

Proxemo:

Documenting Observed Emotions in HCI



Inaugural-Dissertation
zur Erlangung der Doktorwürde der
Fakultät für Humanwissenschaften der
Julius-Maximilians-Universität Würzburg
vorgelegt von Stephan Huber aus Burgau

Würzburg
2022



This document is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0):
<http://creativecommons.org/licenses/by-nc/4.0> This CC license does not apply to third party material (attributed to another source) in this publication.

Erstgutachter: Professor Dr. Jörn Hurtiene
Zweitgutachterin: Professorin Dr. Carolin Wienrich
Tag des Kolloquiums: 29. März 2022

Acknowledgements

This dissertation was conducted at the Chair for Psychological Ergonomics at Julius-Maximilians-Universität Würzburg. I would like to take the opportunity to thank all the people who supported or inspired me on this journey.

Jörn, thank you for being the best imaginable supervisor, providing patience in the beginning and then encouragement and invaluable advice as I forged a connection between two application domains that appear absurdly unrelated at first sight. More importantly, I am thankful for the scent of freedom in research and allowing me to chase research questions beyond my thesis and letting me juggle with as many side projects as I could handle.

Carolin, thank you for assuring your support as the second supervisor for my thesis at an early stage, consulting on ethics and operationalisation in times of COVID and teaching me a thing or two about networking between academia and industry.

Tobi, thank you for the early scientific onboarding, facilitating the full Brissi experience with Penny and then being a role model in structure and scientific clarity.

My special gratitude goes to all **Psyergos** for making my PhD time in Geb82 so good that it became hard working towards putting an end to it. Thank you all for your input and support throughout all stages of this work, as well as great not-work related chitchat in between! With respect to this thesis, I want to thank **Franzi** and **Sara** for final comments on the manuscript and to the graphical heroines **Clara** and **Cordula** for performing visual miracles.

Among the partners and students, who had contact to my research, I first want to thank **Stefanie** for being the best and particularly most thorough experimenter imaginable, while journeying the wild, wild world of physio data. **Alex**, thank you for being the first Proxemo-early-adopter in reminiscence studies, facilitating data collection and even creating the first port for Wear OS, thus signalling me that the demand of observation methods which I had identified was not imaginary. **Jan Erik**, thank you for the quest to master Tizen even when the path led you through South Korean tech forums and the depths of Samsung Support. Furthermore, I want to thank **Beate**, **Patricia** and all **InterMem partners** for early testing of Proxemo, as well as **Eliana** and all air traffic controllers, supervisors, researchers and developers for their feedback and for giving the playful emoji-method a chance in the first place within **FUTURE**. I am thankful to all participants who spent their time formatively evaluating and thus improving Proxemo or using the method in lab studies and therefore making its advancement measurable.

What would the world be without artists? To the team of **Emoji One**: thank you for making your artwork open source and licensed in a way that facilitates its use and adaptations. For unknowingly contributing to this work I thank **Einmusik**, **Apocalyptica**, **The HU** and

many more who facilitated hours of concentrated writing, as well as **Alexandra Elbakyan** who crafted a convincing quick-access experience. For their much more conscious contributions I want to thank the language artists **Iso** and **Stefan**, as well as the typography connoisseurs **Angela & Jens**.

For contributing to my health and sanity over the last four years I want to thank **Johanna** and all WÜ acroyogis. My gratitude goes to all KA acrobats and **Lara & Annca** for diverting me during “writing vacations” with adventurous road trips as well as low and high altitude flights (up to two-and-a-half-persons). In particular, I want to thank **Parzival** for being my interdisciplinary project sparring partner who provided great feedback, ambitious visions and facilitated deep dives into other ponds than literature.

I am thankful to my family for their support over the years and for showing me that there are topics beyond my screen. Finally, and most of all, I want to thank **Hannah** for the loving relationship that served as an emotional safe harbour over all the years that led to this book.

Abstract

For formative evaluations of user experience (UX) a variety of methods have been developed over the years. However, most techniques require the users to interact with the study as a secondary task. This active involvement in the evaluation is not inclusive of all users and potentially biases the experience currently being studied. Yet there is a lack of methods for situations in which the user has no spare cognitive resources. This condition occurs when 1) users' cognitive abilities are impaired (e.g., people with dementia) or 2) users are confronted with very demanding tasks (e.g., air traffic controllers). In this work we focus on emotions as a key component of UX and propose the new structured observation method *Proxemo* for formative UX evaluations. Proxemo allows qualified observers to document users' emotions by *proxy* in real time and then directly link them to triggers. Technically this is achieved by synchronising the timestamps of emotions documented by observers with a video recording of the interaction.

In order to facilitate the documentation of observed emotions in highly diverse contexts we conceptualise and implement two separate versions of a documentation aid named *Proxemo App*. For formative UX evaluations of technology-supported reminiscence sessions with people with dementia, we create a smartwatch app to discreetly document emotions from the categories *anger*, *general alertness*, *pleasure*, *wistfulness* and *pride*. For formative UX evaluations of prototypical user interfaces with air traffic controllers we create a smartphone app to efficiently document emotions from the categories *anger*, *boredom*, *surprise*, *stress* and *pride*. Descriptive case studies in both application domains indicate the feasibility and utility of the method Proxemo and the appropriateness of the respectively adapted design of the Proxemo App.

The third part of this work is a series of meta-evaluation studies to determine quality criteria of Proxemo. We evaluate Proxemo regarding its reliability, validity, thoroughness and effectiveness, and compare Proxemo's efficiency and the observers' experience to documentation with pen and paper. Proxemo is reliable, as well as more efficient, thorough and effective than handwritten notes and provides a better UX to observers. Proxemo compares well with existing methods where benchmarks are available.

With Proxemo we contribute a validated structured observation method that has shown to meet requirements formative UX evaluations in the extreme contexts of users with cognitive impairments or high task demands. Proxemo is agnostic regarding researchers' theoretical approaches and unites reductionist and holistic perspectives within one method. Future work should explore the applicability of Proxemo for further domains and extend the list of audited quality criteria to include, for instance, downstream utility. With respect to basic research we strive to better understand the sources leading observers to empathic judgments and propose reminisce and older adults as model environment for investigating mixed emotions.

Zusammenfassung

Für formative Evaluationen der User Experience (UX) wurden im Laufe der Jahre zahlreiche Methoden entwickelt. Die meisten Methoden erfordern jedoch, dass die Benutzer als Nebenaufgabe mit der Studie interagieren. Diese aktive Beteiligung an der Evaluation kann das untersuchte Erlebnis verfälschen und schließt Benutzer komplett aus, die keine kognitiven Ressourcen zur Verfügung haben. Dies ist der Fall, wenn 1) die kognitiven Fähigkeiten der Benutzer beeinträchtigt sind (z. B. Menschen mit Demenz) oder 2) Benutzer mit sehr anspruchsvollen Aufgaben konfrontiert sind (z. B. Fluglotsen). In dieser Arbeit konzentrieren wir uns auf Emotionen als eine Schlüsselkomponente von UX und schlagen die neue strukturierte Beobachtungsmethode *Proxemo* für formative UX-Evaluationen vor. Proxemo ermöglicht es qualifizierten Beobachtern, die Emotionen der Nutzer in Echtzeit zu dokumentieren und sie direkt mit Auslösern zu verknüpfen. Technisch wird dies erreicht, indem die Zeitstempel der von den Beobachtern dokumentierten Emotionen mit einer Videoaufzeichnung der Interaktion synchronisiert werden.

Um die Dokumentation von beobachteten Emotionen in sehr unterschiedlichen Kontexten zu erleichtern, konzipieren und implementieren wir zwei verschiedene Versionen einer Dokumentationshilfe namens *Proxemo App*. Für formative UX-Evaluationen von technologiegestützten Erinnerungssitzungen mit Menschen mit Demenz erstellen wir eine Smartwatch-App zur unauffälligen Dokumentation von Emotionen aus den Kategorien *Ärger*, *allgemeine Wachsamkeit*, *Freude*, *Wehmut* und *Stolz*. Für formative UX-Evaluationen prototypischer Nutzerschnittstellen mit Fluglotsen erstellen wir eine Smartphone-App zur effizienten Dokumentation von Emotionen aus den Kategorien *Ärger*, *Langeweile*, *Überraschung*, *Stress* und *Stolz*. Deskriptive Fallstudien in beiden Anwendungsfeldern zeigen die Machbarkeit und den Nutzen der Methode Proxemo und die Angemessenheit des jeweiligen Designs der Proxemo App.

Der dritte Teil dieser Arbeit besteht aus einer Reihe von Meta-Evaluationsstudien zu den Gütekriterien von Proxemo. Wir evaluieren Proxemo hinsichtlich der Reliabilität, Validität, Gründlichkeit und Effektivität, und vergleichen die Effizienz von Proxemo und die UX der Beobachter mit der Dokumentation mit Stift und Papier. Proxemo ist reliabel, sowie effizienter, gründlicher und effektiver als handschriftliche Notizen und bietet den Beobachtern eine bessere UX. Proxemo schneidet gut ab im Vergleich zu bestehenden Methoden, für die Benchmarks verfügbar sind.

Mit Proxemo stellen wir eine validierte, strukturierte Beobachtungsmethode vor, die nachweislich den Anforderungen formativer UX Evaluationen in den extremen Kontexten von Benutzern mit kognitiven Beeinträchtigungen oder hohen Aufgabenanforderungen gerecht wird. Proxemo ist agnostisch bezüglich der theoretischen Ansätze von Forschenden und vereint reduktionistische und ganzheitliche Perspektiven in einer Methode. Zukünftige Arbeiten sollten die Anwendbarkeit von Proxemo für weitere Domänen erkunden und die Liste der geprüften Gütekriterien erweitern, zum Beispiel um das Kriterium Downstream Utility. In Bezug auf die

Grundlagenforschung werden wir versuchen, die Quellen besser zu verstehen, auf denen die empathischen Urteile der Beobachter fußen und schlagen Erinnerungen und ältere Erwachsene als Modellumgebung für die künftige Erforschung gemischter Emotionen vor.

