen negativen Einfluss auf das Auftreten von regelbasierten Fehlern. Zusätzlich können fehlerhafte mentale Modelle aufgrund von unvollständigem oder falschem Wissen zu einer Störung von mobiler Kommunikation führen. Ebenso erweist sich das CoConUT-Framework als zuverlässig für die Unterstützung mobiler Feldstudien. Auch wird ein Überblick über die gesammelten Sensordaten gegeben. Handelsübliche Wearables für Endnutzer_innen wie Brustgurte werden als robuste und leistbare Lösung zum Messen des Arousal-Levels der Nutzer_innen während Feldstudien eingeführt. Das Arousal-Level dient hierbei als Indikator für den internen Status.

Zusammenfassend lässt sich sagen, dass die vorliegende Dissertation *wesentliche Beiträge für das Verständnis mobiler Kommunikation im Feld nach der weitreichenden Einführung mobiler Geräte leistet, während sie gleichzeitig einen speziellen Fokus auf das Auftreten verschiedener Fehlerarten legt.*

1

Introduction

The past 20 years have seen a shift from mainly stationary to increasingly mobile communication in Information and Communications Technology (ICT). For instance, electronic communication does not only happen on desktop workstations anymore but increasingly on mobile devices like modern smartphones. Devices have become smaller and smaller while at the same time becoming more and more connected. Mark Weiser's prediction of a **ubiquitous** future has long become a reality, and ubiquitous technology integrates into our everyday life [Wei91]. Even more, in the wake of the Internet of Things, everyday devices become "smart" and interconnected, as well as aim at creating additional values [AIM10].

With this new ecosystem of mobile devices, the **way we interact with mobile devices has changed** in the past ten years. As an example, new possibilities and the following habituation fostered mobile communication to the degree that even meaningful conversations increasingly happen while being on the move. When communicating in a mobile setting, users have to allocate some of their mental resources to monitor their surroundings, especially when moving. Crossing a busy street requires attention, and solely looking at the smartphone screen could quickly become a life-threatening idea. The fact that mental resources are limited forces the user to multitask. **Together with contextual characteristics (noise, many people nearby, et cetera), this can influence the user's internal state and, for example, induce a stress reaction.**

Of course, during mobile communication also **errors can happen**. These errors can be simple typing errors when missing buttons on the software keyboard, applying the wrong routines when small-scale decisions are required, or

even wrong problem-solving strategies due to erroneous knowledge. Thus, from the most straightforward typing error to a profound misunderstanding of the underlying mechanisms, communication can be interrupted by errors in multiple ways. This fact is especially striking when it happens in the field, in situations which seem casual at first (for example during our daily commute), but can have a significant impact on our future communication habits. Especially with **secure instant messengers** like Signal[1], mobile communication has outrun email based communication with regards to security, for example Pretty Good Privacy (PGP)[2]. When usage errors happen during communication over dedicatedly secure messengers, users potentially compromise their communication, no matter how secure the technology design.

The goal of this thesis is to gain a deeper understanding of the interplay of context, mental resources, and mobile interaction in the field. A particular focus is put on the occurrence of several types of errors that potentially disrupt communication in mobile environments. Furthermore, the thesis will examine the interplay of contextual influences, the users' internal state, and their interaction with the mobile device during mobile interaction in the field. In order to examine these interrelations, the mobile field study toolkit CoCoNUT will be developed and applied.

1.1 Research Questions

To systematically investigate the occurrence of errors and the interplay of context, mental resources and mobile interaction in the field, three major research questions with respective subquestions have been formulated, ranging from the methodology, for example the necessary means for assessment in the field, over the influence of contextual factors and the user's state on mobile interaction to the investigation of different types of errors.

1 https://signal.org, last opened on February 9th, 2019
2 https://www.openpgp.org, last opened on February 9th, 2019

Research Question 1: Methodology

How can context, the user's internal state (as an indicator for mental factors) and interaction (especially errors) be assessed in field studies?

1. Which kind of data can be assessed and how (quantitative or qualitative, surroundings, or users themselves)?
2. How accurate is the data?
3. How can the assessed data be visualized and analyzed?

In contrary to the laboratory, where everything is laid out to be as controllable as possible, field studies happen in the real world and are often unpredictable and sometimes "messy" as real life [JM+06]. In consequence, not one field study setup resembles the next, and requirements are diverse. While several frameworks for data assessment during mobile field studies exist [SHR16], to the best of our knowledge, none of them meets the requirements for short-term field-studies assessing context, the internal state of the users, as well as interaction with the device at the same time. This is astonishing, given the fact that the diversity of field studies has to be tackled with maximally versatile, customizable, and expandable solutions. Moreover, a particular focus is put on the usage of Open-Source software and hardware. This research question also explores which methods are best to examine the mentioned dimensions, and which operationalizations are advisable.

Research Question 2: Context

Which kind of contextual factors influence mobile interaction in the field and to which degree?

1. Which role does the user's internal state play?
2. In particular, which kind of contexts have an impact on the user's primary Human-Computer Interaction (HCI) task?

The second research question aims at systematically exploring the interplay of this work's three key dimensions: mobile interaction, context, and the user's internal state (as an indicator for mental resources). A critical aspect of this

research question remains to narrow down, define, and operationalize these dimensions for scaling purposes.

This research question starts with the assumption of a limited capacity of mental resources, which have to be split between monitoring the user's surroundings and the mobile task at hand [Oul+05]. Solely focusing on completing the mobile task at hand could potentially harm the user, for example if they do not pay attention to the surrounding traffic. While paying attention to traffic is a highly conscious action, subconscious reactions can also divert attention from the device to the surroundings, or vice versa. These subconscious reactions might happen due to sudden changes in the surroundings, notifications, or other types of disruptions. In this research question, the interplay of context and mobile interaction occurs, when mental resources have to be split between the two and multitasking ensues.

The third research question focuses on the occurrence of errors during mobile interaction. Since *to err is human* [Sen+69], errors may take place during every stage of conducting an action, be it the planning phase or the execution on the motoric level. These errors can potentially disrupt or even compromise mobile communication.

Research Question 3: Errors

When do which kinds of usage errors happen in the field?

1. Which types of errors occur in the field?
2. How can these types of errors be assessed?
3. What is their potential impact on secure communication?

The following section will briefly highlight the contributions made in this work and summarize the findings which directly address the research questions above.

1.2 Contributions

In this thesis, we present the following key contributions:

As a contribution to **Research Question 1**, the field study framework CoCoNUT has been developed in the course of this work. The framework consists of several software tools: First of all, the Android app CoConUT assesses contextual factors over smartphone sensors. The Android app CoCoQuest probes the user

for quantitative as well as qualitative experience feedback during studies. The visualization dashboard CoCoVis visualizes the collected data and allows for exploration. Different smaller additions (CoCoBoard, CoCoBot) fulfill study-specific purposes. Furthermore, hardware projects have been realized: A chest belt connected to the CoConUT app assesses the user's stress level as an indicator for internal state. The Open-Source wearables CoCoHat and CoCoBand aim at further supporting mobile field studies (see section 4.9). The Open-Source code can be found online[3].

As a further contribution to **Research Question 1**, the CoConUT toolkit has been successfully validated. Overviews over the gathered sensor data by two field studies are presented and give an impression of the accuracy of current smartphone sensors (see section 6.1.4 and section 6.2.6).

The last contribution to **Research Question 1** finds that consumer devices are more suitable to assess internal states during mobile field studies than self-built solutions since they prove to be affordable, robust, and reliable (see section 5.2.6).

Regarding **Research Question 2**, the work at hand shows that users do not slow down for typing any more during walking (see section 6.1.4 and 6.2.6). This contribution has been confirmed in two field studies. Users have constant typing speed across all potential walking speeds, which indicates strong habituation of today's smartphone users.

Regarding **Research Question 3**, a laboratory study showed that incomplete mental models could lead to false mitigation strategies and compromised security after attacks (see section 5.1.7). Surprisingly, users have a false sense of security while having very high trust in secure apps. Bad usability of high-risk security features can lead to non-solvable security problems.

A laboratory study suggests that stress plays a role in the occurrence of rule-based errors: It is assumed that the higher the stress, the higher the error ratio (see section 5.2.6). A field study also showed the impact of stress, to be specific that context and the user's stress level influence typing slips (see section 6.2.6). The more stressed a user is, the higher the error rate of typing slips. When the user has to multitask or engages in physical activity due to context in the field,

3 https://github.com/coconut-framework, last visited March 14th 2019

the error rate increases. These findings contribute to **Research Questions 2 and 3**.

1.3 Published Work

Parts of this PhD thesis have been published and presented at international peer-reviewed conferences and workshops:

S. Schröder, M. Huber, D. Wind, and C. Rottermanner.
"When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging".
In: *European Workshop on Usable Security. IEEE.* 2016.

S. Schröder, J. Hirschl, and P. Reichl.
"CoConUT: Context Collection for Non-stationary User Testing".
In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct. MobileHCI '16.* Florence, Italy: ACM, 2016, pp. 924–929.

S. Schröder, J. Hirschl, and P. Reichl.
"Exploring the Interplay of Context and Interaction in the Field".
In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).* IEEE. 2018, pp. 1–6.

S. Schröder, A. Rafetseder, and P. Reichl.
"Errare Mobile Est: Studying the Influence of Mobile Context and Stress on Typing Errors in the Field".
In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE.* 2019.

1.4 Overview

The remainder of this work is structured as follows: chapter 2 outlines relevant related work about the basic concepts and methods underlying this research,

followed by chapter 3 about the research approach. The following chapter 4 describes the CoConUT framework and all of its components. This chapter points out the requirements for such a framework, relevant related apps, and describes the components of the framework. The subsequent two chapters (chapter 5 and chapter 6) present the field and laboratory studies that were done in the course of this project. Following this, chapter 7 offers a discussion of the studies' results in the context of relevant related work. Finally, in chapter 8, a conclusion is given, and future work is laid out.

2

Related Work

In this chapter, related work concerning the key aspects of this thesis will be presented: First, an overview of the concept of context and its relevance for human-computer interaction in the field and especially field studies will be described. Afterward, a brief delve into cognitive psychology is given (mental conditions and attention), leading to a description of stress and how it can be measured over biophysical signals. This chapter concludes with a definition of interaction, followed by an explanation of different kinds of errors that can happen in human-computer interaction, with a particular focus of different kinds on errors.

One key motivation for this thesis is the work by Oulasvirta, Tamminen, Roto, and Kuorelahti [Oul+05], which investigates mobile smartphone usage in the wild while taking into account the users' mental resources as well as contextual influences. In this research project, the authors conduct a field study to show how the different tasks during mobile interaction (for example interaction task or task to monitor one's surroundings) compete for attentional mental resources. This competition eventually leads to a breakdown of fluent interaction. To describe the allocation of those resources, they created the Resource Competition Framework (RCF) to explain how multiple psychosocial tasks compete for mental resources with limited capacity, for example, interaction and mobility tasks in realistic environments. When the mental capacity limit is reached, resource depletion happens, which leads to tasks being slowed down, postponed, put on hold or terminated. Some of their findings include (among others): Attention to the mobile device reaches from 65% of the time in the laboratory to 20% of the time

on a long quiet street. Also, participants often had to slow down or stop walking to resume interaction with the mobile device.

With this in mind, the rest of the literature can be set into perspective.

2.1 Studying Context in the Field

Originally coming from other scientific disciplines like ethnographic research, field research has become an integral part of computer science research [JD06], not only as a result of the embedment of technology in our everyday lives. Field studies allow the experimenter to collect realistic behavior via interviews in real-life environments, observations, and in-situ assessment. Thus, field studies are especially interesting for highly specific use cases like health care (assisted living, hospitals, et cetera) or safety-critical workplaces (firefighters, security monitoring, et cetera). In social sciences, methods of choice encompass participatory observation or field interviews [JD06].

While testing in the lab certainly makes sense for clearly separable, well-defined hypotheses, testing in the field opens up a wide array of possibilities for deeply understanding human behavior in its multi-faceted nature. Very often, user behavior can only truly be understood by observation or evaluation in the field [PRS15]. Thus, external validity is higher in the field.

Concluding, field studies enable the researchers to get a good grasp of how users will ultimately use software systems, or, more generally, a product in their everyday lives [PRS15].

Depending on the setting, field studies can be more or less biased by the study setup. For example, the operator effect or clunky testing equipment worn on the body can create a bias. Biases can be sought to be minimized, but that bears other implications: Testing or observing participants without consent or releasing them into potentially dangerous situations poses questions about ethical research methods. Early field studies in ethnography, for example, sometimes did not ask for consent for observations, which makes those studies highly questionable with regards to ethical standards [Bur05].

Robson [Rob11] presents a framework for structuring field observations in social sciences:

- ***Space:*** *What is the physical space like and how is it laid out?*
- ***Actors:*** *What are the names and relevant details of the people involved?*

- ***Activities:*** *What are the actors doing and why?*
- ***Objects:*** *What physical objects are present, such as furniture?*
- ***Acts:*** *What are specific individual actions?*
- ***Events:*** *Is what you observe part of a special event?*
- ***Time:*** *What is the sequence of events?*
- ***Goals:*** *What are the actors trying to accomplish?*
- ***Feelings:*** *What is the mood of the group and of individuals?*

This framework already reflects a broad variety of contextual aspects, not only on actors and conducted activities. As general as it is, this framework could be modified to be of use to describe field studies in computer science.

There are certain differences in the outcomes of lab and field studies. While in the field, the outcomes possess a high *ecological validity* and a *low level of control*. In the lab, usually it is vice-versa [KS14]. Thus, for reproducible results or controlled experiments, one should test in the lab, while the field provides a highly realistic testing environment with possibilities to gather interesting side results.

Especially when it comes to research in ubiquitous computing going into the field is crucial due to the embedded nature of interaction [LFH17]. Sometimes design explorations or rather qualitative case studies without a focus on evaluation are favored over strictly empirical studies to explore possibilities. Still, field studies remain relatively expensive and challenging on many levels: starting with the planning over the conduction to the evaluation, the effort of conducting a field study is often higher than with laboratory studies, since potential contextual influences have to be taken into account. Prevalent methods encompass qualitative data like interviews, observations, but also quantitative data over sensors or logging.

One way to achieve a good tradeoff between external validity and controllability in the field is to work with a *quasi-experimental study design* since in the field running a real experiment with full control is usually not manageable [FH02]. In a quasi-experimental study, the independent variable is not under full control by the experimenter. Thus outside a laboratory setting the scientific measurement of cause and effect is difficult.

In 2003, Kjeldskov and Graham did a literature review to examine research methods in Mobile Human-Computer Interaction (Mobile HCI) [KG03]. They

found that the majority of methods being applied were tests in laboratory settings. Consequently, they argued in favor of more realistic testing scenarios like field studies or case studies to create more user and use case oriented systems. After the identification of this gap followed a wide-spread debate in the field of Mobile HCI about whether to test in the laboratory or in the field [Kje+04; KS14; Nie+06; Rog+07].

According to Kjeldskov and Skov, it is not essential *if* or *why* researchers should do studies in the laboratory or field, but *when* and *how* they should do it [KS14]. These crucial questions are only answerable from study to study. Kjeldskov, Skov, Als, and Høegh, for example, conducted a usability study of the same system and the same tasks with six persons in the field and six persons in the lab. According to their results, evaluating in both settings can bring the same list of usability issues, and that recreating some contextual features in the lab suffices. Subsequently, they brought up the question whether field studies are *"worth the hassle"* [Kje+04]. Nielsen, Overgaard, Pedersen, Stage, and Stenild also conducted the same usability study in the field and in the lab. One of their findings is that field studies are indeed *"worth the hassle"*: in the field, they identified more usability problems, as well as cognitive load and interaction style issues specific to the field [Nie+06]. Hence, testing in the field can bring benefits compared to the lab. Rogers et al. as well show why in-situ studies are valuable. They conclude: *"Finally, it is impossible, and nor is it desirable, to capture everything when in situ. The key is to use various methods that reveal both hoped for and unexpected effects of the context of use. Identifying user experience and usability goals also provide a good framing reference from which to analyze the details of certain events"* [Rog+07]. Kjeldskov and Skov *"suggest moving beyond usability evaluations, and to engage with field studies that are truly in-the-wild, and longitudinal"* [KS14]. On a concluding note it can be said that while it is possible to simulate certain contextual factors in the lab, the whole range of context(s) of usage is only accessible "in the wild" [KS14].

2.1.1 *Methods and Best Practices for Field Studies in HCI*

Roto, Vätäjä, Jumisko-Pyykkö, and Vänänen-Vainio-Mattila identified 18 best practices for taking context factors into account during user experience field studies [Rot+11]. This guideline aims at supporting researchers in conducting more reliable user experience studies in the field and will be listed and explained in the following. While some of those guidelines seem to be very basic at first,

they nonetheless provide a sound basis for a structured approach to plan, conduct, and evaluate field studies.

Planning phase:
During the planning phase, the most important steps are to get a good feeling for the contexts to assess and balance the data to gather against expected outcomes. Running a pilot test is crucial:

- *P1. Identify and select realistic contexts for the tasks.*
- *P2. Recruit realistic participants for the selected contexts.*
- *P3. Examine selected contexts in advance.*
- *P4. Identify central context characteristics, and plan how to treat them.*
- *P5. Combine several methods and instruments to collect context data.*
- *P6. Consider the cost-benefit ratio of context data richness.*
- *P7. Prepare for unexpected events and changes in context.*
- *P8. Run a pilot test in the context to ensure fluent capturing of context data.* [Rot+11]

Data collection phase:
During the data collection phase, it is essential to assess context, usage data, as well as the participants' subjective data. This necessity goes hand in hand with the three factors that affect User Experience (UX) in real life (properties of the interactive system, the user's current state and context):

- *C1. Minimize the effects of research setup on participants and the context.*
- *C2. Capture the context with multimedia.*
- *C3. Respect social norms when recording context.*
- *C4. Supplement objective context data with subjective data on participants' context perceptions.*
- *C5. Record the context during use, but collect participants' perceptions of it retrospectively to minimize interference.*

- *C6. Support participants on self-reporting context data.* [Rot+11]

One way to collect contextual data in the field is to use the broad variety of available sensors. Many ready-to-use consumer devices exist (inexpensive action cameras, et cetera), but also Open-Source hardware solutions with inexpensive parts (Arduino, Raspberry Pi, et cetera) are available [LFH17]. However, this sensor data comes with specific challenges: the data has to be saved on the go (often on a remote server), while the corpus of data can get enormous. Preprocessing, filtering, identification of relevant features, classification, and either heuristical or machine learning driven evaluation are just some of the challenges. Simulation or mathematical models can also help in gaining insights from these data sources.

Analysis phase:
Finally, in the analysis phase, it is important to carefully consider both UX and contextual data while paying attention to different context categories:

- *A1. Synchronize context data with collected UX data.*
- *A2. Pay attention to the different context categories when identifying context characteristics that affected UX.*
- *A3. Use context categories to understand context effects, surprising results in particular.*
- *A4. Communicate the context insights to designers, not only the UX.* [Rot+11]

These guidelines are especially helpful for researchers who usually do all their testing in the lab.

2.1.2 *Context*

In research, context plays a major role in a variety of disciplines. The first concepts of context stem from marketing research [JV10; Bel75]. The number and quality of contexts, in which mobile interaction takes place, is unforeseeable. However, categorizing and describing context is crucial during the whole software development cycle for mobile systems: during requirement analysis, different contexts need to be taken into account. For development, contextual factors can heavily influence the usage of sensors of a system (for example a smartphone). Last but not least, testing a mobile system should ideally happen in its natural

contexts. Those natural contexts need additionally to be taken into account during the evaluation of a mobile system.

Reichl et al. [Rei+15] define context for mobile ICT research as follows: *"Context refers to anything that can be used to specify or clarify the meaning of an event. In research settings, context is typically used to illustrate something that complicates a seemingly neutral situation, such as a research laboratory with as few disturbing effects as possible. [...] There is no context-free situation".*

There are many different definitions of the term "context", stemming from different disciplines. Dey, Abowd, and Salber refer to context as *"any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects."* [DAS01]. Another definition describes context as circumstances under which a (mobile) activity takes place [Rot06]. According to the International Organization for Standardization (ISO) standard 13407, the context of use is related to task, user characteristics, as well as physical, technical, and social environment [Sta99]. Jumisko-Pyykkö and Vainio portray numerous additional approaches for defining the term "context" [JV10].

Across all disciplines, Belk was one of the first researchers to describe the impact of context (or, *"situational variables"*, as Belk calls them) on consumer behavior in the field of marketing research [Bel75], since a broader view on realistic consumer behavior was needed to account for the influence of the environment on behavior. Belk distinguishes between five main factors to systematically describe a situation: physical surroundings, social surroundings, temporal perspective, task definition, and antecedent states. While the first four factors are relatively easy to grasp without further definition, the last factor, "antecedent states", describes features that characterize the situation. Antecedent states include all states the individual brings to the situation, rather than states of the individual that directly result from the situation.

The impact of the contextual environment on mobile device usage and the influence of mobile device usage on the context are subject to many studies [Hua+12; Min+11]. Those studies show that not only context has an impact on usage behavior. It also affects the perceived quality of experience [LMP+12]. However, mobile device interaction can also lead to neglection of one's surroundings. This behavior can lead to potentially fatal accidents [Smi+13] - even if the user adapts his/her strategies of dealing with contextual influences during mobile phone usage [Tim+17].

A first working model of systematically describing the different dimensions of context comes from Schmidt, Beigl, and Gellersen [SBG99]. They describe six potential influence factors, which are again associated with two dimensions: human factors (information on the user, the user's tasks, and social environment) and physical environment (infrastructure, physical conditions, and location). Another model, directly building on the one by Schmidt, Beigl, and Gellersen, comes from Reichmuth and Möller: Their model directly refers to mobile devices, to which they added tiredness and stress as influences [RM14].

Reichl et al. distinguish between two context dimensions relevant for communication networks and services [Rei+15], which is especially important to research done in the field of Quality of Experience (QoE): *physical environment* (home, office, commuting, other places, indoors, outdoors, etc.) and *social environment* (alone, with an important person, with friends, in a public place, etc.). Both of those dimensions predictably affect behavior and thus have an impact on the scenario mentioned above. From this starting point, they develop a set of *general-purpose contextual parameters* that indirectly describe contextual influence so that they enable inferring to behavior: *opportunity cost*, *interruption cost*, *social attention*, *time pressure*, *disruptions and distractions*, *pressure to be satisfied or dissatisfied*. An overview can be seen in Table 2.1.

Finally, Jumisko-Pyykkö and Vainio present the CoU-HMCI, based on an extensive literature review of 100+ papers [JV10]. Their model consists of five main context components (physical, temporal, task, social and technical/information) and is designed for exploring general forms of contextual influence, thus allowing to compare specific circumstances of, for example, usage scenarios or field studies (see Figure 2.1). This model goes hand in hand with the classification of factors influencing the UX of a system: properties of the interactive system, the user's current state and the context [Rot+11].

The CoU-HMCI includes significant context components (physical, social, task, temporal, technical and informational) and properties (level of magnitude, dynamism, pattern and typical combinations), which enable researchers to describe and define those contextual dimensions in more detail and in a more dynamic way.

Contextual dimensions (after [JV10]):

Physical context Includes spatial location (for example coordinates), functional places and spaces (for example sports gym, pedestrian areas, coworking space), environmental attributes (for example sensed over sen-

Parameter	*Affects...*	*Effect*
Other opportunities	Behavior	Security Task as secondary task
Interruptions	Behavior	Enhances chance for errors
Social attention	Experience	Enhances chance for errors due to lack of focus and attentional resources. Distraction due to potential privacy and security breach (shoulder surfing, et cetera)
Time Pressure	Experience	Enhances chance for errors
Disruptions and distractions	Experience	Enhances chance for errors
Pressure to be satisfied or dissatisfied	Experience	Not applicable. Mostly those tasks are not experienced to be satisfying since they are necessary, hard, and mostly secondary tasks.

Table 2.1: General-purpose contextual parameters

sors), movement and mobility of the user and artifacts (physical objects surrounding the human-mobile interaction.

Temporal context Includes duration of the interaction session or event where the interaction takes place, time of day, week and year, before - during - after of the user's interaction, relative temporal tensions like hurrying or waiting (actions relation to time) and synchronism (for example asynchronous actions like text messaging or synchronous ones like phone calls).

Task context Includes multitasking, interruptions, and task domain (work-related or entertainment-related).

Social context Includes persons present in the situation, interpersonal interaction (for example turn taking and relations to one another) and culture (values, norms, and attitudes of a specific culture, like organizational or work culture).

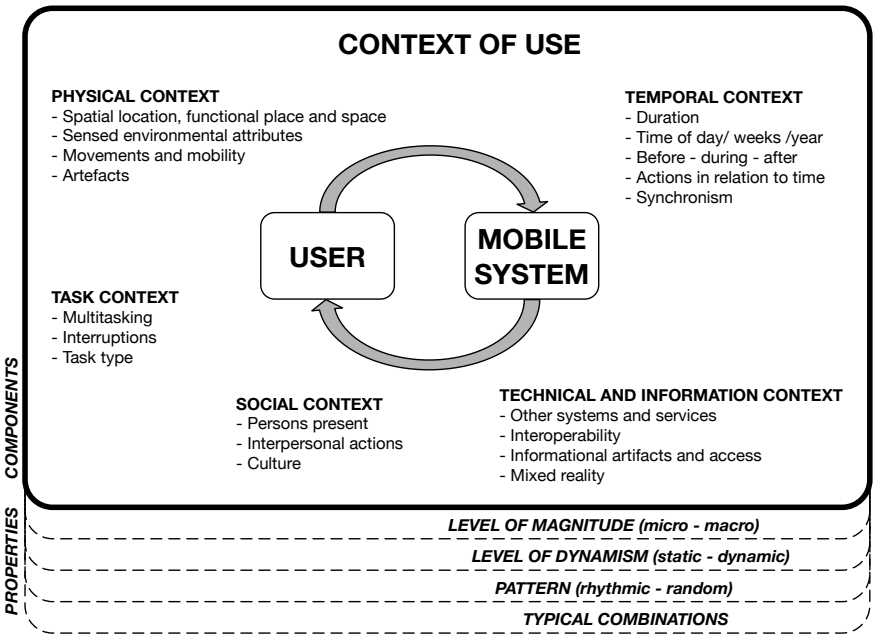


Figure 2.1: Unified model of context in human-mobile computer interaction (CoU-HMCI) by Jumisko-Pyykkö and Vainio [JV10]

Technical and information context Includes other systems and services which are related to the users' system or service (devices, applications, networks, ...), interoperability between and across devices, informational artifacts and access (relevant information provided by other means) and mixed reality systems.

Properties (after [JV10]):

Level of magnitude (micro - macro) The specifics of the environment and contextual dimensions can range from micro (small artifacts near to the user) or macro (the city district in which the interaction is happening). Also, interaction can be described on a micro (single clicks on a touch interface) as well as on a macro scale (interaction types among different groups in the city).

Level of dynamism (static - dynamic) This, for example, relates to the contexts in which a task happens. Some tasks happen in rather static environments (for example always in the evening at home), while others happen in dynamically changing contexts (for example searching for an online route while changing transport means).

Pattern (rhythmic - random) Random patterns are events or interactions that take place only infrequently or once, while rhythmic patterns are events or interactions that repeatedly happen on a predictable basis.

Typical combinations Those are combinations of contextual characteristics or properties that typically occur together (for example group interaction during team meetings, routing during a commute).

As can be seen, context plays a role in different fields of research, including field study research methodology and Mobile HCI research. After having illustrated the most relevant definitions and frameworks for the concept of context in this section, the next section will deal with relevant psychological concepts stemming from cognitive psychology.

2.2 Mental Conditions and Attention

The work at hand takes into account the internal state of the user during mobile interaction. Several facts influence the internal state at this point: When the user is in the field, he/she has to split attention between monitoring of surroundings and interaction task, due to limited mental resources. This way, multitasking ensues and potentially induces stress. This section outlines underlying concepts from cognitive psychology.

As a discipline, psychology has found its way into computer science over the interdisciplinary HCI. In 1983, Norman suggested to establish a discipline called *"cognitive engineering"*, which should provide tools with *"well-established procedures and methods with known benefits and costs, advantages, and disadvantages"*, as well as *"a set of quantitative modeling aids that can be used to assess the performance to be expected from a particular design choice"* [Nor83]. To achieve these goals, he suggested three potential starting points: the underlying psychological mechanisms, the users' mental models, and analyses of the users' performances. In the following years, Norman continued to emphasize the influence psychological factors have on the design, usage, and evaluation of electronic systems [Nor86].

Relevant for this work are concepts from cognitive psychology like mental resources, attention, resource allocation, situational awareness, and memory. This chapter will explain central concepts and focus on the interdependencies of these concepts as well as how they influence each other. A model on the

relationship between attention and mental workload by Vidulich and Tsang can be seen in Figure 2.2 [VT12]. Context heavily influences mental workload, which in turn influences attention and situation awareness. As can be seen, these dependencies are intertwined and cannot easily be separated. Overall, it can be said that workload primarily is a result of limited attentional resources, while perception, memory, and expertise directly influence situation awareness [VT12].

Gopher and Donchin describe workload as a concept *"to account for those aspects of the interaction between a person and a task that cause task demands to exceed the person's capacity to deliver"* [GD86]. The workload of a task is described as having the potential to exceed the user's mental resources. As a consequence, a high workload can lead to the failure of a task to be fulfilled. Vidulich and Tsang define mental workload as *"supply and demand of attentional or processing resources"* [VT12]. The demand can be exogenous (task difficulty, task priority, contextual influences, et cetera) or endogenous (decision making, memory updating, processing, et cetera). The skill level or expertise in the task field moderate those processes. A task is easier for a user if he or she is an expert in the field, and thus, this task demands less mental resources.

Measuring mental workload is of importance to researchers since, via the outcomes, one can make predictions on task performance. For assessing mental workload, several different methodological approaches are possible [VT12]:

- **Subjective experience** can be assessed, either via qualitative interviews or questionnaires, or standardized scales in a quantitative way.
- **Performance** during task execution can be measured, for instance by the time it takes to complete a task, by the quality of the outcome or the errors which are being made during execution.
- **Physiological manifestations** measure the user's internal state over bio-physiological measurements, like an Electroencephalography (EEG) or an Electrocardiogram (ECG).

Load theory explains distractions and the influence of task load [Lav05; Lav10]. Lavie argues that for tasks with high *cognitive load*, we need most of our cognitive processing capacity and tend to be more distracted by irrelevant stimuli. This means we get distracted more quickly because our ability to discriminate between relevant and irrelevant stimuli is diminished. With high *perceptual load*, it is vice-versa: the more a task binds our perceptual capacities, the less quickly we

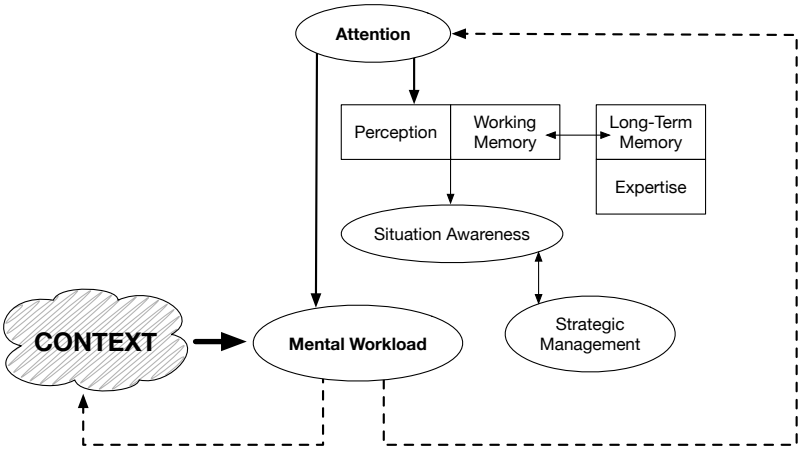


Figure 2.2: Relationship between attention, mental workload and situation awareness according to Vidulich and Tsang [VT12]

get distracted. These effects on attention are presumably independent of each other [EK15].

Most of those theories become interesting when having a look at *multitasking* or divided attention. Interestingly, heavy multitaskers are distributing their attention on several stimuli at once, while low multitaskers have a better top-down attentional control. This fact means that a long history of multitasking is not necessarily beneficial for top-down attentional control. On the other hand, heavy multitaskers were more efficient at task switching than low multitaskers. Furthermore, it seems that we can split our attention during multitasking more effortlessly between tasks in different modalities (auditory, visual, olfactory). The theory of Wickens also supports this fact, for instance by developing the Multiple Resource Model for predicting potential breakdowns of mental resources during divided attention and improving user interfaces based on those predictions [Wic02; WM07]. In this model, multiple dimensions characterize mental demand (see Figure 2.3). For example, focus on the smartphone uses focal vision, which is used in the foveal region for discrimination of small details, and the surroundings are monitored by ambient vision. While no task relies exclusively on one modality of information processing or another, two tasks can still compete for resources from a common resource-defining channel. The 4-D multiple resource model can help in predicting the degradation of performance and help in suggesting improvements in mobile interaction design.

Summarizing, practice is the most crucial factor during multitasking: if a person practices multitasking between different tasks, he/she is becoming in-

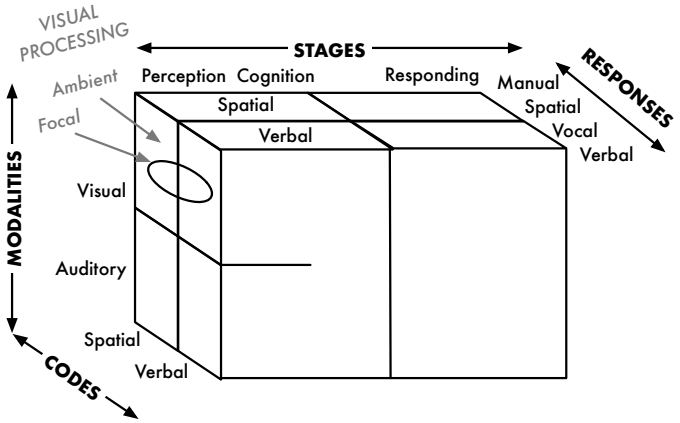


Figure 2.3: The 4-D Multiple Resource Model by Wickens [Wic08]

creasingly better at handling the multitasking [EK15]. Also, *habituation* can play a role [SS12], since habituation lowers the subjective difficulty of the involved tasks and leads to a decreased workload.

Today's media enriched world causes a high workload by sensory flooding, which is why it continuously gets harder to hold the focus of attention for a longer time, especially for the younger generation. According to Ferscha, it *"become[s] difficult for individuals to allocate attention to the right things at the right time"* [Fer14].

Sternberg and Sternberg say that *"attention is the means by which we actively process a limited amount of information from the enormous amount of information available through our senses, our stored memories, and our other cognitive processes."* [SS12] Attention hence allows us to use our limited mental resources economically. If we paid attention to everything which surrounds us, we would remain incapable of action. Additionally, attention steers the memory process, since recalling memories is easier if the circumstances were in the focus of our attention [SS12]. The workflow of the allocation of attention is also depicted in Figure 2.4. Furthermore, Sternberg and Sternberg describe four significant functions of attention: signal detection and vigilance, active search, selective attention, and divided attention [SS12].

One particularly illustrative phenomenon regarding selective attention is the *cocktail party problem*, which describes the ability to attentively listening to one conversation, while there are many other stimuli (other conversations) in the same room [Che53; SS12]. Three factors govern the ability to tend to one conversation despite the distractions: the characteristics of the conversational partner's speech,

sound intensity, and location of the sound origin [SS12]. This prevalent problem from modern psychology describes our ability to willfully allocate our attentional focus despite all distractions and other external influences.

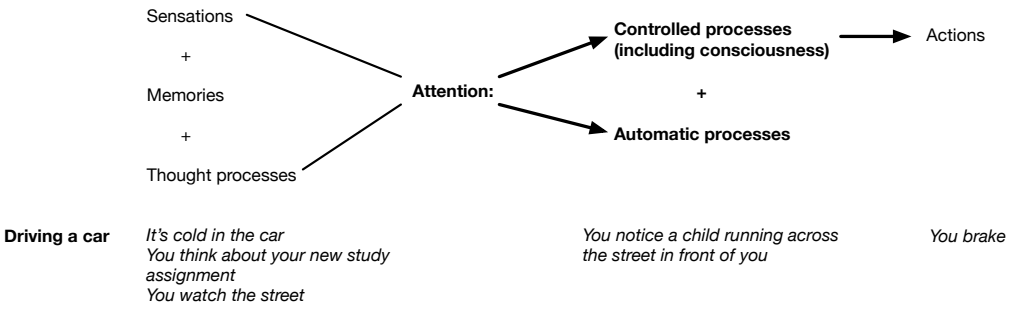


Figure 2.4: Workflow of attention by Sternberg and Sternberg [SS12]

In academic literature, attention is being distinguished in several ways: *active/-conscious attention* is controlled top-down by a person and is led by goals and expectations, whereas *passive/unconscious attention* happens bottom-up and is guided by external stimuli [EK15]. Attention can also be distinguished between *focused* and *divided attention.* While for the former, a person only focuses on one of several stimuli, for the latter attention divides between two or more stimuli at the same time. Finally, there are *internal* (thoughts, memories, cognitive processing, et cetera) and *external stimuli* (sounds, movements, smells, et cetera) [EK15].

Ferscha gives an overview of the concept of attention. Furthermore, he provides a model to illustrate the different factors that influence the allocation of attention, which can be seen in Figure 2.5 [Fer14]. The model is based on Wickens' Multiple Resource Theory [Wic08] and extends the Salience-Effort-Expectation-Value (SEEV) model of attention by an effort-awareness model. In the top left part of the figure, two processes filter external stimuli: the conscious top-down filtering process, which is mainly influenced by endogenous factors from the user (for example willful decision of the user to look at a specific object), and the unconscious bottom-up process, which is influenced by exogenous factors coming from the environment (for example a loud ambulance car passing by on the street). Those stimuli are subsequently assigned mental resources from a limited pool, according to their relevance for the current task. The lower part of Figure 2.5 extends the SEEV model by an effort component, which adds an estimation of expected invested attentional resources and potential attention shifts.