Contents

1	Introduction	1
1.1	Demand for Proxemo	5
1.1.1	Context of Dementia	5
1.1.2	Context of Air Traffic Control	6
1.2	Scope and Research Questions	9
1.3	Overview	11
2	UX, Emotion and Empathy	13
2.1	What is UX?	13
2.1.1	The Roots and Growth of UX	13
2.1.2	Theoretical Perspectives on UX	14
2.1.3	Observable Notions of UX	17
2.1.4	How Much Emotion is There in UX?	18
2.2	Emotion	19
2.2.1	Emotion Categories.	20
2.2.2	Emotion Dimensions.	21
2.2.3	Emotion as appraisal process.	24
2.2.4	Measuring Emotion	26
2.3	Empathy - Feeling into Others	30
2.3.1	A Short History on Empathy	31
2.3.2	Relevance of Empathy for User Research	32
2.3.3	Determining Cognitive Empathy in Observers	33
2.3.4	Emotions and Empathy in Dementia	34
2.4	Wrapping up Theoretical Perspectives	35
3	Review on UX evaluation methods	37
3.1	Evaluation Approaches in Dementia Literature	38
3.1.1	Method	39
3.1.2	Findings	41

3.2	Evaluation Approaches Reported in ATC Literature	49
3.2.1	Method	50
3.2.2	Findings	51
3.3	Need for a New Formative UX Evaluation Method	54
3.3.1	Discussion	57
4	Proxemo in the Dementia Context	59
4.1	A Structured Observation Method for the Dementia Context	60
4.2	Design Solution: the Proxemo App	62
4.3	Feasibility of Proxemo in the Dementia Context	66
4.3.1	Researcher	66
4.4	Study 1: Evaluating Reminiscence Interventions With Proxemo	67
4.5	Study 2: Video Analysis With and Without Proxemo Data	69
4.6	Study 3: Proxemo Usage on Top of Moderation	71
4.7	Study 4: Expert Evaluation in Reminiscence Sessions	73
4.8	Discussion	75
4.8.1	Limitations	76
5	Adapting Proxemo for Air Traffic Control	79
5.1	Design Solution for the ATC Context	79
5.2	Method	84
5.2.1	Context of Simulation	85
5.2.2	Participants	85
5.2.3	Researcher	86
5.2.4	Data Collection and Analysis	86
5.2.5	Procedure	88
5.2.6	Hygienic Measures	90
5.2.7	Data Collection	90
5.2.8	Data Preparation and Analysis	91
5.3	Results	91
5.3.1	Descriptive Statistics	91
5.3.2	Qualitative Insights	92
5.3.3	Downstream Utility	96
5.4	Discussion	97
5.4.1	Contextual Fit of Proxemo for Air Traffic Control	97
5.4.2	Suitability of the Predefined Emotional Set	98
5.4.3	Benefits for Developers	99
5.4.4	Limitations	101

6	Inter-Observer Reliability	103
6.1	Selecting the Appropriate Quality Criteria	103
6.1.1	Objectivity	105
6.2	Determining Measurement Error	106
6.3	Method	107
6.3.1	Design	107
6.3.2	Setup	107
6.3.3	Participants	108
6.3.4	Material	109
6.3.5	Procedure	109
6.3.6	Analysis	109
6.4	Results	112
6.4.1	Agreement Scores	113
6.5	Discussion	114
6.5.1	Limitations & Future Work	116
7	Effectiveness, Efficiency and Observer Experience	119
7.1	Pre-study to Generate Stimulus Material	122
7.1.1	Procedure: Conceptually Replicating Reminiscence Group Sessions	122
7.1.2	Analysis: Extracting and Annotating Meaningful Sequences	123
7.2	Method: Effectiveness, Efficiency and Observer Experience	125
7.2.1	Setup	125
7.2.2	Experimental Design	125
7.2.3	Participants	126
7.2.4	Procedure	127
7.2.5	Data Processing and Analysis	127
7.3	Results	128
7.3.1	Effectiveness	128
7.3.2	Efficiency	129
7.3.3	Observer Experience	130
7.4	Discussion	132
7.4.1	Implications	134
7.4.2	Limitations & Future Work	137
7.5	Post-study to Determine Cognitive Empathy	138
7.5.1	Method	139
7.5.2	Results	140
7.5.3	Eyes Test Score	141
7.5.4	Item Difficulty	141

7.5.5	Discussion of Results for the Eyes Test	144
7.6	Preliminary Conclusion on Effectiveness	147
8	Effectiveness and Intrusiveness	149
8.1	Intrusiveness	150
8.1.1	Intrusive Effect on Workload	150
8.1.2	Intrusive Effect on Performance	150
8.1.3	Intrusive Effect on Experience	151
8.2	Effectiveness	152
8.3	Method	155
8.3.1	Setup	155
8.3.2	Experimental Design	158
8.3.3	Participants	160
8.3.4	Procedure	161
8.3.5	Data Processing and Analysis	162
8.4	Results	163
8.4.1	Effectiveness	164
8.4.2	Intrusiveness	165
8.4.3	Proxemo and Physiological Data	170
8.4.4	Explorative Post-hoc Analysis of Quality Criteria Across Emotion Categories and Intervals	171
8.5	Discussion	172
8.5.1	Theoretical Implications	176
8.5.2	Physiological Data	178
8.5.3	Limitations & Future Work	180
9	Conclusion	183
9.1	Capturing User Emotions	183
9.2	Theoretical Considerations	186
9.2.1	Proxemo Within the Three Paradigms of HCI	186
9.2.2	The Relation Between Proxemo and UX	188
9.2.3	A Short Reflection About Emotion	189
9.2.4	A Glance on Automatic Emotion Detection Approaches	191
9.3	Recommendations for Practitioners	192
9.4	Summary & Outlook	195
	References	199

Appendices	237
A.1 Chapter 4: Interview Guide for Case Study 1	237
A.2 Chapter 5: Interview Guide for Case Study	237
A.3 Chapter 5: ELAN Screenshot	238
A.4 Chapter 8: Game Screenshots	239
A.4.1 Flight Control HD	239
A.4.2 SuperTuxKart	240
A.5 Chapter 8: Additional Detailed Data	241
A.5.1 Descriptive Data on Effectiveness	241
A.5.2 Emotion Frequency by Condition	241
A.5.3 EDA Effectiveness	242
A.5.4 GEQ Items and Factor Consistency	243
A.6 Chapter 9: Perspectives for the Proxemo App	245
A.7 Description of method review criteria as proposed by Stanton (2017)	246

Acronyms

ATC air traffic control

AttrakDiff Attractivity Differential

EDA electrodermal activity

EEG electroencephalography

EWPL Emotion Word Prompt List

facial EMG facial electromyography

GEQ Game Experience Questionnaire

HCI Human-Computer Interaction

IOR inter-observer reliability

ISO International Organization for Standardization

LEMtool Layered Emotion Measurement tool

meCUE Modular Evaluation of key Components of User Experience

OERS Observed Emotion Rating Scale

PREMO Product Emotion Measurement Instrument

Proxemo Proxy documentation of Emotions

SAM Self-Assessment Manikin

UEQ User Experience Questionnaire

UX user experience

Disclosure of Publications

The following list entails posters, demonstrations and position papers which were published with the aim of getting early and multifaceted feedback from experts on Dementia and Human-Computer Interaction. Passages which have been partially published are denoted in this work. This list may also be seen as “further readings” with respect to design and development of the Proxemo App which will not be discussed in exhaustive detail as part of this thesis.

Huber, S., Bejan, A., Radzey, B., & Hurtienne, J. (2019). Proxemo or How to Evaluate User Experience for People with Dementia. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6). CHI 2019, May 4–9, 2019, Glasgow, Scotland UK. <https://doi.org/10.1145/3290607.3313018>

Huber, S., Bejan, A., Radzey, B., Berner, R., Murko, P., & Hurtienne, J. (2018). UX-Evaluationen in der Erinnerungspflege bei Demenz [UX Evaluations of Reminiscence Activities in Dementia]. *Mensch und Computer 2018-Workshopband*. Dresden: Gesellschaft für Informatik e.V..

Huber, S., Preßler, J., Ly-Tung, N., & Hurtienne, J. (2017). Evaluating Interaction-Triggered Emotions in People with Dementia. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2659-2667). ACM. <http://dx.doi.org/10.1145/3027063.3053251>

Huber, S., Berner, R., Ly-Tung, N., Preßler, J., & Hurtienne, J. (2017). Evaluation eines Public Displays für Menschen mit Demenz [Evaluating a Public Display for People with Dementia]. *Mensch und Computer 2017-Workshopband*. Regensburg: Gesellschaft für Informatik e.V..

Huber, S., Preßler, J., & Hurtienne, J. (2017). Proxemo: A Demo-Presentation. In *Proceedings of the 2017 DementiaLab* (pp. 161-164).

Chapter 1

Introduction

Emotions play a key role in our interactions with computers and are a crucial component in the construct of user experience (UX) (Desmet & Hekkert, 2007; McCarthy & Wright, 2004; Thüring & Mahlke, 2007). Knowing about users' experience is the foundation for improving products, services or systems. As support towards gaining this knowledge, there is already a plethora of UX evaluation methods, neatly published in guidebooks for practitioners (Goodman et al., 2012; Hartson & Pyla, 2018). Apart from altruistic motives behind contributing to a positive UX, design for users' positive experience is associated with flourishing business. While it is difficult to track down the impact of each individual design- or research decision, there is a correlation between companies' active investment in UX design and business performance across industries (Sheppard et al., 2018).

Research on UX has long been limited to discretionary use, doubtful of whether enjoyable interactions are appropriate in the workplace (Hollnagel, 1999 as cited in Hassenzahl et al., 2000). In recent years, it has been increasingly considered that positive emotions such as pride and pleasure could improve job satisfaction also in safety-critical domains (Mentler & Herczeg, 2016) and might even improve safety and performance (Grundgeiger et al., 2020). Dukes et al. (2021) lift emotions way above that level, seeing affectivism on the rise, an era in which emotions and other affective states are the key to understanding cognition and behaviour. The common thread across authors is that emotional aspects of UX are gaining relevance in further domains which poses the question whether the existing methodology for measurement of emotions is already fit for all application domains.