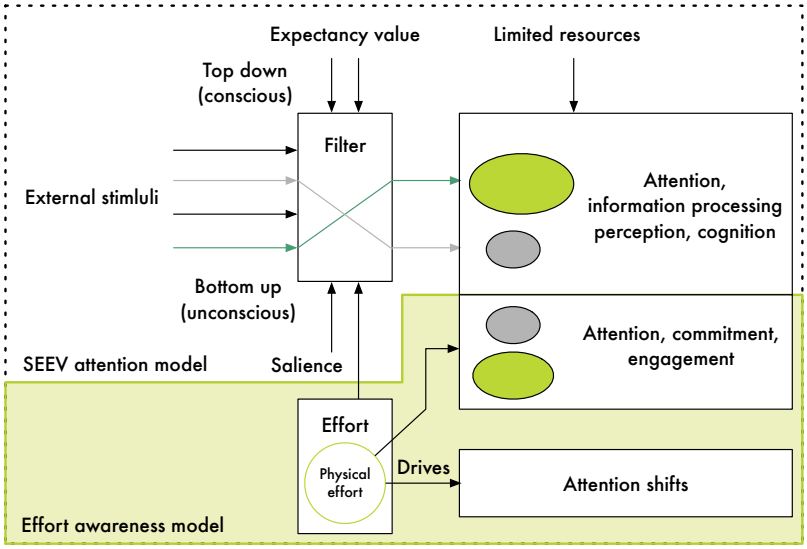


Figure 2.5: "The SEEV (Salience-Effort-Expectation-Value) attention model - extended with an effort awareness model" by Ferscha [Fer14]

2.2.1 *Stress*

The following section gives an overview of the concept of "stress", its research history, and potential influences on the human body and mind.

One of the first researchers to extensively explore the phenomenons we today subsume as "stress" was Hans Selye [CTB07]. In an early article from 1936 he described experiments on rats which were exposed to *"acute non-specific nocuous agents such as exposure to cold, surgical injury, production of spinal shock ([...]), excessive muscular exercise, or intoxications with sub-lethal doses of diverse drugs ([...])"* (stressors) [Sel+36]. He explained that the observed syndrome, which he called "general alarm reaction" had three stages: initial alarm, resistance, and finally exhaustion. The Hypothalamic-Pituitary-Adrenal Axis (HPA) was activated during those reactions. In his later works, he coined the terms "eustress" and "distress" and generally defined stress as the *"non-specific response of the body to any demand"* [Sel76].

Cannon, whose work influenced Selye, brought a psychological perspective to the phenomenon later to be known as stress, stating that also emotional stress can lead to a "fight or flight" response [Can28; CTB07]. Stress, in general, can have a negative outcome on the occurrence of diseases, most widely backed through the link between stress and cardiovascular and infectious diseases, as well as

cancer [CTB07]. Stress can take on chronic and acute forms, where the former has more potential implications on health. Figure 2.6 shows a simplified model of stress-related processes by Cacioppo, Tassinary, and Berntson [CTB07], which could be related to health outcomes.

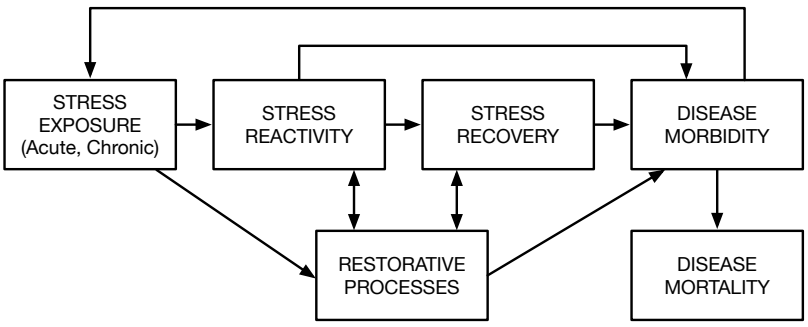


Figure 2.6: Simplified model of stress related processes which could be related to health outcomes by Cacioppo et al. [CTB07]

Bali and Jaggi define stress as a *"state of threatened homeostasis during which a variety of adaptive processes are activated to produce physiological and behavioral changes"* [BJ15]. Due to distress and eustress, which can be roughly described as negative respectively positive stress, merely using the terminus "stress" comes with some drawbacks. What most people colloquially mean by stress in psychology is often referred to as "arousal".

Sternberg and Sternberg define arousal as *"degree of physiological excitation, responsitivity, and readiness for action, relative to a baseline"* [SS12]. Different psychological models exist, for example the circumplex model of valence and arousal by Russell [Rus80] (see Figure 2.7). Low as well as high arousal can therefore be either positively or negatively loaded. Examples would be high positive arousal during a video game (excitement), high negative arousal during a frightening experience (distress), low negative arousal during an unemployment phase (depression) or low positive arousal after a rich and delightful dinner (contentment).

As mentioned before, stress influences our body. The nerve system of the human body can be divided into the central nervous system and the peripheral nervous system. While the central nervous system (or voluntary nervous system) consists of the brain and the spinal cord, the peripheral nervous system denotes all other nerve structures in the body. The somatic nervous system, which controls the sensory and somatosensory system, and the Autonomic Nervous System (ANS), which contains all systems to regulate the body's functions involuntarily,

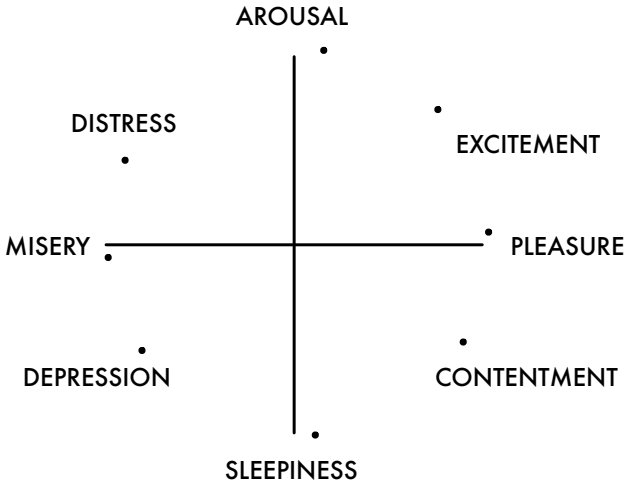


Figure 2.7: Circumplex model of valence and arousal by Russell [Rus80]

are parts of the peripheral nervous system. The autonomic nervous system (formerly called vegetative nervous system) regulates stress and stress reactions in the human body.

The autonomic nervous system contains two antagonists regulating stress reactions (among other functions): the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). Both systems are always active at a certain level to guarantee the body's equilibrium and homeostasis, but according to different stimuli, one or the other can overtake for a while. The sympathetic nervous system responds to external stressors with a "fight or flight" response, while the parasympathetic nervous system is responsible for "rest and digest" phases of the body. Activation of the SNS, the interplay of SNS and PNS and recovery time after high activation of the SNS are indicators for a stress reaction in the body [PRR06].

The concept of stress and stress research have also found their way into current HCI research. Most notably, the field of Affective Computing explores how human emotions can be sensed, modeled, adapted to, expressed and generally taken into account by a technological system during human-computer interaction [Pic99; Pic03]. Many studies are dealing with stress detection for affective computing [GTC16; HKC10; ZB06; SP13]. The section below deals with the assessment of the users' internal states (for instance stress) via bio-signals.

Fairclough and Gilleade define "physiological computing" as *"any technological system where human physiology is directly monitored and transformed into a control*

input. It represents the logical endpoint of convergence between the human nervous system and its silicon-based counterparts." [FG14]. Originally coming from the medical field, those measurements aid behavioral researchers in psychology, HCI, and also QoE [CTB07; Cow+16; Eng+17]. Those systems can deliver representations of emotions, motivations, and cognition of the user. Recent advances in sensor technology render those systems increasingly affordable, mobile, and accurate. This quantified approach can pose a counterpart to qualitative research in HCI, for example in eHealth applications. The recent customer trend of "Quantified Self", or "self-tracking", directly emerged from the increasing number of available solutions on the market [Swa12; Swa13; Lup16].

Cowley et al. provide an overview of current possibilities and methodology to capture bio-physiological measurements for HCI research [Cow+16]: *"The aim is to extract quantitative indices of essentially qualitative cognitive or affective states. [...] there is an issue of establishing ground truth, and the choice of psychological model becomes important"*.

Psychophysiological computing holds many possibilities: Through unobtrusive monitoring of the users' internal states systems can automatically adapt to the users' needs. Also, automatic assessments of emotions or even quality ratings of media can replace extensive and potentially biased questionnaires. Relevant application fields contain workplace related applications (especially high-risk environments), multimedia systems, or eHealth. The most relevant field for this work regarding psychophysiological computing is the field of HCI since some automatic logic in the background can improve the interplay of the system and the human user. In their work, Cowley et al. concentrate on sensors that are *"lightweight, wearable or remotely operable, and application-ready"* [Cow+16]. Concerning signals, they distinguish between *internal*, *external*, and *combined signals*.

Also, biophysiological measurements have arrived in the field of HCI, mainly as a source of insight during testing. With sensors becoming increasingly mobile, researchers can incorporate them into their mobile field studies, or even research in the fields of augmented and virtual reality (Augmented Reality (AR) respectively Virtual Reality (VR)) [CMF18]. In the best case, these measurements are unobtrusive, discrete, and - ideally - objective. However, measuring signals directly from the human body remains challenging, since bodies are not normed, signals are very small, and moving during testing potentially creates artifacts. With increasing miniaturization, biophysical sensors can be utilized during field studies, while so far, the possibilities were limited to a laboratory setting only.

There are different kinds of signals that can be measured on the human body: *bioelectrical signals*, that means signals from muscles and nerves, *electrical conductance*, e.g. Galvanic Skin Response (GSR) on the skin, *bio-acoustic signals*, which are sounds within the body, e.g. heartbeats or air circulation in the lungs, or *bio-optical signals*, which are images or videos taken of the human body, e.g. pupil movements or blood flow beneath the skin. (This list is not exhaustive and focuses on the application in HCI.) [Sch15]. An overview of different biosignals, which can be measured on the human body, can be seen in the appendix (see section A.3).

Measuring electrical signals coming from the body became a medical diagnostic standard in the past 200 years [CTB07]. Electrostimulation, which means stimulating the body through electrical signals to, for example, induce muscle contractions, is frequently used in physiotherapy and rehabilitation [Sch15]. As mentioned before, electrophysiological research has gotten its place in the field of HCI, and interfacing with the human body has become an emerging field of research. *"Understanding how to use electrical connections as new interaction modalities, creating interaction techniques and devices based on biosignals, and using physiological information as a means of evaluation is all clearly within our discipline"* [Sch15].

One particularly interesting strand of research is biofeedback, which aims at bringing unconscious physiological body processes to consciousness, for example by visualizing biophysical signals to the respective user [Yu18]. To accomplish this process, different signals can be used, for example heart rate or breathing patterns [Yu18; Fre+18]. Application areas tackle several psychological phenomena, for instance stress management or anti-anxiety training.

Newer research ideates new input as well as output possibilities with the help of biophysical measurements or signals. One prototype by Lopes and Baudisch, for example, uses induced muscle contractions as output by a video flight game. As a result of muscle stimulation on the lower forearm, participants were forced to tilt the smartphone they played on and had to actively work against this movement by the help of their other muscles [LB13]. This work is prototypical for the newest strands of research which envision new input and output modalities for ubiquitous use cases or gaming, for example.

To evaluate biophysical signals, a supervised classifier can be trained with pre-labeled data from a psychological model (for instance the circumplex model of valence and arousal by Russell [Rus80], see Figure 2.7 from earlier) [Cow+16]. Quite a lot of the user's internal states can be referred through biophysical measurements. For example, Barral, Kosunen, and Jacucci use EEG, Electrodermal

Activity (EDA), and ECG as sources to infer humor appraisal in a realistic environment [BKJ18]. They build predictive models to infer the participants' appraisal of humor successfully.

The next big step for physiological computing will be to focus more on the field. As an example, Barral, Kosunen, and Jacucci also conclude in their study mentioned above that most people view humoristic content on their smartphone while on the go (for instance webcomics, please refer to the study), and that mobile sensors and mobile studies should be conducted to assess more realistic data [BKJ18]. Cassani, Moinnereau, and Falk also describe the use of a novel EEG measurement technique, mounted on a VR glasses set. It is used for non-intrusive QoE assessment [CMF18] on the go, which could be used for building highly immersive VR settings or real-time quality ratings. Hernandez, McDuff, and Picard measure cardiac and respiratory parameters using only accelerometer and gyroscope sensors worn as a wristband [HMP15]. This type of measurement is called *Ballistocardiography (BCG)*. In another study, Frey, Grabli, Slyper, and Cauchard develop a wearable pendant for assessing breathing patterns and sending biofeedback in real-time. Participants could intentionally modify their breathing pattern to match the biofeedback as a technique for understanding the underlying emotion [Fre+18].

As these diverse research projects demonstrate, we can expect a high number of innovative and efficient use cases and systems to measure biophysical signals in the field.

Heart Rate Variability

One particularly informative biosignal to measure is the human heart rate variability:

Heart Rate Variability (HRV) is an umbrella term for several metrics regarding signal from the heart. It indicates the variability between different heartbeats or changes in the rhythm of the beating heart [Cow+16]. While Heart Rate (HR) only indicates the aggregated Beats per Minute (BPM), HRV takes the time intervals between subsequent heartbeats (often denoted in milliseconds) and indicates how much variation there is in-between single heartbeats. These intervals are also called RR intervals. Both the parasympathetic and the sympathetic nervous system influence the heart, and different HRV metrics map different functionalities of the nervous system [Cow+16]. Summed up, HRV can be seen as an easily assessable biometric measure that gives an exceptional all-round indicator of overall health and stress of the body.

Different measures can indicate the HRV of a user. They can be distinguished into time-domain, frequency-domain, and non-linear analysis measures. Depending on the length of the measurement, those measures can be ultra-short term ($<$ 5 minutes), short term ($>$ 5 minutes) and long term ($\backsim$ 24 hours).

In the following some **time domain measures** for HRV are listed [SG17]:

- **RMSSD**: RMSSD stands for *"root mean square of differences between successive R-R intervals"*. It reflects the variance of successive heartbeats and is mostly affected by the parasympathetic nervous system. As has been laid out before, the parasympathetic nervous system (PNS) is responsible for our "rest and digest" mechanisms. The RMSSD can, therefore, be seen as an indicator of the ability of the body to recover. Equation 2.1 defines RMSSD as a measure. It is often used for short-term measurements.
- **SDNN**: The Standard Deviation of the Inter-Beat Interval of Normal Sinus Beats (SDNN) is influenced by both the parasympathetic, as well as the sympathetic, nervous system. It is often used for long term measurements.
- **pNN50**: The *"Percentage of Successive Normalized Sinus (N-N) Intervals that Differ from Each Other by More Than 50 Milliseconds (pNN50)"* is closely correlated with parasympathetic nervous activity [Mie+02]. It is often used for long term measurements.

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N-1}(RR_i - RR_{i+1})^2}{N-1}} \tag{2.1}$$

For **frequency domain measures**, a spectral analysis is performed, usually by applying a Fast Fourier Transformation (FFT). The outcome can be separated into Ultra Low Frequency (ULF), Very Low Frequency (LF), Low Frequency (LF), and High Frequency (HF) power bands. Here, only high and low frequencies are described [SG17]:

- **HF**: This measure denotes the high frequencies (0.15 to 0.4 Hz) in the frequency domain. High frequencies denote parasympathetic activity and heart rate variations due to the respiratory cycle. A low value in the high-frequency range indicates stress or anxiety.
- **LF**: Low frequencies (0.04 Hz to 0.15 Hz) in the power spectrum are mainly influenced by the sympathetic nervous system and the activity of the baroreceptors, which are blood vessel sensors to regulate blood pressure.

- **LF/HF ratio**: The ratio between the frequency bands as mentioned above can help to estimate the ratio between sympathetic and parasympathetic nervous system. It is not an ultimate indicator, however, since the interplay between sympathetic and parasympathetic nervous systems is more complex.

Non-linear measures of HRV also exist but will not be described at this point since they are not relevant for this work.

The measurement length to calculate short-term HRV depends on the nature of each experimental or clinical setup. In 1996, Malik et al. pointed out that while 5 minutes of measurements should be taken to ensure a certain comparability, shorter measurement durations are also possible [Mal+96]. Also, in 1996, The European Society for Cardiology and the North American Society of Pacing and Electrophysiology published a paper with standards surrounding HRV [Cam+96]. They provided reference values, which, unfortunately, not all were based on sufficiently large groups of probands. The task force suggests the usage of RMSSD and pNN50 due to their robustness in mathematical terms. The quasi-standard of 5 minutes as the shortest measurement interval for short-term HRV measurements has been followed mostly until today.

However, HRV measurements shorter than 5 minutes recently have proven to have predictive qualities [HMM04; Bru+99; EF14; MA06; EFN17]. Hamilton, Mckechnie, and Macfarlane [HMM04] argue that HRV over five minutes (short term) or 24 hours (long term) is mostly measured to assess autonomic function and as a reliable predictive indicator in cardiology. Nonetheless, in some settings, even 5 minutes is too long a period. Also, the 5-minute measurements are assumed to be made under stable conditions, which often cannot be guaranteed. They conducted a study in which they took standard ECG measurements of 10 seconds and succeeded in predicting Cardiac Vagal Tone (CVT) [HMM04]. This result indicates that also measurements of 10 seconds can bear predictive qualities.

Plews, Laursen, Stanley, Kilding, and Buchheit monitor training status in endurance sports based on vagal-related indices of HRV [Ple+13]. They prefer the logarithm of the RMSSD, the Natural Logarithm of the Root Mean Square of Differences Between Successive Heartbeat Intervals (LNRMSSD)), as in their eyes it is the most practical HRV measure to apply in this case for several reasons: It proves to be suitable for ambulatory usage, is not influenced by breathing frequency and captures parasympathetic activity even over short time spans, which is useful in cases where probands do not have time for extensive measurements [HMM04] (athletes in their case).

Other studies also have found the LNRMSSD to be the measure of choice. Esco and Flatt, for example, compared 5-minute measurements of HRV with 10-, 30- and 60-second long measurements [EF14]. They conclude that *"the utility of ultra-short-term LNRMSSD measures, especially 60 seconds in duration, within field setting for monitoring athletes at rest and in response to stress appears promising"* [EF14]. McNames and Aboy analyze the accuracy of 11 HRV metrics with differing time recording spans from 10 seconds to 10 minutes compared to 5-minute measurements [MA06]. They conclude that most of those HRV metrics are *"biased estimates"* and that segments of different durations are not comparable. Esco, Flatt, and Nakamura note that LNRMSSD can be accurately measured in a timespan of 60 seconds and a following 60 second stabilization period [EFN17]. However, in a mathematical sense, it still is up to debate how accurate HRV calculations over ultra-short to short timespans are.

Cinaz, Arnrich, Marca, and Tröster demonstrate that RMSSD significantly decreases when mental workload increases [Cin+13; Fal+16]. Thus, RMSSD proves to be a reliable indicator for determining mental workload.

Nunan, Sandercock, and Brodie provide a review of short-term HRV data (gathered from healthy individuals) from publications published since 1996 [NSB10]. They review 44 studies with 21438 participants. A set of agreed normative values for HRV is lacking, which this study seeks to deliver. The cross-study overall mean and standard deviation for RMSSD across 15 studies was 42 ± 15. Umetani, Singer, McCraty, and Atkinson show that HRV and HR decrease during the ageing process, and that HRV is also gender dependent. The authors show that for age 20-29 an RMSSD value of 43 ± 19 and for ages 30-39 a value of 35 ± 11 is the average [Ume+98].

Regarding application areas, first and foremost, HRV is used in medical examination. Of course, the most apparent application area is cardiac diseases. However, also in psychotherapeutic settings, HRV can prove very helpful. Tan, Dao, Farmer, Sutherland, and Gevirtz, for example, examine the usage of HRV with Post-Traumatic Stress Disorder (PTSD) patients and find that they have significantly depressed HRV and that biofeedback could help during treatment [Tan+11].

In HCI research, HRV is often used to measure mental workload or negative and positive valence of an experience [DG11]. For instance, Choi, Kim, Kwon, Kim, Ryu, and Park examine the validity of HRV as a tool to evaluate emotions using the International Affective Picture System (IAPS) [Cho+17]. HRV was evaluated against Self-Assessment Manikin (SAM) ratings. The study suggests that one can assess strong emotions only with HRV. Valderas, Bolea, Laguna, Vallverdú, and Bailón aim at assessing human emotions (relax, fear, and joy)

through HRV and Respiratory Frequency (RF) [Val+15]. Their first results are promising.

Last but not least, in safety-critical working environments, research can be supported by using HRV as a measure. Orsila et al., for example, conducted a pilot study to assess the potential connection of subjectively perceived stress and HRV. With 30 participants and a 1-item scale of perceived stress, they showed a high correlation between the scale and HRV differences between morning and workday measurements [Ors+08]. Despite its severe limitations, the study is interesting in the fact that the increase in HRV-measured stress could be explainable by subjectively perceived stress.

Summarizing, it can be said that mental resources, internal state, attention, stress, and biophysical measurements are broad topics that can only briefly be broached in the scope of this thesis. The next chapter will deal with mobile interaction and particularly different types of errors in human-machine-interaction.

2.3 Mobile Interaction and Errors

Another central concept for this thesis is interaction, especially mobile interaction, and errors during interaction. Oxford University Press defines the term "interaction" as follows [Oxf18]:

> 1 *Reciprocal action or influence.*
>
> 1.1 *Communication or direct involvement with someone or something.*

Interaction is the primary building block of human-computer interaction (HCI) since, without interaction, there would be no connection between humans and computers. As already mentioned, the work at hand mainly focuses on mobile human-computer interaction, which describes the mobile interaction of humans and (mainly mobile) computers or devices. Mobile interaction has quickly risen since modern smartphones were introduced with the first Apple iPhone to the market in 2007 and have reached wide popularity. Due to the vast array of possibilities, mobile interaction can happen in a variety of modalities, situations, and locations [JM+06].

Several large scale studies have explored the way users interact with their mobile devices naturally or communicate on a daily base. They find that social communication apps like instant messengers account for the majority of human-smartphone interaction: Dingler and Pielot investigate how attentive people are

towards their mobile phone messages in a large scale study [DP15]. They show that in 75% of the cases, mobile phone users return to their attentive state within five minutes. They also show that in their sample, WhatsApp is used for 77% of all sent messages. Henze, Rukzio, and Boll developed a typing game in order to investigate users' typing behavior through a large scale study [HRB12]. They made suggestions on how to improve the typing accuracy on modern smartphones by visualizing a small dot on the keyboard, where the user has pressed. In another study, Sahami Shirazi, Henze, Dingler, Pielot, Weber, and Schmidt have a look at 200 million notifications from more than 40000 users [Sah+14]. They find out that 50% of the interaction with notifications happened within the first 30 seconds. WhatsApp dominates the list, which is suggesting that this app is used most widely and most frequently (notifications per day: mean 19.9, SD 64.5). Messaging apps also had the shortest click time. In another large scale study, Böhmer, Hecht, Schöning, Krüger, and Bauer logged detailed usage data of 4100 users with an Android smartphone. [Böh+11]. They show that communication applications are used heavily during the whole day, and are almost always opened as first application after logging into the smartphone. Users also approximately spend about an hour a day using their smartphone.

Of course, errors also happen during mobile interaction. In another large scale study with observing 136 million keystrokes, Dhakal, Feit, Kristensson, and Oulasvirta show that fast typers make fewer mistakes during typing [Dha+18]. This fact goes hand in hand with research showing that expertise can help reduce error rates [VT12]. Practice, in this case, means skills through deliberate practice, but can also mean acquired domain-specific knowledge.

When humans interact with technology, **errors** happen unavoidably. Stress seems to play a role: Ciman, Wac, and Gaggi show that stress affects interaction and that the way that humans interact with their smartphone. Particularly they showed that the number of errors made during writing increases with stress, which they found to be statistically significant [CWG15]. This section will deal with different types of human errors that can happen during the planning and execution of action sequences.

Reason points out that there are two types of views on errors: the system approach and the person approach [Rea00]. While the person approach focuses on description and explanation of errors caused by individuals, the system approach describes how the surroundings and circumstances, under which people work, foster, or prevent errors to happen. He also describes the "the Swiss Cheese Model of System Accidents", which means that several layers of defenses should exist to prevent a single point of failure (for example from the user side).

The fundamental work done by Reason [Rea90] led to the different layers of error which this work systematically assesses. In his *Generic Error-Modeling System (GEMS)*, Reason distinguishes between three main forms of errors: **skill-based slips (and lapses)**, **rule-based mistakes** and **knowledge-based mistakes**: There are "*slips and lapses, in which actions deviate from current intention due to execution failures and/or storage failures*" and "*mistakes, in which the actions may run according to plan, but where the plan is inadequate to achieve its desired outcome*" [Rea90].

Slips and lapses therefore happen during the execution of an action, due to execution or memory storage failures. Mistakes, on the other hand, are based on a flawed plan for action execution, and therefore lead to an unsuccessful action despite correct execution. In a nutshell, mistakes generally happen during intentional, controlled processes, while slips often occur during automatic processes [SS12]. These types of errors are closely based on Rasmussen and Jensen's framework, which describes three levels of cognitive control mechanisms, which are steering human performance: skill, rule, and knowledge [RJ74]:

Skills represent patterns of preprogrammed instructions that are stored in memory. When errors happen at the skill level, they mostly happen due to incorrect force, space, or time coordination.

Rules describe solutions to familiar problems in the form of `if (state) then (diagnosis || remedial action)`. Errors occur most often when wrong rules are applied, or incorrect procedures are recalled.

Knowledge is the highest level, which comes into play when novel situations are met. At this level, Rasmussen and Jensen found eight succeeding stages of decision making for solving a problem: activation (of the whole process), observation, identification, interpretation, evaluation, goal selection, procedure selection and activation (of the action). During the whole process, conscious analytical processes and previously stored knowledge come into action. Errors happen due to resource limitations or incomplete or incorrect knowledge.

Particularly interesting with regards to knowledge-based mistakes is the cognitive concept of mental models: Mental models are an internal representation of concepts and influence cognition, reasoning, and decision-making. Although being incomplete and inaccurate by nature, mental models can provide predictive and explanatory powers for understanding interaction [Joh83; SN93; Jon+11]. Knowledge-based mistakes are often based on erroneous mental models. Especially in the field of software security, this can prove fatal: One possible threat scenario, for example, is for malicious software to take advantage of gaps in the

users' mental models [Was10]. Nevertheless, mental models can help to shed light on users' decisions in novel situations [Bra+10].

Coming back to the categorization of errors, historically seen, an early categorization of errors in HCI discriminates between*mistakes* and *slips* [Nor83; RC02]. While the highest level specification of an action is called an *intention*, a mistake represents an error in the intention itself, while the slip denotes an error in carrying out this intention. Intentions can result from conscious decision making or unconscious processes.

Categorizing errors along the *Goals, Operators, Methods, and Selections rules (GOMS) error modeling system* [Rea90] is prevalent until today [Sch16]. The mechanisms of all levels often go hand in hand and complement each other (see Table 2.2).

Having a look at potential errors during system usage is crucial when wrong actions can lead to security or privacy breaches and are potentially non-recoverable. Think of workplaces with a high risk of potential damage to either the user or potential other stakeholders (airplane piloting, disaster management, health professionals, et cetera). Here, a simple task failure can lead to severe dangers or even death [Cac13].

In terms of usability evaluation, Albert and Tullis [AT13] found three situations in which the assessment of errors makes sense, namely when errors could lead to...

- ...significant loss of efficiency during usage.
- ...especially high costs to the organization or the end user.
- ...task failure.

Usability engineering can reduce error rates, especially for simpler interfaces. Another possible way to reduce error rates is automation, especially for highly complex or safety-critical expert systems [PR97].

2.4 Conclusion

This chapter gave an overview of related work concerning the central topics of this thesis. First, the origins, methodology, and characteristics of field studies were depicted. A particular focus was put on best practices for conducting mobile field studies. Context and its various dimensions were outlined. Furthermore,

	Error Types and Dimensions		
Dimension	***Skill-Based Errors***	***Rule-Based Errors***	***Knowledge-Based Errors***
Type of activity	Routine actions	Problem-solving activities	
Focus of attention	One something other than the task at hand	Directed at problem-related issues	
Control mode	Mainly by automatic processors (schemata, respectively stored rules)		Limited, conscious processes
Predictability of error types	Largely predictable "strong-but-wrong" errors (actions, respectively rules)		Variable
Ratio of error to opportunity for error	Though absolute numbers may be high, these constitute a small proportion of the total number of opportunities for error		Absolute numbers small, but opportunity ratio high
Influence of situational factors	Low to moderate; intrinsic factors (frequency of prior use) likely to exert the dominant influence		Extrinsic factors likely to dominate
Ease of detection	Detection usually fairly rapid and effective	Difficult, and often only achieved through external intervention	
Relationship to change	Knowledge of change not accessed at proper time	When and how anticipated change will occur unknown	Changes not prepared for or anticipated

Table 2.2: Summarizing all three error types and their distinctions after Reason [Rea90]

an outline of fundamental concepts in cognitive psychology led to explanations of psychological concepts relevant to this thesis, like attention, workload, or the influence of different types of stimuli. Subsequently, the concept and physiological fundamentals of stress were explained, followed by a detailed overview of different physiological signals to measure in physiological computing. Heart rate as a physiological signal, as well as heart rate variability and its different measures, are prominently featured. The chapter concludes with a brief introduction of interaction, followed by a delve into the topic of human errors. Thus, all essential fundamentals for researching the role of several types of errors that hinder users from communicating safely in the field are covered.

3

Research Approach

This thesis follows the general methodological approach of user testing to tackle the three research questions raised in section 1.1. Since this work is motivated by investigating behavior in the field, the prevalent method is field study research.

However, not all the research needed for shedding light on the research questions can or need to be realized in field studies. Especially when gathering biophysical measurements on the human body, or conducting extensive qualitative research, the quiet setup, immediacy, and recording possibilities of a laboratory are more appropriate. As we already learned by Kjeldskov and Skov [KS14], the question is not *why* or *if* researchers should do studies in the laboratory or field, but *when* and *how*. Thus, the studies presented in this work have been conducted in the laboratory or the field, depending on the underlying research question.

Another essential aspect for the studies conducted in this work is their relatively short duration from around one hour each. Especially when logging user behavior in the field, many approaches perform long-term studies over days or even weeks. However, long-term assessment is only feasible when battery drainage is taken into account, and the data collection is unintrusive and reliable. Since in this work, a lot of contextual and other sensor data are collected in short measurement intervals, we chose to run *short-term studies*, to gain a detailed insight into human smartphone interaction. To reach a good tradeoff between controllability and external validity, we mainly chose quasi-experimental study designs for our field studies, which has proven to be useful when some control should be retained [FH02].

In the following, methodological approaches and operationalizations concerning central aspects of this thesis will be explained. These explanations will flow

together into a big picture in the form of a model structuring the conducted research.

3.1 Development of a Field Study Framework

Research Question 1 is of methodological nature and directly addresses the need for a framework to support mobile field studies. For the conduction of field studies, appropriate software, as well as hardware, has to be chosen in order to guarantee a successful assessment of the needed data. Frameworks for supporting field studies exist but are not sufficient for the requirements posed by the underlying research questions [SHR16]. Due to this reason, after extensive requirement analysis and evaluation, the CoConUT framework for supporting mobile field studies was developed[1]. The CoConUT framework consists of several Open-Source Android apps for supporting mobile field studies, and several hardware projects complementing these apps. An overview of the most important parts can be seen in Figure 3.1. The framework makes use of multiple smartphone sensors since this is an easily accessible and widespread resource most of us carry with us daily. Also, the framework provides the necessary means for quickly assessing the gathered data and further analysis.

A particular focus is also put on the development of additional hardware modules, which section 4.8 will describe.

3.2 Measuring Context and Internal State

Research Question 2 aims at exploring the interplay between mobile interaction, contextual characteristics, and the user's internal state. Contextual characteristics, as well as the user's internal state, can be assessed both by quantitative as well as qualitative means. While qualitative methods work by direct user feedback in the form of free text questionnaires, interviews, or other forms of feedback, quantitative methods seek to measure, log, or otherwise assess quantifiable data. Since the work at hand aims at assessing as unbiased and as realistic behavior in the field as possible, a focus is put on quantitative measuring.

1 https://coconut.cosy.wien, last visited February 13th, 2019

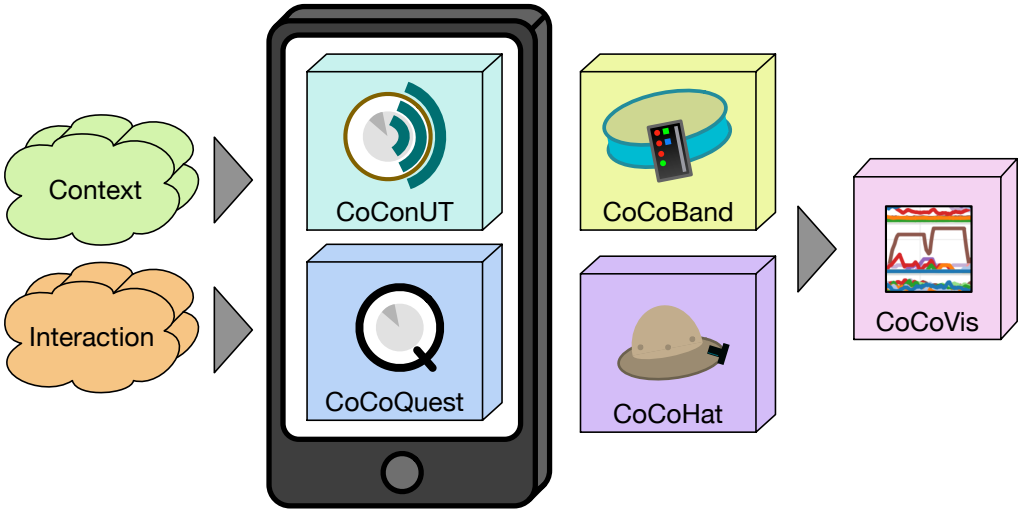


Figure 3.1: The most important parts of the CoConUT framework: sensing app CoConUT, the study guide app CoCoQuest, the biofeedback wearable CoCoBand, the qualitative recording wearable CoCoHat and the visualization dashboard CoCoVis.

Regarding contextual features, the sensor functionalities of modern smartphones are used. Since modern smartphones are relatively well-equipped with all kinds of sensors, nearly everybody today is carrying a highly capable sensing device in their pocket. Users can use their mobile technology in an infinity of potential contexts. To tackle this infinite number of mobile contexts, a pre-given set of contexts is analyzed according to their sensor measurements, not vice-versa.

Measuring mental resources directly is impossible since at most indicators can be gathered. To estimate their availability or distribution, indicators for availability or depletion of mental resources for different tasks can be assessed, like attentional focus, subjective workload, or objective biophysical measurements. This work aims at assessing the user's internal state mainly by biophysical measurements of stress, for example the states valence and arousal. While much research in this area is still to be done (which cannot be covered by this thesis alone), hardware and evaluation approaches of heart rate variability (HRV) are already widely used for stress detection. This thesis takes advantage of this fact and measures the user's internal stress level by heart rate variability.

By these means, the work at hand aims at creating a way of ubiquitously assessing context and internal state without disturbing the user in their natural behavior.

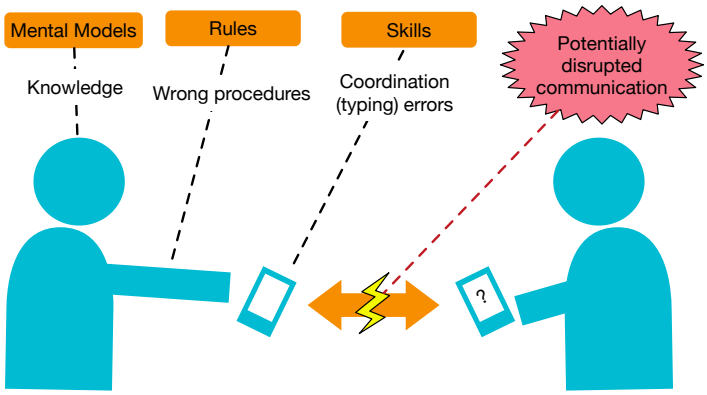


Figure 3.2: Different errors that can occur during chatting

3.3 Differentiating Types of Errors in Communication

One starting point of this thesis was to investigate behavior in the field, or more narrowed down, interaction. However, since interaction still is a broad concept, we chose to examine one particular phenomenon during interaction, namely errors. **Research Question 3** focuses on the occurrence of different types of errors. While errors are potentially fatal for many applications, one major and particularly interesting use case with implications for errors is mobile communication. Thus, this thesis chooses to assess the influence of different kinds of errors on mobile chat communication.

As we already discussed in section 2.3, Reason distinguishes between three fundamental forms of errors: *skill-based slips (and lapses)*, *rule-based mistakes* and *knowledge-based mistakes* [Rea90]. These three major error types can also happen during mobile communication (see also Figure 3.2). Users need fitting and well-built mental models, have to apply the appropriate rule-based actions during Mobile HCI, and need skills for typing on mobile keyboards to not compromise their communication.

However, not all kinds of errors mentioned above need to or can be assessed directly in the field. Especially mistakes based on erroneous knowledge cannot be measured in the classical sense, but need to be assessed through qualitative methods. Qualitative methods on a small scale can be cumbersome to assess in the field. Asking why every single step was taken directly after the action in the field would be too intrusive and bias the outcome. Luckily, mental models

can also be assessed in the lab, since the underlying mechanisms are learned and applied no matter which context. Rule-based decision strategies are based on internalized knowledge and, therefore, can also be measured in the lab. While available mental resources do influence the occurrence of mistakes, the influence is not as high as with slips. Slips are heavily influenced by context and can easily be measured since they are based on dexterity. Subsequently, measuring slips is predestined for field studies.

3.4 Model and Research Plan

Figure 3.3 shows an overview of the research depicted in this work. Based on the considerations outlined in the previous section, we decided on the mixture of laboratory and field tests. For optimally supporting the planned field studies, the CoConUT framework was developed. Requirement analysis, conceptualization, and development of this framework are described in Chapter 4, which tackles Research Question 1. The two subsequent chapters aim at answering Research Questions 2 and 3 and present two laboratory and two field studies. The laboratory studies aim at evaluating aspects that cannot be addressed in the field. In Chapter 5, two laboratory studies about knowledge-based and rule-based mistakes are presented. The first study explains how flawed mental models during secure instant messaging can lead to knowledge-based mistakes. Insights are based on qualitative data. In the second laboratory study, stress and their influence on rule-based mistakes are explored. The participants' heart rate is assessed while they solve mental arithmetic tasks (MATs). Due to the nature of the assessment, the study takes place in the laboratory as well. The field studies are presented in Chapter 6. The first study explores mobile interaction in the field and aims at giving some first meaningful insights into mobile behavior in different contexts. The subsequent study builds upon the first and brings together different contexts, stress measurements, and mobile interaction as skill-based typing slips. The results of this research are discussed in Chapter 7.

As can be further seen in Figure 3.3, the main application case for assessing mobile interaction, respectively mobile communication are instant messaging apps. In the studies presented in this thesis, several instant messengers for Android are drawn on, depending on the use case (Signal[2], Zom[3], and Telegram[4]).

2 https://signal.org, last visited February 13th, 2019
3 https://zom.im, last visited February 13th, 2019
4 https://telegram.org, last visited February 13th, 2019

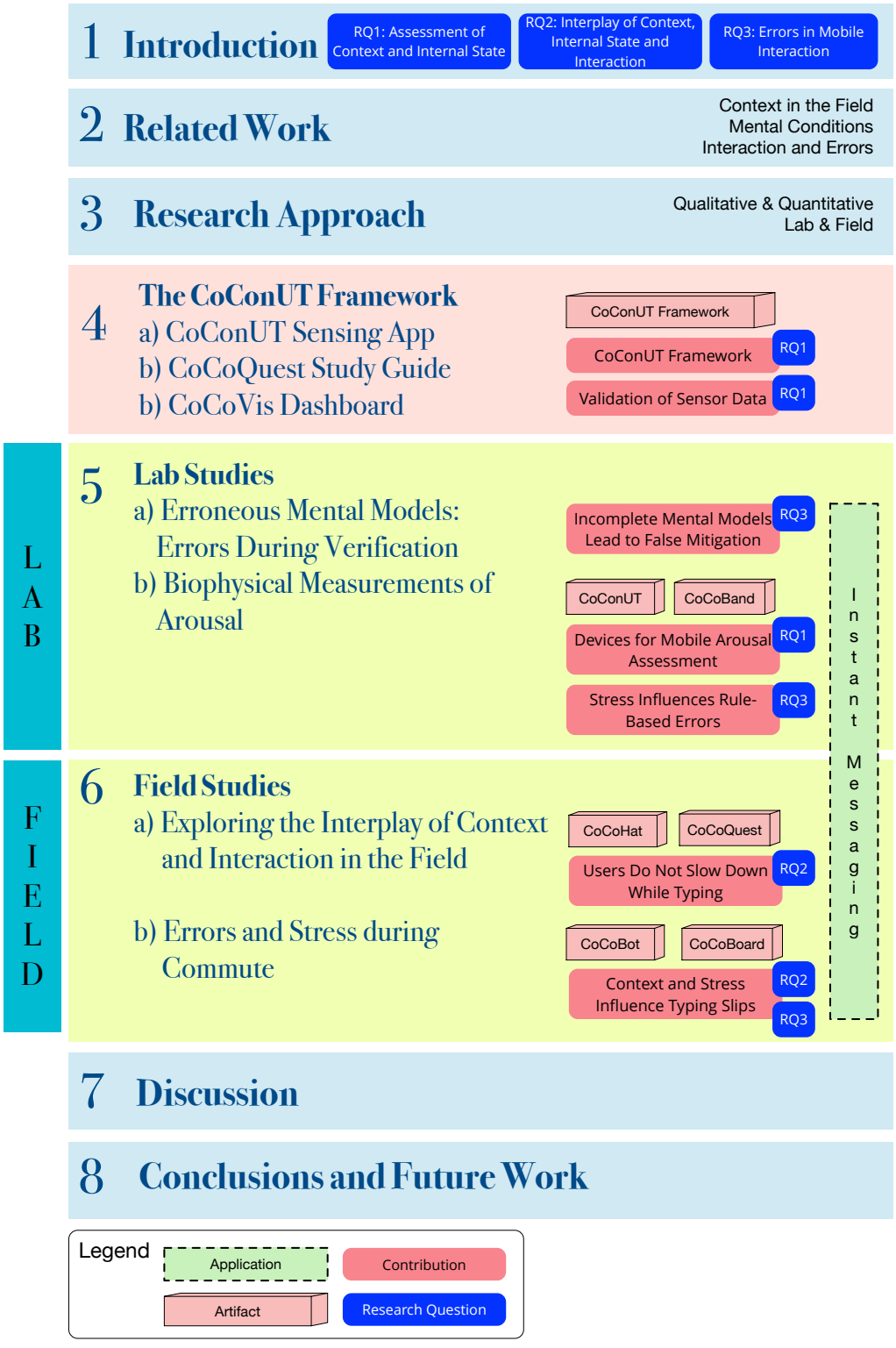


Figure 3.3: Model and structure of the thesis at hand

Please note that machine learning mechanisms to determine different categories in context, mental states, and error occurrence are not in the scope of this thesis. Since such an approach would need a large body of data to train a machine learning system, and the studies in this thesis are mostly exploratory, machine learning remains inapplicable. It remains subject to future work. Additionally, long-term study designs are not part of this thesis, since the planned extensive sensor measurements would render long-term studies unfeasible due to battery drainage.

After having laid out the research model for this thesis, the next chapter will describe the requirement analysis, conceptualization, and development of the CoConUT framework, including all its modules.

4

The CoConUT Framework

The work described in this chapter has partly been published in the following papers [SHR16; SHR18]:

S. Schröder, J. Hirschl, and P. Reichl.
"CoConUT: Context Collection for Non-stationary User Testing".
In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct. MobileHCI '16.* Florence, Italy: ACM, 2016, pp. 924–929.

S. Schröder, J. Hirschl, and P. Reichl.
"Exploring the Interplay of Context and Interaction in the Field".
In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).* IEEE. 2018, pp. 1–6.

Furthermore, Jakob Hirschl, Kaspar Lebloch, Sebastian Dumbs, Jan de Wilde, Christoph Seebacher, Stephan Plank and Christine Julius kindly contributed with their projects to the framework [Hir16; Leb17; Dum17; de 18; See18; Pla18; Jul19].

Within this chapter, we address the following aspects from the research questions: How can context, the user's internal state and interaction (especially errors) be assessed in field studies? How can the assessed data be visualized and analyzed? (RQ1)

In order to do so, the CoConUT framework (*"**Co**ntext **Co**llection for **N**on-stationary **U**ser **T**esting"*) and all of its components are described[1]. Since none of the existing field study frameworks were suitable to match the requirements, the different modules of the CoConUT framework were conceptualized and developed.

4.1 Introduction

There are many apps to gather data during mobile field studies, but since every mobile field study is different, scopes and functionalities of these apps differ widely. As could be seen in section 2.1, field studies vary in length (short-term vs. long-term), their prevalent research methodology (quantitative vs. qualitative vs. mixed) and level of control (observation of naturally occurring behavior vs. field replica in the lab).

Since the existing apps and frameworks cannot fulfill all requirements for the work at hand (see section 4.2), the CoConUT framework to support short-term mobile field studies was built. To foster further collaborations and development, other researchers' requirements were incorporated, and it was decided to build on open hardware and software entirely. In some cases, the usage of solely open hardware did not make sense, so some consumer devices for measurements are incorporated into the framework. The lessons which were learned during the process will be shared later in this thesis. All the apps in this work are developed for Android[2].

Please note that the conceptualization of the CoConUT framework as a whole started after the development of the CoConUT sensing app (which is described in detail in [SHR16]). After the app was created, a requirement analysis was launched to assess the need for a holistic field study framework beyond this app.

The remainder of this chapter is structured as follows: First, related software frameworks will be described. Then, the requirements gathered from an online survey and expert interviews are presented, followed by the concept for the CoConUT framework. Subsequently, all apps and other modules are described, which in sum form the CoConUT framework. This chapter concludes with a conclusion and a list of how CoConUT addresses the posed requirements.

1 https://coconut.cosy.wien, last visited March 18th, 2019
2 https://android.com, last visited March 13th 2019

Figure 4.1: The LiLiPUT prototype, which is a "wearable lab environment" for user tests in the form of a hat [Rei+07]

4.2 Related Software

Context-aware systems and apps are an essential research field in HCI, Internet of Things (IoT), and sensing in general [SAW+94]. With modern smartphones, context-aware systems [HSK09] have gained wide popularity. Since modern apps can access the smartphones' sensors and user information, context awareness surrounds us everywhere and all the time. Hoseini-Tabatabaei, Gluhak, and Tafazolli give a good overview of related context recognition apps and techniques for context classification [HGT13]. The most prevalent issue with context awareness through sensors is the energy consumption of the device. The more sensors are used, the quicker the battery drainage. Additionally, the collected raw sensing data has to be preprocessed, and contextual characteristics have to be inferred. As a consequence, the goal of building context-aware systems should be to maximize the amount of sensed and recognized contextual information through minimum sensor usage. Sensors can be furthermore classified as inertial, positioning, and ambient [HGT13]. Building generalized models for the infinite number of contexts one can be in proves to be impossible, even with sophisticated machine learning methods [Lan+10]. While there are several frameworks for measuring mobile context during field studies, most of them are somewhat outdated and cannot be used on modern systems ([SHR16; Rei+07], see also Figure 4.1). This section will present a selection of context-aware apps and frameworks for supporting field studies and sensor collection and will shed some light on the available apps and frameworks.

Some of the first context-awareness frameworks have been built before the appearance of modern smartphones. With increasing sensing capabilities and ex-

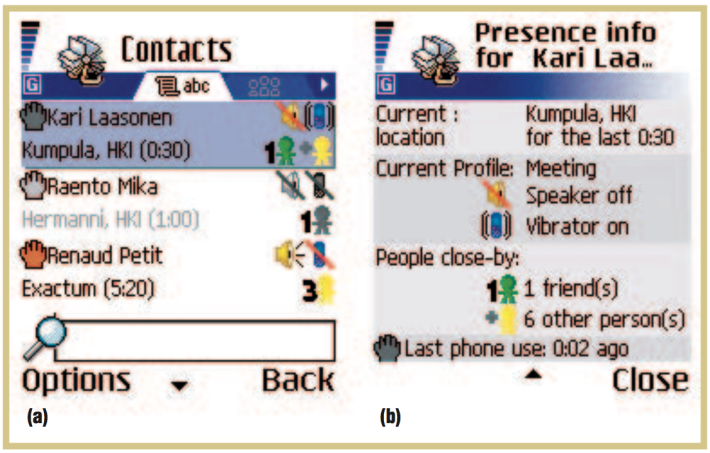


Figure 4.2: Screenshots of the app ContextPhone

tensibility, mobile phones of earlier generations have been capable of supporting context sensing.

For instance, presented by Raento, Oulasvirta, Petit, and Toivonen in 2005, ContextPhone is a prototyping platform for SymbianOS and Nokia Series 60 Smartphone platform that enables developers to build context-aware applications [Rae+05]. With ContextPhone, developers can access context as a resource, incorporate existing applications, and offer fast interaction to their users as well as unobtrusiveness, among other possibilities (see Figure 4.2).

MyExperience, presented by Froehlich, Chen, Consolvo, Harrison, and Landay in 2007, is an application developed for Windows Mobile 2005 [Fro+07]. It can collect subjective and objective in-situ data like user context through sensor reading, subjective user experience through feedback and objective device usage through logging in the background (see also Figure 4.3).

Recent research regarding context awareness puts more focus on battery efficiency, data privacy, and combining several resources to gain additional value:

Xu and Zhu present the privacy-aware sensor management framework SemaDroid [XZ15]. SemaDroid allows users to restrict sensor usage by apps in a fine-grained way and provides mock sensor data to apps if necessary. Van Wissen et al. describe ContextDroid, a framework for context-aware apps on Android [Van+10]. They lay their focus on energy-saving sensor usage as well as usability, efficiency, extensibility, and portability. Another app presented by Rawassizadeh et al. is the long-term lifelogging app UbiqLog [Raw+15]. This app focuses more on the logging of user-centered data regarding the user. Contextual

Figure 4.3: The model of the app MyExperience (left) a screenshot of the app (right)

information and context awareness play a secondary role. Thus, sensor information is recorded, but also application usage, phone calls, text messages, and more.

Regarding mobile field studies, there are several frameworks available for modern smartphones. In the following, three popular frameworks will be presented.

4.2.1 *Funf Open Sensing Framework*

Unfortunately, the Funf Open Sensing Framework is no longer available[3], but it was one of the first frameworks to explicitly support mobile field studies (see Figure 4.4) [Aha+11]. Built for Android, it was bought by Google in 2013 and is still available as Open-Source project, but currently not actively developed. As a framework, it provides a wide range of functionalities to collect, upload, and configure data signals accessible on modern smartphones.

3 http://www.funf.org, last accessed January 25th, 2018

Figure 4.4: Input and output capabilities of the Funf framework (Source: Funf website)

4.2.2 *AWARE Framework*

One of the most popular frameworks for supporting mobile field studies is the AWARE framework[4], which is an Open-Source app for Android and iOS that is "dedicated to instrument, infer, log and share mobile context information, for application developers, researchers and smartphone users"[5]. It offers the possibility to record, save, and process data recorded by smartphone sensors and additional services and plugins to, for example, assess the user's context (also see Figure 4.5). Researchers can run studies online and gather their participants' data over a web service. Experience Sampling Method (ESM) questions can be triggered remotely. Personal information from the participants' phones is not assessed [FKD15]. Figure 4.6 shows several screenshots of the app.

Due to the vast possibilities AWARE offers, it requires many access rights. Since Android's developer guidelines permit some of the required access rights, for Android, the AWARE framework is currently only available directly from the website. As Open-Source solution, developers can directly develop new plugins. Unfortunately, AWARE as a framework is quite heavy-weight, and it does not allow visualization of multiple sensor readings at a time directly on the smartphone.

4 https://awareframework.com, last visited March 14th 2019
5 http://www.awareframework.com/what-is-aware, last visited March 14th 2019

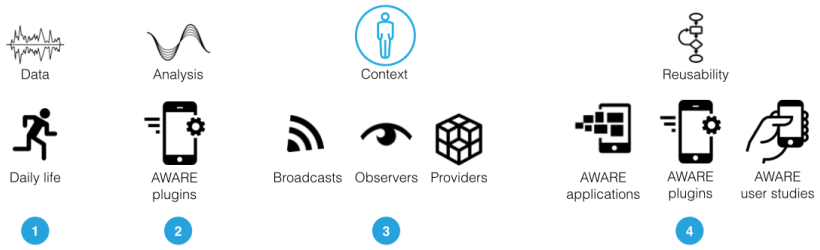


Figure 4.5: Input and output capabilities of the AWARE framework (Source: AWARE website)

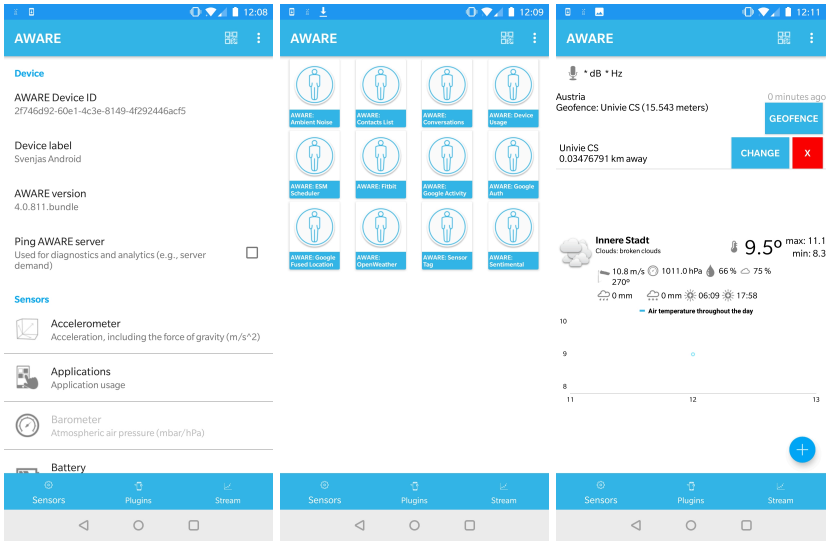


Figure 4.6: Screenshots of the AWARE framework in action. The left screen shows the starting screen with base information regarding the device. The screen in the middle lists all out-of-the-box available plugins. On the right screen, the feed shows the current measurements of activated sensor plugins.

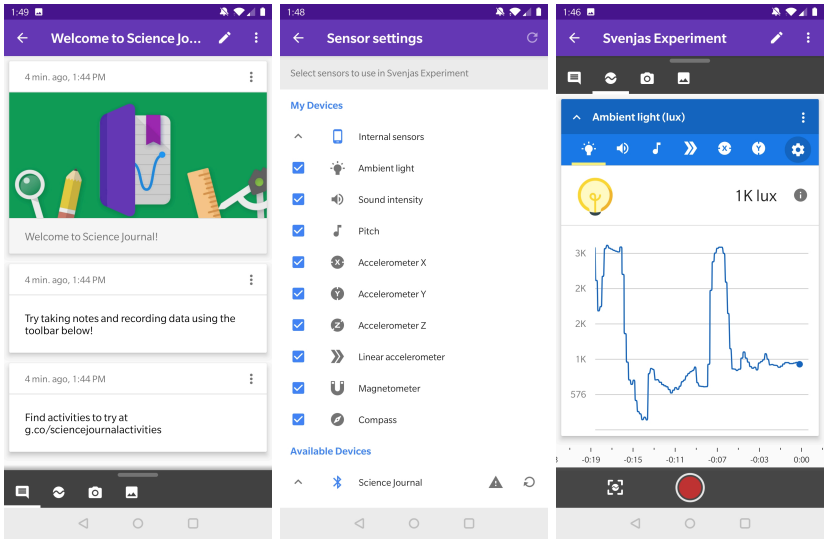


Figure 4.7: Screenshots of Google's Science Journal. The left screenshot shows an open experiment. Science Journal structures the collected data into separate experiment folders, to which sensor measurements can be added. A list of potential sensors to collect is depicted in the next screenshot. The screenshot on the right shows a sensor measurement, in this case lux by the built-in light sensor.

4.2.3 *Science Journal*

Google's Science Journal[6] (which can be seen in Figure 4.7) is an app for Android and iOS, which allows pupils, students, and other researchers to conduct small studies using sensor logging, taking notes, and enriching the study notebook with images and videos. It became Open-Source in 2016[7] and allows connecting external devices via Bluetooth, for example Arduino boards[8]. Functionalities and an overview can be seen in Figure 4.7. However, when this project was started, Science Journal was not Open-Source yet and collected data streams could only be exported in a closed source format.

6 https://sciencejournal.withgoogle.com, last visited March 14th 2019
7 https://github.com/google/science-journal, last visited March 14th 2019
8 https://github.com/google/science-journal-arduino, last visited March 14th 2019

4.3 Requirements

In this section, we describe our requirements for a mobile field study framework.

For the work at hand, several requirements have to be met: A potential framework should be Android-based since Android is the platform of choice for modern smartphone development. Other prerequisites for the development of our field study framework are Open-Source code and low cost of the required components. An aim of this work is giving the framework out to a broad variety of potential field researchers to create additional benefits. Additionally, the framework will aim at supporting short-term field studies, since extensive sensor measurements on the smartphone drain the smartphone battery and are not feasible for long-term studies.

To assess general requirements regarding research surrounding field studies, we additionally did two expert assessments: an online survey, followed by expert interviews. A particular focus was put on evaluation and visualization of results since the analysis of related apps has shown that their focus is on sensing and recording. Quick evaluation and visualization of recorded data are missing so far.

4.3.1 *Online Survey*

To assess requirements by researchers conducting mobile field studies, we carried out an online survey and expert interviews with HCI researchers in 2017. The online questionnaire can be seen in the appendix (see section A.1).

Out of the ten subjects, five were male, four female, and one unspecified. Their ages ranged from 24 to 38, with a mean of 30.8 and a standard deviation of 4.5. Seven of them had a Master's degree, two had a PhD and one a Bachelor's degree. Nine worked in research, and one studied in their Master studies. All of them worked in Computer Science with a focus on HCI.

All of them had at least done a few user studies, most of them conducting them on a regular base. While only one person did not have any experience with field studies, the majority had done both short-term as well as long-term field studies. One person had only done short-term field studies. Their primary motivation was to test their solutions under realistic conditions. A list of methods used in field studies is depicted in Table 4.1.

Regarding the tools used for evaluation, nine persons said they use Microsoft Excel[9], six said they use SPSS[10], four use Python[11], and three use R[12]. Respectively one person stated they use Matlab[13], atlas.ti[14], Java[15], the video annotation software Chronoviz[16] or Apple Numbers[17] (multiple entries were possible). Furthermore, they indicated that they use two tools at least for evaluation. Regarding the statistical methods they apply, predominantly ANOVA (4 mentions) and t-test (4) were mentioned and stated as useful. The variety among single mentions was wide: descriptive statistics, inferential statistics, significance, linear mixed models, regression, qualitative methods ("ethnography or discourse analysis"), classification and "group comparisons".

Method	# (out of 10) of researchers using it
Questionnaires	10
Data collection on the device (logs)	9
Interviews	8
Experience sampling	6
Video and sound recording	5
Screen recording	3
Thinking aloud, focus group	1

Table 4.1: List of methods used in field studies

Regarding features the participants miss in their current evaluation tools, four said they miss a more straightforward way to create diagrams and visualizations of their data. Two stated that they do not miss anything so far. Furthermore, one person mentioned the steep learning curve of existing tools, and another one wished for data filtering and cleaning possibilities.

9 https://products.office.com/en/excel, last visited March 16th 2019
10 https://www.ibm.com/analytics/spss-statistics-software, last visited March 16th 2019
11 https://python.org, last visited March 16th 2019
12 https://www.r-project.org, last visited March 16th 2019
13 https://www.mathworks.com/products/matlab.html, last visited March 16th 2019
14 https://atlasti.com, last visited March 16th 2019
15 https://www.java.com, last visited March 16th 2019
16 http://chronoviz.com, last visited March 16th 2019
17 https://www.apple.com/lae/numbers, last visited March 16th 2019

When being asked which studies they recently have conducted, that are prototypical for their work, answers were quite diverse:

- *"Snoozing notifications"*
- *"Week-long experience sampling study"*
- *"Give different types of devices to people, log what they are doing, use experience sampling"*
- *"Usage of a specific software in mobile setting"*
- *"Text input study on smartphones. Researching if haptic feedback influences the typing performance."*
- *"Placed cameras in homes to record participants watching TV. Also logged devices. The study ran for three nights."*
- *"Workshops with experts and potential users within the field testing and idea gathering; street workshop with passers-by participants."*
- *"*[...] *We install an app on the user's device; this is being used over several days/weeks; depending on the study there are daily, weekly or only one questionnaire at the end or contextual dependent ESM questions over the course of a day* [...] *"*

4.3.2 *Expert Interviews*

Following the online questionnaire, we recruited experts and conducted interviews with them. In the online questionnaire described before, participants could leave their email address optionally to participate in the expert interviews. Further experts were recruited over the researcher's extended network. The interview guideline can be seen in the appendix (see section A.2). In this subsection, the outcome of the interviews will be summarized.

Expert 1

This expert has not done field studies on his/her own, but only studies in the laboratory, comparing touch interactions on smartphones with the users' usability ratings. Also, gyroscope and acceleration sensors are logged. In the future,

expert 1 wants to conduct field studies to assess "how users interact with their phone at the bus station". He/she uses statistical methods to have a look at the data, primarily linear regression, and assesses the users' states and opinions by standardizing questionnaires (AttrakDiff Mini, SEA, SAM, NeoFFI, et cetera). Regarding the preprocessing of data, expert 1 removes outliers, extracts features, and brings the data in a suitable form for further evaluation. For evaluation, he/she primarily uses R and applies machine learning methods, but also tests like ANOVA and t-test. Expert 1 states that "R is great", but he/she would like to be able to "explore the data haptically", for example to have all the data points mapped to a 3-dimensional visualization for exploration purposes. R's ggplot2[18] is excellent for visualization and exploration but remains quite static and cumbersome in terms of quick flexibility. There are many analysis dashboards for mobile apps, like Amplitude[19], Tablytics[20], or Optimizely[21].

Expert 2

The research group of expert 2 conducts few to no field studies and works more in the laboratory. They mostly reduce their research questions so that testing them in the lab becomes possible. If that is not possible, they try to avoid the studies altogether. They claim that stressors and contextual factors in the field would increase the complexity of their already complex studies and reduce the validity. For example, in one study, they work on an AR system, where the smartphone is a "lens" into an AR world. For the rendering of the AR world on the smartphone, they have also to measure the user's head position and not only the back camera, because otherwise, the image would not scale correctly. This fundamental research would be hard to realize in the field. Additionally, to the head tracking measurements, they interview their participants, use the NASA Task Load Index (NASA-TLX) questionnaire to assess workload, log everything they can log and record task completion time and errors. All this data is logged so that it is directly processable, and other data like the questionnaire are transcribed by the researchers ("human preprocessing"). Evaluation happens in R, where they only test whether their hypotheses could be verified, for example utilizing significance tests or distributions. They do not explore their data. Hence, they do

18 https://ggplot2.tidyverse.org, last visited March 16th 2019
19 https://amplitude.zendesk.com, last visited March 16th 2019
20 http://tablyticsmr.com, last visited March 16th 2019
21 https://www.optimizely.com, last visited March 16th 2019

not need exploration tools and so far could find everything in the tools they use, which was necessary for evaluation purposes.

Expert 3

The group of expert 3 conducts mobile field studies in several varieties. For example, they develop Android apps, put them into the Play Store and log the usage data remotely; or they directly pose tasks for participants and assess error rates, or trigger ESM questionnaires on the participants' devices. In mobile field studies, they always log device data (activity recognition, interaction data, sensors), use questionnaires (Likert scales) or qualitative interviews. When field studies are not feasible, they test in the lab. Important are correct timestamps since the reliability of the data can be tricky, mainly when assessed remotely. Using self-developed software can drastically increase reliability, however. For evaluation, they extract the relevant data and preprocess them for evaluation (for example database → Comma-Separated Values (CSV) → Excel / Python). They use Excel for preprocessing and SPSS, Python or R for statistics. Excel, Matlab and R are used for "fancy diagrams". In this workflow, expert 3 criticizes that the switching between different tools is cumbersome.

Expert 4

Expert 4 describes her/himself not as a Mobile HCI researcher and only occasionally does Mobile HCI research. He/she is more interested in qualitative data like interviews, observations, design workshops, or random sampling of participants on the street. This interest is reflected in the data he/she collects: Logging (over Flurry[22]), photos, observation notes, or interviews. Not recorded are metadata from smartphones, demographics, or sensors. Most valuable are qualitative data from interviews, he/she finds. For analysis of the quantitative data, he/she uses R, especially for visual inspection. Outliers are not removed but treated as interesting cases, which should be explored and understood. Preprocessing of raw data could lead to losing interesting data characteristics. Qualitative data is processed analogously by hand, with a printer, paper, markers, and pens. The output of such a process is transcribed summarized data in an Evernote[23] note, or a Grounded

22 https://www.flurry.com, last visited March 16th 2019
23 https://www.evernote.com, last visited March 16th 2019

Theory notebook. Missing in the current evaluation methods is the possibility to save and reuse workflows, for example, in R.

4.3.3 *List of Requirements*

The fact that no study resembles the next is one very central conclusion from the survey as well as the expert interviews. Depending on the research question, available resources, prevalent methods in the respective research field, established research group practices as well as personal taste, field study practices vary heavily. Nonetheless, a list of the most important requirements for a field study toolkit was extracted from the online survey as well as the expert interviews. It has to be noted that a lot of current toolkits are designed for long-term studies, run robustly, and battery efficiently in the background. Of course, in this case, a fine-grained data collection is not possible. Because this thesis focuses on short-term studies, only short-term compatible requirements are taken into account. Extensive qualitative methods that require a face-to-face situation like interviews and the Thinking Aloud protocol are also not incorporated.

- **Predominantly used methods and practices:**
  - Collection of quantitative data directly on the test device via sensors
  - Gathering qualitative data via questionnaires/experience sampling
  - Posing tasks for participants
  - Video/sound recording
  - Screen recording
- **Missing in current solutions:**
  - Easy way to create diagrams and visualizations
  - Interactive exploration of the data
  - Correct and synchronized timestamps
  - All-in-one solution to avoid switching between tools
- **Relevant for the work at hand:**
  - Up-to-date
  - Android-based
  - Support for short-term field studies

	List of Requirements		
Requirement	*Aware*	*Funf*	*Google Journal*
Quantitative data collection	√	√	√
Questionnaire / ES	√	√	×[24]
Conveying tasks	√	√[25]	×
Video recording	×	√	×[26]
Audio recording	×	√	×
Screen recording	×	√	×
Easy visualizations	√	√[27]	√
Interactively exploring data	×	?[28]	×
Correct timestamps	√	?[29]	×
All-in-one solution	√	√	×
Up-to-date	√	×	√
Android-based	√	√	√
Support for short-term field studies	√	√	√

Table 4.2: Requirements for the CoConUT framework.

As can be seen in Table 4.2, none of the apps presented in section 4.2 fulfill all the requirements poses by experts to a field study framework. Due to this shortage, the CoConUT framework was designed and implemented. The concept of the framework will be presented in the next chapter.

24 It is possible to take textual notes
25 Notifications sent by the operator
26 Photos can be taken, but no videos
27 Data can be uploaded to a back-end and visualized there
28 No data since the app and backend cannot be tested
29 No data since the app and backend cannot be tested

4.4 Concept

Based on these requirements and as an answer to Research Question 1, the CoConUT framework was conceptualized. Research Question 1 deals with the way contextual factors, the user's internal state and interaction (especially errors) can be assessed in field studies. Especially the kind of data which can be assessed (quantitative or qualitative, context or users themselves) and its accuracy are of importance. The last subquestion raises the issue of analyzing and visualizing the gathered data.

This research question leads to the following conceptualization steps:

- *Overview of contextual factors to record during mobile field studies*
- *Overview which factors of the user's internal state can be assessed*
- *Approach to record the user's interaction and errors during mobile field studies*
- *Analysis of which quantitative data can be measured*
- *Analysis of which qualitative data can be assessed*

Furthermore, the resulting framework has to prove itself during several field studies, whose *data has to be checked for accuracy* and an appropriate way of *visualizing and evaluating the data* has to be found. The subsequent paragraphs describe how the CoConUT framework was conceptualized following those steps. Afterward, a description of the different modules of the framework ensues.

Regarding **contextual factors** and *quantitative data*, the concept of the CoConUT framework closely follows the unified model of context in human-mobile computer interaction (CoU-HMCI) by Jumisko-Pyykkö and Vainio [JV10], which can be seen in Figure 2.1 (see section 2.1.2). The model has five major dimensions, as well as user, mobile system, and their interaction. Figure 4.8 shows how the CoU-HMCI model dimensions can be mapped to assessable measures.

Since modern smartphones come with a wide variety of built-in sensors, the decision fell to make the best use of this fact. Additionally, external sensors can also be connected to a smartphone via Bluetooth, for example. The main component for recording quantitative data in the field is the CoConUT sensing app, which will be presented in section 4.5, including a detailed list of which sensors are being recorded on the smartphone.

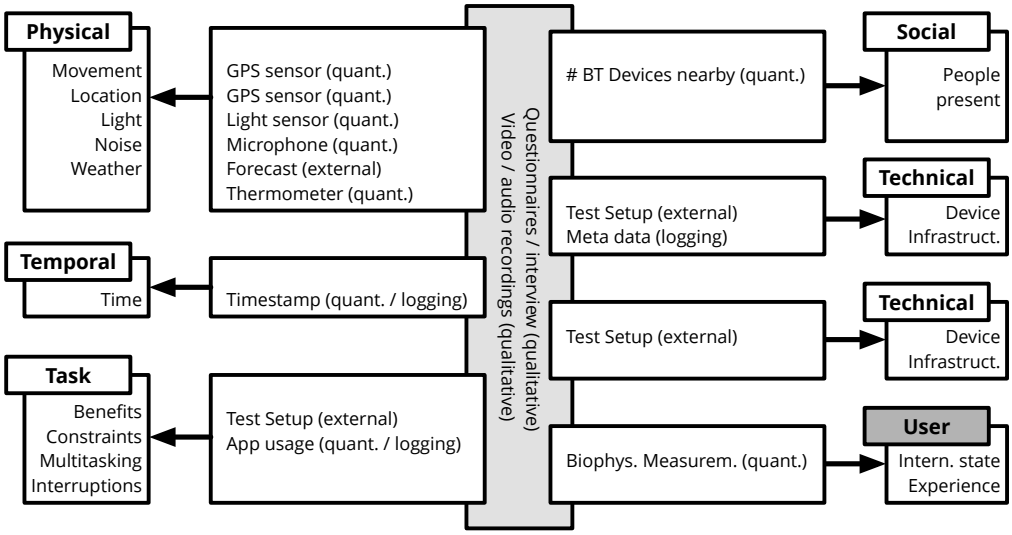


Figure 4.8: Assessable context dimensions according to the CoU-HMCI model by Jumisko-Pyykkö and Vainio [JV10]. The dimensions of the CoU-HMCI model are described on the outer boxes with their sensor and external representations in CoConUT in the inner boxes. Since most dimensions can also be assessed via qualitative means, a longitudinal box cuts the CoConUT representations.

Since the CoConUT app is supposed to run in the background of the study smartphone and only be used by the operator, in a next step we conceptualized the CoCoQuest study guide app, to deliver task descriptions, short questionnaires, and experience probing via media content to the user. Through the CoCoQuest app, quantitative data in the form of questionnaires, for example standardized scales, can be assessed by the user directly in the field. This data can be later matched with the passively collected data from the CoConUT app.

Furthermore, the CoCoHat wearable records the user's surroundings and sound by a camera, and a microphone mounted on the head.

To assess the **user's internal state**, the stress level is measured over the user's heart rate. This measurement can be easily obtained via a wearable that measures the user's heart rate in the field. Since mental resources cannot be measured in the classical sense and measuring attention is cumbersome (for example through eye-tracking or gaze recognition, for which specialized hardware would be necessary), this work relies on the assumption that multitasking in the field leads to higher workload because mental resources have to be split. This splitting goes hand in hand with a higher stress level. When in this state, the assumption is that the user is more likely to produce more errors.

Of course, workload and stress can also be assessed by standardized questionnaire scales like the NASA-TLX [HS88] or the Perceived Stress Scale with 10 items (PSS-10) [CKM83; CKM+94]. However, the disadvantage remains that those scales can only be assessed retrospectively back in the laboratory since filling out questionnaires in the field would disrupt and bias the field experience. Thus, the unobtrusive measurement of heart rate with wearables was chosen as a method of choice.

The CoConUT framework allows several possibilities to capture the user's internal state: First of all, wearables can be connected to the CoConUT sensing app, enabling the assessment of the user's heart rate. The CoCoBand can measure heart rate in beats per minute, on-skin temperature, and galvanic skin conductance. Additionally, a smartwatch and chest belt can be connected to the app and measure heart rate by beats per minute and heart rate variability by successive RR intervals, respectively.

Additionally, the CoCoQuest app delivers short questionnaires to participants in-situ, which enables the capturing of user-input quantitative data in a limited form. For example, the users can answer questions about their internal states on the go, or conclusions over the NASA-TLX questionnaire or other scales can be drawn.

Interaction can also be assessed in either quantitative as well as qualitative way. Video recordings of the interaction with the device as well as screencasts allow operators to observe user interaction in detail and also detect interaction patterns. Qualitative analysis is the most versatile, but also the most time-consuming way to analyze human mobile interaction.

Another way to measure interaction is quantitatively, for example by logging touches on the touch screen, including touch location, duration, intensity, or usage of gestures. Pressing on buttons and the status of the screen (on/off) can deliver information about the interaction with the device.

Also, the CoConUT app offers the possibility to record the user's face over the front camera and simultaneously a screencast of the device. With the help of this data, it can be easily analyzed when the user is interacting with the device. Furthermore, CoConUT counts the touches in the touch screen per measuring interval and logs the screen status. This number of touches gives a quantitative approach to measuring interaction. Furthermore, concerning communication, detailed information on typing behavior can be collected over the CoCoBoard software keyboard. For example, typing slips can easily be measured.

Finally, when it comes to **visualizing and evaluating the data**, the CoCoVis dashboard for visualization comes into play. While a first glimpse at the data can

already be taken right within the CoConUT app, a more elaborate exploration can be gained by using the CoCoVis dashboard. Of course, data collected by all CoConUT components can also be exported in a compatible JavaScript Object Notation (JSON) format and further analyzed in environments like R.

Summarizing, Table 4.3 depicts the way CoConUT addresses the requirements included in Table 4.2.

4.5 CoConUT Sensing App

As described in [SHR16], CoConUT is an Android app for collecting mobile context dimensions, metadata about interaction and the user's stress level during mobile field studies using sensor data on the test smartphone itself. For evaluation and exploration, the recorded user session data can be visually explored on the smartphone. This exploration facilitates an assessment of the user's behavior, route, and internal state.

In its first version, the CoConUT app only supported some environmental sensors [SHR16] and got more features throughout further development cycles, following the list of requirements. Figure 4.8 shows an overview over the captured data, which is according to the CoU-HMCI model presented in Table 2.1.2. The current version of CoConUT (2.1) supports the following sensing and recording functionality:

- **Physical context:** Speed and location are measured over the GPS signal. CoConUT tracks latitude, longitude, and the deduced speed value provided by Google's services. Furthermore, GPS signal accuracy is saved to estimate the quality of the collected values. Lighting conditions are sensed over the light sensor and saved in Lux. In this case, the smartphone is measuring the Lux value of the object the sensor is pointing to. The accuracy of the light sensor is saved as well. The smartphone's microphone can record ambient noise (if the ambient soundscape is not recorded as a sound file). To record ambient noise, the smartphone records small snippets of sound, evaluates their noise level in decibel and discards the sound snippets afterward. An API retrieves the current weather and temperature and conditions are saved.

- **Temporal context:** The CoConUT app records timestamps of the sensed data with an accuracy of milliseconds. Also, the day of the week and the distinction between weekday and weekend was added for convenience.

- **Social context:** The level of crowdedness in the surrounding area is estimated through the number of visible Bluetooth devices nearby. Sensing other devices nearby over Bluetooth has been used in several social intelligence applications [HGT13].

- **Task context:** Interruptions and multitasking can only be roughly estimated by the number of interactions (touches) on the screen. Also, the status of the smartphone screen (on/off) can give indications of usage behavior. (Android's security features make it impossible to log access to other apps from within our app. Otherwise, this could be logged as well.)

- **Technical context:** The battery status of the phone is saved to indicate battery consumptions and preconditions for the test. Another variable indicates whether the user is using headphones or not.

- **User:** The app measured several metrics regarding the user. First and foremost, interaction with the devices through touches on the screen is recorded. The screen status (on/off) is another indicator of interaction. The sensor data is also sent to Google's location services to detect the user's current activity [31]. Finally, to assess the user's internal state, heart rate (HR / HRV) can be measured over either smartwatch or chest belt. The delay data of the connected devices is saved as well to preprocess the resulting data more accurately. From the user's heart rate data, several metrics can be automatically calculated (RMSSD, LNRMSSD, SDNN, pNN50).