In order to gain an overview of existing UX evaluation methods, we cluster them along two dimensions, 1) the role of the evaluation and 2) the person primarily interpreting and reporting the users' emotion. Evaluations in general can serve two purposes (Scriven, 1972). *Formative* evaluations identify aspects of a system that can be improved. *Summative* evaluations identify the (partial) superiority of one system over another and support decision-making about the

implementation or application. In UX research, formative methods are diagnostically deployed between design iterations, and summative evaluations are deployed after design or development processes to assess their success (Hartson & Pyla, 2018).

Another way of systematically classifying UX evaluation methods is by the person performing the interpretative step. In figure 1.1 we depict a dichotomy distinguishing between self-report and proxy ratings. This simplification is arguable as engineers who work on consolidating physiological measurements (e.g., variations in electric activity of the brain, heart, skin, or muscles), consider these automated measurements of stress and emotion a category of its own (e.g., Husain et al., 2018). However, viewing physiological measurements through our dichotomous lens, users or participants are merely the donors of raw data. The interpretative step of physiological data mostly falls to the researchers in that they decide, for example, upon criteria for inclusion and exclusion of data, the application of filters, transformations and thresholds for peaks. A recent review on UX evaluation methods indicates that self-report is most prevalent in published UX evaluation studies (95%), followed by “observation” (37%; here primarily referring to interaction records quantifying performance and only few instances where behaviour or facial expressions are observed) and physiological metrics (14%, Nur et al., 2021).

While physiological methods are mostly deployed as summative performance measures, in theory they could serve to identify peaks during task execution (Reinhardt, 2020) which are then discussed with users, similar to the valence method (Burmester et al., 2010). Additionally, physiological data is generally conceived as objective measure but can actively be influenced by users to a certain degree (Kox et al., 2014), thus potentially rendering it a self-report measure. To be fair, achieving this level of momentary control over one’s own sympathetic nervous system requires cognitive resources competing with resources required for the interaction task under evaluation (Wickens, 2008), making the co-occurrence an impractical means of self-report. Other methodological clusters can be allocated more clearly in the cross table of *evaluation role* \times *person reporting* in figure 1.1:

- Summative proxy ratings of emotion and behaviour are commonly used in the dementia context for assessing persons’ wellbeing in residential settings. Popular examples for so-called *Quality of Life* tools are the Dementia Care Mapping (Kitwood & Bredin, 1992; Sloane et al., 2007) or the Observed Emotion Rating Scale (OERS) by Lawton et al. (1999a).
- Detailed video analysis using the Facial Action Coding System (FACS, Ekman & Rosenberg, 2005) to identify emotions is rarely applied in UX research and conducted either automatically or with specialised software (e.g., Noldus, 2021). Emotions and UX can also be inferred from behavioural coding in video data (e.g., Gowans et al., 2004).
- In summative UX evaluations, self-report questionnaires are widespread, including the User Experience Questionnaire (UEQ) by Laugwitz et al. (2008), the Modular Evaluation of key

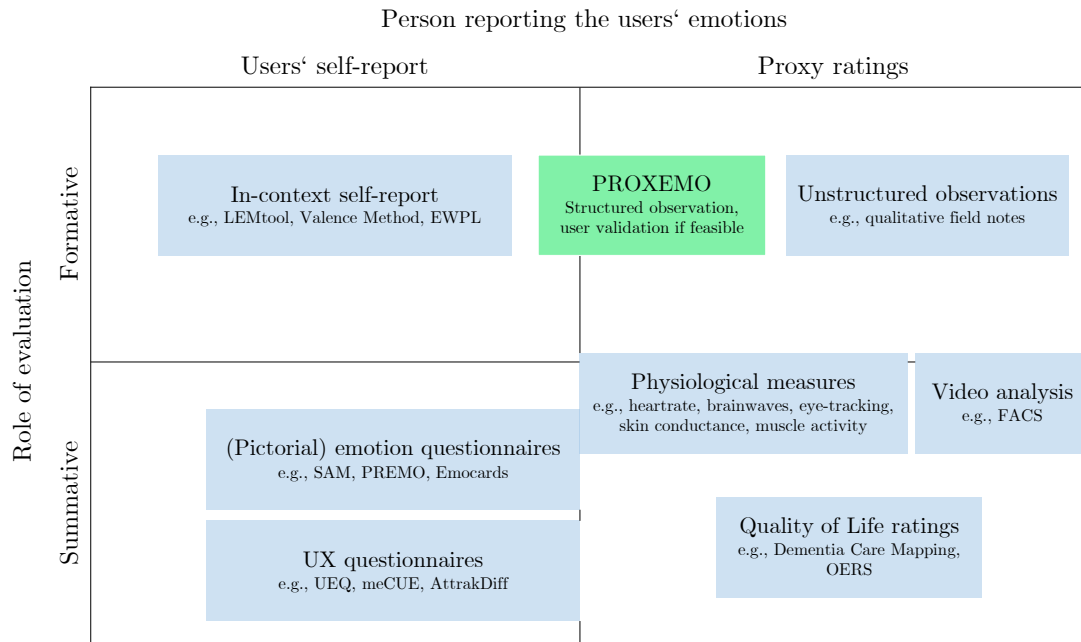


Figure 1.1: Cross table overview of existing methods clustered by their role or purpose of deployment in the design process (formative vs summative) and the person who primarily reads, interprets and reports the users' emotional reactions. Green highlight lies on the research gap covered in this thesis.

Components of User Experience (meCUE) by Minge et al. (2017), or the Attractivity Differential (AttrakDiff) by Hassenzahl et al. (2003). All three examples either consist of modular subscales and/or offer short versions which made them popular across domains.

- For the summative self-report of emotions and other affective experiences, specialised verbal and pictorial questionnaires exist. The Positive and Negative Affect Scales (PANAS) by Watson et al. (1988) and the pictorial Self-Assessment Manikin (SAM) by Bradley and Lang (1994) are rooted in psychological stimulus-reaction experiments and have been applied in user research before specialised scales were developed. Examples for evolutions of the pictorial SAM are Emocards (Desmet et al., 2001) and Product Emotion Measurement Instrument (PREMO) by Laurans and Desmet (2012). The Emotion Word Prompt List (EWPL) supports users with the vocabulary to describe their emotions. Note that the emotion statements during interactions, extracted from the EWPL could also be utilised as a basis for formative evaluations, but we are only aware of utterances being reported in a quantified, summative manner (Aizpurua et al., 2016; Petrie & Precious, 2010). There also may be an interpretation bias between how self-reporting users and researchers understand items of a questionnaire due to how their differing awareness about the questionnaire's

purpose changes the framing of individual items (Lavrakas, 2008). For instance, researchers are usually familiar with the dimensional structure of a questionnaire – a perspective that is (intentionally) hidden from participants and may contribute to a different interpretation of items.

- Embedding emotional pictorials similar to PREMO directly into the interface makes a perfect tool for self-report of emotions during the interaction and facilitates the direct linkage to the emotional trigger. Exactly this was accomplished for web-interfaces by Huisman et al. (2013) in the Layered Emotion Measurement tool (LEMtool). While the LEMtool apparently was discontinued, interestingly, the idea of linking emotional annotations to triggers has made the transition from a user research tool to a feature in the user interface of many communication platforms. In so-called “quick reactions”, users of Signal (version 5.27.12), Rocket.Chat (version 4.22.0.27017), Microsoft Teams (version 1.4.00.31569), or Zoom (version 5.8.4) can share their emotional reaction as emoji during a live feed or bind reaction emoji directly to messages of other users in the chat. Threema (version 4.6) does not provide a large set of reaction emoji for the same purpose but restricts quick reactions to thumb-up and thumb-down pictograms conveying valence. Throwing the bridge back to user research, Burmester et al. (2010) proposed the “valence method” with a comparable core-idea for formative evaluations. Following the valence method, users communicate their perceived emotional valence in the dichotomous categories good or bad by pressing respective buttons on a remote control during usage which sets a timestamp. These valence markers then serve as a foundation in retrospective interviews where researchers apply questioning techniques to gain a deeper understanding about why the user liked or disliked a particular interaction. Instead of categorical markers, peak values in continuous self-report measures for UX operationalised as emotions on one or both dimensions of valence and arousal could serve as a basis for retrospective interviews. Note that accuracy for arousal and valence reports varies over time (Lourties et al., 2018) and the continuous self-report might not be optimal for UX evaluations but is better suited for lab studies with exclusively receptive cues in which participants have the time to fully concentrate on their inner feelings (e.g., Dan-Glauser & Gross, 2015).
- Finally, looking at proxy ratings during formative evaluations, observations of user behaviour resulting in qualitative field notes are common practice.

We are, however, not aware of any structured observation method that facilitates the systematic rating of user emotions in formative settings, or even studies comparing observation techniques regarding their quality criteria. In this work, we therefore propose and evaluate a method that facilitates the Proxy documentation of Emotions (Proxemo).

1.1 Demand for Proxemo

Of course, the lack of a method alone does not necessarily motivate its creation. However, in this case, there are two scenarios where users cannot self-report their emotions, constituting a demand for Proxemo. Self-report during use depicts a secondary task that takes cognitive resources of its own. According to Wickens’s (2008) Multiple Resource Model, self-report should be feasible without impediment to the interaction experience, as long as the self-report methods exclusively rely on resources from a modality (visual or auditory) the primary task does not fill (e.g., thinking aloud while playing a visual only game, or documenting pictorial emotions via touch-input during an auditory experience). Yet, even assuming an interactive system communicating only on one channel, the hypothetical separation of visual and auditory tasks may still require shared resources (Wickens, 2008). Therefore, concurrent self-report biases either the primary task which is here the performance in using the system of interest, or the secondary task which is here the measurement of emotional experience through self-report. Both biases are undesirable. In interaction situations where time does not matter concurrent self-report is feasible and preferable to retrospective methods (Alshammari et al., 2015) as it provides a more thorough insight. However, when cognitive resources are exhausted by the primary task already, self-report is not feasible at all and proxy ratings are a promising approach to still capture users’ emotions during use. This occurs when either 1) the users’ cognitive abilities are limited and overstrained by self-report alone, regardless of primary task simplicity (e.g., people with dementia), or 2) the primary task involving system usage continuously demands the users’ attention and leaves few cognitive resources for self-report (e.g., safety-critical surveillance tasks). For both cases, structured methods for formative evaluation have not been deployed so far or emotional experience has even been neglected entirely. We will introduce the application domains where Proxemo might be beneficial in the following and review the methods applied so far in those contexts later in this work.

1.1.1 Context of Dementia

People with dementia experience cognitive impairments that exceed normal ageing including the most commonly known memory loss, but also executive functions, language, attention and social abilities (6D80-6D8Z1A, ICD11 2018)¹. Memory loss also affects the autobiographic memory and consequently self-identity (Rose Addis & Tippett, 2004). One way to non-pharmacologically slow down this development and alleviate the decline of Quality of Life is training the brain through actively invoking autobiographic memories (Astell et al., 2018). So-called *reminiscence activities*² are popular in dementia care and applied on a daily basis in many facilities. Under

¹Part of this section on dementia and reminiscence has been published in Huber, Preßler, Tung et al. (2017).

²Reminiscence activities are also referred to as *reminiscence therapy* by some researchers. However, we avoid this term since activities are not always accompanied by therapists.

this term, all kind of interventions are pooled which help people with dementia to actively reminisce. Following the person centred care by Kitwood (Kitwood & Bredin, 1992), caregivers should choose from the broad array of possible activities based on the personal experiences and preferences of the participating persons with dementia. There is evidence indicating short term improvements of cognition and a probable slight benefit on quality of life through reminiscence activities for persons with dementia in residential care settings (Woods et al., 2018). Another advantage of reminiscence and all other activities in care settings is diversion which may prevent residents from “dying of boredom” (Wood et al., 2009).

Examples of activities include crafting sessions (Pöllänen & Hirsimäki, 2014), visiting art exhibitions (Algar et al., 2014), creating life-story books with youth volunteers (Chung, 2009), or elaborate reminiscence programs for baseball fans (Wingbermuehle et al., 2014) as well as co-design of individualised jewellery (Wallace et al., 2013) or extendable multimedia albums and picture frames (Edmeads & Metatla, 2019). Extensive ethnographic work in two facilities (Huber et al., 2016) revealed to us that typical reminiscence sessions consist of less costly and extraordinary activities. Showing around printed pictures of formerly popular politicians and artists or pointing towards a relatable figure or caption in the local newspaper often times suffices to trigger reminiscence. However, starting with the CIRCA project (Alm et al., 2003), the HCI community has produced a variety of technological support systems that directly facilitate reminiscence or enable the caregiver in moderating reminiscence activities. Within the trans-disciplinary research project *InterMem* (Interactive Memories) we explored how technology can enrich the way people with dementia reminisce which sparked our awareness of the demand for the Proximo method.

When people shall have a reminiscence experience that is uninterrupted by concurrent self-report, the question arises whether a retrospective analysis of the experience is feasible. For people with dementia, loss of short-term memory hinders retrospective questions already at early to moderate stages (Gibson et al., 2016). With a progression of their disease, people with dementia lose their self-awareness, including the ability to empathise, evaluate their own situation and finally their self-consciousness accompanied by a decline of the ability to perceive certain emotions (Yokoi & Okamura, 2013). Therefore, while older adults with age-typical cognitive functioning can log their own emotions (Gooch et al., 2020), for people with dementia the interpretation and documentation of emotions by proxy is a promising way to still capture their emotions in formative evaluations.

1.1.2 Context of Air Traffic Control

In contrast to the context of residential dementia care facilities where technology facilitating diversion and reminiscence can be designed as simple as possible because no (productive) tasks exist, in air traffic control (ATC) the tasks are extremely challenging. Hence, applicants undergo

a rigid selection process prior to being trained as an air traffic controller (also referred to simply as “controller” in the following). For each shift, controllers are assigned a sector in which they constantly monitor traffic, direct and document aircraft movement and communicate with the controllers of adjacent sectors.

From gate to gate and start to landing, an aircraft traverses the responsibilities of air traffic controllers from ground control, tower, departure control, several en-route sectors (also known as area control) as it travels in high altitude before descending through approach control, tower and finally ground control again (see a vivid explanation from a pilot’s perspective in Lufthansa Services, 2020). Controllers communicate to the pilots all directs such as changes of heading, speed and altitude via radio and document the given clearance on a physical or digital flight strip representing the aircraft. To transfer responsibility on each sector border, controllers share the radio frequency of the next controller on the aircraft’s flightpath with the pilots and hand the respective flight strip to the next responsible controller (Cook, 2007). Within the transdisciplinary research project *FUTURE*, our focus lies on exploring support systems for approach control that is made up of the *pick-up* controller, who gathers aircraft from higher altitude in adjacent sectors and channels it to the *feeder* controller, who further reduces speed and altitude of aircraft while optimising the separation between aircraft before feeding them towards the airport. While the setup of workstations varies internationally between control centres and between positions, systems must enable the controller to monitor, communicate and document. Redundant communication channels as well as complementary information on the weather or traffic situation on the ground leave the controller with a cluster of up to seven screens on which they operate with multiple input devices (touch, pen and up to four mice, Huber et al., 2020).

Despite these working conditions being considered as “usability challenges” (Maybury, 2012, p. 2), so far the focus in ATC has remained on performance measures with little concern for controllers’ emotional experience. Performance in approach control is typically operationalised objectively by aircraft landed within an interval and subjectively by workload measures. One possibility to sample subjective workload frequently — if not continuously — is prompting controllers to rate their momentarily perceived workload in fixed time intervals, for example every two minutes (Sanderson et al., 2007). However, emotion documentation follows unforeseeable trigger events instead of predefined discretised time intervals. Therefore, an obligation to announce emotions would impose an additional event-based prospective-memory task on air traffic controllers. As research indicates that prospective-memory demands incur performance costs in ATC (Loft, 2014), such measurements would stress controllers and therefore bias the resulting experience. Thinking aloud — a formative usability evaluation method — replaces this event-based prospective-memory task with the demand to constantly verbalise actions and strategies, complemented with intervening questions by the researcher. Consequently, the constantly high load of the secondary thinking aloud task leads to less focussed behaviour, longer task completion times and higher workload (Hertzum et al., 2009). For minimising the disruption caused by an

evaluation during workload measures, proxy ratings are not entirely new to ATC and descriptively referred to as “over-the-shoulder” ratings by Averty et al. (2002, p. 1). In this work, we will test the feasibility of extending those over-the-shoulder ratings to include proxy documented emotions.

Continuos proxy ratings and retrospective self-report. When prototypes of ATC workstations shall be evaluated without interruptions to the controllers’ workflow, retrospective interviews come in focus again. In contrast to people with dementia, air traffic controllers are very capable of communicating. Yet, retrospective self-report is prone to several biases. For instance, Eggemeier et al. (1983) asked students in a lab study to state their workload after the completion of a memory task and restate the workload after a 15-minute delay. Two thirds of the participants changed their ratings after 15 minutes with an overall upward tendency in their ratings. Psychologists found several biasing effects influencing retrospective ratings of affective experience in that “retrospective evaluations appear to be determined by a weighted average of ‘snapshots’ of the actual affective experience, as if duration did not matter” (Fredrickson & Kahneman, 1993, p. 45). One of these biases is the *peak-end* effect in which snapshots of the most extreme events or events at the end of an episode overshadow the entire experience (Fredrickson & Kahneman, 1993; Kahneman et al., 1993). Free recall of situations may also be vulnerable to *primacy-recency* effects where participants remember the beginning and end of episodes but struggle to recall events in between (Murdock Jr, 1962). While most effects were studied in psychological lab experiments, effects of recency and peak-end effects appear to play a role in Human-Computer Interaction (HCI) as well (Cockburn et al., 2015, 2017). For air traffic controllers, so far only peak effects in workload have been shown in studies (Qiao et al., 2018). Minge and Thüring (2009) showed that UX judgments are affected by *halo-effects* (hedonic aspects influencing the pragmatic dimension and vice-versa) as well as recency-effects and mere-exposure effects (that means, ratings increase over time and the effect is more relevant for summative evaluations).

Following J. Nielsen’s (1994a) usability heuristic #6 *recognition rather than recall*, it is better to revisit emotional situations of interest in a structured manner during retrospective interviews through triggering users’ recognition of the events than trusting the users to recall every situation together with their emotional experience. Such memory triggers need to be identified first, for example, through situations rated as relevant by proxies during data collection. In ATC, Proxemo does not merely document emotional situations for researchers but aids to prepare triggers for retrospective interviews thus mitigating memory-effects.

There are previous occurrences of continuous ratings during use, some even served as foundations for debriefings. For instance, Rokicki (1987) used plots of pilots’ heart rate data from

evaluation flights as a memory aid for further subjective evaluation of specific events. If air traffic controllers did not have to switch between input devices frequently, workload could even be directly and non-obtrusively retrieved from entropy measurements in movement of the input device in space (Reinhardt et al., 2019; Reinhardt et al., 2020). Taking into account how input devices are touched, pressed and twisted, even few basic emotions might be retrieved for further discussion (Niewiadomski & Sciutti, 2021). However, in both cases the detection and documentation of emotions is bound to a specific interface with a multitude of sensors or requires explicit baseline data with which a system needs to be trained. How simple an emotion documentation interface may be built has been shown in psychological experiments on emotion recognition where Kirouac and Doré (1983) gave participants a seven buttoned device to document identified emotions in stimulus material. In our approach, we will follow in the footsteps of human capability for observation and technical simplicity for documentation.

Simplicity is desirable but not the only virtue of a novel method. John and Marks (1997) argue in their proposal of an effectiveness tree that usability issues which are not discovered will not lead to effective change of the system. The same applies to critical moments of UX which — if not observed and documented during the experience — can not be addressed later on and consequentially not lead to iterative improvements. For the detection of critical moments in UX, carefully evaluated methods are crucial. Salmon et al. (2020) lay the focus of meta-evaluations on reliability and validity. They point out the adverse effects of reliability on creativity methods in which maximum diversity of outcomes is sought. However, capturing users' emotional experience is an analytical method where high values in reliability are cherished. Salmon et al.'s (2020) main argument from a human factors and ergonomics perspective is that evaluation methods need to be classified regarding reliability and validity in order for the whole discipline to be taken seriously. Furthermore, for practitioners, insights from formative UX evaluations gained with reliable and valid methods result in improved requirements of an iterated prototype or even the shipped version. Documenting all steps in detail forms the grounding for measurements of downstream utility (Hartson et al., 2001) or even calculations of specific design decisions and their return of investment which is often demanded yet rarely followed through (Chawana & Adebessin, 2021). To provide a grounding for all that, next to the proposal of a novel method, the scope of this work comprises the thorough evaluation of Proxemo as outlined in the following section.