Further essential features are regarding qualitative research: CoConUT allows to record the user's face through the front camera and the interaction on the screen over screen capturing. Audio can additionally be recorded over the smartphone's microphone if the ambient noise is not measured (the smartphone's microphone can only be used by one resource at a time). Since those recording capabilities consume a lot of storage space on the device and drain the battery, they are only feasible for short-term recordings. Also, the users have to be notified that they are being recorded, and have to consent.

31 https://developers.google.com/android/reference/com/google/android/gms/location/DetectedActivity, last visited March 18th 2019. Potential activities include the constants: IN_VEHICLE, ON_BICYCLE, ON_FOOT, RUNNING, STILL, TILTING, UNKNOWN, WALKING.

4.5.1 *Implementation*

The first version of the CoConUT app was built in Android Studio 1.4 with a minimal Software Development Kit (SDK) 17 (Android 4.2) and Java Version 1.8.0 25, after a requirement-driven prototyping process. This first version was tested in a preliminary study for technical feasibility [SHR16]. The current version of the CoConUT app is version 2.1 and can be downloaded from Google's Play Store[32]. A class diagram of the most important classes can be seen in the appendix (A.4).

4.5.2 *Final Prototype*

The final prototype of the CoConUT sensing app can be seen in Figure 4.9. For conducting a study, the app has to be installed on the smartphone. When opening first, users have to grant a series of access rights, for example for the data storage, the GPS sensor, the microphone, et cetera. When the rights are not granted, the app will proactively ask for the rights to be granted again. Afterward, users can create a new study session, to which they can add new recording sessions (for example one session per participant).

Over the top menu, the settings can be accessed. Here, sensors and recordings can be enabled or disabled, depending on the study. The settings menu is divided into "Session recording", "Heartrate" and "Advanced". In "Session recording", different sensors and services can be set, as well as the measurement interval. The measurement frequency is per default set to once per second, but depending on the study, the frequency can be adapted. For measuring heart rate either a smartwatch or a chest belt have to be connected over Bluetooth. Devices can be added under a distinct page accessible from the top menu ("Devices"). Once a device is added, several measures can be set under on the "Heartrate" settings page. Under "Advanced", camera, screen, and audio recording can be set, as well as a few other more advanced and experimental settings.

Single recording sessions or whole studies can be exported as JSON. Additionally, for each session, there is a little folder icon in the app, which directly grants access to the smartphone's storage where recorded videos and audio files are stored for this session.

When clicking on a single session, the mobile visualization opens and displays the recorded data in an overview. The mobile visualization is especially helpful to

32 https://play.google.com/store/apps/details?id=at.ac.univie.cosy.coconut, last visited March 17th 2019

gain a first overview of the gathered data, for example to quickly check whether the recording was successful or to ask the participant specific questions about certain parts of a route. Of course, for further analysis, the data has to be exported to a more capable evaluation application, but this also requires more time and more effort. The mobile visualization offers the following screens:

- **Normalized overview of all sensors:** All recorded sensor data is displayed on this overview timeline in a normalized way. For normalizing, the data points are mapped between zero and one, where one corresponds to the highest measured value.
- **Single sensor:** For a regular view, single sensors can be displayed here on this timeline.
- **Ratio:** In this graph, two values can be displayed: one on the x-axis and one on the y-axis.
- **Map:** Single sensor data can be displayed according to their geolocation in this map view. Red points display the highest values, where yellow indicates values of medium value and green of low value.

Last but not least, the app offers the possibility to send feedback to the developers directly and offers help over a dedicated help page.

4.5.3 *External Sensors*

As already mentioned before, external sensors can be connected over Bluetooth. Currently, Huawei's smartwatch Watch 2[33] and Polar's H10 chest belt[34] via Bluetooth Low Energy (BLE) are supported. While the smartwatch can only measure aggregated heart rate in beats per minute (BPM), the chest belt can sense single heartbeats and can sample the heart's activity in milliseconds. With the later one calculations of different HRV measures are possible. The recording CoConUT app with a chest belt and smartwatch can be seen in Figure 4.10.

33 https://consumer.huawei.com/en/wearables/watch2, last visited March 18th 2019
34 https://www.polar.com, last visited March 18th 2019

4.6 CoCoQuest App

Since gathering qualitative and in-situ feedback was also deemed necessary in our requirement analysis (see section 4.3), we developed CoCoQuest, which is a study guide app for Android. CoCoQuest was conceptualized as a distinct app since we wanted to leave CoConUT running in the background untouched by the participants, who would have to interact with CoCoQuest regularly in the field. Furthermore, the usage of existing apps was not feasible since they do not log user feedback with timestamps. For synchronizing data gathered by CoConUT and CoCoQuest, timestamps from the same Android system are vital. CoCoQuest was built on the base of the Apache Cordova framework, which ensures cross-platform compatibility (see Figure 4.11).

The app itself can load pre-generated questionnaires in JSON format. Once the questionnaire is imported, it can be completed several times, for example by different participants on the same phone. The idea is that the participants are guided through the whole study by displaying task descriptions, providing questionnaire parts and giving the possibility to record images, videos and sound snippets as qualitative experience sampling while being on the move (for further information on experience sampling also see [Rob11]). Going back in the questionnaire is not possible since, for each completed page, timestamps are saved. After the completion of the study, the operator can see when each task was solved and when which feedback was given. Furthermore, through recorded images, videos, and sound the in-situ experience can be better understood. The results can be exported as JSON. Figure 4.12 shows the CoCoQuest app in action.

CoCoQuest offers the free combination of the following questionnaire parts:

- Likert scale
- Single choice
- Multiple choice
- Free text entry
- Display of text (reading only)
- Taking a photo / video / audio note

4.7 CoCoVis

While CoConUT offers a first in-built mobile visualization to explore data quickly, for further analysis, the data has to be exported and evaluated on a device with a bigger screen. For first exploration purposes and as a design study, the CoCoVis dashboard was conceptualized in a student project with the help of Tableau[35]. Figure 4.13 shows the final version of this CoCoVis prototype. A chart diagram on the left gives an overview of the single participants and completion time for the study. Furthermore, a scatterplot matrix gives first insights into possible correlations between gathered sensor and interaction data. Here, a correlation coefficient with according p-value is given via mouse overlay. On the right side, a map shows the routes the participants take[36]. Finally, sensor plots on the bottom display the values of the measured data throughout the study.

Since visualizations in Tableau are quite static and cannot easily be enhanced, an ongoing project aims at realizing CoCoVis in an easily accessible form via web access and Open-Source code. The first efforts are promising.

An Open-Source version of the dashboard is currently under development as a web app using R[37] and Shiny[38].

4.8 Other Components

The CoConUT framework consists of more components, namely the CoCoHat headpiece for qualitative data assessment, the CoCoBand wearable for measuring biofeedback, the software keyboard CoCoBoard, and the chatbot CoCoBot. Since most of those components were only used in one study or specifically designed just for one study, they will be described in this section.

4.8.1 *CoCoHat*

The CoCoHat enhances the data collected by the CoConUT app by video and sound recordings of the environment as well as video of the mobile device's screen. Its core is a Raspberry Pi 3 Model B V1.2, which is mounted on a stable hat which

35 https://www.tableau.com, last visited March 18th 2019
36 In the study depicted in Figure 4.13 two participants take a detour, shown in brown and beige
37 https://www.r-project.org, last visited May 27th 2019
38 https://shiny.rstudio.com, last visited May 27th 2019

the participants can wear during the study. It features one native Raspberry Pi camera (PiCam) for recording user interaction, one USB microphone for recording sound and one USB webcam for recording a video of the surroundings. Figure 4.14 shows the first prototype of CoCoHat, which is powered by two accumulators. The goal of the CoCoHat was to create a wearable made out of open hardware which runs open software.

4.8.2 *CoCoBand*

The CoCoBand (see Figure 4.15) was built only using Open-Source hardware according to an Open-Source hardware plan[39]. Its components approximately cost 85 Euro in sum. According to the hardware plan, the following three sensors were incorporated: optical heart rate sensor for measuring heart rate in BPM on the finger, a thermometer for measuring temperature on-skin and a GSR sensor using two electrodes applied to distinct fingers on one hand.

Dooren, Janssen, et al. compare several locations on the human body to assess galvanic skin response [DJ+12]. While unquestionably measuring on two distinct fingers yields the best results, they find out that the foot sole, the shoulders, and fingers are optimal for GSR collection, while armpit, back, and arm are locations that provided the worst results. Electrodes for measuring GSR should have constant contact with the skin to guarantee a steady measurement [BPS11].

4.8.3 *CoCoBoard*

CoCoBoard is a modified fork of the Android app Simple Keyboard[40], which itself is a fork of the official Android Open-Source Projekt app LatinIME[41]. In order to measure typing errors on a fundamental level, CoCoBoard records every letter on the soft keyboard with a timestamp. Suggestions for autocorrection of words remain hidden and mistyped words are not automatically corrected, which forces the users to correct mistypes with the backspace button. As a result, errors during typing can easily be referred through the number of backspace taps.

39 https://makezine.com/projects/the-truth-meter-2, last visited March 30th 2019
40 https://github.com/rkkr/simple-keyboard, last visited February 14th 2019
41 https://android.googlesource.com/platform/packages/inputmethods/LatinIME, last visited February 14th 2019

4.8.4 *CoCoBot*

The chatbot **CoCoBot** is a Python script running on a web server. It is based on the pip package python_telegram_bot[42] which abstracts most of the work the bot is doing. With these simple means, CoCoBot can simulate online communication to a certain degree. The messages the chatbot is sending come from a static list of messages on the server, which are always sent in the same order. The message delay can be individually specified.

4.9 Conclusion

As can be seen in section 4.4 and Table 4.3, the CoConUT framework successfully fulfills all the posed requirements both by Research Question 1 as well as by independent experts. The following contribution summarizes the work presented in the chapter:

Contribution

Development of the field study framework CoConUT

1. Android app CoConUT assesses contextual factors over smartphone sensors
2. Android app CoCoQuest probes the user for quantitative as well as qualitative experience feedback during studies
3. A chest belt connected to the CoConUT app assesses the user's stress level as an indicator for internal state
4. CoCoVis visualizes the collected data and allows for exploration
5. The hardware components CoCoHat and CoCoBand
6. Different smaller additions (CoCoBoard, CoCoBot) to fulfill study-specific purposes

In this section, the conceptualization and development of the CoConUT framework have been described. With the finished framework, the studies addressing

42 https://python-telegram-bot.org/, last visited February 14th 2019

Research Questions 2 and 3 could be realized. The subsequent two chapters depict two laboratory respectively two field studies, which have been conducted with the help of the CoConUT framework.

Requirement	CoConUT Approach
Quantitative data collection	Sensor collection with the CoConUT app
	Questionnaire data with the CoCoQuest app
Questionnaire / ES	Questionnaires and Experience Sampling methods with the CoCoQuest app
Conveying tasks	Study instructions with the CoCoQuest app
Video recording	Over the front camera with the CoConUT app
	Over the cameras with the CoCoHat wearable
Audio recording	Over the microphone with the CoConUT app
	Over the microphone with the CoCoHat wearable
Screen recording	The screen can be recorded with the CoConUT app
	Over the interaction camera with the CoCoHat wearable
Easy visualizations	Over the CoCoVis visualization dashboard
	Over the mobile visualization in the CoConUT app
Interactively exploring data	Over the CoCoVis visualization dashboard
Correct timestamps	In all apps and tools provided by the CoConUT tools
All-in-one solution	Yes
Up-to-date	All components up-to-date and mostly provided as Open-Source code[30]
Android-based	CoConUT was developed for Android
Support for short-term field studies	CoConUT was specifically designed to support short-term field studies

Table 4.3: Requirements and the way CoConUT meets them

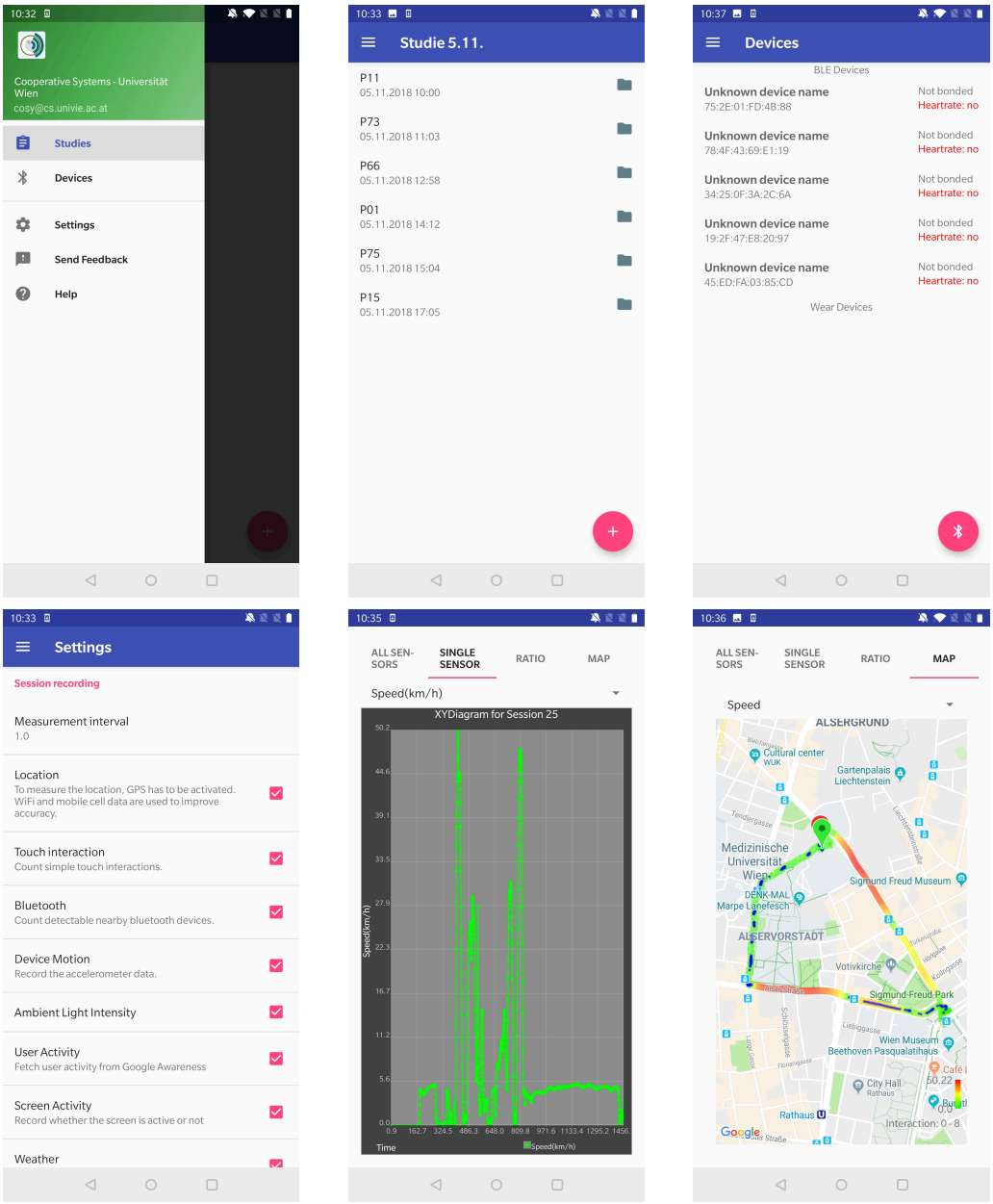


Figure 4.9: CoConUT screenshots, explained from the top left to the bottom right: 1) top menu of the app, 2) single user recording sessions within a study folder, 3) list of Bluetooth wearables nearby which can be connected, 4) settings menu where sensors can be activated and deactivated, 5) built-in visualization with first overview over gathered data of one sensor, 6) map visualization of gathered data, in this case data speed in km/h

Figure 4.10: Recording screen of CoConUT, together with a chest belt and a smartwatch

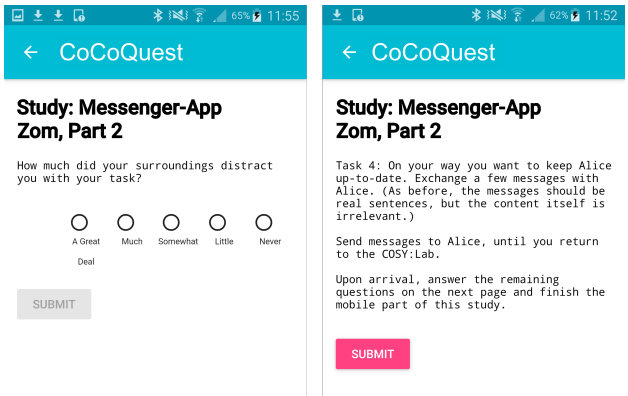


Figure 4.11: The CoCoQuest app with a loaded questionnaire. On the right a task description is displayed while on the left a question for rating is shown.

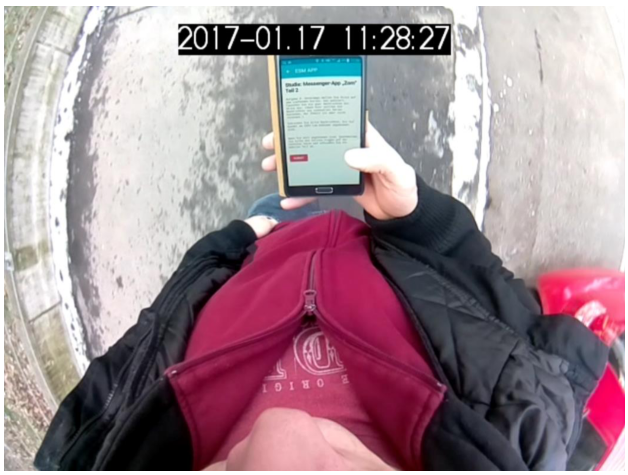


Figure 4.12: The CoCoQuest app in action during a mobile field study.

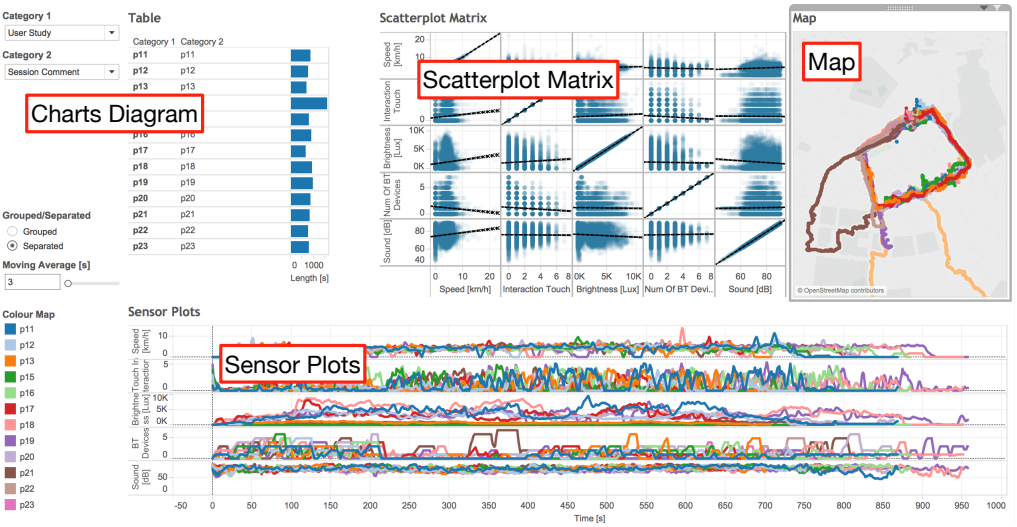


Figure 4.13: The CoCoVis dashboard visualizing the data gathered in a study.

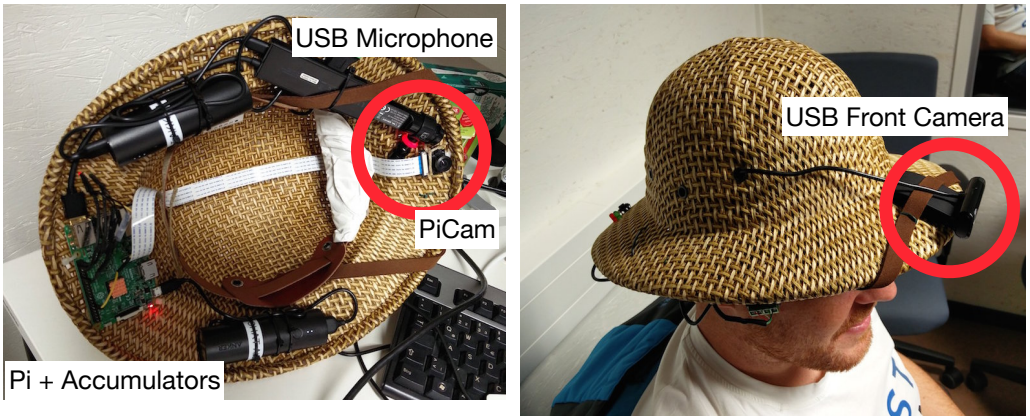


Figure 4.14: The CoCoHat with its different components: The core is a Raspberry Pi plus two accumulators. A PiCam films the user's interaction on the smartphone (see left). A USB microphone and a USB front camera record the surroundings (see right).

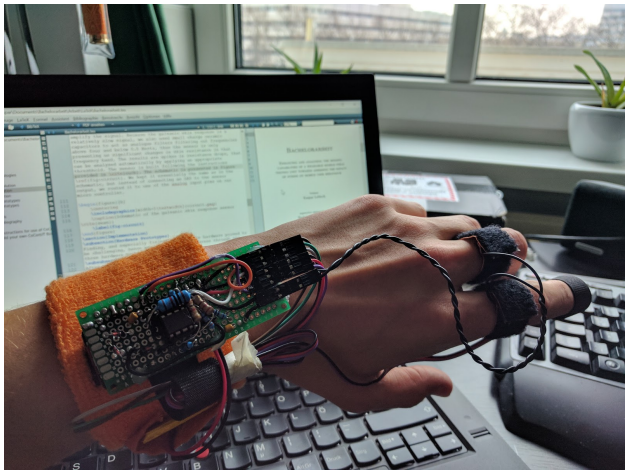


Figure 4.15: The CoCoBand wearable: optical heart rate sensor and galvanic skin measurement electrodes are applied on the fingers, while microboard and battery are attached to a sweat band [Leb17]

5

Laboratory Studies

After having presented the CoConUT framework, we will outline our research addressing different types of errors. First, we have a more in-depth look into two error types based on cognitive processes, namely knowledge-based mistakes and rule-based mistakes. This chapter presents the outcomes of two laboratory studies addressing these types of errors.

The first study aims at providing detailed insight into the knowledge-representing mental models. Getting a grasp of users' mental models based on their acquired knowledge is a challenge since there are no easy tests or acquisition methods for mental models [GS14]. Operators have to work with qualitative data they carefully assess to understand the underlying mental model to a sufficient extent. Acquiring this amount of qualitative data in this level of detail poses questionable in the field since methods like the Think Aloud protocol are only suitable to a certain extent. Additionally, the impact of stress on memory recall is not well-researched in cognitive psychology [Men99]. Thus, testing the impact of the user's internal state on retrieval of mental models would require further experiments. Hence, for this study, the impact of context and internal state are left disregarded.

In the second study, the occurrence of rule-based mistakes was tested under different levels of stress. Mental arithmetic tasks (MATs) are used as an operationalization of rule-based decisions since procedural mathematical knowledge relies on the retrieval of if-then procedures [DD05]. During this study, we worked with our self-built wearable CoCoBand to retrieve biophysical measurements of stress indicators. Due to the nature of these measurements, participants were

not allowed to move their hands and had to remain seated. As a consequence, in this study, the influence of context is being left out as well.

To sum up, both studies were conducted in the laboratory because the study setups could not have been realized in this form in the field.

5.1 Erroneous Mental Models

The work described in this section has partly been published in the following paper [Sch+16]:

S. Schröder, M. Huber, D. Wind, and C. Rottermanner.
"When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging".
In: *European Workshop on Usable Security. IEEE.* 2016.

The first laboratory study is described in this chapter and addresses the question when knowledge-based mistakes happen during mobile communication and how they can be assessed (RQ3). It furthermore has a more in-depth look at their impact on secure communication (RQ3).

Mistakes based on faulty knowledge and erroneous mental models are the most complex form of errors that can happen in human-computer interaction (HCI). Errors, in general, can have different grades of severity, and while some are trivial, others are hard to recover from. In the majority of cases, problems on the level of mental models are hardly recoverable from, since users have to understand first that they lack essential pieces of information, in order to then be able to acquire the necessary additional knowledge.

A particularly interesting use case for knowledge-based mistakes is the field of usable security. Being on the crossroads of human-computer interaction (HCI) and security engineering, usable security begins at the premise that secure systems are only as secure as the user can use them well. Merely blaming the user as the weakest link in the interaction is a too simplified approach [SBW01], since the user has to be taken into the loop during the development of secure systems [Cra08].

A large body of research in usable security deals with secure mobile communication. Especially secure instant messengers like Signal[1] are currently in use

1 https://signal.org, last visited May 25th 2019

by the vast populace. Those end-to-end encrypted instant messengers offer a feasible tradeoff between security and usability. Such end-to-end encrypted communication technologies have been available for decades. For most of those tools, poor usability and other obstacles have hindered them from reaching a broad coverage [Abu+17]. Especially the end-to-end encryption tool par excellence PGP[2] (*"Pretty Good Privacy"*) has been in usage since 1991 but never became a widely accepted standard due to poor usability [WT99; Gar+05; RVR14; FCS12]. Recent research shows that once trust between two mobile messaging parties is successfully established, no further steps by the users are necessary to ensure a secure and private conversation [Ung+15]. This good security drives more and more users to become convinced adopters of secure instant messaging apps.

For this study, the instant messenger Signal was chosen. Signal [Ope16] originated from two separate mobile applications [Sys15]: TextSecure (encrypted instant messaging) and RedPhone (encrypted phone calls). Due to its strong encryption protocols and the availability of its source code under an Open-Source license, Signal has become an important tool for users who face surveillance [The16b]. In April 2016, the currently most popular messenger app WhatsApp [Wha16] rolled out end-to-end encrypted messaging, based on Signal's protocol, to more than one billion users [The16a]. Signal's encryption protocol thus became the de facto standard for end-to-end encrypted mobile messaging.

In this laboratory study, a simulated Man-in-the-Middle (MITM) attack[3] was launched on a secure communication channel over Signal to find out how participants noticed and reacted to the attack and whether they were able to take countermeasures to ensure their safety. As we know from section 2.3, knowledge comes especially into play when we have to deal with novel situations. Using this novel and unusual situation of an unexpected attack, we hoped to gain some deeper insights into the users' mental models regarding mobile communication and to understand potentially ensuing knowledge-based mistakes better.

Findings show that incomplete mental models can lead to fatal knowledge-based mistakes, in this case, false mitigation strategies after a security-related attack. Surprisingly, users keep their false sense of security and continue to trust the compromised app.

2 https://pgp.com, last visited March 19th 2019

3 In a man-in-the-middle attack, an attacker secretly intercepts the communication between two parties and potentially alters it [Sta11].

5.1.1 *Background*

To better understand the study setup, including its technical prerequisites, this section provides some background on secure instant messengers and particularly SIGNAL's underlying security mechanisms.

From a security perspective, state-of-the-art mobile messengers can be split into two categories: messengers that provide *client-to-server encryption* and messengers with *end-to-end encryption*. The first category of messengers allows service providers to read exchanged messages, while the second group ensures that service providers cannot read messages. State-of-the-art end-to-end encrypted mobile messengers only require users to authenticate via their mobile number; the generation and exchange of cryptographic keys are handled transparently by the applications. The transparent end-to-end encryption of messages makes strong encryption accessible to the masses but also creates new security challenges. As compared to PGP, state-of-the-art secure mobile messenger applications rely on centralized services to provide the cryptographic identities of its users. This modus operandi results in the following security challenge: if the key-exchange service is tampered with, either willingly or by an attacker, the overall security of systems is subverted. In order to account for the compromise of the key-exchange service, mobile messaging apps, therefore, offer the possibility to verify the cryptographic identities of other users ultimately to establish the trust of exchanged encryption keys.

Some related work has already dealt with the challenges of and attacks on different kinds of secure messengers. Unger et al. have published the most comprehensive work on secure messaging [Ung+15]. Their survey provides a current view on challenges for secure messaging, especially regarding technical means to verify users and the mitigation of MITM attacks. Onwuzurike et al. [OD16] provide a taxonomy of security features on smartphone messaging apps and identify several gaps between the claims and reality of their promises. Regarding SIGNAL, Frosch et al. [Fro+16] provide a detailed analysis of the underlying cryptographic protocol of SIGNAL. Furthermore, Schrittwieser et al. [Sch+12] discuss the different attack vectors like account hijacking, sender ID spoofing, enumeration, and several other security issues of early mobile messengers. This study has been complemented by Rottermanner et al. [Rot+15], who focused on the unique privacy challenges posed by mobile messengers.

SIGNAL offers forward secrecy at the same time as asynchronous message exchange. As such, SIGNAL combines the PGP-like asynchronous messaging with the security properties of the Off-the-Record Messaging (OTR) protocol [BGB04].

The Signal protocol is divided into three phases (registration, session setup, and message exchange), while a central service is used to exchange the public encryption keys. This service is critical for Signal's security and a potential point for compromise. Frosch et al. [Fro+16] provide a detailed analysis of Signal's protocol.

In the following, the Signal protocol and the establishment of trust between two clients will be explained: *Alice* and *Bob* want to use Signal to exchange end-to-end encrypted messages. Alice installs Signal and verifies her mobile number at the Signal Server with either a verification text message via Short Message Service (SMS) or a voice call. Once verified, Alice creates different sets of keys: a longtime asymmetric key-pair called `Identity Key Pair`, 100 ephemeral key pairs called `One-Time Pre Keys` as well as one `Signed Pre Key` which is signed with the `Identity Key`. Signal automatically uploads Alice's `Signed Pre Key` as well as the 100 `One-Time Pre Keys` to its server. Alice attempts to establish a session with Bob and therefore requests a `Pre Key Bundle` for Bob and Bob's `Identity Key` from Signal's central service. The `Pre Key Bundle` consists of a single public `One-Time Pre Key` and the `Signed Pre Key` of Bob. Based on the `One-Time Pre Key` and the `Signed Pre Key`, Alice derives a symmetric `Master Key` for future communication, and stores Bob's `Identity Key`. Based on the `Pre Key Bundles` of each other, both Alice and Bob derive the same `Master Key`, which is used to create ephemeral `Message Keys` for the actual message exchange.

The unique long-term Identity Key pair remains the same as long as the user does not delete it by, for example, re-installing the Signal application. These Identity Keys are essential to verify the identity of communication partners. The Signal application, therefore, stores the Identity Keys of other users as soon as a secure session has been successfully established. Signal allows users to view this Identity Key within the application. In order to make sure that communicating parties received the correct Identity Keys, both parties have to verify the public Identity Keys via an out-of-bound channel (for example meeting in person or via phone). This verification can be done by comparing the hexadecimal representation of the key byte per byte or by scanning the QR code of each other's Identity Keys in person.

Threat Model

To install security countermeasures, companies and associations often plan according to realistic or potential attack scenarios. The threat model for the MITM attack, which is launched in this study, is described in this subsection.

The threat model chosen for this study accounts for the compromise of Signal's central services. This compromise can be the result of targeted attacks on Signal's service infrastructure or assistance of Signal's team to a subpoena request. The compromise of Signal's key server results in two different possible attacks:

Attacks on the first session setup do not result in direct user feedback. This attack can only be detected by **manually verifying**, for example over the phone or face-to-face via scanning the QR codes. Consider Bob wants to initialize a secure session with Alice, and Bob receives the attacker's Identity Key (Mallory's Identity Key) instead of Alice's Identity Key which is then stored by Signal as Alice's identity.

The second possible attack scenario are **attacks on established sessions**, where Bob has previously established a secure session with Alice and stored Alice's correct Identity Key. An attacker (Mallory) could force both parties to re-negotiate a new communication session. In this scenario, the compromised Signal server would respond with the attacker's Pre Key Bundle including the Signed Pre Key of the attacker, and thus establishes a man-in-the-middle (MITM) attack.

Signal accounts for both of the attack scenarios of this study's threat model. First, Signal provides a feature to manually verify established Identity Keys, as outlined in Figure 5.1. Second, Signal warns users when it detects that long-term keys of users change, see Figure 5.2. In this study, both of these two countermeasures of Signal are addressed.

Of course, during these countermeasures, numerous knowledge-based mistakes can occur, since users need to have suitable mental models during different steps: they have to understand how a secure instant messenger functions, where attacks could happen, how to recognize such attacks and which mitigation strategies could be taken. Ideally, good usability of the security mechanisms would support users during all steps to prevent knowledge-based mistakes.

5.1.2 *Study Setup*

The study consisted of two parts: a usability study of the Signal app with a focus on Signal's instant messaging and security features, and the execution of an

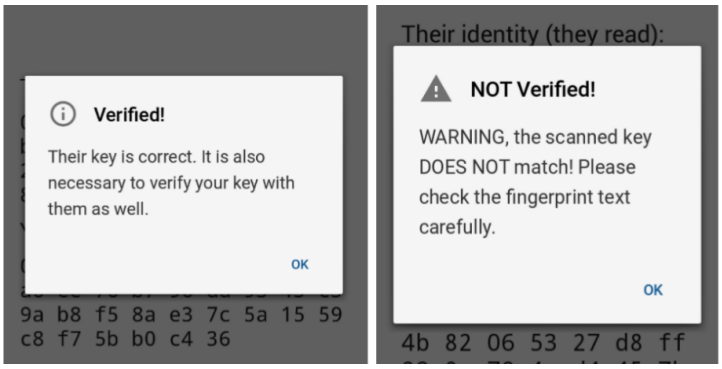


Figure 5.1: Verification of identity keys by scanning the each other's QR codes. On the left: a successful verification. On the right: warning because identity keys did not match.

actual MITM attack with a subsequent assessment of the users' reactions, primarily to assess potential knowledge-based mistakes. The results of the usability evaluation are moved to the appendix and will not be included in this section.

To gain insights into the participants' motivations, strategies and goals they were asked to comment aloud on their actions with the Think Aloud method constantly [Lew82], which facilitated to understand the users' mental models. User interaction and voice were recorded with a camcorder. Participants had to fill in a consent form before the start of the study, as well as a short questionnaire including demographics and general attitude towards privacy and security regarding smartphones and especially messaging apps. The study took place in the COSY:Lab at the University of Vienna, which provides two lab rooms for usability experiments and an operator room. Two tests were conducted in parallel. Thus, four operators (two in the operator room and two in the respective test rooms) had to be present to conduct the study in parallel.

At the beginning of the study, participants received a set of instructions, including all tasks and questionnaires, as well as an Android device with Signal pre-installed. Each phone (Alice) had a contact entry for the conversation partner (Bob), handled by an operator. The detailed technical set-up is described in the next subsection. In the following, we describe the tasks participants had to complete as part of our study.

The **first part** of the study focused on Signal's general usability related to messaging and security features, which will not be described in detail here. This part additionally had the function to accustom the participants to the usage of Signal.

In the **second part**, participants had to exchange messages with Bob. Shortly before this task, the MITM attack of the simulated compromised Signal server was launched, which triggered an error message about Bob's mismatching key (see Figure 5.2). The task description also asked users to verify Bob's identity after the message exchange. The study instructions informed participants that they could ask their chat partner Bob into the room at any time. Bob (simulated by an operator) was instructed to play a completely passive role and not to reveal any information on the verification task. Following the verification task, the participants had to fill in a debriefing questionnaire aiming at assessing the users' mental model of the MITM attack, as well as possible mitigation strategies, by using quantitative and qualitative questions.

5.1.3 *Technical Setup*

In order to conduct the study with two persons in parallel, two identical setups were used, which were each administered by one operator. One working setup consists of three smartphones and one computer which was responsible for intercepting the traffic and for creating a Wireless LAN (WLAN) hotspot for the smartphone's internet connectivity. All smartphones were rooted and had Cydia Substrate [Sau16] and SSLTrustKiller [Bla16] installed in order to eliminate the Secure Sockets Layer (SSL) certificate pinning protection of Signal. For traffic interception and manipulation, we used mitmproxy [Cor16] in combination with a custom script to automatically intercept Signal messages. Two client smartphones (Android 4.4.4) and one attacker smartphone (Android 4.4.4) were used. The attacker smartphone (Mallory) was preloaded with a modified version of Signal to handle intercepted messages and to forward intercepted messages to the original recipient. The two client smartphones had the latest version of Signal installed (3.15.2). One client smartphone was given to the study participant (Alice). The other client smartphone was used by the operator (Bob) in the operator room. Finally, because all smartphones shared the same network, the smartphones connected to our attack proxy via a ProxyDroid [Lv16] configuration. For each study participant, the devices were reset and re-registered with Signal.

5.1.4 *Pilot Study*

First, a pilot study with six participants from the local COSY research group was conducted to refine the study design before the actual study. In the pilot

study, users were being asked to "verify" their communication partner. This request led to confusion as the participants never reached Signal's verification features and had widely diverging understandings of the term "verification". Thus no user managed to compare keys successfully. Based on these results, a brief explanation of Signal was included, to point participants towards Signal's technical verification features. Furthermore, the decision fell in favor to include a "hint": the instructions told the participants that they could ask for their communication partner (Bob) to enter the room at any time. Since participants of the pre-study were unsure whether Bob is a real person or a pre-scripted Bot, this information was crucial to include.

5.1.5 *Results*

Participants

Overall, 28 participants took part in our study (7 female, 21 male), which lasted about 30-45 minutes. All of the participants were computer science students at the University of Vienna, the majority of whom were enrolled in an HCI course and recruited via that course. The only requirement for participation in the study was experience with the Android operating system. The students got a reward in the form of extra points for the HCI course. Two of the participants were 26 to 35 years old. The remaining people were in the age between 18 and 25.

While 22 participants stated in the questionnaire to actively use the Android system, ten specified to have iOS devices, followed by two Windows Surface users and one Blackberry user (multiple choices allowed in the question). Nearly all of the participants actively use text messaging/SMS (27) and WhatsApp (26) as instant messaging apps, followed by Telegram (18), Viber (8), Facebook Messenger (4) and Kakaotalk (2). One participant each used Line, Andchat, Skype, Signal, Threema, and Tango. Regarding self-assessment of computer security knowledge, most of the participants said they had no or some knowledge about privacy and security mechanisms (7 respectively 17), while four stated to have a lot of knowledge. None of the participants claimed to be an expert in computer security. Privacy and security on smartphone apps are of importance to the participants, and they care about third parties reading their messages. Confidentiality of text messages and active security/privacy measures were weighted to be of average importance.

Users' Reactions to the Attack

Shortly before the second part of the study, the MITM attack was launched. After the launch of the MITM attack, messages sent through Signal were not delivered since Signal's protocol needs mutual keys to send messages. As a consequence, all of the users noticed the attack because of an error notification next to the undelivered message (see Figure 5.2), and clicked on the notification icon to open the error dialogue.

At this point, the error dialogue already confronted the users with the task of verifying Bob. While 24 out of 28 users read the text in the subsequent dialogue, the remaining four directly chose the "Accept" option while skipping the text. These participants seemed to follow "the flow" of the dialogue to quickly reestablish messaging functionality.

Even if the participants were able to access the key comparison page, whether from the error dialogue or later in the task (eight users never did), the key verification page of Signal's Android application did not provide any instructions on how to perform the actual verification. As Figure 5.3 shows (picture on the right), Signal displays the Identity Keys of both communication partners, but no further instructions are provided. The participants of the study, therefore, faced problems on how to use the displayed keys. One participant, for example, stated: *"Ok, those are keys, but what am I gonna do with them?"*.

In total, 13 users asked Bob into the room during this task for verification. However, less than half of those users managed to to match keys with Bob successfully (seven users). When they compared keys correctly, a message about verification failure was raised due to the MITM attack. The error message, however, did not provide any information on consequences, further mitigation strategies, or strategy changes. One participant thus said: *"Well great, and now what?"*, while another participant stated: *"To be honest* [...] *I have no idea what to do now"*.

Also, five users made use of the "Reset Secure Session" functionality offered by Signal, which deletes the saved identity of the chat partner (Bob). The confirmation message, which appears before the session is reset, did not provide sufficient information regarding the result of this operation.

Mental Models of the Attack

Preferably, Alice and Bob compare their keys in person for verification purposes of confirming their mutual identity. If Mallory launched a MITM attack on their conversation, Alice and Bob ideally recognize this type of attack, stop communicating over Signal and uninstall the app. As previously stated, successful

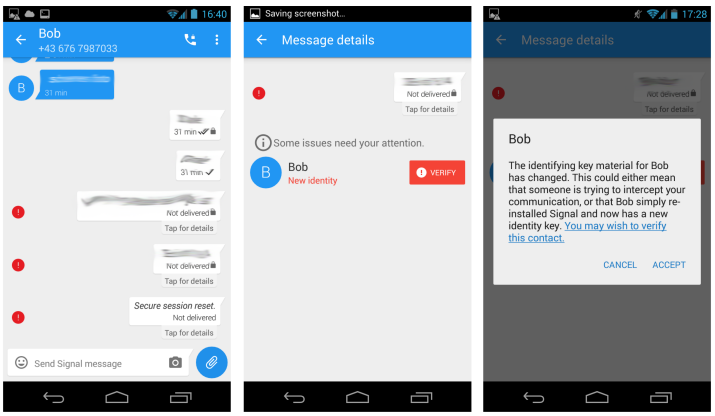


Figure 5.2: Message delivery failure (1), notification about Bob's new identity (2) and new identity dialogue (3)

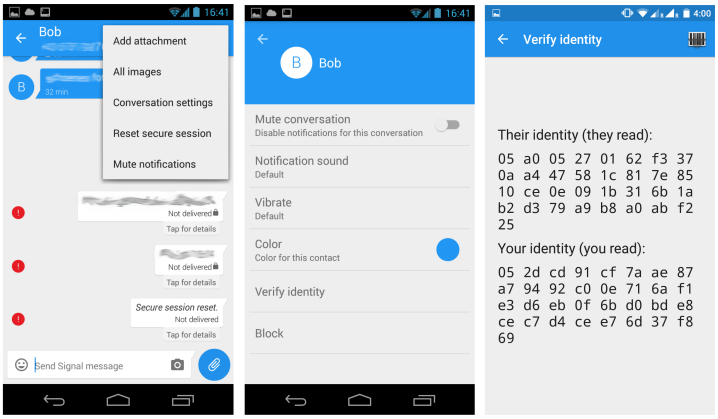


Figure 5.3: "Verify identity" option in the conversation settings (1 & 2). Key comparison page displaying Bob's key at the top and Alice's resp. the user's key at the bottom (3)

MITM attacks on Signal result from their central key exchange services being compromised, Alice and Bob thus need to stop using Signal. In consequence, the successful verification of Bob with matching keys was at no point possible in the setup of this study due to the MITM attack. However, 13 participants assumed that they had successfully verified Bob in the final questionnaire, while they failed to compare keys with Bob correctly. Therefore, they accepted Bob's new identity and would likely have continued to communicate over an insecure connection since they assumed it to be secure. Those users had different (false) verification strategies, which we discuss in the following. This type of knowledge-based

mistake is the worst scenario that can happen after a successful MITM attack and proves to be especially fatal.

Seven users successfully matched keys with Bob. Only three of those assumed some sort of attack but did not mention MITM in particular. Two of those users assumed they were not chatting with Bob, but with the attacker Mallory. Three other users thought that the app malfunctioned. Thus, matching the keys did not necessarily lead to the correct assumptions. Participant assumptions are discussed below.

The rest of the participants (eight users) did not manage to compare keys with Bob and were unsure about having verified Bob or knew they had not. Five of those participants explicitly assumed a MITM attack took place. Subsequently, not all users picked correct mitigation strategies. An overview of strategies that users would have chosen is outlined below.

Verification Strategies

Out of the 13 participants who thought to have verified Bob, but did not manage to do so by comparing the keys, 12 came up with different verification strategies. Six assumed that accepting Bob's new key in the error dialogue following the attack successfully verified Bob. Four "verified" Bob by either meeting him in person or by asking him questions about messages he received and his identity via chat or via phone calls. One person assumed that the presence of the keys on the key comparison page proves the authenticity of Bob's identity, while another person attempted to verify the authenticity of the chat by asking Bob whether he thought the chat was secure.

Verification Strategy	**No.**
Accepting key in error message dialogue	6
Proving that Bob is a real person	4
Keys present in the system	1
Asking Bob	1

Table 5.1: False verification strategies ($n = 12$)

Assumptions about the Attack

In order to assess the users' assumptions about the attack, an open question about the "unexpected events" was included in the final questionnaire. Spoken remarks in the Think Aloud protocol were also taken into account. Overall, 14 participants made remarks about possible explanations for the unforeseen events (multiple mentions could be made). Seven participants speculated or stated that a MITM attack could have taken place, although only one of those participants compared keys correctly. As already stated, not all the participants who successfully compared keys made the right assumptions about the events during the MITM attack. Several other incorrect assumptions were drawn: Four participants stated that an attacker attempted to impersonate Bob. Thus, they assumed that they had compared keys with Mallory instead of Bob. Furthermore, three participants speculated that Bob could have reinstalled Signal as suggested in the error message. Another three users assumed that the app was malfunctioning. Two participants finally stated that an attack could have happened, but did not specify the type of attack.

Assumptions about the Attack	**No.**
MITM attack	7
Malicious conversation partner	4
App reinstalled	3
App malfunctioning	3
Attack (not further specified)	2

Table 5.2: Assumptions about the attack

Participants also had to answer several questions regarding their mental model of the unforeseen events, as well as perceived security, privacy, and trust of the Signal app after the attack. The most notable observation was that in most cases, users did not lose trust in the Signal app after we triggered the attack.

Mitigation Strategies

The final questionnaire contained another open question about participants' possible mitigation strategies after the unexpected events. The type of attack was deliberately not revealed to not bias answers. Also, the users' actions and remarks

Question	Mean μ	SD σ
I'm having a clear image in my mind of what happened exactly.	2.79	1.03
I'm feeling extremely insecure regarding the app.	2.36	1.06
I always had the feeling to be completely informed of what's happening through the app's notifications.	2.71	0.90
The app has damaged my trust in it.	1.71	0.85
I had the feeling that I had full control over the app the whole time.	2.79	1.03

Table 5.3: Mental model of the app (questions were answered on a Likert scale from 1 to 5 where 1 is completely disagree and 5 is completely agree)

during the last study task were considered. Several possible mitigation strategies (not necessarily referring to MITM attacks in particular) arose from the answers: 11 participants would simply uninstall the app (the only valid mitigation strategy against compromise of the server), although it was not clear whether they wanted to avoid further hassle and would simply use another messaging app, or whether they knew it was the recommended mitigation strategy. Other strategies aimed at gathering more information, such as contacting Bob on another channel via other apps, phone or face-to-face meetings (eight participants), searching for information on the Internet (six participants) or asking friends (four participants). Three participants would inform the developers or read license agreements and policies (three respectively one participants). Another branch of strategies involved problem solving: restarting the app (two participants), disconnecting the phone from the Internet (two participants) or a virus scan (one participant). The results of the different mitigation strategies we observed are summarized in Table 5.4.

While the key page had the purpose of Alice and Bob both reading their keys aloud, Signal also offers a QR code functionality for comparison. Some of the users used the barcode functionality, but the resulting error message did not provide sufficient information for a strategy change.

Mitigation Strategy	No.
Uninstalling the app	11
Contacting Bob on another channel	8
Searching for information on the Internet	6
"Reset Secure Session"	5
Asking friends	4
Informing developers	3
Disconnecting phone from the Internet	2
Reading license agreements and policies / virus scan	1

Table 5.4: Possible mitigation strategies as expressed by the participants

5.1.6 *Discussion*

To the best of our knowledge, we were the first to study the security, as well as usability challenges of end-to-end-encrypted messengers up to the point of the study (2016). The central services used to exchange user keys pose a major security risk of today's end-to-end encrypted messengers. Therefore, in the study, a compromised key service was simulated by performing an active MITM attack. Hence, the usability of Signal's security features in the case of active attacks was assessed.

However, like any user study, this work has some limitations: First, the participants recruited for the study were homogeneous since all were students of computer science and shared the same age group. Similar experiments with different groups of participants might, therefore, lead to different outcomes. Second, the extent of information provided to participants on Signal's encryption/verification features had to be balanced. We decided to explicitly ask users to verify each other in order to asses the usability of this core-security feature of Signal. The initial study design tested in the pilot study showed that none of the six participants used the verification feature in the face of the simulated attack. Similar experiments with participants without a computer science background or a focus on a security subtask would likely result in even less successful key verifications.

Overall, the outcome of the study is considered surprising, especially given the fact that the participants had a computer science background. The results suggest

that the "verification" process and therefore the overall security of end-to-end encryption on mobile instant messaging faces serious usability obstacles, since 21 of 28 participants failed to compare keys with their conversational partner accurately. Especially surprising in the study was the high number of participants who thought they had successfully verified while in reality, they failed to compare keys.

Signal, as easy-to-use end-to-end encryption enhanced app, should support struggling users to overcome common knowledge-based mistakes by good design and achieve security in the sense of increased usable security. Usability problems, in terms of missing support, can lead to fatal errors and ensuing severe security breaches, for example aborting the reestablishment of a secure connection after an attack.

The gaps between self-assessment, mental models of differing correctness respectively level of detail as well as actual outcome (un/successful defense) could be explained in several ways: Either participants lacked the required knowledge, the app failed to support the users, they had a different understanding of what "verification" meant or the effort for successful defense was too high.

During the MITM attack, Signal was explicitly hinting at the fact that the connection could have been compromised. The fact that only seven participants assumed the possibility of a MITM attack and only three thought that Bob reinstalled the app seem quite surprising. Either those users ignored, or did not read, the informational error message or excluded the possibility of an attack/reinstallation while remaining under the false illusion of security.

The different strategies for verification and mitigation hint at flawed mental models: users seem to lack an understanding of end-to-end encryption in general, possible attack scenarios and risk potentials. The findings from section 5.1.5 also indicate a great trust by the users in the app to deal with security issues in the background, therefore assuming that the app's dialogues could be trusted.

Based on our findings on the usability of Signal's error handling of actual attacks, we found that these features led to more problems than to actual attack mitigations. Under these circumstances, it is not surprising that WhatsApp has disabled all encryption-related error messages by default. If users want to get feedback on mismatching Identity Keys or alike, they explicitly have to enable the error messages in the preferences. Since reactions to non-comprehensible error messages (due to the interplay of potentially missing information on the app's side and incomplete mental models on the user's side) range from uninstalling the app, contacting the developers, or a definitive feeling of insecurity in general, we assume the developers of WhatsApp made a compromise between usability and

security due to economic reasons. Since communication via WhatsApp was only encrypted between the client and the server recently, messages on a changed Identity Key might lead to confusion, ultimately angry users and eventually uninstallation.

5.1.7 *Conclusion*

In this chapter, a user study on the occurrence of knowledge-based mistakes based on erroneous mental models is presented. In this study, Signal for Android is used, which is a secure mobile messenger that provides a promising solution for widely adoptable end-to-end encrypted conversations. First, the unique security challenges and threats today's secure mobile messengers face were discussed. Second, a comprehensive user study on mental models regarding a man-in-the-middle attack was conducted, and knowledge-based mistakes were assessed. It was shown that the majority of users failed to detect and deter such attacks.

Our results show that the majority of users made fatal knowledge-based mistakes by failing to correctly compare keys with their conversation partner for verification purposes due to usability problems and incomplete mental models. Hence, users are very likely to fall for attacks on the essential infrastructure of today's secure messaging apps: the central services to exchange cryptographic keys. Those potential knowledge-based mistakes, from which the users can hardly recover and lead to compromised security, need to be addressed by the usability site. Since the users do not notice their erroneous mental models and take the wrong actions, secure messaging apps need to support the users in making the right decisions.

A series of studies conducted after the publication of the study at hand addresses the authentication mechanisms of Signal in more detail. Vaziripour et al. find several critical usability issues in a study on Signal [Vaz+17]. For example, they point out that the duration required to find and complete the verification procedure is too long from a usability standpoint. In a subsequent study, they successfully implement a call for action (*"Action Needed!"*) in Signal and prove that this usability improvement helps the users to find and complete the authentication process more easily [Vaz+18]. Finally, they integrate various social media services into the Signal app to establish a social authentication process [Vaz+19]. This procedure leads to good overall usability and user satisfaction, but some users dislike the usage of social media since they do not deem them to be private

enough. These studies are promising with their results, but still, much work to improve secure end-to-end encryption in instant messengers needs to be done.

Summarizing, the following contributions have been made:

Contribution

Incomplete models can lead to false mitigation strategies and compromised security after attacks

1. Surprisingly, users have a false sense of security.
2. Users have very high trust in secure apps.
3. Bad usability of high-risk security features can lead to non-solvable security problems.

The setup of the study at hand has made it clear that this study design could not have been realized in the field in this level of detail. This necessity for testing in the laboratory also applies to our second laboratory study on rule-based mistakes under stress, which will be described in the next section.

5.2 Biophysical Measurements of Arousal

The study in this section was conducted in close collaboration with the Medical University of Vienna. Furthermore, Kaspar Lebloch kindly contributed with the outcomes of his practical programming course.

The study described in this section explores rule-based mistakes, how to assess them, and their potential impact on secure communication (RQ3). It further has a look at the role the user's internal state, particularly stress, plays on the occurrence of rule-based mistakes (RQ2). Finally, it deals with the question of how the user's internal state can be assessed, especially in field studies, and the accuracy of the gathered data (RQ1).

As could be seen in section 2.3, rule-based mistakes are errors that happen during the execution steps of a composed action plan. During the different steps of the action plan, decisions about medium-scale tasks have to be made, and most of these decisions are made following simple if-then-rules and heuristics. During these decisions, rule-based mistakes can happen. In this case, mental arithmetic

tasks function as an operationalization of rule-based decisions, since procedural mathematical knowledge relies on the retrieval of if-then procedures [DD05].

The main goal of this study was to test the impact of stress on rule-based decisions and potential errors in the field. Since another aim of this thesis was to build a field-study framework entirely made out of Open-Source components, the wrist-worn wearable CoCoBand (see subsection 4.8.2) was built in the course of this study to retrieve biophysical measurements of stress indicators. During data assessment with the CoCoBand, participants were not allowed to move their hands and had to remain seated, because measurements of galvanic skin response are prone to movement artifacts. In order to test the suitability of the CoCoBand for field studies and to compare it against heart rate measuring consumer devices (in this case a chest belt and a smartwatch), the decision fell to hold the study in a laboratory, namely the COSY:Lab at the University of Vienna.

Our results indicate that stress plays a role in the occurrence of rule-based errors. While the data acquisition with CoCoBand demonstrated to show accurate results in comparison with the consumer devices, we show that consumer devices (chest belt and smartwatch) produced the same quality of data while having a more robust design. GSR measurements proved to be highly influenced by movement artifacts, which would potentiate in a field setting.

5.2.1 *Related Hardware*

To find a suitable wearable for measuring biophysical feedback during field studies, we had a look at related hardware. Potential choices encompassed wearables for eHealth, sports-based applications, research and Open-Source hardware in general.

Cheap and multifunctional eHealth wearables are becoming a trend among consumers. Fitness, diets, and self-optimization, in general, are reasons why consumers buy health-related electronic devices and especially wearables.

Consumer devices from vendors like FitBit[4], Garmin[5], or Polar[6] mostly come with a limited range of sensors, draw conclusions over aggregated data processed in a cloud and lack the standard of approved medical electronics. Nonetheless, conclusions can be drawn from the data, and users can track their continuing

4 https://www.fitbit.com, last visited February 22th, 2019
5 https://www.garmin.com, last visited February 22th, 2019
6 https://www.polar.com, last visited February 22th, 2019

health progresses. It can be said that those devices have a high potential for general well-being [Bax16].

Devices like the "Meta" sensor series from MBIENTLAB[7] or the Empatica E4[8] offer a variety of functions for researchers and direct data access. Unfortunately, they are closed source, and additional sensors cannot be added. The E4 is also quite expensive and not affordable for low to mid-priced field study setups. On the other hand, existing Open-Source solutions like the HeartyPatch[9] or the MobileECG[10] either are not available yet or only provide a limited array of sensors.

All of the solutions mentioned above might be suited for usage during mobile field studies, but all of them have certain drawbacks. While the Open-Source solutions certainly offer the best access to raw sensor data readings, most of the time, they are not widely available. Therefore practical experience with and extensive testing of the hardware are not given. The manufacturers of consumer devices, on the other hand, seldom grant access to a developer SDK. If access is given, sometimes only aggregated data from the manufacturer's cloud is provided without allowing to access the raw sensor data. This aggregation mostly prohibits the calculation of self-chosen algorithmic measures and requires an active internet connection at all times. Also, the acquisition of data from the manufacturer's cloud is not real time.

Summarizing, there is a need for an Open-Source, reliable, widely available, and cheap wearable for assessing all kinds of biophysical signals in the field.

5.2.2 *Study Setup*

As already described in section 2.2.1, there is a variety of stressors in everyday life, and what one perceives as a stressor is highly subjective. For clinical testing and laboratory experiments, ethical and reproducible stress tests have been developed to work for the majority of test persons under laboratory conditions. Bali and Jaggi give a good overview and describe different stress induction methods (tests et cetera) and which factors can be measured to assess stress [BJ15], including biochemical markers, physiological and behavioral changes as well as cardiovascular changes. Stress can be induced by either physical (environmental or physiological) or psychological (cognitive or emotional) stressors. The gold standard Trier

7 https://mbientlab.com, last visited February 22th, 2019
8 https://www.empatica.com/research/e4, last visited February 22th, 2019
9 https://www.crowdsupply.com/protocentral/heartypatch, last visited February 22th, 2019
10 https://github.com/peterisza/mobilecg, last visited February 22th, 2019

Social Stress Test (TSST), for example, builds on social evaluation pressure and lets participants speak publicly and calculate MATs in front of present evaluators. In another test protocol called the Cold Pressor Test (CPT), the participants immerse their hands up to the wrist into ice-cold water (between $0-2\,°C$). Essential factors in standard psychosocial stress tests are social evaluative threats, uncontrollability, or unpredictability [BJ15].

MATs (mental arithmetic tasks) are a usual task to evoke stress in laboratory experiments [BJ15; KPH93; KMY12; Jer+91]. Best combined with a social evaluative situation, MATs do not require a complicated setup, are language neutral, and probands do not have to move much, which is relevant for avoiding movement artifacts in the sensor data.

In our case, mental arithmetic tasks bear a certain similarity with rule-based decisions, since for calculating mathematical tasks mentally, stored sets of separate rules are evoked from long-term memory and applied to get the result. Research outcomes also underline this: for example, Ryu and Myung find that MATs engage memory processes for retrieval of arithmetic facts from long-term memory [RM05].

Based on these prerequisites, a customized stress test to be able to fit each participant into a time slot of one hour was developed, allowing the participants to sit still and apply separate measurement devices. Due to the nature of the measurement, GSR, participants ideally should not move at all. This necessity to remain seated is the reason why this experiment was conducted in a laboratory environment.

Participants were recruited from a university course at the University of Vienna. The students got points that they could use to gain better grades if positively passing the course. We invited them under the pretense of calculating mathematical tasks mentally while being connected to different biophysical measurement devices.

After arriving at the lab, participants were given a very brief introduction with an overview of the whole experiment, including the duration of the measurement part. They were not told that their stress level would be assessed, only that they will have to solve mathematical tasks mentally. They had to fill in a consent form that data would be recorded (video, audio, biophysical signals). After that, they had to fill in the first half of a questionnaire, which assessed basic demographics and base stress level. Additionally, some questions about exercise per week, and medical conditions were asked to be able to understand variations in the

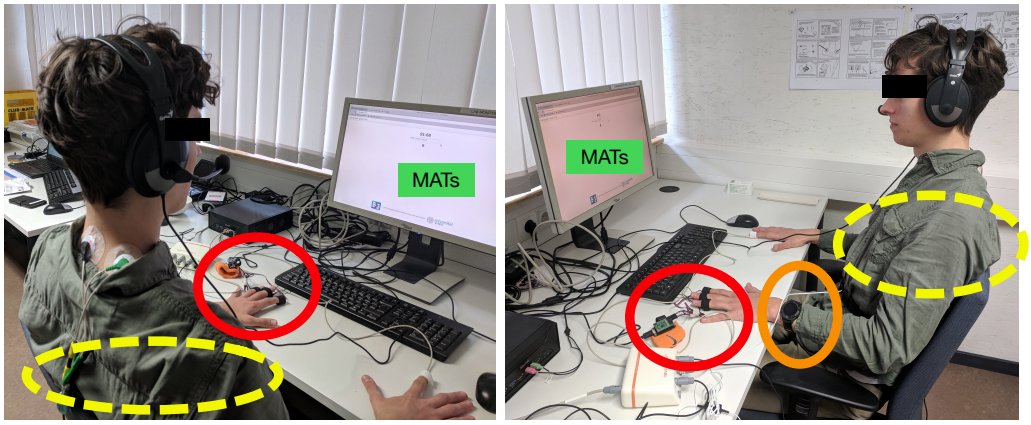


Figure 5.4: Setup of the experiment. The participant is instructed to sit still in front of the computer. Both hands are laying flat on the desk. On the left hand, the smartwatch (orange circle) and the CoCoBand (red circle) are applied. The chest strap is worn under the clothing (yellow dotted circle). On the screen, the MATs are displayed (green box), and voice input is given over the headset.

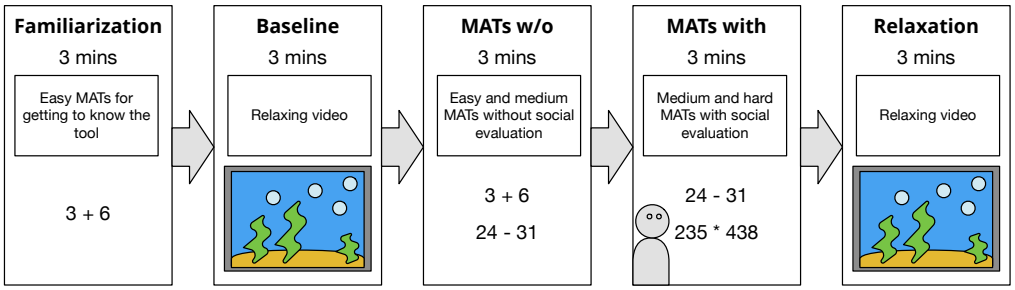


Figure 5.5: Storyboard of the test procedure

recorded biophysical data[11]. Following the questionnaire, participants were aided with attaching all required biophysical measurement devices: They had to wear the Huawei Watch 2 smartwatch on the left wrist, with the CoCoBand being attached to the fingers of the same hand. The Polar H10 chest belt was worn under the clothing directly on the skin (also see Figure 5.4). The participants were instructed to remain seated and to sit still during all times of the experiment, and to only give input over voice into the microphone. When they moved during the experiment, they were asked to remain still. After the connection to the devices, the stress experiment began.

The phases were as following (as can also be seen in Figure 5.5):

11 Nonetheless, it would have been impossible to inquire for all potential interfering influences.

- **Familiarization:** Three minutes of MATs (mental arithmetic tasks) for getting to know how the online tool
- **Baseline:** Three minutes of a relaxing video
- **MATs w/o:** Three minutes of MATs with low and medium difficulty
- **MATs with:** Three minutes of MATs with medium and high difficulty, with the operator standing behind their back watching to increase stress
- **Relaxation:** Three minutes of a relaxing video to calm down.

After the phases were completed, participants were relieved from the instruments and had to fill out the second, concluding part of the questionnaire.

5.2.3 *Technical Setup*

An affordable, ready-to-use Open-Source solution, which could be integrated into an existing test setup, as well as easily be enhanced by new features, does not exist up to this point. This lack of a solution is why the self-made wrist-worn wearable CoCoBand was developed, which was realized with the motivation of creating an Open-Source wearable for assessing biophysical feedback in field studies. It is entirely assembled from Open-Source hardware and software and was built following an instructional manual from the internet. It measures the user's surface skin temperature, heart rate (HR/BPM), and skin conductance (GSR). For the development of the CoCoBand, please refer to subsection 4.8.2.

The consumer devices of our choice were a Huawei Watch 2[12] and an H10 heart rate sensor by the brand Polar[13], which were both connected to a smartphone running the CoConUT app. While new wearables like the Apple Watch or Android Smartwatches allow a very rough estimation of the user's inner state, chest belts made for training in- as well as outdoors offer the possibility to gain more insights over heart rate measurements. As already mentioned in subsection 4.5.3, both devices can be connected to the app.

To minimize movement artifacts, a web application for mental arithmetic tasks with voice input possibility was developed. The web application was developed

12 https://consumer.huawei.com/en/wearables/watch2, last visited February 22th, 2019

13 https://www.polar.com/en/products/accessories/H10_heart_rate_sensor, last visited February 22th, 2019

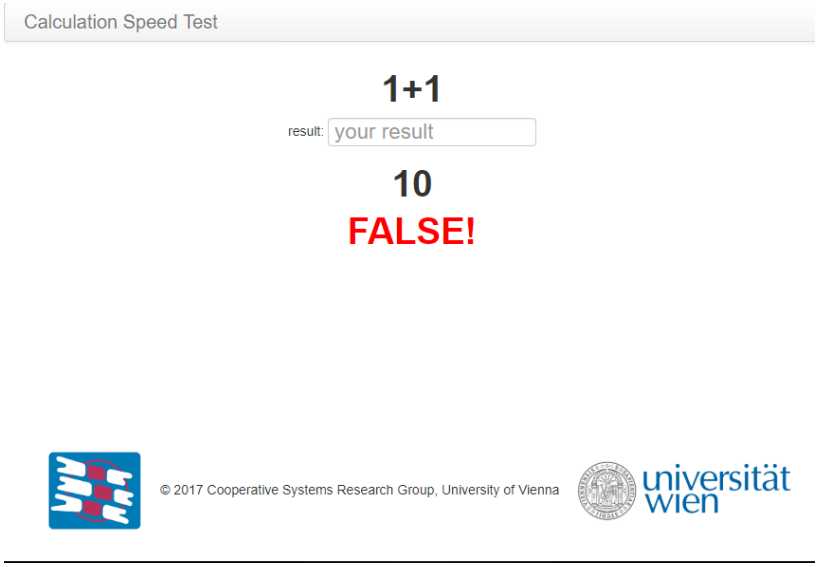


Figure 5.6: The MATs tool poses an arithmetic task, evaluates it and indicates whether the solution is correct or wrong

in Python based on flask[14]. This MATs tool (see Figure 5.6) enabled the operators to specify a customized MATs test by setting difficulty level and durations for easy, intermediate, and hard arithmetic tasks. For speech recognition, Google's Web Speech API was used[15]. The software was installed on a local computer in order to avoid network latencies.

5.2.4 *Results*

Overall, 31 participants took part in the study between March and June 2018. Because the experimental setup had to be adapted during the execution of the trials, 21 participants proceeded with the setup described above. The data sets comprised galvanic skin response, heart rate in BPM and on-skin temperature from the CoCoBand, heart rate in BPM from the smartwatch and subsequent heartbeat intervals in milliseconds (for calculating HRV) from the chest belt. Since the setup was complex, not all gathered datasets could be gathered in a sufficient way[16]. In the end, 19 data sets could be evaluated.

14 http://flask.pocoo.org, last visited April 3rd 2019
15 https://w3c.github.io/speech-api, last visited April 3rd 2019
16 In two cases measurements were not complete since the measurement devices failed to record or the assessed data was not accurate enough.

Out of these 19 participants, five were female and 14 male. The mean±SD of age was 23.37±2.73. Thus, the participants were roughly in the same age group, according to Umetani, Singer, McCraty, and Atkinson [Ume+98]. All participants were students at the University of Vienna, most of them studying Computer Science. According to the questionnaire, none of them had a heart disease, and two of them were regularly taking psychotropic drugs (anti-depressants, anti-anxiety medication).

In the following sections, we will focus our evaluation on the heart rate data and the errors regarding mental arithmetic tasks.

CoCoBand

Overall the CoCoBand worked as intended and successfully gathered sensor data. However, the CoCoBand had unforeseen hardware issues, which led to inaccuracy of the produced data. The process of data transmission led to unequal sampling intervals of the sensors, due to characteristics of a serial USB connection or the Bluetooth device via BLE. While equidistant measurement intervals would have been necessary, this underlying hardware problem rendered the skin conductance readings unusable. Also, this complicated obtaining reliable heart rate estimates using an inter-beat-interval detection algorithm. A higher sampling frequency and equidistant measurement intervals would have needed to be implemented or must be implemented in a future prototype.

Nevertheless, after some data cleaning, the BPM could successfully be extracted from our recorded data. The intercorrelation between BPM values of each chest belt, smartwatch, and CoCoBand was very high and ranging between 0.9 and 1.

Heart Rate across Phases and Stress

Among the tested consumer and self-built devices, the chest belt was the most reliable device regarding heart rate. While CoCoBand and smartwatch both only measured the aggregated value of beats per minute (BPM), the chest belt was able to calculate both BPM and subsequent RR intervals, which allows for a calculation of different HRV measures (for instance RMSSD). In the following, only the data from the chest belt will be taken into account.

As can be seen in Table 5.5 and Figure 5.7a, the beats per minute (BPM) varied across the different phases. While during the phases calculating the MATs ("familiarization", "MATs without social pressure" and "MATs with social pressure") the mean of the BPM is over 80, for the relatively quiet phases ("baseline" and

"relaxation") the BPM remains under 80 in the mean. While BPM is a quite vague measure, some first insights can be drawn from the values.

Since we wanted to compare values on a normalized base, the values of the relaxation phase were taken as baseline values, and relative changes in percent were calculated. The relaxation phase was taken as a baseline since it was noticed that participants were a bit tense during the original baseline phase since they anticipated the upcoming MATs. Figure 5.7b presents the relative changes in percent compared to the relaxation values (as baseline). Here the highest deviations are noticeable during the familiarization and the MATs with pressure phases.

Phase	**Familiariz.**	**Base**	**MATs w/o**	**MATs with**	**Relax.**
BPM Mean:	87.87	78.52	82.96	83.19	75.15
BPM SD:	22.52	21.21	17.17	15.34	14.98
RMSSD Mean:	39.11	47.37	42.97	44.21	54.09
RMSSD SD:	24.63	30.39	26.36	26.60	35.12

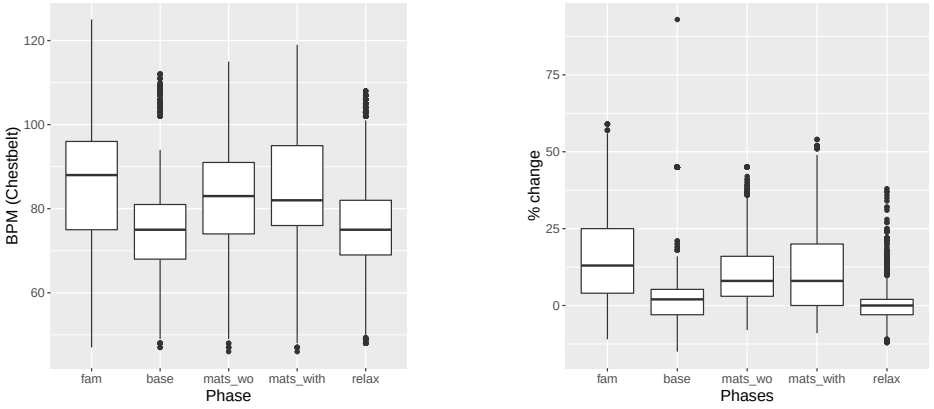
Table 5.5: Heart rate values across phases

The RMSSD values, an indicator for arousal (in this case: stress level), were in line with the BPM values: for the base and relaxation phase, the mean was clearly below the other values (see Table 5.5 and Figure 5.8a). Also, for the RMSSD, in Figure 5.8b, one can see relative changes in percent compared to the relaxation phase as the baseline.

MAT Errors

As can be seen in Figure 5.9b, for the last MATs phase (medium to hard MATs with social pressure) the ratio of correct answers drops drastically. During the first two MATs phases the mean and sd are relatively even (mean$\pm$SD of 63.91%$\pm$22.64% of correct answers for the familiarization phase and a mean$\pm$SD of 66.30% $\pm$20.95% for the phase with easy and medium MATs and without social pressure). In contrast, the mean drops to 32.63%$\pm$16.69% correct answers during the last phase with medium and hard MATs plus social pressure.

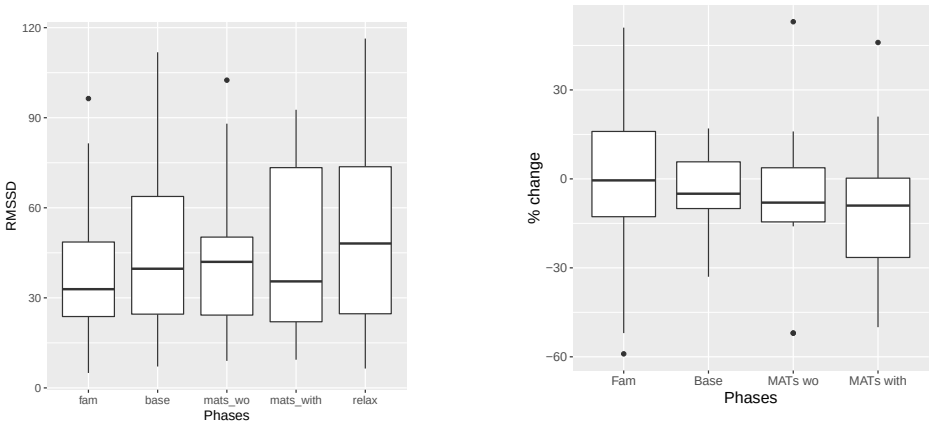
When being brought into relation, a medium, but non-significant correlation can be seen between stress level (in RMSSD) and the percentage of correct answers: For the *Familiarization* phase, Pearson's product-moment correlation yields a



(a) BPM

(b) BPM normalized after relaxation phase

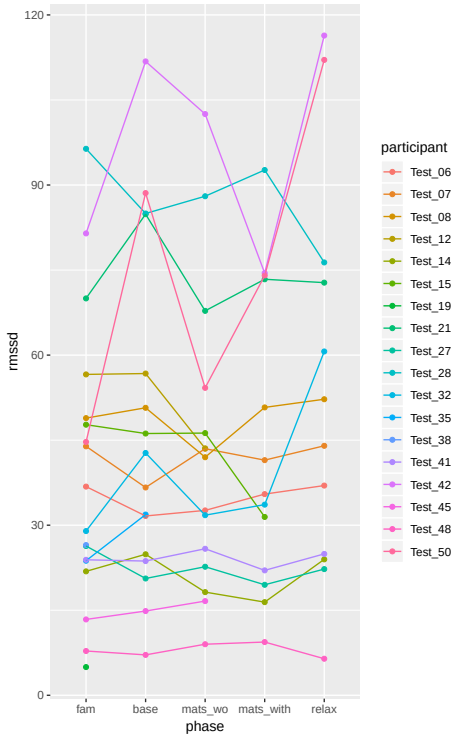
Figure 5.7: BPM normalized in relation to relaxation phase



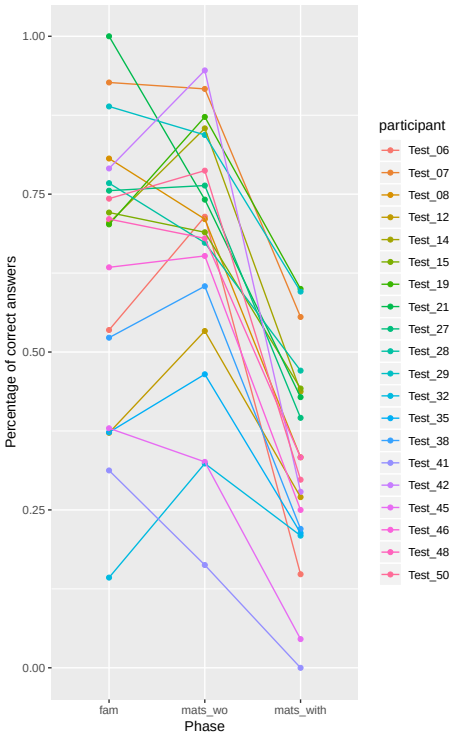
(a) RMSSD

(b) Percentual changes of the RMSSD values normed after the relaxation phase

Figure 5.8: Boxplots of RMSSD across phases



(a) RMSSD values per phase per participant (please note that two datasets are incomplete)



(b) % of correct answers for all participants

Figure 5.9: RMSSD and errors per participant across phases

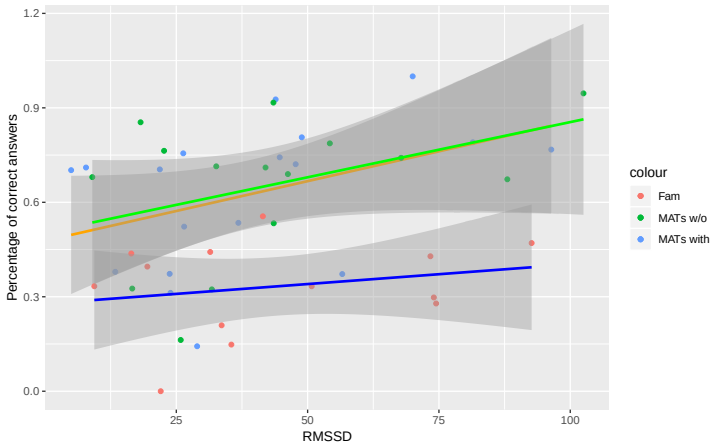


Figure 5.10: Relation of RMSSD to the percentage of correct answers for the different phases. The x-axis denotes the RMSSD value: The lower the value, the higher the participant is stressed. The y-axis denotes the percentage of correct answers. Each line represents a different phase of the MA tasks. It can be seen that the percentage of correct answers decreases when stress increases

positive correlation of 0.41 with a significance of $p = 0.09$. For the *MATs w/o social pressure* phase the correlation is 0.41 with $p = 0.13$ and for *MATs with social pressure* a correlation of 0.22 with $p = 0.47$ was calculated. The distribution of the points can also be seen in Figure 5.10.