1.2 Scope and Research Questions

The goal of this work is the development and evaluation of Proxemo — a method to support the proxy documentation of emotions. Proxemo's presumed value lies in specific formative evaluation situations with users whose cognitive resources are restricted permanently (people with dementia) or temporarily (air traffic controllers). We keep referring to the person documenting observed

emotions by proxy as “observer” throughout this work, to highlight their passive role when using this formative evaluation method. As emotion is the best — if not only (see chapter 2) — observable notion of UX, we use the terms “emotion” and “UX” interchangeably within the context of observations. Foundations for our work are that

- emotional reactions in users are observable,
- observers are able to recognise emotional reactions (cognitive empathy),
- the trigger and the emotional reaction are in temporal proximity which allows us to re-establish the link between emotional markers and video recorded interaction after a synchronisation.

With these preconditions established we follow a design research process (Blessing & Chakrabarti, 2009) towards Proxemo, contributing the following:

- We propose a *method* that allows the efficient documentation of observed emotions during actual use. Subsequently, these highlights of critical situations aid further analysis and guide retrospective interviews.
- We conceptualise and develop context-appropriate *apps* that facilitate the implementation of the method in two application domains respectively where we evaluate the feasibility and usefulness of the method.
- We evaluate the method in lab studies regarding its *quality criteria* reliability, validity, thoroughness, efficiency and observer experience.

Along that process we first answer exploratory research questions and design research questions based on literature and qualitative studies before addressing descriptive and comparative research questions regarding Proxemo’s performance. The overarching questions are:

RQ1 How suitable is existing methodology for formative UX evaluations with users who have no spare cognitive resources?

RQ2 How can we enable evaluators to document observed emotions by proxy?

RQ3 Is Proxemo as a tool and method feasible for formative evaluations in the contexts of dementia and air traffic control?

RQ4 How does Proxemo perform regarding the most relevant quality criteria reliability, validity, thoroughness, efficiency and observer experience?

The studies within this thesis involve people with dementia – a vulnerable target group – and air traffic controllers representing users from a prototypical safety-critical domain. We strive

to protect participants by acting on the ethical maxim of data economy in that descriptions of participants' demographics are kept to a minimum. To maximise future benefits to users and increase the generalisability of our research, we strive for ecological validity in the study design.

1.3 Overview

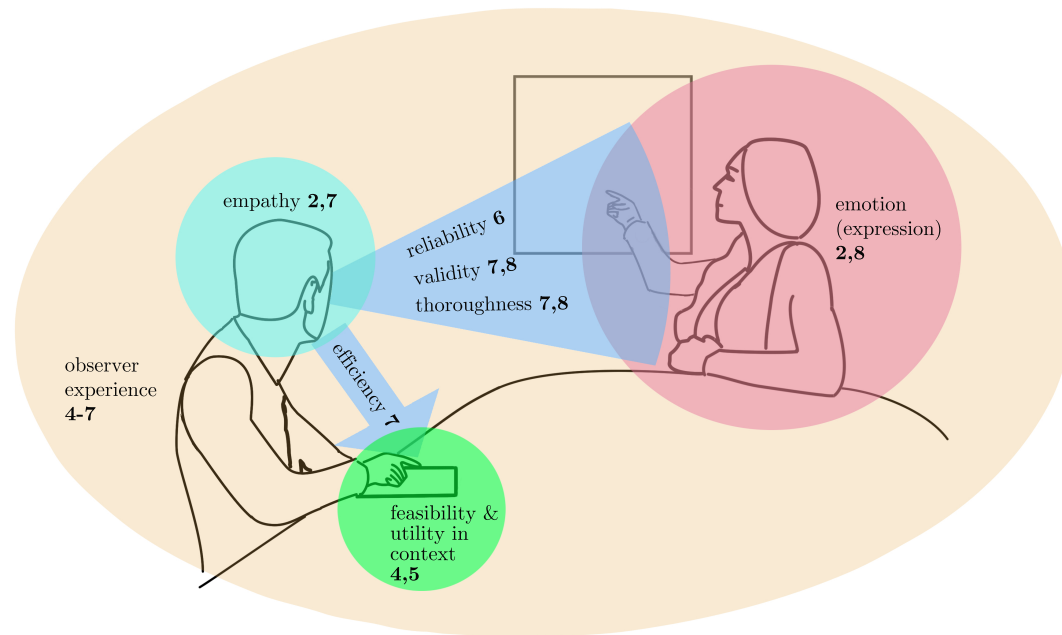


Figure 1.2: A structured observation scenario (observer on the left, user on the right) and how chapters of this work relate to it.

As depicted in figure 1.2, the contributions and chapters of this work are derived from aspects of structured observation scenarios and address the above listed research questions as follows.

Chapter 2 reviews research and theories on user experience, emotions and empathy. We provide an overview of theoretical views on the concepts, list possible measures and clarify our own perspectives. This chapter establishes a basis of criteria that facilitate observations.

Chapter 3 reviews formative evaluation approaches in the contexts of dementia and ATC (RQ1). We identify criteria for formative evaluations that meet the constraints of both domains and propose the Proximo pipeline as method (RQ2).

The next two chapters dive into the application domains and iteratively develop Proxemo as well as qualitatively investigate Proxemo's feasibility with a series of case studies (RQ2, RQ3) in the fields of dementia (chapter 4) and ATC (chapter 5). For both domains, we first work out an appropriate set of emotions and implement an application that supports the documentation of said emotions. Subsequently, we test the Proxemo method during formative evaluations of technology in the respective context and investigate the feasibility, utility and observer experience.

The next three chapters encompass lab studies aimed at determining the most relevant, yet measurable quality criteria of observations with Proxemo (RQ4). As a start, chapter 6 lists candidates of quality criteria, discusses objectivity of observational methods and measures Proxemo's inter-observer reliability based on video material from reminiscence sessions in residential groups. Chapter 7 primarily measures documentation efficiency, validity, thoroughness and observer experience in re-enacted reminiscence sessions and as a by-product determines the level of empathy in student-participants. Chapter 8 conceptually replicates aspects of the ATC workflow in a gaming study taking a different approach to measuring intrusiveness, validity, thoroughness and explores the utility of physiological data.

Chapter 9 wraps up the insights gained during this work and links them to HCI theory, summarises implications for practitioners and gives future directions for Proxemo.

Chapter 2

UX, Emotion and Empathy

Before we dive into the creation and evaluation of a novel UX method in the following chapters we need to explain three concepts that are crucial to formative UX evaluations of technology. In this chapter we pursue definitions of UX, emotion and empathy and point out how they are interrelated¹ and which role they play in our application domains.

Most will agree that *user experience* is unseparable from users' *emotions*. However, *empathy* plays an important role as well, for instance when researchers or practitioners attempt to interpret users' emotions or even claim to design the users' experience.

2.1 What is UX?

User experience or simply UX is a term that is comparatively young but well established in academic and industry research.

2.1.1 The Roots and Growth of UX

One of its first documented mentions is in an organisational overview of Apple at the CHI conference in 1995 where Norman et al. introduced it as a synonym for anything that affects research or application of the human interface. The seed worked well, and the buzzword *user experience* struck roots in academic research as well as in practitioners from design over development to marketing. As a consequence of Norman et al.'s broad (or non-) definition, each research group associated own ideas with the topic. In order to structure the variety of research approaches, Law et al. (2007) conducted a workshop that aimed at creating a picture of contributors' principles, policies and plans of UX. Reviewers classified workshop contributions to bipolar scales of five predefined dimensions which resulted in a colourful picture of diversity ranging from reductive to

¹Excerpts of our thoughts on the interrelation of UX, Emotion and Empathy reported in this chapter have been accepted for publication in Huber and Rathß (in press)

holistic theories, from quantitative to qualitative methods, from work domains to leisure, from personal to social applications and focussed on development, evaluations or both. In sum, one could say that since its coining, UX has lost none of its overarching meaning, spanning anything related to the user but has gained dimensions to specify different viewpoints. And this was all before the widespread use of smartphones and social networks which increased our daily interaction time and hence instances of UX. As technology became a crucial part of everyone's lives, UX research shifted from efficient interactions via hedonism towards meaningful relationships. The following section provides an overview of prevalent UX theories.

2.1.2 Theoretical Perspectives on UX

Theoretical views in HCI are commonly divided into three paradigms or waves. While the three waves are not perfectly distinct and researchers' perspectives on these waves vary in detail, the evolution described in the following is widely accepted. HCI originates from an engineering and human factors background with a focus on performance (first wave), then developed over cognitivist/information processing theories (second wave) before focussing on meaning-making (Bødker, 2015) and phenomenology (Harrison et al., 2007) in the third wave. During this evolution, the context of use shifted from the workplace towards discretionary use during leisure time which blend into each other as the third wave rolls on (Bødker, 2015). Instances of intermingled interaction situations are ubiquitous already, include checking emails at home as well as private messaging during work and may have gained in complexity through home office arrangements during the COVID-19 pandemic. As a second important development over the three waves, the focus of design processes and evaluations shifted from mere performance over context considerations to rethinking the role of emotions during technology use (Harrison et al., 2007). The growing importance of dynamic context, emotions and non-task oriented computing marked the need for a paradigm shift and the beginning of the third wave (Harrison et al., 2007) and resulted in a higher diversity of theoretical perspectives fuelled by various disciplines.

Embracing multiplicity of perspectives. With a fast-growing technology market that drives the application of UX methods and design practice in the most diverse application domains, it is not easy for theorists to keep up. The resulting plurality of theories and vocabularies may not be unifiable which can even be seen as beneficial as it allows researchers to view situations through different theoretical lenses and promote insight (Baumer & Tomlinson, 2011). Yet, it is helpful to be aware of all the lenses one can choose from. Theories, for example, describe the relationship between the user and technology (Engeström, 2015), the users within their physical environment (Dourish, 2004) or the users within their social context (Suchman, 2007). To gain an overview of theoretical foundations of UX research and application, Obrist

et al. (2012) collected responses from 70 participants during a Special Interest Group² at the CHI conference in 2011. They found that among theoretical perspectives which describe the design rationale, social aspects, the artefact-user-environment-relationship or which are artefact centred, the most prevalent theoretical focus in UX is centring on the individual user and rooted in psychology (Obrist et al., 2012). Kaasinen et al. (2015) title this user centred focus also the *empathy* approach.