5.2.5 *Discussion*

In summary, it can be said that both self-built and consumer devices measured the beats per minute to a sufficient degree, as reflected by the high correlations. However, the characteristics of the CoCoBand make some of its functionalities unusable in the current version: Connection issues prevent a sufficient sampling rate and movement artifacts during GSR acquisition on the fingers would require a new approach. Subsequently, we focused on the heart rate provided by the chest belt, which proved to be the most reliable and provided the best data quality.

From the measured heart rates, BPM and the HRV measure RMSSD offered some insights into the stress level of the participants. BPM and RMSSD show an increased respectively decreased mean during the phases with higher stress potential. For further insights, more testing would be required.

Regarding error rates, the participants made more errors during phases with higher stress potential (see section 5.2.4). It can be assumed that stress has an

impact on the occurrence of rule-based mistakes. On the other hand, of course, during the last phase of MATs, the mathematical tasks were more difficult and could have led to a higher error rate. More experiments to assess the impact of stress on small-scale decisions during mobile interaction have to be conducted.

Of course, our study has some limitations. First of all, we only ran the experiment with university students. While usually a heterogeneous user group is seen as negative, in this case, the heterogeneity of the user group is useful regarding the comparison of heart rate values [Ume+98]. Additionally, a custom stress test protocol was developed, based on the limited study time slots and additional special requirements like stationary seating. This test protocol has yet to be tested and verified with a larger user group.

5.2.6 *Conclusion*

Overall, our results from this exploratory study are quite promising. First of all, the results indicate that stress could play a potential role in the occurrence of rule-based errors. Furthermore, consumer devices prove to be more reliable to assess the internal state of the users during mobile field studies, since they prove to be more robust, cheaper and yet versatile: While the data acquisition with CoCoBand worked well, we show that chest belt and smartwatch produced a better quality of data while having a more robust design. The hardware price of about 90 euros for the CoCoBand and the higher effort for assembly leave space for the discussion whether it is worth the effort to prefer the self-built solution over the consumer devices, even if we would build an enhanced prototype. While the Open-Source design of the CoCoBand makes it easier to add additional sensors, most of those sensors could also easily be acquired in a pre-assembled and ready-to-use form. For assessing GSR though, another location than the fingers on one hand would have to be chosen, since obviously, participants in the field would need their hands, which would produce movement artifacts. To our knowledge, an affordable consumer device for assessing GSR on the go does not exist.

Summarizing, the following contributions can be concluded from this study:

Contribution

Stress could play a role in the occurrence of rule-based errors: It is assumed that the higher the stress, the higher the error ratio.

Contribution

Consumer devices are more suitable to assess internal states during mobile field studies than self-built solutions.

1. Consumer devices prove to be affordable, robust, and reliable.

Future work will improve the CoCoBand with regards to hardware and software, to ensure a sufficient sampling rate and decrease the effort for incorporating new sensors. The next experiment will adapt the experimental plan for the stress test: the baseline phase would happen in another room before the participants could build up anticipation. Also, each phase will last at least five minutes to ensure a reliable HRV measurement in the medical sense.

Overall, in this chapter, the outcomes of studies conducted in the laboratory have been described. While the first study explored the occurrence of knowledge-based mistakes during secure communication, the second study led to a more in-depth insight into rule-based mistakes under stress. With the insights gained in the laboratory studies, a series of field studies could be planned. In the next chapter, two subsequent field studies conducted in the course of the thesis at hand will be described.

6

Field Studies

After some insights into erroneous mental models causing knowledge-based mistakes and rule-based mistakes under stress gathered in laboratory settings, several field studies were planned and conducted. These studies aimed at tackling the influence of contextual factors on mobile interaction in the field, and the role mental resources, respectively the internal state of the user play (RQ2). Since we already covered errors based on cognitive processes in laboratory studies, we sought to test skill-based slips in the field.

6.1 Exploring the Interplay of Context and Interaction in the Field

The work described in this chapter has partly been published in the following paper [SHR18]:

S. Schröder, J. Hirschl, and P. Reichl.
"Exploring the Interplay of Context and Interaction in the Field".
In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX). IEEE.* 2018.

The study described in this section explores which contextual factors do influence mobile interaction in the field and to which degree, while seeking out

which kind of contexts have an impact on the user's primary HCI task (RQ2). Furthermore, the accuracy of the gathered data is evaluated (RQ1).

To address the aspects mentioned above, we conducted an exploratory study in the wild with 25 participants to gain first insights into the influence of context on mobile interaction. We present and discuss the results of this study in this chapter.

6.1.1 *Study Setup*

To gain a deeper understanding of communication behavior in the wild, and to collect sensor data via CoConUT to explore contextual factors, an initial study in the field was planned. Participants had to fulfill tasks with the secure Android instant messaging app Zom[1]. Zom is a free and open-source instant messenger which is an UX-centered fork of the Extensible Messaging and Presence Protocol (XMPP) client ChatSecure[2]. For task descriptions and subsequent questions about the tasks, the participants had to use the CoCoQuest app.

The field study consisted of two sessions on two different days with predefined routes to be walked along. Both routes of approximately 1 km were located in the university district and took approximately 10 minutes to walk. In the first part, participants were led along the sidewalk on streets with moderate traffic, while the second part went along sidewalks on streets as well, but partially also through a relatively quiet area of the university campus without traffic. Participants had to cross one street with low traffic during the study. Before the start of each session, an introductory questionnaire assessed demographic data, Zom was set up, and participants were instructed about the used apps. The route was given to them in the form of an annotated map image on the smartphone. Meanwhile, the operator remained in the lab and acted as a chat partner. The CoConUT app was running in the background of the smartphone to assess contextual and behavioral data via sensors. It was explicitly stated before the start of the study that participants should put their safety first and take care not to be harmed in traffic. Participants had to work on two tasks: A security-relevant task asked them to set a password for the app, while the second, open-ended task instructed them to exchange chat messages with the operator for the rest of the route. A subsequent questionnaire, which was filled out following the field part, asked the participants about their experiences in the field. Additionally, participants had to

1 http://www.zom.im (accessed January 7th, 2018)
2 https://chatsecure.org (accessed January 7th, 2018)

wear the CoCoHat (see subsection 4.8.1) to assess qualitative data in the field (audio recording over the microphone, video recordings of both the street and the user interaction with the screen).

Participants were recruited from an HCI course at the university and received credits for the course as compensation.

6.1.2 *Results*

Participants

The study took part in December 2016 and January 2017. During both parts of the study, temperatures outside were rather cold and ranging between 1 °C and 7 °C in December 2016 and −1 °C and −4 °C in January 2017. Note that the CoConUT app failed to record two participants' data sets due to sensor malfunctions, but worked reliably in all other cases. CoCoQuest worked in all cases and successfully guided the participants through the study.

Overall, 25 users participated in the field study (11 in the first and 14 in the second), which lasted about 45-50 minutes per participant. Eight of the students were female, while 17 were male. All of the participants owned a smartphone or tablet and were experienced with chatting on mobile devices. Out of the 25 participants, 23 completed the study. One participant could not go into the field due to heavy rain, another one due to technical difficulties during the overall study setup. While two data sets had to be omitted due to recording problems, the resulting 21 data sets were complete and could be analyzed (nine complete data sets from December and 13 from January). Out of those 21 participants, two lost their way during the second study and made significant detours (see Figure 4.13).

Sensor Data

The 21 complete data sets correspond to a total of approximately five hours of collected data. For a first overview, we examined the data in the CoCoVis dashboard (see Figure 4.13 for a first impression). Thus, several features of the collected data were already detected. For example, the scatterplot matrix shows that brightness varied heavily in-between participants (most probably due to different times of day the study sessions took place), while the noise level seemed to be quite even, without allowing to differentiate between soundscapes. Table

			Sensor Data	
Value	*mean μ*	*sd σ*	*Min*	*Max*
Light (lux)	1516.01	1880.14	0	23360
Noise (dB)	77.93	7.55	43.75	90.31
# Bluetooth Dev.	1.23	1.51	0	7
Interaction	0.77	1.30	0	7
Speed (km/h)	4.02	1.77	0	22.28
Accuracy (m)	12.14	16.04	3	145

Table 6.1: Table with summary about assessed sensor values of the CoConUT app.

6.1 shows some characteristics of the different sensor data as we describe them in further detail in the following:

Location and Speed: Table 6.1 shows an average speed of 4.02 km/h with a standard deviation of 1.77 km/h, and an accuracy of the GPS measurements of 12.14 m with a standard deviation of 16.04 m. Remember that, according to a U.S. government statement, GPS-enabled smartphones typically have an accuracy of 4.9m under a completely open sky, although accuracy decreases near buildings, trees, and other obstacles[3].

Brightness: The mean value of *Brightness* (measured in lux) was 1516 with a maximum of 23360. As a comparison: An overcast day typically has a lux value of about 1000, while full daylight lies around 10000 - 25000 lux[4]. In the data sets different maxima of brightness among the individual data sets were observed.

Number of Nearby Bluetooth Devices: On average, around one nearby Bluetooth device could be detected, with a minimum of 0 and a maximum of 7.

Sound Level: As can be seen in Table 6.1, the standard deviation of the recorded noise level was (7.55 dB) with a mean of 77.93 dB. Again, for comparison: A very calm room has a sound pressure of 20-30 dB, while a regular conversation has 40-60 dB and traffic on a busy roadway 80-90 dB[5].

Touch Interaction: Touch interaction had a maximum of seven touches per second. The mean lies at 0.77 with a standard deviation of 1.3, which is averaged

3 https://www.gps.gov/systems/gps/performance/accuracy (accessed January 13th, 2018)

4 http://stjarnhimlen.se/comp/radfaq.html (accessed January 26th, 2018

5 http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm (accessed January 23rd, 2018)

over input and non-input phases. For reference: The current Guinness World Record for 'Fastest time to type a text message (SMS) on a touch-screen mobile phone' lies at 160 characters in 17 seconds, which makes for a mean of 9.41 touches per second[6].

Influence of Context on Interaction

In order to increase data quality, for the following evaluation, all data points with a GPS accuracy of >10m were discarded. 65,01% of the original data points remained. To gain a first insight into the diversity of the data and influence of contextual factors on interaction, correlations between different data dimensions among the individual data sets were calculated and visualized as violin plot in order to show the distribution of correlations through the different data sets (see Figure 6.1). The plot thereby presents the influence of the contextual variables *speed*, *light*, *nearby Bluetooth devices* and *sound level* on *interaction*). For further analysis, we had a more in-depth look at the joint influence of nearby Bluetooth devices, and speed on the interaction parameter (see Figure 6.2). In more detail, Figure 6.3 and Figure 6.4 depict the impact of different typing speeds on walking speeds and vice versa.

Experiences in the Field

For gaining further insights into the experiences the participants made in the field, we examined the answers from the accompanying questionnaire. When being asked whether their attention was more on the surroundings or the smartphone, participants stated after the study that their attention was more likely on the smartphone (on a Likert scale from 1 to 7, where 1 meant "almost exclusively on the smartphone" and 7 meant "almost exclusively on the surroundings", the mean was $\mu = 2.33$, with a standard deviation of $\sigma = 0.92$). Regarding difficult situations in the surrounding traffic, almost all answered that they had not found themselves endangered, while one person stated to have been in a dangerous situation. Later, in an informal conversation, the person indicated that she almost bumped into a car while crossing the street. Participants were being asked what distracted them the most outside. Five of them mentioned they had to take care of pedestrians while walking: *"I had to watch out, so I don't run into other people"*.

6 http://www.guinnessworldrecords.com/news/2014/5/fastest-touch-screen-text-message-record-officially-broken-with-fleksy-keyboard-57380 (accessed January 23rd, 2018)

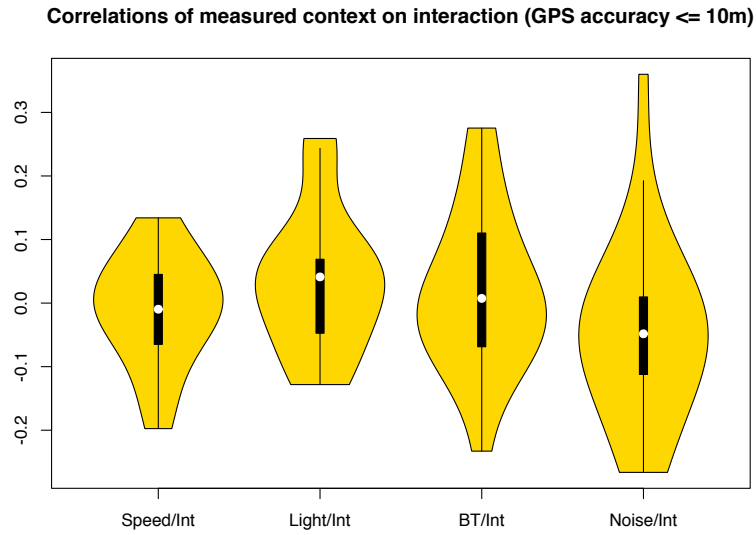


Figure 6.1: Violin plot of correlation distribution between different dimensions of context (speed, light, Bluetooth devices, sound level) and interaction

One person mentioned she had to look out for cars while she crossed the street. Since the second route did not require to cross a single street, this factor did not apply there. Finding the right way was a source of distraction that three participants mentioned. In the chat messages, participants stated that it was quite cold, and due to low temperatures, it was hard to type on the screen.

Some participants made remarks that the CoCoHat was quite eye-catching and drew the attention of passers-by. Two persons explicitly stated that the CoCoHat distracted them and mentioned the funny looks they got for it. One participant wrote it was *"a bit awkward to walk around with the hat"*.

6.1.3 *Discussion*

In our field study, we successfully gathered preliminary insights about the influence of contextual factors on interaction in the field, and our tools proved their high reliability.

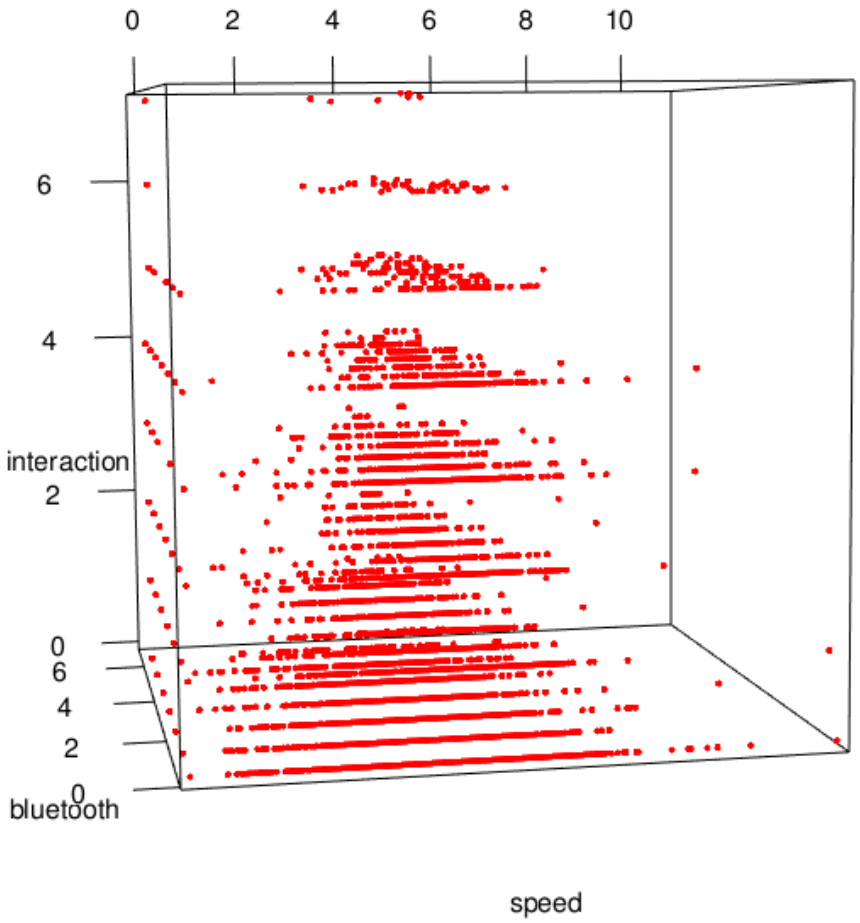


Figure 6.2: 3D plot of speed (x-axis), interaction (y-axis) and nearby Bluetooth devices (z-axis)

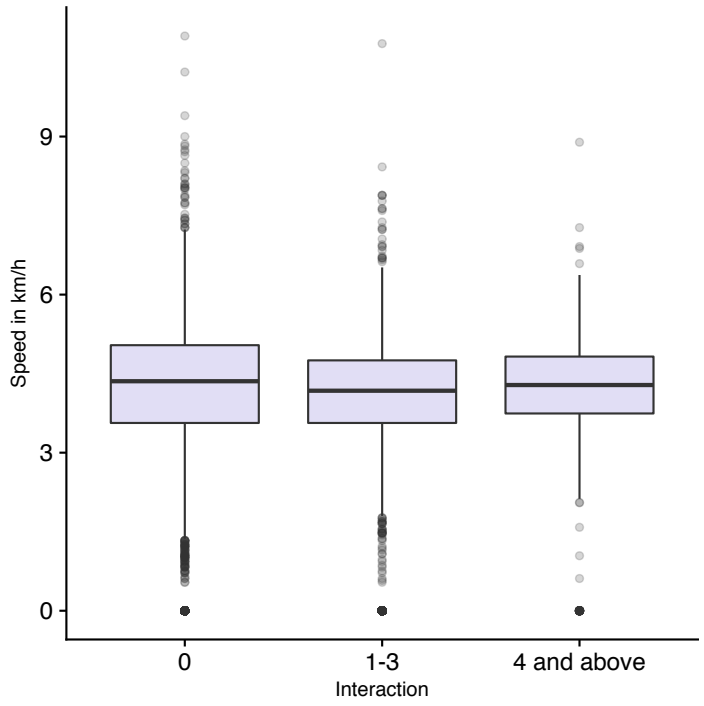


Figure 6.3: Boxplot of different typing behaviors (none, slow and fast)

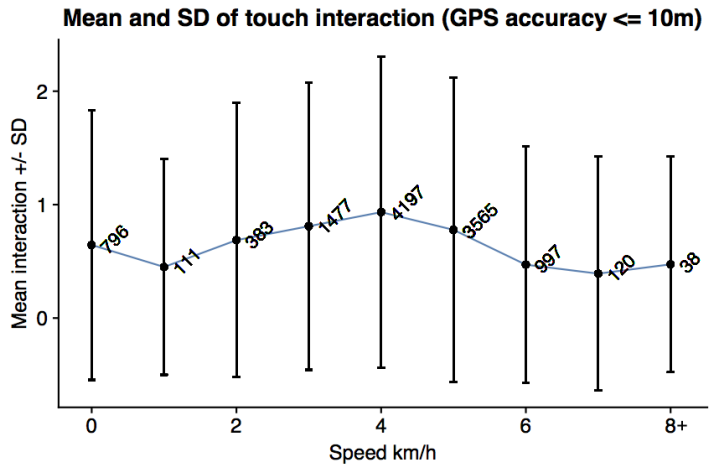


Figure 6.4: Mean and SD of touch interaction (GPS accuracy $\leq$ 10m). Labels on the means indicate the number of data points in this range.

Sensor Data Reliability

Some contextual data categories seem to allow more profound insights into context and interaction in the wild, while others seem to have limitations. Hence, we first discuss the reliability of the sensor data gathered by CoConUT.

Location and Speed: Since the study took place open-air, measuring location and speed was reliable, although GPS accuracy could have been higher. Except for a few data points, participants maintained average walking speed, while the exceptions might either be due to a sprinting participant or a measurement inaccuracy.

Brightness: The mean value of approximately 1516 does not surprise since the study took part in winter, and it was mostly cloudy. The different maxima of brightness among the individuals can be explained by the fact that some of the users participated in the late afternoon, and due to the season, it was relatively dark. Also, participants wore clothes of different colors. Comparability in-between the different individual data sets can, therefore, be regarded as low.

Number of Nearby Bluetooth Devices: The number of Bluetooth devices proved to be a good indicator for people nearby (also compare to [Min+11]) since on the university campus the number of measured Bluetooth devices increased near large lecturing halls and coffee shops. However, a large number of Bluetooth devices nearby does not necessarily mean that the participants had to walk through crowds, as the example of closely passing by a coffee shop shows. Also, the measurements highly depend on whether or not surrounding Bluetooth devices were configured to be visible. Additionally, discoverable Bluetooth devices must not necessarily be human-worn.

Sound Level: The relatively low variation in the data set can be explained by the noise regulating function of the built-in microphone. Given the fact that the participants walked alongside a road with moderate traffic and on the relatively quiet university campus, the flatness of the data set seems surprising, but the average value does not.

Touch Interaction: The maximum of seven indicates a high touch frequency, probably during text input. Although participants stated that their focus was mostly on the smartphone, they did not write messages or interacted with the smartphone the whole time. Generally, a higher value indicates a high focus on the screen and probably complex input, but the value itself does not say much about potential gestures and the usage of other input possibilities (buttons, voice, movement).

Context and Interaction

As seen in section 6.1.2, interaction and walking speed seem to be vaguely related. The mean of typing input frequency appears to be relatively steady, while the standard deviation reaches its peak around 4 km/h (see Figure 6.4). A potential interpretation could be that during regular walking speed (speed $\leq$ 5 km/h) there is a trend among the participants to either type a lot or not type at all, while participants who had a high degree of interaction walked more steadily in average walking speed, with fewer outliers (see Figure 6.3). The speed values for fast-typing data points did not vary as much as the data for the non-typers or slow-typers. All of these tendencies can be seen in Figure 6.2, in which the data points center around the plane of 5 km/h walking speed while thinning out towards the top and the sides of the plot. These results show evidence for different degrees of typing proficiency among the participants: there seem to be none-typers, slow typers, and fast typers. This distinction is especially interesting in the light of related research: smartphone usage is more widespread now than it used to be a few years ago. Since smartphones are still used mostly for communication, strong habituation of using them in all kinds of situations seems likely [Böh+11]. Thus, the users probably have developed adaptive strategies [Tim+17], leading to an increase in automation. The circumstance that the participants' attention was mostly on the smartphone also supports this hypothesis. Note that this (slightly speculative) conclusion is in contrast to [RM14], where only a few years ago it seemed likely that users stopped for entering text on a smartphone. Having said that, as a key result, it turns out that the overall impact of contextual variables on user interaction with the device in the field seems to be surprisingly low according to our data (see Figure 6.1). Of course, certain changes in the surroundings would inevitably lead to a change of behavior, for example the extreme case of an ambulance passing by. However, it is difficult to collect enough data on such rare events. Hence, more diverse data on varying types of contexts with different characteristics will be desirable, as related work shows that smartphones are used differently in different contexts [Min+11; RM14]. The same is valid concerning the reliability of the data since not all sensors proved to be reliable enough (while others definitely were). Further potential limitations concern the fact that nearly all participants were students of computer science at the university and thus the group was relatively homogenous. However, since smartphone usage has permeated all layers of society, we assume to have gathered sufficiently representative data. Finally, the quality and reliability of the sensor data gathered

are dependent on the used device and its capabilities - in our study, we consistently used a Samsung Galaxy Note 4 as a test device across all participants.

6.1.4 *Conclusion*

Summarizing, in this chapter, we presented the outcomes of our initial exploratory study examining the interplay of context and interaction. Our findings are discussed with a particular focus on the gathered data. We show that the CoConUT toolkit proves to be reliable for supporting field studies and provide initial evidence for strong habituation of smartphone usage compared to results from related work conducted only a few years ago.

In the end, the following contributions can be listed:

Contribution (RQ1)

Accuracy of current smartphone sensors

1. CoConUT proves as reliable, and an overview of the gathered sensor data is presented.

Contribution (RQ2)

Users do not slow down for typing anymore

1. Indicates strong habituation of today's smartphone users.

In the next chapter, a subsequent field study with a similar setup, but a more extended scope will be presented, which directly builds upon the findings gathered in this study.

6.2 Errors and Stress during Commute

The work described in this chapter has partly been published in the following paper [SRR19]:

S. Schröder, A. Rafetseder, and P. Reichl.
"Errare Mobile Est: Studying the Influence of Mobile Context and Stress on Typing Errors in the Field".

In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE.* 2019.

The study described here directly builds on the exploratory study from the previous section. Its first and foremost goal is to investigate skill-based slips in the field. In the course of doing so, the study deals with how these slips can be assessed, when they happen in the field and what their potential impact on secure communication is (RQ3). It furthermore investigates how the user's internal state and which contextual factors influence mobile interaction in general (RQ2) and has a look at the accuracy of the gathered data (RQ1).

Design, conduction, and findings of this study will be presented in the following section. The CoConUT toolkit tracks contextual and behavioral variables with apps and specialized hardware. Prior results are confirmed in detail that people do not slow down for typing while walking, despite producing more typing slips. Furthermore, the extent to which increased stress leads to an increase in typing errors is shown.

6.2.1 *Introduction*

In this study, we conduct a semi-realistic field study in a commuting setup to assess typing behavior, indicators for stress, and contextual characteristics. The CoConUT toolkit allows tracking of contextual variables via smartphone sensors, indicators for stress via a Bluetooth chest belt, and typing behavior via a modified software keyboard. We focus on the occurrence of typing errors and provide a first quantitative description of the influence of context and arousal on typing behavior in mobile environments.

This field study builds on the results from the previous field study described in section 6.1. The previous study shows that participants keep their walking speed constant during interaction with the mobile device. In our study, we use and extend the CoConUT framework to address Research Question 2 further: Which kinds of contextual factors do influence the occurrence of mobile typing errors? Which role does stress play?

The following section is structured as follows: In subsection 6.2.2, we give a brief overview of relevant literature that has not yet been covered in the previous chapters. After that, in subsection 6.2.3, the setup used in this study is described, which also includes additional software that was developed. In subsection 6.2.4,

the results of the study are laid out, which are further discussed in subsection 6.2.5. This section ends with a conclusion in subsection 6.2.6.

6.2.2 *Related Work*

Previous studies have shown that smartphone behavior in the field has changed since modern smartphones have been introduced with the first version of Apple's iPhone in 2007. As indicated by prior results, people do not slow down for typing anymore [SHR18], as they did before widespread smartphone coverage [Oul+05]. On the one hand, this can be risky, since not paying attention to one's surroundings can lead to severe accidents [Smi+13]. On the other hand, studies have shown that users can adapt their strategies according to context while at the same time using their smartphone [Tim+17]. Not as severe as accidents, but also potentially dangerous are errors during usage. Reason [Rea90] distinguishes between three primary types of errors during technology usage: skill-based slips (and lapses), rule-based mistakes and knowledge-based mistakes. Based on a large scale observation of 136 million keystrokes, Dhakal et al. show that fast typers make fewer slips during mobile typing, and that error correction mean in percent is 6.31 with a standard deviation of 4.48 [Dha+18].

6.2.3 *Study Setup*

In this subsection, the setup of the field study will be described. First, we will explain the general study setup, followed by preceding technical developments. The process of recruiting participants will be covered, as well.

General Setup

In the study presented in section 6.1, participants had to solely walk across secure sideways and answer chat messages, which brought some interesting explorative insights into mobile communication behavior. To dig a bit deeper into mobile behavior and especially errors, we designed this study as a semi-realistic study, in which we strove to control some factors. Semi-realism, in this case, fosters reproducibility and comparability, while at the same time enables realistic behavior in the field. Since field studies can pose an unlimited amount of contexts, in this study, the contexts were pre-given. Our scenario copies a typical commute, consisting of walking on sideways, waiting on the station and taking

the tramway. To assess the users' stress level, their heart rate was recorded via a non-intrusive chest belt. To emulate realistic chat communication, participants received a standardized series of 16 messages over the instant messaging app Telegram[7] sent by the chatbot CoCoBot (see subsection 4.8.4). The messages were sent in intervals of 90 seconds (after an onset of 240 seconds) under the pretense of being sent by the operator. Participants did not know beforehand they would receive messages. To assess errors, a software keyboard was modified to log every character input. Thus, the number of typing slips and error rate could easily be calculated.

The route consisted of: waiting on a station (*station1*), taking a tramway (*tram1*), waiting on another station (*station2*), taking a second tramway (*tram2*) and walking back to the lab (*walking*). The route the participants had to walk and commute can be seen in Figure 6.5[8]. They only had to wait for the tramways on the first two stations, since on the last one they only got out to walk back to the COSY:Lab.

In-between study sessions, the chest belt was disinfected, and the chat history in Telegram was cleared. The participants used the OnePlus 5 Android smartphone of the COSY research group.

Technical Setup

To assess context via smartphone sensors, the CoConUT framework[9] [SHR16; SHR18] and an Android-based smartphone were used. The CoConUT (sensor collection) app was extended to connect to a Polar H10 chest belt[10] via Bluetooth Low Energy (BLE) to record heart rate. The app measured GPS (location, speed, accuracy) and subsequent heartbeat intervals in milliseconds (among other sensors).

For measuring typing slips, the software keyboard app CoCoBoard was used (see subsection 4.8.3). All single character touches were logged with a timestamp. Auto-correction, as well as word suggestions, were disabled. Thus, the number of typed characters, including the usage of the backspace key for corrections of skill-based typing slips could be calculated.

7 https://telegram.org, last visited March 4th 2019
8 http://maps.stamen.com/toner-lite/#16/48.2188/16.3565, last visited February 14th 2019
9 https://coconut.cosy.wien, last accessed March 4th, 2019
10 https://www.polar.com, last visited March 4th 2019

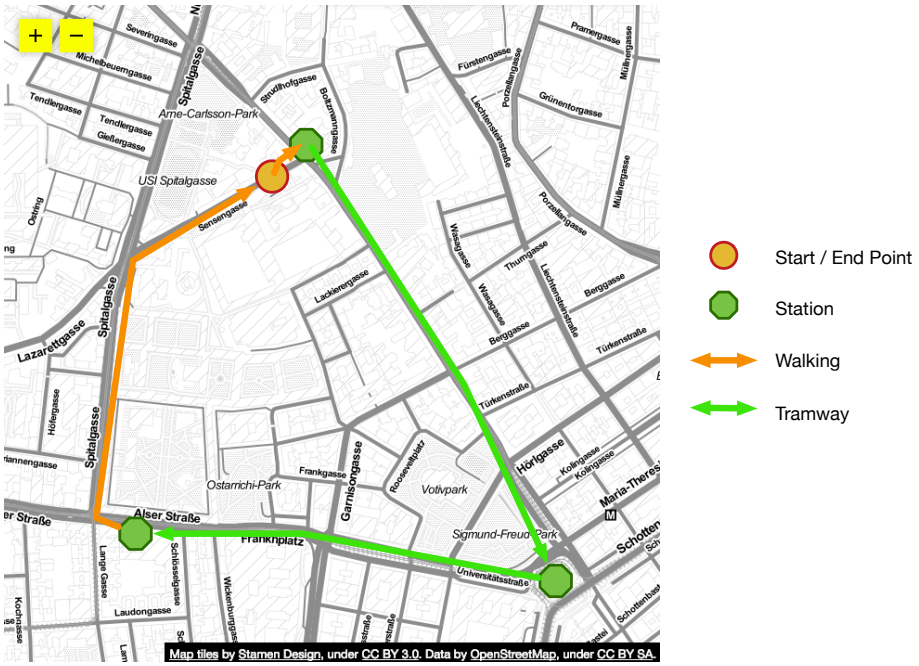


Figure 6.5: Route the participants had to take. It consisted of walking on secure sideways, standing on two stations and waiting for the tramways, and taking two tramways.

Study Setup

Participants were recruited via university courses and bulletin notes. As compensation, either points for their university course or 10€ were offered. Participants were invited for a CoConUT (sensor app) feasibility study of one hour. The COSY:Lab of the COSY research group was the starting point for all participants. They were invited separately into the lab and given a brief introduction. Also, they had to fill in a General Data Protection Regulation (GDPR) compliant consent form about data usage. Subsequently, they were asked to put on a chest belt (an H10 heart rate sensor by the brand Polar[11]). The operator then prepared the study smartphone by starting the CoConUT app, connecting the chest belt and starting a new recording session. Participants were told to use the CoCoQuest app for instructions on which way to go. During the route, CoCoQuest delivered short texts with directions. Additionally, participants could use Google Maps. Shortly before the participants went outside, the operator started the chatbot, which

11 https://www.polar.com/en/products/accessories/H10_heart_rate_sensor, last visited February 14th 2019

started delayed by four minutes. Even if the participants took some moments to get on their way, there was enough time to pretend realistic communication. After they returned, participants were asked to fill in a subsequent questionnaire.

6.2.4 *Results*

In this subsection, the results of the study will be described. Data evaluation and visualization were done by using R[12] and RStudio[13].

On average participants took about 15 minutes to complete the outdoor part, with an sd of 1.8 minutes. The single phases had the following durations (indicated as mean $\pm$ sd, rounded to seconds): Station 1 (162 $\pm$ 76), Tram 1 (192 $\pm$ 35), Station 2 (212 $\pm$ 137), Tram 2 (201 $\pm$ 41), Walking (520 $\pm$ 57).

Participants

Overall, 44 participants (35m / 9f) took part in the study in November and December 2018. They were between 18 and 38 years old ($\mu = 26.64$ and $\sigma = 4.91$) and most of them studied computer science at the university, while the rest were university staff members and externals. All of them were experienced with smartphones as well as instant messaging apps and noticed the messages. However, only about 50% realized they were communicating with a bot. 27.3% were unsure, and 22.7% did not realize. When being asked how much the chatbot messages stressed them, the trend was going towards not being stressed (mean $\mu = 3.95$, sd $\sigma = 1.18$ on a 5-point Likert scale where 1 meant "very stressed" and 5 "not stressed at all").

Participants were tested during the daytime between 9 am and 7 pm. Due to winter, some of the participants had to commute after sunset. Temperatures were moderately cold and ranging between 0 °C and 12 °C.

When being asked whether they stopped for typing on their smartphone, or also typed during walking, 41 out of 44 said that they also type during walking. Out of those 41, some indicated limitations: Two participants wrote that they only walked when it was safe. One person said they typed during walking, but did not have a look at the screen. Other limitations were: many people nearby, crossing a street, or if the quality of the typed text or the conversation was important.

12 https://www.r-project.org, last visited February 15th 2019
13 https://www.rstudio.com, last visited February 15th 2019

			Sensor Data	
Value	*mean μ*	*sd σ*	*Min*	*Max*
Speed (over GPS)	6.95	9.42	0	52.56
Light (lux)	851.8	1841.26	0	32767
# Bluetooth Dev.	12.56	12.27	0	76
Interaction	0.29	0.89	0	9
GPS Accuracy (m)	7.96	17.86	3	700

Table 6.2: Table with summary about assessed sensor values of the CoConUT app.

Another question asked about if there are any situations in which the participants preferred to write text messages via SMS or Instant Messenger. Nine persons wrote that they like to write text messages in public transportation. Reasons were increased privacy due to an avoided voice call, the idling time, or the seated position. A few others said they wrote text messages during waiting or when they were bored.

Data Collection

Unfortunately, not all 44 data sets recorded in the study were complete. In one case, CoConUT (sensor app) failed to record. For five participants, the data had to be excluded since the HR recordings were incomplete or missing. Two other participants did not use the keyboard at all, and one changed from the modified keyboard to another software keyboard, so that in three cases typing data is missing.

The smartphone was measuring sensor data every second. Table 6.2 shows the mean and standard deviation of some sensors measuring the users' surroundings. Other sensor data, like HRV and BPM measurements over the chest belt, will be described later on.

For reference values, those sensor levels can be compared with section 6.1.

To distinguish between the different categories, data point ranges were allocated manually according to their geospatial location and the speed patterns. The sensor data had to be split into the three context categories "walking", "station", and "tramway". To achieve splitting the points, geospatial polygons were laid over a map and data points were sorted according to their coordinates matching

the respective polygons. As can be seen in Figure 6.6, there were five polygons the data points could fall into two tramway paths, two stations, and one walking route (the original route can be seen in Figure 6.5). Since the calculation of HR and errors requires ranges of data points, start and end points of the different categories in the data set were determined manually, referring to the polygon matches, speed values and knowledge of the route. For example, if the speed pattern in the data indicated that a participant had passed two stations by tram, it was likely that he/she reached the destination.

Figure 6.7 shows the differences between specific sensor readings or calculations in-between the different contexts.

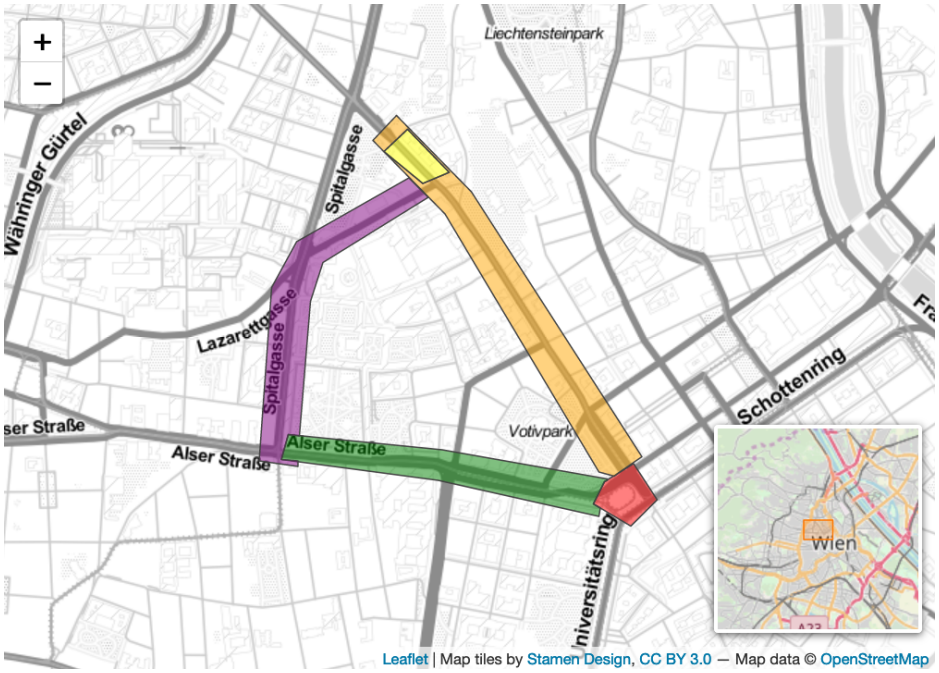
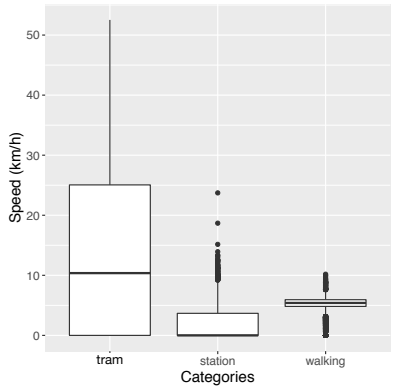


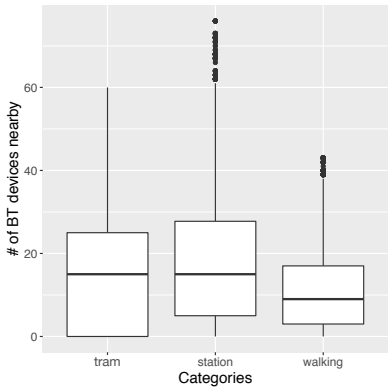
Figure 6.6: The polygons on the map used for sorting the data points into categories. The orange and green polygons are the parts of the routes to be taken with tramways, while the yellow and red forms are the two stations the participants had to wait on. Finally, the purple polygon was the part of the route to be walked by the participants.

HRV and Stress

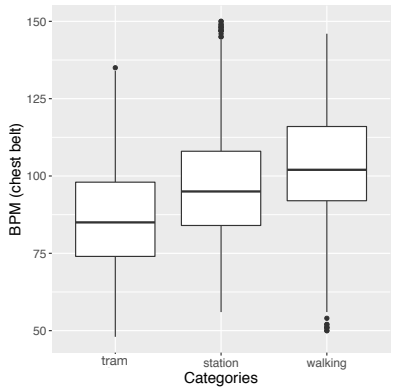
The following results base on 37 data sets with HRV measurements from the chest belt. For evaluation of the HRV, the RR intervals were taken from the CoCoNUT



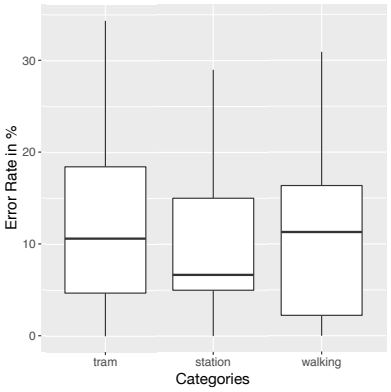
(a) Speed in km/h measured via GPS



(b) # of visible Bluetooth devices nearby

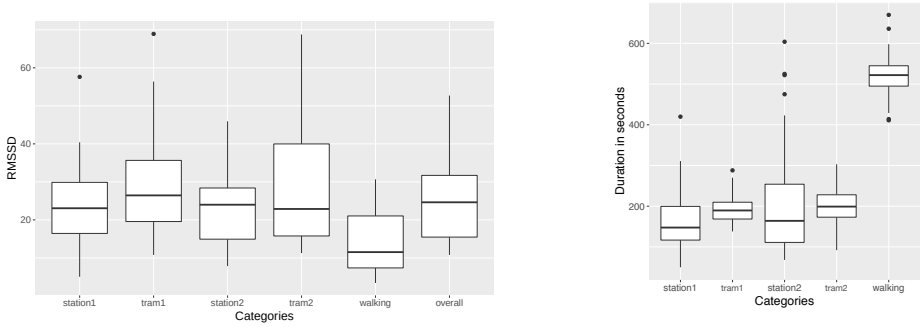


(c) Beats per minutes (BPM) from the chest belt across categories



(d) Mean and sd of error ratio in percent across the categories.

Figure 6.7: Boxplots of sensor measurements across the different categories



(a) RMSSD values (mean and sd) across categories.

(b) Mean and sd of the underlying measurement durations.

Figure 6.8: RMSSD values and their underlying measurement durations across categories.

file. Due to delays in the connection between the chest belt and the smartphone, some data points were doubled in the data set. Reproducibly, this only happened for delays $> 1000ms$, so the doubled data points had to be deleted. The resulting list of RR intervals was then analyzed using the RHRV library for R[14]. As already has been described in subsection 6.2.4, one outlier had to be removed. The RMSSD value of this participant was clearly outside the apparent distribution of the data, having an overall value of 100, whereas all other values had a mean of $\mu = 26$ and $\sigma = 11.12$.

The chest belt also provided aggregated data about beats per minute (BPM). Results regarding BPM can be seen in Figure 6.7c.

An overview of the different RMSSD values across the different categories can be seen in Figure 6.8a. Since measurement lengths are important for the calculation of the RMSSD, Figure 6.8b depicts the mean and sd of the underlying lengths of data ranges. For obtaining the best accuracy for calculation, a measurement duration of five minutes (300 seconds) for the RMSSD is advised. A lower RMSSD value indicates a higher level of stress. As a reference for the RMSSD, users aged 20 to 29 have an average RMSSD value of approximately 43 ± 19 [Ume+98].

Interaction and Typing Errors

Due to the onset of the bot after four minutes, few participants received their first message on the first tramway station. Most of the participants started responding to the bot in the first tramway. Table 6.3 shows statistics about error rates during the different categories. The mean and sd of interaction (measured as overall

14 http://rhrv.r-forge.r-project.org/, last visited February 20th, 2019

Category	*mean*	*sd*	*median*	*min*	*max*
Station	9.13	7.16	6.67	0	29.00
Tramway	12.04	9.30	10.64	0	34.36
Walking	10.55	8.91	11.30	0	30.92

Table 6.3: Mean, sd, median, min and max of error rates in percent for different parts of the route

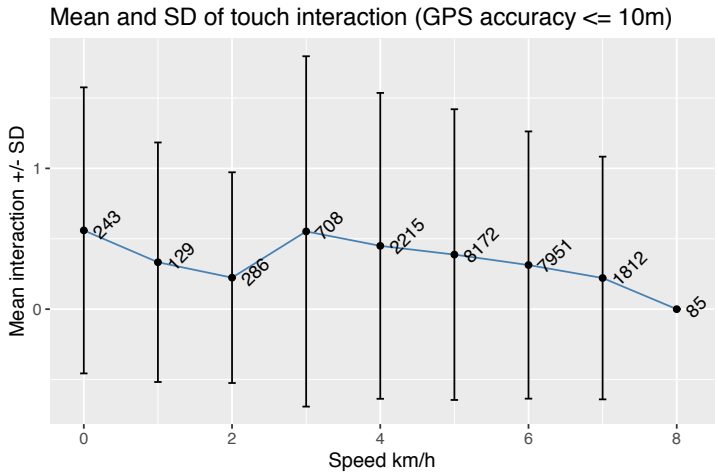


Figure 6.9: Mean and sd of screen touch interactions per second over walking speed for the *Walking* phase. The numbers of touches for each speed are specified in the graph. Only GPS values with a sufficient accuracy of less than 10 meters were taken into account

touches on the screen) can be seen in Figure 6.9. For calculation, the speed values obtained through the GPS were summed up according to their rounded value. There were no statistically significant differences between the means as determined by a one-way ANOVA ($F(2,103) = .899$, $p = .41$).

Figure 6.10 finally shows the connection between RMSSD and error ratio throughout the whole experiment. Pearson's product-moment correlation yields a negative correlation of -0.37 with a significance of $p = 0.03$.

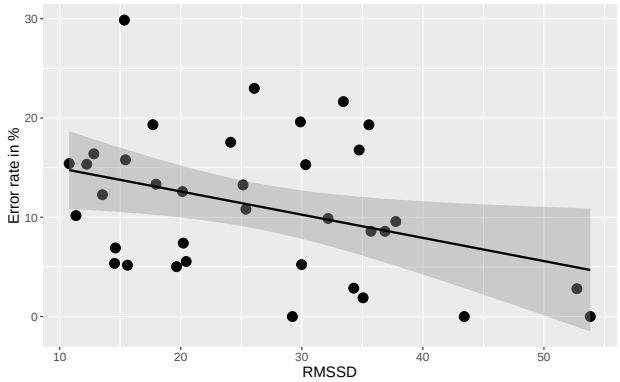


Figure 6.10: In this figure the relation between RMSSD and error ratio in % over the whole span of the experiment is shown.

6.2.5 *Discussion*

In this chapter, the results from the field study are discussed. Overall, nearly all participants answered the chat messages, despite not being told to do so beforehand. Though, after a while, a majority noticed that a bot was sending the messages. Still, nearly all continued to answer. Thus, semi-realistic chatting behavior could be assessed.

Interaction and Errors

According to our results, users do not slow down to type on their smartphones during walking, indicating that interaction frequency is steady across potential walking speeds (see Figure 6.4), which is in line with a previous study [SHR18]. Also, participants subjectively stated that they rarely stop for typing and instead type during walking, which indicates a particular typing expertise. Surprisingly, typing during walking did not significantly raise the error ratio above the other categories (see Table 6.3). The slightly higher error rates than those presented in [Dha+18] (6.31% $\pm$ 4.48) can be attributed to the unfamiliar test smartphone the participants had to type with.

Stress

Users that are more stressed tend to cause more typing errors (see Figure 6.10). This assumption can be explained with a higher overall workload and fewer mental resources for the typing task. Also, walking seems to cause more stress

than the other categories. Although users do not slow down for typing, even on sidewalks, they still have to pay attention to their surroundings. This multitasking probably causes a higher level of stress, which can also explain the lower level of stress during the tramway phases: once users are in the tramway, they do not have to pay attention to their surroundings anymore, only when approaching a station. Surprisingly, the error rate during the tramway phase was the highest (albeit not being statistically significant).

Of course, our results have certain limitations, since our user group was quite homogeneous. A heterogeneous group might yield different findings. However, for the calculation of stress levels, user homogeneity is positive [Ume+98]. The time spans for calculating short-term RMSSD have not always had the suggested five minutes length, which is required for comparability in medical settings. Also, ECG segments of differing lengths are usually not comparable [MA06]. Since in the field, standardized intervals are not always feasible, we put up with slightly less precise outcomes.

Sensors and Context

As can be seen in subsection 6.2.4 and Figure 6.2.4, the data recordings provided by CoConUT were reliable in nearly all cases but failed in few. Especially the chest belt sometimes lost its connection, which is unsurprising due to the field setting demanding a great deal of the equipment. The data could be used as-is, with only minor cleaning.

6.2.6 *Conclusion*

In this section, we presented a semi-realistic field study to assess the influence of context and stress on typing errors in the field. Our results show that people do not slow down for typing on their smartphone during walking and to which extent people are more likely susceptible to typing slips when they are walking. Finally, the users' base stress level influences the number of typing errors, which increases when stress is increasing.

Contribution (RQ1)

Accuracy of current smartphone sensors

1. An overview of the gathered sensor data is presented in a second field study.

Contribution (RQ2)

Users do not slow down for typing anymore

1. Results from the previous section have been confirmed.

Contribution (RQ2)

Context and the user's stress level influence typing slips

1. The more stressed a user, the higher the error rate.
2. When the user has more to multitask or is engaged in physical activity (for example during walking) due to the context, the error rate increases.

Overall, in this chapter, the outcomes of two subsequent field studies have been described. While the first study was exploratory, the second study managed to take up approaches and findings from the first study and go deeper into the research subject. While some findings could be reproduced in the second study, also new insights could be gained.

In the next chapter, the outcomes from all four laboratory and field studies from chapters 5 and 6 will be discussed with regards to the related work presented in chapter 2.

7

Discussion

This thesis addresses three research questions concerning human behavior in the field regarding mobile human-device interaction (see section 1.1). While these research questions aim at assessing mobile device interaction in the field using field studies, some subquestions had to be addressed in the laboratory due to methodological reasons. However, as a first main result, a framework for supporting mobile field studies was developed, and several studies in the field as well as in the lab were conducted.

In this chapter, the conducted research is reflected upon. The discussion runs along three major strands: lessons that can be learned for mobile field study researchers and developers, habituation of mobile device usage, and a reflection on errors in mobile interaction. This chapter concludes with some remarks about the assessment of biophysical measurements in the field.

7.1 Supporting Mobile Field Studies

This section reflects on the conceptualization and development of the CoConUT framework. Remember that the CoConUT framework was developed to address the Research Question 1, which asks how context, the user's internal state (as an indicator for mental factors) and interaction (especially errors) can be assessed in field studies. It puts particular foci on the kinds of data that can be assessed, the accuracy of data, as well as how the collected data can be visualized and analyzed.

The CoConUT framework could successfully demonstrate that a combination of Open-Source software and support of affordable consumer devices designed for field applications is the ideal support for mobile field studies.

Although some Open-Source frameworks for supporting mobile field studies, or also for collecting sensor data with the smartphone, exist (for example the AWARE Framework or Google's Science Journal, as can be seen in section 4.2), none of those fully fulfilled the requirements posed by this thesis (as can be seen in Table 4.2).

In chapter 4, the process of conceptualizing and developing the CoConUT framework to assess context (mainly by smartphone sensors), the internal state (by measuring biophysical data and collecting qualitative feedback from the user) and interaction (by measuring touches on the screen or typing behavior) during field studies were described. A particular focus was put on the usage of Open-Source and open hardware.

As has been shown in two field studies (see section 6.1 and section 6.2), the data collected by our framework demonstrates to be accurate and reproducible (see Table 6.1 and Table 6.2). With the timestamps collected by all CoConUT modules, the gathered data from all apps and wearables could retrospectively be merged for evaluation purposes after a study. When assessed directly on the smartphone, data collection was very reliable. Measurement problems only occurred when connecting to external devices via Bluetooth caused problems.

For visualization purposes, data can either be visualized on the small screen directly during or after a study in the CoConUT sensing app. The CoCoVis dashboard allows for a more detailed exploration and visualization of the study data after a study.

Among the range of realized CoConUT modules, two open hardware projects for making wearables to support mobile field studies were evaluated as not successful, namely the CoCoHat and the CoCoBand. While the idea of open hardware powered by open software is intriguing, neither device lived up to the posed expectations. Self-built devices proved to be less reliable, less accurate, and even more expensive than their consumer device counterparts. The camera and microphone recording capabilities of the CoCoHat could easily be realized with a modern wearable action camera. The interaction with the device itself can be recorded via screencasting, which the CoConUT sensing app supports. The biophysical measurements with the CoCoBand also did not prove to live up to their consumer counterparts. Modern chest belts for athletes even do a better job in measuring heart rate.

Referring back to the best practices mentioned by Roto, Vätäjä, Jumisko-Pyykkö, and Vänänen-Vainio-Mattila in subsection 2.1.1 [Rot+11], it can be summarized that the CoConUT framework supports each of the mentioned steps. Especially during the data collection phase, CoConUT addresses most of the posed best practices with great care: CoConUT is very unintrusive (best practice *C1*) while collecting most of its data directly on the smartphone. It is possible to record a broad variety of multimedia data about the context (*C2*), as well as seamlessly integrating all recorded data into one dataset (*C4*). Also, using questionnaires and experience probing, subjective data can easily be assessed (*C4*). Collecting subjective data also helps participants to give their feedback in a rather discreet way directly in the field or shortly after that (*C5* and *C6*). Best practice *C3* is not applicable here, because it can only be addressed by study design.

The approach by Möller, Westermann, Beyer, and Reichmuth can be fully confirmed: quantitative data should be augmented with user opinion data to cover the broadest amount of knowledge that can be gathered in the field [Möl+14]. All in all, it can be said, that the approach of blending rich qualitative with detailed quantitative data gathered in the field from an all-in-one solution proves to be reliable and suited for a broad range of field study purposes.

7.2 Habituation of Mobile Device Usage

For addressing interaction, context, and internal state in the field, two field studies were conducted. The underlying Research Question 2 asked, which kinds of contextual factors influence mobile interaction in the field. Furthermore, a subquestion referred to the kinds of contexts that have an impact on the user's primary interaction task.

Our studies showed that people do not slow down anymore for typing on their device. Quantitative as well as qualitative data could confirm this finding: While people did not slow down in their walking speed and kept a steady typing rate according to the sensor measurements, they also stated explicitly that all of them type during walking outside of the study (see subsection 6.2.4). This fact indicates a strong habituation that has happened in the past ten to 15 years since smartphone usage has become widespread.

In the studies, contexts were pre-chosen to be able to control the number of potential contexts. Since mental resources are only indirectly measurable, the users' heart rates were assessed to be able to calculate their stress levels in the field. Since in the field, the users can be in an indefinite amount of contexts, and every direct inquiry (for example by an in-situ workload questionnaire) can bias the outcomes, the pre-selection of relevant contexts (here: commute) and the indirect measurement of heart rate were chosen as operationalizations.

Regarding different contexts in the field, behavior changes depending on the surroundings, and the available attention for the interaction task. While on empty sidewalks, only a few disturbances are expected, users still have to walk actively, which negatively impacts their focus and subsequently rises their typing error rate. In the tramway, no action has to be taken by the users, which can (almost) completely concentrate on typing on their smartphone. In the station, all different kinds of actions have to be taken, and there is a high number of people nearby. Surprisingly, probands in our second field study had a relatively low rate of errors in the station setting. Probably multitasking was not possible, and they only answered when they were in a standing and calm position.

Although users do not slow down anymore, let alone stop completely, it still stresses them to type during walking (as stress measurements could show by a higher RMSSD). Apparently, today, people can multitask during their commute, but it still stresses them out to take other actions during typing. This conclusion also goes hand in hand with the findings by Vidulich and Tsang, who state that the workload of the secondary task is more meaningful for assessing overall mental workload. During multitasking, it is not the primary task that stresses us out, but the secondary task that keeps sliding into the foreground periodically [VT12].

Of course, typing during walking can be relatively safe, when, for example, done on a quiet, long sidewalk without much interference. Much research has been done on more risky use cases like, for instance, driving and aviation [EK15]. However, also everyday situations can be dangerous and remain relevant for the question at hand, as more people are crossing roads than experts piloting airplanes.

As could be seen in Figure 2.2, the role of habituation during multitasking has been well-researched. Persons become increasingly better at handling multitasking as they become better and more habituated in the different tasks [SS12; EK15; VT12]. This circumstance can directly be applied to our multitasking scenario of interaction with the device and monitoring of surroundings in the field. It remains a challenge to *"allocate attention to the right things at the right time"* [Fer14] due to sensory flooding in today's multimedia world, but at the same time, habituation

makes it gradually easier for us to choose the right stimuli to focus our attention on. With our focused attention breaking up into more and more divided attention, at the same time, training and expertise in mobile device usage seem to make up for the more rapid nature of today's networked world. Training can reduce the required resources and lead to a reduced mental workload [VT12], but still – the need for multitasking increases, and it remains cumbersome.

Summarizing our findings, it can be said that in the past 10-15 years a strong habituation of mobile device usage has happened. In the following, less mental resources are required nowadays for multitasking in the field during, for example, textual communication. This contrasts with the findings by Oulasvirta, Tamminen, Roto, and Kuorelahti, who in 2005 found out that people paid attention to their mobile device on a long quiet street only 20% of the time, and that they often had to slow down or completely stop to walking to continue interaction with their mobile device [Oul+05] (see chapter 2).

7.3 Understanding Errors in Mobile Interaction

A particular focus in the work at hand was put at the occurrence of errors on different levels. Research Question 3 asks when errors of different types do happen in the field and how these types of errors can be assessed. Furthermore, a subquestion addresses their potential impact on secure communication. The work at hand can successfully demonstrate the impact of different kinds of errors on mobile communication.

As noted in the introduction, this research is based on the human error classification by Reason [Rea00], who distinguishes between knowledge-based mistakes, rule-based mistakes, and skill-based slips. All of these errors can happen during mobile human-device interaction, but not all of them make sense to assess in the field. Skill-based slips are relatively easy to assess in detail in the field because they happen on a small scale level. Regarding rule-based mistakes, we also decided to test them in the laboratory, because we wanted to better understand the impact of stress on rule-based mistakes. While skill-based slips on smartphones are well-researched in the field (in the form of typing errors on smartphones' software keyboards), rule-based errors are lacking this broad background. Since knowledge-based mistakes need a broad extent of qualitative data (interviews,

detailed video analysis), these large-scale errors are best to be assessed in a laboratory setting as well.

Knowledge-based mistakes often stem from erroneous mental models. This circumstance can be especially fatal with regards to system security: While mental models are incomplete and inaccurate by nature, several attack scenarios exist that exploit these inaccuracies directly [Was10]. These exploits happen for a reason since our research showed that even persons with a technical background and moderate knowledge of security failed to take up the right mitigation strategies, even after detecting an attack. Those users had high trust in secure software and, at the same time, a false sense of security. While our attack explicitly aimed at exploiting a piece of the software (in this case the mutual key verification in the messaging app Signal), this shows that bad usability of high-risk security features can lead to the ultimate compromise: users noticing an attack, taking up mitigation strategies, failing unknowingly and lulling themselves in a false sense of security.

One level below, our findings indicate that stress could play a role in the occurrence of rule-based errors. It is assumed that the higher the stress level is, the higher the error ratio gets. Further research in that direction has yet to be done. Finally, skill-based slips are influenced by the user's stress level as well, as also by the context in the field. This type of error also occurs more frequently when the stress level rises. Additionally, heavy multitasking and physical activity seem to increase the error rate. Regarding rule-based errors, we are expecting the error ratio to rise in straining contexts in the field as well.

As demonstrated above, contextual factors and the user's internal state (in this case, stress) influence the occurrence of errors on the go. The impact of the different errors on the user's security in this case varies. While of course at first glance, knowledge-based mistakes seem to have the most impact, very small scale touch interaction slips can have a considerable impact as well, for example, while typing a password or clicking the wrong button. This fact is also supported by literature: stress affects the interaction with the smartphone, and the error rate rises [CWG15]. A certain level of skill expertise as well as a sufficient knowledge level can be gained by training and generally reduce error rates [Dha+18], but under stress the error rate per se increases, no matter which type of error is involved.

Reason also states that there are the system approach and the person approach when looking at errors [Rea00] — solely blaming the user as "the weakest link in the security chain" [SBW01] for making errors has never paid off so far in the long run. On the other hand, the "system" of human-smartphone interaction

in the field is extremely diverse and completely unpredictable during software planning and design, especially when planning which errors could occur. There is a need for better inclusion of this volatility during mobile system design, for example using certain security precautions depending on context and the user's current activity, while at the same time not restricting the user in their interaction to prevent errors. This work contributes to a better understanding of errors in mobile interaction, but more research is urgently needed.

7.4 Assessing Biophysical Signals in the Field

On a concluding note, the matter of assessing biophysical signals in the field will be discussed. While this has not been explicitly included in one of the research questions in section 1.1, measuring signals from the human body to infer the internal state of a user has been a particularly interesting aspect of this thesis.

Our attempts to build open hardware to measure biophysical signals have neither proven to be reliable nor accurate. Additionally, the costs of hardware and development time do not justify the effort in comparison to consumer devices. While the preferable type of devices for the best accuracy would be medical grade devices, like those being used in doctor's practices or hospitals, these devices remain unattainable with regards to costs and are slightly technically outdated due to the required certification processes (for example with the Technischer Überwachungsverein (TÜV)[1] in Germany and Austria). Consumer devices, like the chest belts used in this work for athletes, remain the most affordable and feasible solution at the moment, especially for low-cost Open-Source projects.

Collecting biophysical signals in the field remains a challenge. First of all, signals measured from the human body are quite small. Measuring itself is a challenge with certain signals, and additionally, all kinds of interferences and biophysical differences from person to person have to be taken into account. Even assessing those signals in the laboratory can be challenging, despite static contexts and without the test person moving around. Thus, measuring biophysical signals in the field is a broad field yet to explore, with many electrophysical and evaluative challenges to solve. With augmented measuring techniques, which are applicable in the field, new input and output modalities could become available, stimulating further applied research in the fields of biofeedback or self-optimization.

1 https://www.vdtuev.de, last visited May 4th, 2019

Last but not least, there are certain algorithmic considerations concerning the work at hand. In the evaluative parts, HRV plays a significant role, particularly the measure RMSSD, which is defined as the *"Root Mean Square of the Successive Differences"*. While all HRV measures are clearly *"biased estimates"* [MA06], some seem to be more meaningful for certain bodily functions than others. According to its formula, the RMSSD is a measure that increases when the time intervals between successive heartbeats highly deviate. The more homogeneous the time spans, the lower the RMSSD gets. A certain amount of literature suggests using the natural logarithm of the RMSSD, namely the LNRMSSD, as a measure of choice for ultra short-term HRV calculations, especially in settings with professional athletes [Ple+13; EF14; MA06; EFN17]. While the application of a natural log happens to shift the range of potential values to a more easily understandable range, the conclusion prevails that simply the RMSSD is a meaningful measure for ultra-short term measures as well. Still, the RMSSD remains a meaningful measure for subjective stress, also in safety-critical working environments [Ors+08].

After having discussed the findings of our research, the next chapter will conclude this thesis and outline future work.

8

Conclusions and Future Work

The motivation for this thesis was to revisit the fundamentals of mobile interaction on mobile devices in the field with a particular focus on different kinds of errors. Since the introduction of modern smartphones nearly a decade ago, the way users interact with their mobile devices has changed. Even important conversations are increasingly happening while being on the move, for example during the commute. In such situations, users cannot solely focus on the mobile task at hand but have to monitor their surroundings as well. While crossing a busy street, for example, just exclusively attending to the smartphone would be potentially fatal. Our limited mental resources force us to multitask, which potentially leads to the internal state of stress. Summarizing, contextual influences, and the user's internal state influence mobile interaction.

For assessing this interplay between mobile interaction, context, and internal state, field studies have been chosen as the prevalent method. None of the existing frameworks for supporting mobile field studies could fulfill the posed requirements, which leads to **Research Question 1: How can context, the user's internal state (as an indicator for mental factors) and interaction (especially errors) be assessed in field studies? Particularly, what kind of data can be assessed and how (quantitative or qualitative, surroundings, or users themselves)? How accurate is the data? Moreover, how can the assessed data be visualized and analyzed?**

The CoConUT framework, which is described in detail in chapter 4, addresses this research question. In this chapter, the shortcomings of related software are listed, requirements for a field study framework are gathered and condensed into

a concept. At the heart of the framework, the CoConUT sensing app gathers contextual information through sensors and through the help of connectable wearables. The CoCoQuest study guide app guides the probands through a field study and allows assessment of contextual questionnaires and qualitative experience probing in-situ. The assessed data can be visualized and explored in the CoCoVis visualization dashboard. Further components have either been discarded or were solely developed for a single study: The CoCoHat is a wearable in the form of a hat, offering the possibility to record the surroundings and the user's interaction via video and audio. The CoCoBand is a wrist-worn wearable that measures the user's heart rate in beats per minute, on-skin temperature, and galvanic skin response. Finally, the CoCoBot is a chatbot which emulates a conversation over instant messaging, and the CoCoBoard is a modified software keyboard, that enables the logging of every keypress on the keyboard.

Furthermore, in section 5.1.7, the self-built hardware for measuring biophysical signals CoCoBand is compared with consumer devices. The CoCoBand, as well as the CoCoHat, were both discarded in favor of more affordable, robust, and reliable consumer wearables. Our studies have shown that in our case consumer wearables for supporting mobile field studies outperform self-built open hardware.

Overall, the CoConUT framework has been successfully validated in several studies and proves to be reliable. In this work, two overviews of gathered sensor data with the CoConUT sensing app give an impression of the data quality and accuracy (see chapter 6).

Based on this framework, a particular focus is put on the occurrence of errors. This work examines errors on three levels: knowledge-based mistakes, rule-based mistakes, and skill-based slips. All of these types of errors can disrupt mobile communication. **Research Question 3** addresses the occurrence of errors: **When do which kinds of usage errors happen in the field? In particular, how can those types of errors be assessed? Moreover, what is their potential impact on secure communication?**

In chapter 5, two laboratory studies are described, which explore the occurrence of knowledge-based and rule-based mistakes. Both studies address aspects that cannot be examined in the field. In the first study, the amount of qualitative data could not have been assessable in the field, while for measurements, participants had to remain seated in the second study.

In the first laboratory study (section 5.1), knowledge-based mistakes are explored using an extensive qualitative analysis of man-in-the-middle attack mitigation with the Android instant messaging app Signal. Probands were invited

under the pretense of participating in a usability analysis of Signal, during which a secret MITM attack was launched. Qualitative analysis revealed erroneous mental models and therefore, incomplete or merely false knowledge about the underlying mechanisms so that the majority of users chose false mitigation strategies following the attack. This behavior led to compromised security after the attack, while the users found themselves in a false sense of security. Apparently, users have very high trust in secure apps. Coupled with bad usability of high-risk features, this can lead to non-solvable security problems.

Furthermore, the occurrence of rule-based mistakes under stress is tested in the laboratory (see section 5.1.7). Here, the goal was to examine how the error-rate of rule-based mistakes changes under varying degrees of stress. Since several measurement methods were tested as well in this trial, probands had to remain seated and ideally not move at all. This necessity is the reason why the study took place in the laboratory. Probands underwent a customized stress test, in which they had to mentally calculate easy to hard mathematical tasks, while partially being monitored by the operator to induce stress. The mental arithmetic tasks emulated rule-based decisions. The outcomes of the study show signs that the higher the stress, the higher the error ratio (see subsection 5.2.6). Outcomes of the comparison of stress measurement devices have already been described above.

Skill-based slips are examined in section 6.2. Since slips during typing are relatively easy to assess via quantitative data, a field study was conducted. Additionally, the results from section 5.1.7 identified consumer devices for biophysical measurements and heart rate as an efficient means for mobile assessment of stress indicators. The outcomes of this study are described below in more detail.

Finally, **Research Question 2** looks at the interplay of different aspects in the field: **Which kind of contextual factors do influence mobile interaction in the field and to which degree? Which role does the user's internal state play? In particular, which kind of contexts have an impact on the user's primary HCI task?**

Two consecutive field studies are presented in chapter 6. During both field studies, participants used a provided smartphone on which the apps CoConUT and CoCoQuest ran.

In the first, exploratory field study, participants had to walk a pre-given route and chat with the operator over the Zom instant messenger.

In the second, more extensive field study (see section 6.2), probands had to take a commute-like, pre-given route (walking, taking the tramway, with a provided

smartphone. During the route, a chatbot started to chat with them over the instant messaging app Telegram, which led to semi-realistic chatting behavior. Stress was measured using a chest belt.

Both studies show that today, users do not slow down for typing during walking outdoors anymore (see chapter 6). Their typing ratio stays constant irrespective of their walking speed. This fact indicates that, over the past decade, apparently a strong habituation with smartphones has happened and the expertise of the users has increased.

Furthermore, findings of the second field study show that context and the stress level influence typing slips (see subsection 6.2.6): the more stressed a user is, the higher the error rate of typing slips. When the user is engaged in physical activity (for example steady walking) or has to multitask, the error rate rises as well. Despite the increased error rate, users still keep on moving while typing on their mobile devices. These outcomes of the second field study contribute to both Research Questions 2 and 3.

Summarizing, the work at hand provides *contributions to gain an insight into mobile interaction in the field after the widespread adoption of mobile devices in the past decade. In doing so, it puts a particular focus on the occurrence of different types of errors and highlights their potential influence on mobile communication.*

The findings of the research described so far suggest multiple potential directions for future work.

First, the CoConUT sensing app will be further developed and enhanced with several new features. A new mobile visualization will provide new ways to explore data on the go visually, and potentially also offer built-in statistical features to aid the operator. Operators could, for example, gain a more profound overview of the gathered data and make decisions depending on these insights. Furthermore, machine learning features could be integrated to enable the system to learn about specific environmental characteristics. The activity or the state of the user, and certain contextual features could lead to the automatic provision of in-situ notifications or triggered questions. Of course, machine learning would require a large body of gathered and tailored data, which would have to be assessed first.

Concerning errors, integrated studies that investigate the co-occurrence of different types of errors are necessary, especially in the field. While mental models lead to task-solving strategies, those strategies are carried out in different steps of smaller-scale if-then-rules ("*if* this occurs *then* do this"), whose execution relies on skills like typing skills. How these different forms of layers interact during the occurrence of human-device interaction errors remains yet to be seen. Especially in the field, under different contextual influences, this has to be investigated.

Another interesting point for further research is the fact that users do not stop for typing anymore in the field. While this obviously can lead to fatal outcomes due to decreased attention on the surroundings, it also denotes the need for electronic communications means while being on the move. Since texting while being on the steering wheel is forbidden, being on the phone is possible as long as both hands can remain at the steering wheel. According to Wickens, attention can be split when not being in the same attentional dimension [WM07]: hearing a phone conversation while watching the road is not as risky as splitting visionary attention between the road and the device screen. For communicating while walking more effective means could be striven for, like better text-to-speech recognition, or more usable tools for voice mail recording.

Overall, it can be said that communication has changed drastically over the last decades, and will continue to change even more in the future. Human-Computer Interaction as a discipline has to keep up and provide usable and safe tools for mobile communication to keep humans connected.

A

Appendix

A.1 CoConUT Online Survey

Expert Survey about Evaluation Methods in Mobile Field Studies

Welcome and thank you for participating in this study

In this questionnaire we want to assess how researchers in the field of Human Computer Interaction (HCI) evaluate data gathered in mobile field studies.

This study will take approximately 5-10 minutes. All gathered data will be handled anonymously and only be used for scientific purposes.

Thank you very much for participating!

If you have any questions concerning this questionnaire, please feel free to contact:

Svenja Schröder M.Sc.
Email: svenja.schroeder@univie.ac.at
Phone: +43-1-4277-79422

Demographic Questions

1. Age
   _____
2. Gender
   O Male
   O Female
   O Other: _____
3. What is your highest degree?
   O High school (or equivalent)
   O Bachelor (or equivalent)
   O Master (or equivalent)
   O PhD/Doctorate
   O Habilitation
   O Other: _____
4. I work in…
   O Research
   O Industry
   O Other: _____
5. Which subject(s) is(are) your expertise in?
   _____
6. What's your position?
   _____
7. Do you conduct user studies? If yes, how many have you already conducted?
   O No
   O Yes, I've done a handful
   O Yes, on a regular basis
   O Other: _____

Questions about Mobile Field Studies

1. Do you have experience with short--term or long--term field studies?
   O Short-term (several minutes up to several hours)
   O Long-term (several hours up to several weeks)

O Both
O Other: _____

2. Why do you conduct mobile field studies?
   O Gathering big-picture insights
   O User group only accessible in the field
   O Testing under realistic conditions
   O Ethnographic research in groups
   O Other: _____

3. Which methods do you use?
   ☐ Questionnaires
   ☐ Data collection on the device (Logs)
   ☐ Experience Sampling
   ☐ Video and sound recording
   ☐ Screen Recording
   ☐ Interviews
   ☐ Other: _____

4. Which tools do you usually use to evaluate your study data?
   ☐ SPSS
   ☐ R
   ☐ Matlab
   ☐ atlas.ti
   ☐ Mathematica
   ☐ Python + Libraries
   ☐ SageMath
   ☐ Excel
   ☐ Other: _____

5. How many tools do you usually use?
   O 1
   O 2
   O 3
   O 4
   O 5
   O 6
   O 7
   O 8
   O more than 8

6. Which statistical methods are you employing?
   _____
   _____
   _____
7. Which of them do you find extremely insightful?
   _____
   _____
   _____
8. Please describe a prototypical mobile field study you recently conducted (if applicable, short outline is enough)
   _____
   _____
   _____

Improvements in User Study Evaluation

1. What do you miss in the evaluation software you usually use?
   _____
   _____
   _____
2. Does your workflow include interactive visual analysis, e.g. dashboard interfaces?
   O Yes
   O No
   Other: _____
3. If yes, which tools do you use that provide such dashboards?
   ☐ Your own solution
   ☐ Tableau
   ☐ Microsoft Power Bi
   ☐ Other: _____

Just one final question...

Would you be interested in participating in an iterative design process for a new evaluation tool regarding mobile field studies? If yes, please leave your email address here, or directly write an email to the coordinator of the study.

1. Email address:

   _____

Thank you!

Thank you for participating in this study.

If you have any questions concerning this questionnaire, please feel free to contact:

Svenja Schröder M.Sc.
Email: svenja.schroeder@univie.ac.at
Phone: +43-1-4277-79422

A.2 CoConUT Expert Interview Guideline

- Which kind of mobile field studies do you conduct / are conducted in your group?
  - Which methods do you use?
- Which kind of data do you collect?
  - Always / sometimes / never? (Logs, video, audio, user input, sensors, …)
  - How reliable is this data? Is this important?
- Do you preprocess your data before you start your evaluation?
  - Which steps are necessary and why?
- With how many and which tools do you perform your evaluation? Which function does each tool fulfill?
  - How do you combine your data? How do you detect correlations? (Get specific over data and evaluation)
- Which statistical methods are you employing?
  - Which do you find insightful?
  - ANOVA?
- What do you miss in the evaluation software you usually use?
  - Specific needs / wishes?
- Does your workflow include interactive visual analysis, e.g. dashboard interfaces?

A.3 Overview over different Biophysical Signals to Measure

Cowley et al. define a signal as a *"real-time data stream supplied by a sensor"* and differentiate into *internal* (autonomic nervous system), *external* (ocular system, remote) and *combined* signals to measure [Cow+16]. There are different ways of measuring these signals [Sch15] and some properties of human bodies can be measured in multiple ways. For example, the human heart beat can be captured either by listening on the chest with a stethoscope, by optical signals directly on the skin or by measuring the electrical signals coming from the heart with an ECG. Schmidt lists types of biosignals which are relevant in the field of HCI [Sch15]:

Bio-electrical: Signals from muscles and nerves

Electrical conductance: Conductance in or on the body, e.g. on the skin (e.g. galvaniv skin response)

Bio-impedance: Resistance while applying an harmless, alternating current to tissue

Bio-acoustic: Sounds that originate in the body (heartbeat, lung ventilation, etc.)

Bio-optical: Changes to the body that can be captured by special cameras (blood flow, changes to skin color, etc.)

When it comes to specific biosignals to measure, Cowley et al. describe the most common biosignals, which are being used in HCI [Cow+16]:

Electrocardiogram (ECG) An ECG measures the electrical signals that originate from the cardiovascular system, which means the activity of the human heart. Measurements usually are non-invasive and relatively easy to obtain.

Galvanic Skin Response (GSR) GSR measures electrodermal activity, which means changes in the skin's electrical properties. Changes in the autonomic nervous system cause the sweat glances to be more or less activated which changes the conductivity of the skin. These changes in conductivity can be measured.