Holistic for understanding. So what is experience from a user centred point of view? The philosopher Dewey (1934) suggested distinguishing the continuous stream of humans *experiencing* their environment and themselves from having *an experience*. Dewey defines an experience as a clearly demarcated period with a beginning and an ending — a whole episode that forms a closed story. Experience on the other hand is the raw material that may contribute to an experience, a stream of thoughts, desires, perceptions and reflections that may be tangled or interrupt each other. Dewey (1929) describes this experience as *holistic*, including subjects, behaviour, artefacts and environmental influences. Dewey’s thoughts were picked up by a number of HCI researchers who advocated a holistic, situated and constructed perspective on experience as well but contributed distinctive terms to facilitate communication about experience. For example, Wright et al. (2003) propose referring to experience as four intertwined threads, namely the compositional, the sensual, the emotional and the spatio-temporal thread. Forlizzi and Battarbee (2004) added the definition of *co-experience* as a further complexity that occurs when the fluent stream of experience or a specific experience is shared with other social actors.

Reductionist for measurement. The opposing *reductionist* perspective cuts experience into measurable and manipulable dimensions. Researchers seek patterns in UX and identify dimensions or influencing factors such as experience categories (Zeiner et al., 2018), qualities (Desmet & Hekkert, 2007; Hassenzahl et al., 2000), components (Thüring & Mahlke, 2007) or generic user needs behind experiences (Desmet & Fokkinga, 2020; Hassenzahl et al., 2010). Reductionist researchers believe that these factors affect the core of any experience and can, hence, universally inform design decisions and facilitate detailed summative evaluations. Wurhofer (2018) analysed reductionist and holistic conceptions on UX more deeply and proposes an integrative model. She suggests that instead of conceiving the two perspectives as mutually exclusive, researchers and practitioners should switch between perspectives depending on the stage of their research process. Pursuing a holistic approach to understand the field and, when clear themes (dimensions, factors) are identified, gradually adopting a reductionist perspective to postulate and test hypotheses is in fact a common approach in HCI research. Wurhofer’s process describes for the context of UX how Blessing and Chakrabarti (2009) more generally advise the approach of any structured design research project.

²Workshop format; details given in Obrist et al. (2011)

UX and Usability. From a reductionist point of view, it is not only important to model what aspects the concept of UX contains but also to discriminate UX from other concepts. According to the International Organization for Standardization (ISO), usability consists of effectivity, efficiency and satisfaction (ISO 9241-11:2018, ISO, 2018) and we are not aware of any author arguing with this definition. Of UX, however, multiple conceptions exist and have changed over time. Today, the only prevailing opinion is that UX contains more factors than usability. Compared to usability, the ISO standard definition of UX is more vague, encompassing “user’s perceptions and responses that result from the use and/or anticipated use of a system, product or service [... including] the users’ emotions, beliefs, preferences, perceptions, comfort, behaviours and accomplishments that occur before, during and after use” (ISO 9241-11:2018, ISO, 2018, p. 3.2.3). Most important, the ISO standard definitions draw no clear line between UX and usability. They seize Roto’s (2007) idea of satisfaction somehow being on the intersection of UX and usability, belonging to both. In contrast, Hassenzahl (2001) includes aspects of usability entirely in UX as *ergonomic qualities* (later labelled *pragmatic qualities*). Here UX extends the construct of usability by adding an equally important emotional dimension as a second strand titled *hedonic qualities*. Morville’s (2005) *user experience honeycomb* proposes usefulness, usability, accessibility, desirability, credibility, findability and value as seven equally important qualities of UX.

Both the intersecting and the inclusive perspectives lack notions of meaning — a topic that some authors saw as a necessary part of UX early on (e.g., Roto, 2007) and that has gained in importance over recent years under the term *eudaemonia* (e.g., Mekler & Hornbæk, 2016). Anderson (2011) even went one step further proposing a hierarchical structure of experiences where meaning is not only part of UX but the very top of it. According to Anderson, *functionality/usefulness* and *reliability* build the basis for *usability* on which then *convenient*, *pleasurable* and *meaningful* experiences can bloom. Kamp and Desmet (2014) see no such dependency between the pragmatic, hedonic and eudaemonic attributes. They argue by instantiating products with mostly hedonic (e.g., an aesthetically pleasing art piece) or eudaemonic attributes (e.g., personally meaningful jewellery) that have only minor practical use and thus no need to be usable. Where researchers see the greatest potential in UX may be given away already by their choice of words. For example, *enchanted technology* (McCarthy & Wright, 2003) satisfies hedonic needs of stimulation and *warm technology* (Ijsselsteijn et al., 2020) strives for meaning.

So far, we have learned that the scope of factors entailed within UX is growing. Yet, there is no consensus about whether usability is part of UX, is entirely disjunct from UX or if the two constructs share the common factor satisfaction. To overcome this deadlock with the “overused [...] buzzword” (Obrist et al., 2013, p.2433), Sauer et al. (2020) call for dropping the term *user experience* altogether and instead propose the construct *interaction experience* of a particular user group, thereby integrating the definition of accessibility.

And what position do we take? The short answer is: we encompass all user behaviour and

reactions, advancing in a mostly reductionist way by concept with holistic aspirations. We believe it is important to acknowledge that UX is highly context dependent and influenced by environmental and social factors. In this, we strive towards keeping the role of the researcher as an additional actor as small as possible during the interaction. To allow researchers to choose their own set of theoretical lenses a method best would be entirely theory agnostic. Our aspired method is independent of theoretical presumptions regarding UX, motives, needs and other factors underlying observable emotions. On the emotion theory side, however, our method requires that emotions can be correctly inferred by others and shortly reduced to their category before being explored in more detail. As our method heavily depends on the observability of emotion as a consequence of experience, a more detailed answer is given in the next section.

2.1.3 Observable Notions of UX

When capturing the users' experience through observation only, we must ask ourselves what threads (holistic) or factors (reductionist) of UX are observable. Following the thread terminology (Wright et al., 2003), the three threads *spatio-temporal* (how we perceive space and time), *sensual* (how we engage sensorily) and *compositional* (how we make sense of an experience) are encapsulated in the users' perception unless explicitly communicated to an observer. Only the *emotional* thread can be observed from the outside — given that emotions are strong enough to for instance manifest in mimic expressions. Since Wright et al. (2003) state that the threads in their model are intertwined, observable emotions may be the key to all other experiential notions.

Here lies a bridge to reductionist literature. For example, Desmet and Hekkert (2007) see *emotions*, *aesthetics* (similarly defined to sensuality) and *meaning* as key components. However, they explicitly draw a causal connection suggesting that on top of their discrete occurrence, emotions can be elicited by aesthetics or meaning. Meaningful experiences rely on cognitive processes such as association, retrieval of memories or interpretation (Desmet & Hekkert, 2007). Therefore, similar to aesthetic experience, meaning is only observable from the outside if it manifests in emotional reactions.

The same applies to the temporal component of UX. A common conception of UX is that it does not only occur during actual use but already begins with users' expectations of use and may be reflected later on (Desmet & Hekkert, 2007; Hassenzahl & Tractinsky, 2006; Roto, 2007). Expectations certainly have an influence on UX but are hardly observable unless they are manifested in emotional reactions or behaviour (e.g., a child smiling and nervously dribbling when waiting for something expected to be perceived as fantastic). Our focus therefore lies on what is referred to as *UX during (actual) use* (ISO, 2018; Roto, 2007) or *instant UX* (Wurhofer, 2018).

An even more detailed breakdown of UX is given in the CUE-Model (Components of User Experience, Thüring & Mahlke, 2007). It presents emotional reactions as mediating factor between

perceived instrumental qualities (usefulness, usability), non-instrumental qualities (visual aesthetics, status, commitment) and consequences of use, which include the overall evaluation and the intention to use a product in the future (Minge et al., 2017). Their perceived quality dimensions correlate highly (Minge et al., 2017) with the respective pragmatic (instrumental) and hedonic (non-instrumental) qualities of the AttrakDiff self-report scale based on the needs model by Hassenzahl (2001). It is therefore not surprising that in line with the CUE-Model, Hassenzahl et al. (2010) come to the conclusion that positive affect plays a central role. In their study, positive affect was an outcome of need fulfilment that predicted hedonic quality and mediated between need fulfilment and pragmatic quality. Another experimental approach to UX components supports the assumption that emotions are a consequence of need fulfilment (Jung et al., 2017). Regarding the measurability of these models' subcomponents, some aspects of usability and usefulness can be determined through performance measures. However, apart from self-report questionnaires, the *perceived* pragmatic qualities, needs and need fulfilment are as hard to capture as status and commitment. Therefore, when observing interactions, emotional reactions are the key component again.

2.1.4 How Much Emotion is There in UX?

In the last paragraphs, we argued that emotional responses are likely the sole observable notion of UX and in some models, emotion represents a consequence of other UX factors. Revisiting the ISO 9241-11:2018 (ISO, 2018), users' emotional perceptions and responses are included and even listed on the first position in the definitions of both, UX and satisfaction. In the following we will take a look at empirical evidence on the role emotions take in UX.

A review of UX literature revealed that among the authors who break down UX into dimensions, *emotions and affect* (24%) are evoked most frequently followed by *fun and enjoyment* (17%), *aesthetics* (15%) and *hedonic quality* (14%, Bargas-Avila & Hornbæk, 2011). Considering that emotions can be observed directly and *fun and enjoyment* most likely trigger an emotional reaction, the most prevalent categories can be covered through observation. Empirical evidence for emotions as key-indicator of UX has been brought by Agarwal and Meyer (2009). In their study, usability metrics revealed no difference between interfaces but users' emotional responses differed.

Wurhofer (2018) distinguishes *instant UX* as the experience during the interaction from *remembered UX* as the memories of the interaction and warns that memory effects may distort the reproduction of UX. This may particularly affect longer periods of interaction or pauses before recall. For short interactions and instant debriefings, literature indicates that users are well capable of reproducing their emotions. For instance, in an adaptation of the think aloud protocol, Petrie and Precious (2010) asked participants to verbally express their emotions while they used a website versus after the usage. They found that participants produced twice as

many emotions in retrospective think aloud than in concurrent think aloud. When participants were offered a list with emotional word prompts this did not change the number of emotional expressions, neither did the overlap of verbal expressions with words on the list increase. There are several important insights in this study: a) users are aware of their own emotions and, b) know how to express them without support. Unfortunately, c) users cannot express all their emotions during use but d) are able to reproduce them after a short task. On the other hand, e) when support is offered, it does not bias the results. Interactions that last longer than two tasks on a website possibly decrease the amount of recalled emotional situations.

For such situations, it is important to support recall of emotions and their triggers by documenting the context of use. In that way, emotions can act as a valuable anchor and may unveil richer experiential insights through debriefings with the users themselves or analyses by experts. Thorough analyses of events with particularly prominent valence or meaning build the foundations of the valence method (Burmester et al., 2010) and the critical incident method (e.g., Mekler & Hornbæk, 2016) which have successfully been used in several UX studies to analyse positive or negative experiences. From our perspective, emotional reactions are the ultimate UX dimension in observational approaches. In being just one part of the experience, documented emotions are reductionist in themselves but utilising emotional instances of interaction offers the key towards a holistic understanding of users' experience.

In sum, holistic and reductionist conceptions agree that emotions are a crucial facet of UX. Nevertheless, documenting exclusively emotions during evaluations represents a reduction of UX to one thread or dimension which in many cases may not suffice to depict the users' multifaceted experience. Therefore, documented emotions are not solitary representative of UX. However, the key-question is what happens after observed emotions have been documented. Proxemo, our intended method, is not thought to be a mere emotion counter. Proxemo shall support the capturing of emotions and their triggers in critical instances of interaction that can later serve as starting point for a detailed analysis. When the users are capable of articulating their remembered UX, it lies within the questioning technique of the researcher to gain reductionist or holistic insights. Applying the UX laddering technique (Abeele & Zaman, 2009) may reveal users' goals, needs and the importance of specific interactions (reductionist). Inviting the user to revisit a moment during the interaction and unfolding with them the dimensions of the experience as part of a micro-phenomenological approach (Prpa et al., 2020) may provide more holistic insights. The foundation for all these opportunities are emotional responses to system behaviour captured during use. Hence, an engagement with emotions seems worthwhile.

2.2 Emotion

Emotions are complex mental states that are not yet fully understood. As summarised by Schirmer (2014), neuroimaging studies show great overlaps between the constructs of emotion,

mood and affect. Nevertheless, psychologists provide definitions that allow to clearly distinct the concepts. According to Scherer (2005), emotions consist of the five components *cognitive appraisal*, *subjective feeling*, *bodily symptoms*, *action tendencies* and *facial and vocal expressions*. Furthermore, Scherer classifies emotions as being highly event focused, quickly changing and of short duration. This distinguishes them from other affective states such as *mood* which is defined to be longer lasting and having internal (e.g., hormones) or cumulated external causes (e.g., a chain of events) that are potentially unknown to or forgotten by the experiencing person (Russell, 2003). *Affect* is referred to as the feeling component of emotions that lacks the reflective aspects of emotion (Russell, 2003) or due to this uniting component used as an umbrella term for all affective phenomena (Scherer, 2005).

Philosophers and scientists have shared their thoughts on emotions for hundreds of years. In the following section, we confine ourselves to briefly summarising three modern psychological views on emotions all of which root back centuries or even to antiquity (Schirmer, 2014). While theorists differ in the gravity they assign to each of the emotions' components listed above, they agree on core components, three of which are central to this work:

- Emotions are object-related which means they can be attributed to an identifiable trigger.
- Emotions are expressed by the experiencing person and hence observable from facial expression (e.g., Ekman & Rosenberg, 2005) or body posture (e.g., Kleinsmith & Bianchi-Berthouze, 2013) as well as audible from vocal parameters of the voice (e.g., Schirmer, 2017).
- Similar to other affective states, emotions can affect decision-making (Schwarz, 2000) or behaviour independent of whether they are directly utilitarian or aesthetic (Scherer, 2005).

2.2.1 Emotion Categories.

Most prominent across disciplines and historically dominant is the idea of emotion categories. Already the ancient Greeks saw emotions as categories (e.g., pity, anger, fear, love and jealousy). However, they focussed on human-human interaction and put emotions in second place behind ratio (Konstan, 2015). Over two millennia later, Tomkins (1962, 2008) proposed eight basic affects as biologically rooted: surprise-startle, distress-anguish, anger-rage, enjoyment-joy, interest-excitement, fear-terror, shame-humiliation and *dissmell*³-disgust. Building on Darwin's (1872) ideas, he considered affect as functional for defence and reproduction and hence evolved.

Ekman shifted the focus of emotions' function from survival instincts to social communication. He firstly proposed six (happiness, sadness, anger, surprise, disgust and fear) *basic emotions* (Ekman & Friesen, 1971), later added *contempt* as a seventh category and now strongly believes that evidence will be found for ten further enjoyable emotions (Ekman & Cordaro, 2011): sensory

³Dissmell is a neologism that describes bad smells triggering us to reject inappropriate food (Tomkins, 2008).

pleasures, amusement, relief, excitement, wonder, ecstasy, pride⁴, Schadenfreude⁵ and rejoicing. Ekman and Cordaro (2011) acknowledge the existence of further emotions but point out that not all of them meet the list of 13 requirements for *basic emotions*. Taking a glance beyond theory, the variety of emoji used from unicode alone implies that users tend to identify and express more nuances in their feelings than 6 (or 17) basic emotions (The Unicode Consortium, 2019). Between eastern and western cultures, use of emoji differs with only small correlations in emoji displaying smileys and people (Guntuku et al., 2019). This indicates how fixating on cross-culturally consistent basic emotions restricts researchers in capturing the richness and variety of emotional experiences. Furthermore, Ellis and Tucker (2020) highlight that the universal recognisability of emotions does not necessarily imply a universal experience of emotions. They propose to speak of *versions of emotion*. Other critics argue that the situational context in which the emotion is experienced may be even more important than the emotion itself (Barrett, 2006). In fact, some more granular lists of emotion classifications contain for instance 48 emotion categories (Petta et al., 2011) or present even 154 words describing partly highly context dependent emotions (Watt-Smith, 2015).

More recent large scale self-report studies indicate 27 emotions with disjunct categories such as confusion, craving or nostalgia (Cowen & Keltner, 2017). Machine learning approaches on internationally sourced video material indicate that 16 facial expressions occur worldwide in similar contexts (Cowen et al., 2021). Some of these categories are included identically or similarly in Ekman’s list (amusement, anger, awe, contempt, elation, interest, sadness, surprise, triumph) while others differ (concentration, confusion, contentment, desire, disappointment, doubt, pain). In their study, Cowen et al. (2021) fed image data into a neural network to learn emotion categories. In this process, the network finds a vector representation of the emotional stimulus. Since similar input leads to a similar vector representation, distances between emotions in vector space can be calculated. For visually conveying the proximity of emotion categories, the high-dimensional embeddings were reduced to 2D resulting in a presentable map of emotions. This visualisation of emotion categories reminds of the two-dimensional mapping of affect or emotions (e.g., Yik et al., 2011) which we discuss in the following paragraphs with the difference that the two dimensions resulting from a vector projection have not been explicitly named.

2.2.2 Emotion Dimensions.

According to dimensional theorists, emotions can not only be described through categories but are additionally distinguishable through their value on two or more dimensions. Best known is the circumplex model of affect by Russell (1980). Whereas originally developed for the construct of core affect, the circumplex model can be used to map mood and emotions since core affect is “a key ingredient in both” (Yik et al., 2011, p. 723). It locates emotions on the two dimensions

⁴subdivided in pride of one’s own achievements (*Fiero*) and pride of one’s offspring achievements (*Naches*)

⁵enjoyment about opponents’ failure

of valence (pleasant — unpleasant) and arousal (activated — deactivated). Plotting emotions on a Cartesian coordinate system, they can be conveniently located through the angle of a ray from the origin (figure 2.1).

Other dimensional theorists suggest the use of three dimensions of emotion (activation, pleasantness and attention, Schlosberg, 1954) or three factors in language describing emotional events (evaluation, potency and activity/dynamics, Osgood et al., 1975). Note that Osgood et al. (1975) only found those three factors to be stable across their studies but identified additional factors restricted to certain languages or cultures.

Plutchik (2001) maps eight basic “primary” bipolar emotion categories (joy-sorrow, anger-fear, acceptance-disgust, surprise-expectancy) to a circumplex model depicted as a colour wheel (figure 2.1). Most prominent in his model are the opposing categories — allegorically represented by complementary colours. Similar to the metaphorical colour palette, primary emotions can be mixed to form emotional dyads or even triads (Plutchik, 1991). Note, however, that the colour wheel allegory has limitations since three primary colours suffice to mix any other colour whereas Plutchik’s wheel requires eight primary emotions for the same purpose. This becomes apparent in an example: Green can be mixed from yellow and blue. However, *trust* is not an amalgam of *joy* and *surprise* (see figure 2.1).

In contrast to other dimensional theorists, Plutchik does not name the two dimensions in which his colour wheel unfolds but uses the term *dimensions* as a synonym for the eight categories. Additionally, the wheel uses distance from the centre as a third dimension to indicate intensity (e.g., acceptance < trust < admiration).

A subdivision of basic emotions regarding their intensity is a dimension even Ekman accepts, whom we quoted as a categorical theorist in the previous section. According to the Atlas of Emotions⁶ — a collaborative project by Paul Ekman and the Dalai Lama (Stamen, 2020) — basic emotions are not associated by dimensions but subdividable by the single dimension of intensity.

Despite decades of emotion research, there is no unified dimensional model but instead two critical aspects that divide dimensional theories in two groups. The first group contains all above-mentioned dimensional theorists who name abstract dimensions subsuming more than one emotion. Each of them identified two dimensions that represent valence and arousal (Osgood et al., 1975; Schlosberg, 1954; Yik et al., 2011) even though some authors complemented their model with a third or more dimensions. This was also noted by Schirmer (2014) who poses the question whether emotional qualities that do not map to either dimension are irrelevant.