Eyetracking Eyetracking, or ocular tracking externally measures movements and dilation of the human pupils. Usually this bio-optical signal is recorded by a camera and interpreted by software. Eye movements allow to reconstruct the course of the gaze across for example a website. Attention and intention can be inferred. Micro-movements, also called micro-saccades, permit the researchers to draw conclusions about cognitive activity.

Breathing Breathing or respiration is an interesting biosignal to measure because it can be either consciously controlled by the central nervous system, but is also automatically influenced by the autonomous nervous system. Also speech and movement create artifacts during measurements. Breathing can be measured either non-invasive with a chest strap (which is imprecise) or by invasive techniques like breathing through a tube. Sometimes in HCI it is used in biofeedback methods [Fre+18].

Electroencephalography (EEG) EEG denotes the monitoring of brain activity through non-invasive electrodes connected to the head's skin. It measured summarized activity in clusters of neurons. There is a broad variety of EEG systems, which differ in their scope and quality: from one-channel customer solutions to high-end medical devices with high resolution many systems are available. The user's cognitive processes and internal states can be assessed relatively reliably with the right tools, but most high-end systems require a certain set-up time (for example lubricating and placing the electrodes under the user's hair on the skin).

Electromyography (EMG) EMG non-invasively measures electronic signals from the muscles. Electrodes on the skin surface detect contractions of muscles and deliver an according signal. Since contractions of certain muscles highly correlate with certain internal states, certain emotions can be estimated over EMG (for example contracted muscles in the forehead are a sign for anger).

A.4 CoConUT Class Diagram

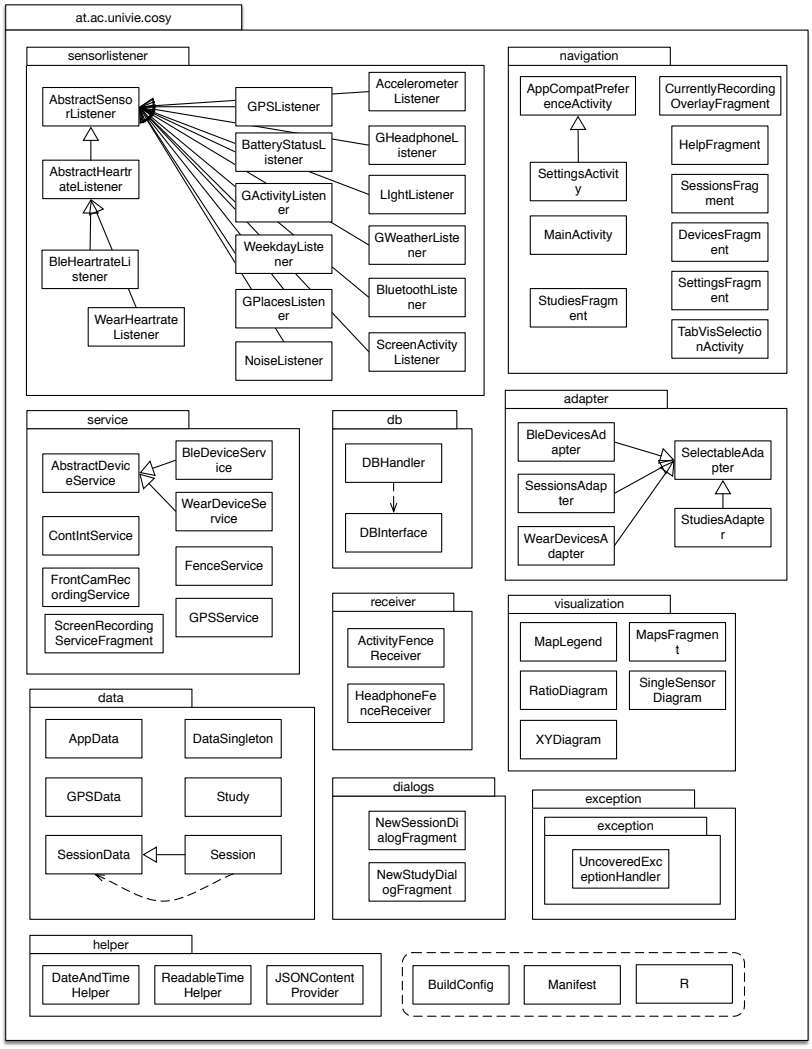


Figure A.1: Class diagram of the CoConUT sensing app

A.5 Usability Evaluation of Signal

The study conducted in section 5.1 also contained a usability evaluation of the Signal app with focus on Signal's instant messaging and security features. The following usability improvements to contribute towards an enhanced usable security experience for Signal can be suggested:

Awareness on security status of conversations: Conversations can only be assumed to be properly end-to-end encrypted once Alice's (the user's) and Bob's (the conversational partner's) Identity Keys were successfully verified. Signal does not remember the verification status — only point-in-time verifications are possible and the user has to remember whom of their partners they already verified. Signal thus lacks mechanisms to quickly assess the security status of a conversation. Such a security status should be directly visible in the corresponding conversation.

Comprehensible instructions for recommended actions: In order to avoid risky behavior, especially in the verification and attack mitigation process, users should be provided with clear instructions respectively suggestions for actions. On the key comparison page users with no exact knowledge of asymmetric encryption mechanisms failed to act on the displayed information. In our opinion, a brief instructional message combined with optional further information would have led to a higher verification success rate (e.g. *"Please contact your partner outside the app to compare your Identity Keys. If the Identity Keys do not match, please consult the FAQ or contact the developers."*). We found that this issue is most pressing for the Android version of Signal. The iOS version of Signal provides brief information on how to verify users: *"Compare both fingerprints to verify your contact's identity and the integrity of the message"*. However, no information is provided on how to proceed in case of failure (fingerprint mismatch).

Clear risk communication: On the other hand Signal should inform users of the possible consequences of their actions. E.g. during the process of accepting Bob's identity after the attack the denomination of the buttons ("Verify" and "Accept") was misleading. Under the false assumption that the mitigation process would lead to a verification of Bob, users failed to have a clear understanding of the risks.

Easily accessible verification: The verification options should be easily accessible in the menu. A suggestion would be to add a shortcut for the verification mechanism directly to the conversation in order to maximize visibility.

With the implementation of improvements as suggested above, Signal would instantly gain a huge factor of both usability as well as security.

Ich versichere an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Svenja Schröder, MSc BSc, Vienna
Saturday 29th June, 2019

Bibliography

[Abu+17] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith.
"Obstacles to the Adoption of Secure Communication Tools".
In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 137–153.

[Aha+11] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. "Social fMRI: Investigating and Shaping Social Mechanisms in the Real World".
In: *Pervasive Mob. Comput.* 7.6 (2011), pp. 643–659. ISSN: 1574-1192. DOI: 10.1016/j.pmcj.2011.09.004. URL: http://dx.doi-org.uaccess.univie.ac.at/10.1016/j.pmcj.2011.09.004.

[AIM10] L. Atzori, A. Iera, and G. Morabito.
"The Internet of Things: A Survey".
In: *Computer Networks* 54.15 (2010), pp. 2787–2805.

[AT13] W. Albert and T. Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Newnes, 2013.

[Bax16] G. Baxendale. "Health Wearables".
In: *ITNOW* 58.3 (2016), pp. 42–43.

[Bel75] R. W. Belk. "Situational Variables and Consumer Behavior".
In: *Journal of Consumer Research* 2.3 (1975), pp. 157–164.

[BGB04] N. Borisov, I. Goldberg, and E. Brewer.
"Off-The-Record Communication, Or, Why not to use PGP".
In: *Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society*. ACM. 2004, pp. 77–84.

[BJ15] A. Bali and A. S. Jaggi.
"Clinical Experimental Stress Studies: Methods and Assessment".
In: *Reviews in the Neurosciences* 26.5 (2015), pp. 555–579.

[BKJ18] O. Barral, I. Kosunen, and G. Jacucci. "No Need to Laugh Out Loud: Predicting Humor Appraisal of Comic Strips Based on Physiological Signals in a Realistic Environment". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 24.6 (2018), p. 40.

[Bla16] M. Blanchou. *Android-SSL-TrustKiller*. https://github.com/iSECPartners/Android-SSL-TrustKiller. 2016.

[Böh+11] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. "Falling Asleep with Angry Birds, Facebook and Kindle: A Large-Scale Study on Mobile Application Usage". In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM. 2011, pp. 47–56.

[BPS11] J. Bakker, M. Pechenizkiy, and N. Sidorova. "What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data". In: *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE. 2011, pp. 573–580.

[Bra+10] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. "Bridging the Gap in Computer Security Warnings: A Mental Model Approach". In: *IEEE Security & Privacy* 2 (2010), pp. 18–26.

[Bru+99] M. C. d. Bruyne et al. "Both Decreased and Increased Heart Rate Variability on the Standard 10-Second Electrocardiogram Predict Cardiac Mortality in the Elderly: the Rotterdam Study". In: *American Journal of Epidemiology* 150.12 (1999), pp. 1282–1288.

[Bur05] R. G. Burgess. *In the Field: An Introduction to Field Research*. Taylor & Francis e-Library, 2005.

[Cac13] P. C. Cacciabue. *Guide to Applying Human Factors Methods: Human Error and Accident Management in Safety-Critical Systems*. Springer Science & Business Media, 2013.

[Cam+96] A. Camm et al. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation and Clinical Use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology". In: *Circulation* 93.5 (1996), pp. 1043–1065.

[Can28] W. B. Cannon.
"The Mechanism of Emotional Disturbance of Bodily Functions".
In: *New England Journal of Medicine* 198.17 (1928), pp. 877–884.

[Che53] E. C. Cherry. "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *The Journal of the Acoustical Society of America* 25.5 (1953), pp. 975–979.

[Cho+17] K.-H. Choi, J. Kim, O. S. Kwon, M. J. Kim, Y. H. Ryu, and J.-E. Park.
"Is Heart Rate Variability (HRV) an Adequate Tool for Evaluating Human Emotions?–A Focus on the Use of the International Affective Picture System (IAPS)".
In: *Psychiatry Research* 251 (2017), pp. 192–196.

[Cin+13] B. Cinaz, B. Arnrich, R. Marca, and G. Tröster.
"Monitoring of Mental Workload Levels During an Everyday Life Office-Work Scenario".
In: *Personal and Ubiquitous Computing* 17.2 (2013), pp. 229–239.

[CKM+94] S. Cohen, T. Kamarck, R. Mermelstein, et al. "Perceived Stress Scale".
In: *Measuring Stress: A Guide for Health and Social Scientists* (1994), pp. 235–283.

[CKM83] S. Cohen, T. Kamarck, and R. Mermelstein.
"A Global Measure of Perceived Stress".
In: *Journal of Health and Social Behavior* (1983), pp. 385–396.

[CMF18] R. Cassani, M.-A. Moinnereau, and T. H. Falk.
"A Neurophysiological Sensor-Equipped Head-Mounted Display for Instrumental QoE Assessment of Immersive Multimedia".
In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2018, pp. 1–6.

[Cor16] A. Cortesi. *mitmproxy*. https://mitmproxy.org/. 2016.

[Cow+16] B. Cowley et al.
"The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human-Computer Interaction".
In: *Foundations and Trends® in Human-Computer Interaction* 9.3-4 (2016), pp. 151–308.

[Cra08] L. F. Cranor.
"A Framework for Reasoning About the Human in the Loop."
In: *UPSEC* 8 (2008), pp. 1–15.

[CTB07] J. T. Cacioppo, L. G. Tassinary, and G. Berntson.
Handbook of Psychophysiology. Cambridge University Press, 2007.

[CWG15] M. Ciman, K. Wac, and O. Gaggi. "iSenseStress: Assessing Stress Through Human-Smartphone Interaction Analysis".
In: *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*.
ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2015, pp. 84–91.

[DAS01] A. K. Dey, G. D. Abowd, and D. Salber.
"A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-aware Applications".
In: *Hum.-Comput. Interact.* 16.2 (Dec. 2001), pp. 97–166.
ISSN: 0737-0024. DOI: 10.1207/S15327051HCI16234_02.
URL: http://dx.doi.org/10.1207/S15327051HCI16234_02.

[DD05] F. Domahs and M. Delazer.
"Some Assumptions and Facts About Arithmetic Facts".
In: *Psychology Science* 47.1 (2005), pp. 96–111.

[de 18] J. de Wilde.
"Using Wearable Device Sensors to Determine Psycho-Physiological States During Short-Term Mobile Field Studies".
Master Thesis at the University of Vienna.
MA thesis. University of Vienna, 2018.

[DG11] A. C. Dirican and M. Göktürk. "Psychophysiological Measures of Human Cognitive States Applied in Human-Computer Interaction".
In: *Procedia Computer Science* 3 (2011), pp. 1361–1367.

[Dha+18] V. Dhakal, A. M. Feit, P. O. Kristensson, and A. Oulasvirta.
"Observations on Typing from 136 Million Keystrokes".
In: *Conference on Human Factors in Computing Systems-Proceedings*.
Vol. 2018. 2018.

[DJ+12] M. van Dooren, J. H. Janssen, et al.
"Emotional Sweating Across the Body: Comparing 16 Different Skin Conductance Measurement Locations".
In: *Physiology & Behavior* 106.2 (2012), pp. 298–304.

[DP15] T. Dingler and M. Pielot. “I’ll be there for you: Quantifying Attentiveness towards Mobile Messaging”. In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM. 2015, pp. 1–5.

[Dum17] S. Dumbs. *Conceptualization and Implementation of a Framework for Assessing Contexual Experience During Field Studies*. Bachelor Thesis at the University of Vienna. 2017.

[EF14] M. R. Esco and A. A. Flatt. “Ultra-Short-Term Heart Rate Variability Indexes at Rest and Post-Exercise in Athletes: Evaluating the Agreement with Accepted Recommendations”. In: *Journal of Sports Science & Medicine* 13.3 (2014), p. 535.

[EFN17] M. R. Esco, A. A. Flatt, and F. Y. Nakamura. “Agreement Between a Smartphone Pulse Sensor Application and Electrocardiography for Determining lnRMSSD”. In: *The Journal of Strength & Conditioning Research* 31.2 (2017), pp. 380–385.

[EK15] M. W. Eysenck and M. T. Keane. *Cognitive Psychology: A Student’s Handbook*. Psychology Press, 2015.

[Eng+17] U. Engelke et al. “Psychophysiology-based QoE assessment: A survey”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2017), pp. 6–21.

[Fal+16] M. Fallahi, R. Heidarimoghadam, M. Motamedzade, and M. Farhadian. “Psycho Physiological and Subjective Responses to Mental Workload Levels during N-Back Task”. In: *J Ergonomics* 6.181 (2016), p. 2.

[FCS12] A. Fry, S. Chiasson, and A. Somayaji. “Not Sealed But Delivered: The (un) Usability of S/MIME Today”. In: *Annual Symposium on Information Assurance and Secure Knowledge Management (ASIA’12), Albany, NY*. 2012.

[Fer14] A. Ferscha. “Attention, Please!” In: *IEEE Pervasive Computing* 13.1 (2014), pp. 48–54.

[FG14] S. H. Fairclough and K. Gilleade. *Advances in Physiological Computing*. Springer, 2014.

[FH02] A. Field and G. Hole. *How To Design and Report Experiments*. Sage, 2002.

[FKD15] D. Ferreira, V. Kostakos, and A. K. Dey. "AWARE: Mobile Context Instrumentation Framework". In: *Frontiers in ICT* 2 (2015), p. 6.

[Fre+18] J. Frey, M. Grabli, R. Slyper, and J. R. Cauchard. "Breeze: Sharing Biofeedback Through Wearable Technologies". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 645.

[Fro+07] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. "MyExperience: A System for In Situ Tracing and Capturing of User Feedback on Mobile Phones". In: *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services*. ACM. 2007, pp. 57–70.

[Fro+16] T. Frosch, C. Mainka, C. Bader, F. Bergsma, J. Schwenk, and T. Holz. "How Secure is TextSecure?" In: *2016 IEEE European Symposium on Security and Privacy*. IEEE. 2016, pp. 457–472.

[Gar+05] S. L. Garfinkel, D. Margrave, J. I. Schiller, E. Nordlander, and R. C. Miller. "How To Make Secure Email Easier to Use". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2005, pp. 701–710.

[GD86] D. Gopher and E. Donchin. "Workload: An Examination of the Concept". In: *Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance*. Oxford, England: John Wiley & Sons, 1986, pp. 1–49.

[GS14] D. Gentner and A. L. Stevens. *Mental Models*. Psychology Press, 2014.

[GTC16] S. Greene, H. Thapliyal, and A. Caban-Holt. "A Survey of Affective Computing for Stress Detection: Evaluating Technologies in Stress Detection for Better Health". In: *IEEE Consumer Electronics Magazine* 5.4 (2016), pp. 44–56.

[HGT13] S. A. Hoseini-Tabatabaei, A. Gluhak, and R. Tafazolli. "A Survey on Smartphone-Based Systems for Opportunistic User Context Recognition". In: *ACM Computing Surveys (CSUR)* 45.3 (2013), p. 27.

[Hir16] J. Hirschl. *Bursted Attention: Collecting Contextual and Interaction Data in Mobile Field Studies.* Bachelor Thesis at the University of Vienna. 2016.

[HKC10] S. A. Hosseini, M. A. Khalilzadeh, and S. Changiz. "Emotional Stress Recognition System for Affective Computing Based on Bio-Signals". In: *Journal of Biological Systems* 18.spec01 (2010), pp. 101–114.

[HMM04] R. M. Hamilton, P. S. Mckechnie, and P. W. Macfarlane. "Can Cardiac Vagal Tone be Estimated from the 10-Second ECG?" In: *International Journal of Cardiology* 95.1 (2004), pp. 109–115.

[HMP15] J. Hernandez, D. McDuff, and R. W. Picard. "Biowatch: Estimation of Heart and Breathing Rates From Wrist Motions". In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015 9th International Conference on.* IEEE. 2015, pp. 169–176.

[HRB12] N. Henze, E. Rukzio, and S. Boll. "Observational and Experimental Investigation of Typing Behaviour Using Virtual Keyboards for Mobile Devices". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM. 2012, pp. 2659–2668.

[HS88] S. G. Hart and L. E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in psychology.* Vol. 52. Elsevier, 1988, pp. 139–183.

[HSK09] J.-y. Hong, E.-h. Suh, and S.-J. Kim. "Context-Aware Aystems: A Literature Review and Classification". In: *Expert Systems with Applications* 36.4 (2009), pp. 8509–8522.

[Hua+12] K. Huang, C. Zhang, X. Ma, and G. Chen. "Predicting Mobile Application Usage Using Contextual Information". In: *Proc. Int'l Conf on Ubiquitous Computing (UbiComp)* (2012), pp. 1059–1065. DOI: 10.1145/2370216.2370442.

[JD06] B. Jürgen and N. Döring. *Forschungsmethoden und Evaluation für Human und Sozialwissenschaftler, 4. überarb. Aufl.* Heidelberg: Springer, 2006.

[Jer+91] S. Jern, M. Pilhall, C. Jern, and S. G. Carlsson. "Short-Term Reproducibility of a Mental Arithmetic Stress Test". In: *Clinical Science* 81.5 (1991), pp. 593–601.

[JM+06] M. Jones, G. Marsden, et al. *Mobile Interaction Design.* West Sussex, England: John Wiley & Sons, 2006.

[Joh83] P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* 6. Harvard University Press, 1983.

[Jon+11] N. Jones, H. Ross, T. Lynam, P. Perez, and A. Leitch. "Mental Models: An Interdisciplinary Synthesis of Theory and Methods". In: *Ecology and Society* 16 (Jan. 2011), pp. 46–46.

[Jul19] C. N. Julius. *Stressanalyse und Stresserkennung mittels HRV Biofeedback anhand einer Feedback Applikation.* Bachelor Thesis at the University of Vienna. 2019.

[JV10] S. Jumisko-Pyykkö and T. Vainio. "Framing the Context of Use for Mobile HCI". In: *International Journal of Mobile Human Computer Interaction* 2.4 (Oct. 2010), pp. 1–28.

[KG03] J. Kjeldskov and C. Graham. "A Review of Mobile HCI Research Methods". In: *International Conference on Mobile Human-Computer Interaction.* Springer. 2003, pp. 317–335.

[Kje+04] J. Kjeldskov, M. B. Skov, B. S. Als, and R. T. Høegh. "Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field". In: *Mobile Human-Computer Interaction - MobileHCI 2004: 6th International Symposium, MobileHCI, Glasgow, UK, September 13 - 16, 2004. Proceedings.* Ed. by S. Brewster and M. Dunlop. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 61–73. ISBN: 978-3-540-28637-0. DOI: 10.1007/978-3-540-28637-0_6. URL: http://dx.doi.org/10.1007/978-3-540-28637-0_6.

[KMY12] P. Karthikeyan, M. Murugappan, and S. Yaacob. "A Study on Mental Arithmetic Task Based Human Stress Level Classification Using Discrete Wavelet Transform". In: *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*. IEEE. 2012, pp. 77–81.

[KPH93] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer. "The 'Trier Social Stress Test'–A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting". In: *Neuropsychobiology* 28.1-2 (1993), pp. 76–81.

[KS14] J. Kjeldskov and M. B. Skov. "Was It Worth the Hassle?: Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations". In: *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*. MobileHCI '14. Toronto, ON, Canada: ACM, 2014, pp. 43–52. ISBN: 978-1-4503-3004-6. DOI: 10.1145/2628363.2628398. URL: http://doi.acm.org/10.1145/2628363.2628398.

[Lan+10] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. "A Survey of Mobile Phone Sensing". In: *IEEE Communications Magazine* 48.9 (2010).

[Lav05] N. Lavie. "Distracted and Confused?: Selective Attention Under Load". In: *Trends in Cognitive Sciences* 9.2 (2005), pp. 75–82.

[Lav10] N. Lavie. "Attention, Distraction, and Cognitive Control under Load". In: *Current Directions in Psychological Science* 19.3 (2010), pp. 143–148.

[LB13] P. Lopes and P. Baudisch. "Muscle-Propelled Force Feedback: Bringing Force Feedback to Mobile Devices". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 2577–2580.

[Leb17] K. Lebloch. *Enhancing and Analysing the Sensing Capabilities of a Wearable Mobile Field Testing Unit Towards Assessing the Impact of Stress on Mobile User Behaviour*. Bachelor Thesis at the University of Vienna. 2017.

[Lew82] C. Lewis. *Using the "Thinking-Aloud" Method in Cognitive Interface Design*. IBM TJ Watson Research Center, 1982.

[LFH17] J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction.* Morgan Kaufmann, 2017.

[LMP+12] P. Le Callet, S. Möller, A. Perkis, et al. "Qualinet White Paper on Definitions of Quality of Experience". In: *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)* 3 (2012).

[Lup16] D. Lupton. *The Quantified Self.* John Wiley & Sons, 2016.

[Lv16] M. Lv. *ProxyDroid.* https://github.com/madeye/proxydroid. 2016.

[MA06] J. McNames and M. Aboy. "Reliability and Accuracy of Heart Rate Variability Metrics versus ECG Segment Duration". In: *Medical and Biological Engineering and Computing* 44.9 (2006), pp. 747–756.

[Mal+96] M. Malik et al. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use". In: *European Heart Journal* 17.3 (1996), pp. 354–381.

[Men99] M. Mendl. "Performing Under Pressure: Stress and Cognitive Function". In: *Applied Animal Behaviour Science* 65.3 (1999), pp. 221–244.

[Mie+02] J. Mietus, C. Peng, I. Henry, R. Goldsmith, and A. Goldberger. "The pNNx Files: Re-examining a Widely Used Heart Rate Variability Measure". In: *Heart* 88.4 (2002), pp. 378–380.

[Min+11] T. Minh, T. Do, J. Blom, and D. Gatica-perez. "Smartphone Usage in the Wild : A Large-Scale Analysis of Applications and Context". In: *ICMI '11 Proceedings of the 13th international conference on multimodal interfaces* (2011), pp. 353–360. DOI: 10.1145/2070481.2070550.

[Möl+14] S. Möller, T. Westermann, J. Beyer, and R. Reichmuth. "Exploring User Behavior and Preferences in the Wild with Mobile Apps: Lessons Learned from Four Case Studies". In: *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS).* IEEE. 2014, pp. 532–538.

[Nie+06] C. M. Nielsen, M. Overgaard, M. B. Pedersen, J. Stage, and S. Stenild. "It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field". In: *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles*. NordiCHI '06. Oslo, Norway: ACM, 2006, pp. 272–280. ISBN: 1-59593-325-5. DOI: 10.1145/1182475.1182504. URL: http://doi.acm.org/10.1145/1182475.1182504.

[Nor83] D. A. Norman. "Design Rules Based on Analyses of Human Error". In: *Communications of the ACM* 26.4 (1983), pp. 254–258.

[Nor86] D. A. Norman. "Cognitive Engineering". In: *User Centered System Design* 31 (1986), p. 61.

[NSB10] D. Nunan, G. R. Sandercock, and D. A. Brodie. "A Quantitative Systematic Review of Normal Values for Short-Term Heart Rate Variability in Healthy Adults". In: *Pacing and Clinical Electrophysiology* 33.11 (2010), pp. 1407–1417.

[OD16] L. Onwuzurike and E. De Cristofaro. "Experimental Analysis of Popular Anonymous, Ephemeral, and End-to-End Encrypted Apps". In: *Proceedings of the 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. Darmstadt, Germany, 2016.

[Ope16] Open Whisper Systems. *Signal Messenger*. online. https://whispersystems.org. 2016.

[Ors+08] R. Orsila et al. "Perceived Mental Stress and Reactions in Heart Rate Variability—A Pilot Study Among Employees of an Electronics Company". In: *International Journal of Occupational Safety and Ergonomics* 14.3 (2008), pp. 275–283.

[Oul+05] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti. "Interaction in 4-Second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2005, pp. 919–928.

[Oxf18] Oxford University Press. *Interaction - Definition of Interaction in English by Oxford Dictionaries*. 2018. URL: https://en.oxforddictionaries.com/definition/interaction (visited on 02/02/2018).

[Pic03] R. W. Picard. “Affective Computing: Challenges”. In: *International Journal of Human-Computer Studies* 59.1-2 (2003), pp. 55–64.

[Pic99] R. W. Picard. “Affective Computing for HCI”. In: *HCI (1)*. Citeseer. 1999, pp. 829–833.

[Pla18] S. Plank. *Konzeption und Entwicklung einer Dokumentations- und Stresserkennungs-App für den ÖFAST im Rahmen der Feuerwehrausbildung*. Bachelor Thesis at the University of Vienna. 2018.

[Ple+13] D. J. Plews, P. B. Laursen, J. Stanley, A. E. Kilding, and M. Buchheit. “Training Adaptation and Heart Rate Variability in Elite Endurance Athletes: Opening the Door to Effective Monitoring”. In: *Sports Medicine* 43.9 (2013), pp. 773–781.

[PR97] R. Parasuraman and V. Riley. “Humans and Automation: Use, Misuse, Disuse, Abuse”. In: *Human Factors* 39.2 (1997), pp. 230–253.

[PRR06] G. Pocock, C. D. Richards, and D. A. Richards. *Human Physiology*. Oxford University Press, 2006.

[PRS15] J. Preece, Y. Rogers, and H. Sharp. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, 2015.

[Rae+05] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. “ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications”. In: *IEEE Pervasive Computing* 2 (2005), pp. 51–59.

[Raw+15] R. Rawassizadeh, E. Momeni, C. Dobbins, P. Mirza-Babaei, and R. Rahnamoun. “Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices”. In: *Journal of Sensor and Actuator Networks* 4.4 (2015), pp. 315–335.

[RC02] M. B. Rosson and J. M. Carroll. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. Morgan Kaufmann, 2002.

[Rea00] J. Reason. “Human Error: Models and Management”. In: *Bmj* 320.7237 (2000), pp. 768–770.

[Rea90] J. Reason. *Human Error*. Cambridge university press, 1990.

[Rei+07] P. Reichl, P. Froehlich, L. Baillie, R. Schatz, and A. Dantcheva. "The LiLiPUT Prototype: A Wearable Lab Environment for User Tests of Mobile Telecommunication Applications". In: *CHI'07 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2007, pp. 1833–1838.

[Rei+15] P. Reichl et al. "Towards A Comprehensive Framework for QoE and User Behavior Modelling". In: *7th International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE. 2015, pp. 1–6.

[RJ74] J. Rasmussen and A. Jensen. "Mental Procedures in Real-Life Tasks: A Case Study of Electronic Trouble Shooting". In: *Ergonomics* 17.3 (1974), pp. 293–307.

[RM05] K. Ryu and R. Myung. "Evaluation of Mental Workload with a Combined Measure Based on Physiological Indices During a Dual Task of Tracking and Mental Arithmetic". In: *International Journal of Industrial Ergonomics* 35.11 (2005), pp. 991–1009.

[RM14] R. Reichmuth and S. Möller. "Classification of the Context of Use for Smart Phones". In: *International Conference on Human-Computer Interaction*. Springer. 2014, pp. 638–642.

[Rob11] C. Robson. *Real World Research: A Resource for Users of Social Research Methods in Applied Settings (3rd Edition)*. West Sussex: John Wiley & Sons, 2011.

[Rog+07] Y. Rogers et al. "Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp". In: *International Conference on Ubiquitous Computing*. Springer. 2007, pp. 336–353.

[Rot+11] V. Roto, H. Vätäjä, S. Jumisko-Pyykkö, and K. Vänänen-Vainio-Mattila. "Best Practices for Capturing Context in User Experience Studies in the Wild". In: *In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (2011), pp. 91–98.

[Rot+15] C. Rottermanner, P. Kieseberg, M. Huber, M. Schmiedecker, and S. Schrittwieser. "Privacy and Data Protection in Smartphone Messengers". In: (2015).

[Rot06] V. Roto.
Web Browsing on Mobile Phones: Characteristics of User Experience.
Helsinki, Finland: Helsinki University of Technology, 2006.

[Rus80] J. A. Russell. "A Circumplex Model of Affect".
In: *Journal of Personality and Social Psychology* 39.6 (1980), p. 1161.

[RVR14] K. Renaud, M. Volkamer, and A. Renkema-Padmos.
"Why Doesn't Jane Protect Her Privacy?"
In: *Privacy Enhancing Technologies*. Springer. 2014, pp. 244–262.

[Sah+14] A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt. "Large-Scale Assessment of Mobile Notifications".
In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 3055–3064.

[Sau16] L. SaurikIT. *Cydia Substrate*. http://www.cydiasubstrate.com. 2016.

[SAW+94] B. N. Schilit, N. Adams, R. Want, et al.
Context-Aware Computing Applications.
Xerox Corporation, Palo Alto Research Center, 1994.

[SBG99] A. Schmidt, M. Beigl, and H.-W. Gellersen.
"There is More to Context Than Location".
In: *Computers & Graphics* 23.6 (1999), pp. 893–901.

[SBW01] M. A. Sasse, S. Brostoff, and D. Weirich.
"Transforming the 'Weakest Link'— A Human-Computer Interaction Approach to Usable and Effective Security".
In: *BT Technology Journal* 19.3 (2001), pp. 122–131.

[Sch+12] S. Schrittwieser et al. "Guess Who's Texting You? Evaluating the Security of Smartphone Messaging Applications." In: *NDSS*. 2012.

[Sch+16] S. Schröder, M. Huber, D. Wind, and C. Rottermanner.
"When SIGNAL Hits the Fan: On the Usability and Security of State-of-the-Art Secure Mobile Messaging".
In: *European Workshop on Usable Security. IEEE*. 2016.

[Sch15] A. Schmidt. "Biosignals in Human-Computer Interaction".
In: *Interactions* 23.1 (2015), pp. 76–79.

[Sch16] M. Schulz. "Simulation des Interaktionsverhaltens von Senioren bei der Benutzung von mobilen Endgeräten".
PhD thesis. Berlin, Germany: Technical University Berlin, Feb. 2016.

[See18] C. Seebacher. *The Mechanics of Digital Paper Chase: GeoCaching under the Microscope*. Bachelor Thesis at the University of Vienna. 2018.

[Sel+36] H. Selye et al. "A Syndrome Produced by Diverse Nocuous Agents". In: *Nature* 138.3479 (1936), p. 32.

[Sel76] H. Selye. "Stress Without Distress". In: *Psychopathology of Human Adaptation*. Springer, 1976, pp. 137–146.

[Sen+69] L. A. Seneca et al. *Epistulae Morales ad Lucilium*. Vol. 210. Penguin UK, 1969.

[SG17] F. Shaffer and J. Ginsberg. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (2017), p. 258.

[SHR16] S. Schröder, J. Hirschl, and P. Reichl. "CoConUT: Context Collection for Non-stationary User Testing". In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '16. Florence, Italy: ACM, 2016, pp. 924–929. ISBN: 978-1-4503-4413-5. DOI: 10.1145/2957265.2962658. URL: http://doi.acm.org/10.1145/2957265.2962658.

[SHR18] S. Schröder, J. Hirschl, and P. Reichl. "Exploring the Interplay of Context and Interaction in the Field". In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2018, pp. 1–6.

[Smi+13] D. C. Smith, K. M. Schreiber, A. Saltos, S. B. Lichenstein, and R. Lichenstein. "Ambulatory Cell Phone Injuries in the United States: An Emerging National Concern". In: *Journal of Safety Research* 47 (2013), pp. 19–23.

[SN93] N. Staggers and A. F. Norcio. "Mental Models: Concepts for Human-Computer Interaction Research". In: *International Journal of Man-machine studies* 38.4 (1993), pp. 587–605.

[SP13] A. Sano and R. W. Picard. "Stress Recognition Using Wearable Sensors and Mobile Phones". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE. 2013, pp. 671–676.

[SRR19] S. Schröder, A. Rafetseder, and P. Reichl. "Errare Mobile Est: Studying the Influence of Mobile Context and Stress on Typing Errors in the Field". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2019.

[SS12] R. J. Sternberg and K. Sternberg. *Cognitive Psychology*. Wadsworth, Cengage Learning, 2012.

[Sta11] M. Stamp. *Information Security: Principles and Practice*. John Wiley & Sons, 2011.

[Sta99] I. O. for Standardization. *ISO 13407:1999, Human-Centered Design Processes For Interactive Systems*. Tech. rep. 1999.

[Swa12] M. Swan. "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0". In: *Journal of Sensor and Actuator networks* 1.3 (2012), pp. 217–253.

[Swa13] M. Swan. "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery". In: *Big data* 1.2 (2013), pp. 85–99.

[Sys15] O. W. Systems. *Open Whisper Systems Blog: Just Signal*. Nov. 2015. URL: https://whispersystems.org/blog/just-signal/.

[Tan+11] G. Tan, T. K. Dao, L. Farmer, R. J. Sutherland, and R. Gevirtz. "Heart Rate Variability (HRV) and Posttraumatic Stress Disorder (PTSD): A Pilot Study". In: *Applied Psychophysiology and Biofeedback* 36.1 (2011), pp. 27–35.

[The16a] The Electronic Frontier Foundation. *WhatsApp Rolls Out End-To-End Encryption to its Over One Billion Users*. online. https://www.eff.org/deeplinks/2016/04/whatsapp-rolls-out-end-end-encryption-its-1bn-users. Apr. 2016.

[The16b] The Intercept. *With Facebook No Longer a Secret Weapon, Egypt's Protesters Turn to Signal*. online. https://theintercept.com/2016/04/26/facebook-no-longer-secret-weapon-egypts-protesters-turn-signal/. Apr. 2016.

[Tim+17] M. A. Timmis, H. Bijl, K. Turner, I. Basevitch, M. J. D. Taylor, and K. N. van Paridon. “The Impact of Mobile Phone Use on Where We Look and How We Walk When Negotiating Floor Based Obstacles”. In: *Plos One* 12.6 (2017), pp. 1–20. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0179802. URL: https://doi.org/10.1371/journal.pone.0179802.

[Ume+98] K. Umetani, D. H. Singer, R. McCraty, and M. Atkinson. “Twenty-Four Hour Time Domain Heart Rate Variability and Heart Rate: Relations to Age and Gender over Nine Decades”. In: *Journal of the American College of Cardiology* 31.3 (1998), pp. 593–601.

[Ung+15] N. Unger et al. “SoK: Secure Messaging”. In: *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE. 2015, pp. 232–249.

[Val+15] M. T. Valderas, J. Bolea, P. Laguna, M. Vallverdú, and R. Bailón. “Human Emotion Recognition Using Heart Rate Variability Analysis With Spectral Bands Based on Respiration”. In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE. 2015, pp. 6134–6137.

[Van+10] B. Van Wissen, N. Palmer, R. Kemp, T. Kielmann, H. Bal, et al. “ContextDroid: An Expression-Based Context Framework for Android”. In: *Proceedings of PhoneSense* 2010 (2010).

[Vaz+17] E. Vaziripour et al. “Is That You, Alice? A Usability Study of the Authentication Ceremony of Secure Messaging Applications”. In: *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. 2017, pp. 29–47.

[Vaz+18] E. Vaziripour et al. “Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal”. In: *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*. 2018, pp. 47–62.

[Vaz+19] E. Vaziripour et al. “I Don’t Even Have to Bother Them!: Using Social Media to Automate the Authentication Ceremony in Secure Messaging”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 93.

[VT12] M. A. Vidulich and P. S. Tsang. "Mental Workload and Situation Awareness". In: *Handbook of Human Factors and Ergonomics* 4 (2012), pp. 243–273.

[Was10] R. Wash. "Folk Models of Home Computer Security". In: *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM. 2010, p. 11.

[Wei91] M. Weiser. "The Computer for the 21st Century". In: *Scientific American* 265.3 (1991), pp. 94–104.

[Wha16] WhatsApp Inc. *WhatsApp*. online. https://whatsapp.com. 2016.

[Wic02] C. D. Wickens. "Multiple Resources and Performance Prediction". In: *Theoretical Issues in Ergonomics Science* 3.2 (2002), pp. 159–177.

[Wic08] C. D. Wickens. "Multiple Resources and Mental Workload". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50.3 (2008), pp. 449–455.

[WM07] C. D. Wickens and J. S. McCarley. *Applied Attention Theory*. CRC press, 2007.

[WT99] A. Whitten and J. D. Tygar. "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0." In: *Usenix Security*. Vol. 1999. 1999.

[XZ15] Z. Xu and S. Zhu. "SemaDroid: A Privacy-Aware Sensor Management Framework for Smartphones". In: *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*. CODASPY '15. San Antonio, Texas, USA: ACM, 2015, pp. 61–72. ISBN: 978-1-4503-3191-3. DOI: 10.1145/2699026.2699114. URL: http://doi.acm.org/10.1145/2699026.2699114.

[Yu18] B. Yu. *Designing Biofeedback for Managing Stress*. PhD Thesis. Department of Industrial Design, Eindhoven University of Technology, 2018.

[ZB06] J. Zhai and A. Barreto. "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables". In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2006, pp. 1355–1358.