The second group advocates for an intensity dimension. As most theorists, Ekman and Cordaro (2011) as well as Plutchik (2001) root their theories in Darwinian ideas and see emotions as functional link between an environmental trigger and an appropriate reaction. In this, the intensity dimension serves as regulator for the adequate amount of emotion that prevents under-

⁶available on <http://atlasofemotions.org/>

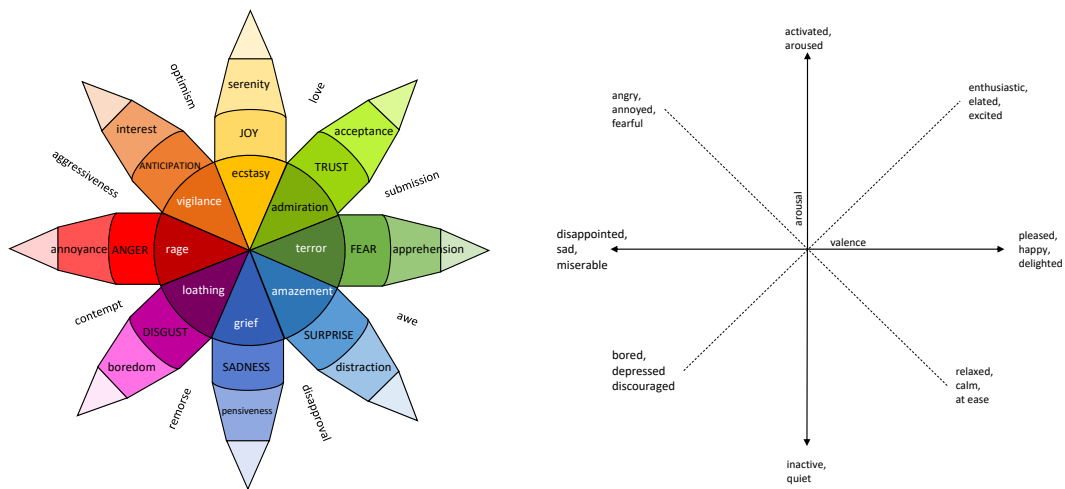


Figure 2.1: The illustrations visualise replications of dimensional emotion models. On the left is Plutchik’s (2001) three-dimensional circumplex model of a colour wheel. It shows eight basic emotions in complementary pairs of two. Distance from the centre indicates intensity as a third dimension. Further emotions can be explained as an amalgam of the basic dimensions. For example, *optimism* is a dyad formed of *joy* and *anticipation/expectancy*. On the right is a schematic drawing of Russell’s (1980) circumplex model with the two dimensions *valence* and *arousal*. In this illustration, we prototypically filled in the location of emotions as found by Remington et al. (2000).

or overreacting. For this purpose, Plutchik’s wheel of emotions distinguishes three degrees of intensity embedded in the eight categories (“dimensions”) and the Atlas of Emotions (Stamen, 2020) contains five⁷ basic emotions with up to thirteen labels for reaction intensities. Here, the Atlas of Emotions is more concrete and explicitly depicts for each basic emotion a mapping between emotional reactions of varying intensity and possible resulting decision tendencies that acknowledge the situational dependency. For instance, *contentment* (medium intensity of joy) may lead to the experiencing person’s reaction of maintaining the state or connecting with people to share their joy. *Euphoria* may motivate experiencing persons to maintain this state as well and fully immerse themselves into indulgence, appreciating the experience. *Schadenfreude* on the other hand — the joy about another person’s mishap — triggers malice and the need for connecting with others. As a medium intense state of joy, we strive to maintain this emotion savouring it through connecting with others and sharing our joy. Such context specificity of emotional reactions already seizes the idea of an appraisal process.

⁷sic! — surprise is missing in this publication

2.2.3 Emotion as appraisal process.

One point of criticism against the idea of universal basic emotions had been their lack of context sensitivity (Barrett, 2006). Similar to the categorical and dimensional perspective, no unified appraisal model has emerged so far that would be acknowledged by all theorists. They agree, however, on the five components involved in an emotional episode of which the *appraisal* component is most prominent and eponymous for the theories (Moors et al., 2013, pp. 119–120):

“Appraisal theories are componential theories in that they view an emotional episode as involving changes in a number of organismic subsystems or components. Components include

- an appraisal component with evaluations of the environment and the person–environment interaction;
- a motivational component with action tendencies or other forms of action readiness;
- a somatic component with peripheral physiological responses;
- a motor component with expressive and instrumental behavior;
- and a feeling component with subjective experience or feelings.”

The main difference between appraisal theories and the former listed perspectives is, however, that appraisal theorists see emotions not as a mental state but as a cognitive process (Moors et al., 2013, p. 120):

“The emotion process is continuous and recursive. Changes in one component feed back to other components. For example, changes in appraisal may lead to changes in physiological and behavioral responses. These may, in turn, lead to changes in appraisal, either directly or indirectly (via a change in the stimulus situation). As a consequence, several emotional episodes may run in parallel.”

Thereby, the dimensions supposedly considered during the appraisal process are (1) *relevance* with the components *novelty*, *valence*, *goal relevance* and (2) *implication* with the components *agency*, *outcome probability*, *goal conduciveness*, *urgency* (e.g., Grandjean & Scherer, 2008) as well as further criteria not shared by all theorists (Ellsworth & Scherer, 2003). One major difference of these appraised emotional episodes affects observers. According to appraisal theories, facial expressions are not derivatives of a discrete emotion category (compare e.g., Ekman & Rosenberg, 2005) but rather put the entire appraisal process of an event to display. Consequently, the facial display of emotion does not resemble a single prototypical output but rather resembles multiple dimensions of appraisal inference, mainly *novelty*, *valence* and *control* (Scherer et al., 2021). While appraisal theorists argue for the existence of emotional episodes, they acknowledge the

possibility of a more template like categorical appraisal of emotions in certain situations (Scherer et al., 2021).

Another important difference to categorical and dimensional theories is the number of available emotions. Due to the variability of events that run through the appraisal process, the complexity of resulting emotional expressions is rather unlimited than bound to few categories. Yet, prototypical, categorical patterns may describe the most frequent emotions (Scherer et al., 2021).

A good example for this is Scherer's (2005) approach to cluster 80 emotion words on two dimensions that are derived from appraisal theory (goal conduciveness | coping potential) but then making the map more intuitive by breaking the clusters down to 16 categories ("emotion families"). Interestingly, the resulting mapping ("Geneva Emotion Wheel") displays an intensity dimension. This intensity dimension is inverted to the wheel proposed by Plutchik (2001, figure 2.1) in that intensity increases with distance to the origin and, additionally, substages of emotion intensities are not labelled.

Finally, the emotional episodes of appraisal theory comprise more complex emotions that belong to more than one category/family, or to be more precise encompass more than one emotion within the episode. Such instances are named *emotion blends* by the dimensional theorists Watson and Stanton (2017) if all emotions of one event have the same valence polarity or *mixed emotions* if they are composed of both positive and negative affect. Within their observation data⁸, participants reported for most instances no affect (42.2%), followed by 30.3% blended emotions, 22.4% "pure emotion" and 5.1% mixed emotions. This indicates that affective states that can be described by one emotion alone are least frequent. One of the most relevant mixed emotions with respect to the application domain of reminiscence in our work is the bittersweet emotion of *nostalgia* which is overall considered positive but combines happiness and sadness (Baldwin et al., 2015; Watson & Stanton, 2017).

Coming back to appraisal processes, one reason for the lack of a unified model is its complexity. Scherer et al. (2021, p. 76) summarise:

"All of these components, appraisal results, action tendencies, physiological changes and motor expressions are centrally represented and constantly fused in a multimodal integration area in the brain (with continuous updating as events and appraisals change). Parts of this centrally integrated representation may then become conscious and subject to assignment to fuzzy emotion categories, as well as being labeled with emotion words, expressions, or metaphors."

In short, emotion appraisal is a complex process that cannot easily be grasped or even described in real time while experiencing or observing emotion. Therefore, humans seek emotion words approximating their feelings. As a consequence, when applying appraisal theory to field

⁸mood samples, measured with the PANAS-X

studies, the cognitive processes are broken down to the emerging *appraisal patterns* (happiness-joy, contentment-satisfaction, anger-irritation, disappointment-dissatisfaction), that are again categories (e.g., Demir et al., 2009).

What implications can we derive for emotion measurement in user research? In the tradition of HCI research, we cater for the user and observer by designing in accordance with their mental model (Loeffler et al., 2013) — here clearly categories — albeit it may not perfectly depict the state of knowledge on neurophysiological processes. In doing so, we strive to capture the emotional influence conscious to the user and observer.

2.2.4 Measuring Emotion

Temporal dynamics of emotions. Emotion appraisal is a complex process, but happens comparatively fast (figure 2.2). Brain wave analyses indicate that novelty, valence/pleasantness and goal relevance are appraised in picture stimuli within 200 ms (Grandjean & Scherer, 2008). Further lab studies with electroencephalography (EEG) on participants viewing enacted emotions indicate that neutral cues can be already distinguished from emotional cues⁹ after 100 ms, but it took up to 1000 ms to distinguish the negative emotion fear from anger in neuroimaging (Jessen & Kotz, 2011). Similarly, a study with pictures of emotional faces as stimuli showed general effects of emotion in brain waves¹⁰ after 90 ms, differences¹¹ between emotions after 140 ms and distinct patterns¹² between emotions after 330-420 ms (Batty & Taylor, 2003). More precisely, the *insula* seems to be activated by emotional stimuli after 200 ms with distinctive activity for happiness and disgust after 350 ms (Chen et al., 2009). When using emotional and neutral video material of persons instead of static images, emotion-unspecific effects appeared in brain waves¹³ only after 200-350 ms with first emotion-specific effects¹⁴ after 350-500 ms (Recio et al., 2014), indicating that data from studies with pre-selected picture based stimuli overestimate the accuracy and velocity with which emotions can be recognised from participants' brain waves. Furthermore, participants in this study were good at identifying happiness but depending on the emotions' intensity confused disgust with anger, fear with surprise, or sadness with any negative emotion (Recio et al., 2014).

Once emotions are being formed in the brain, their expression through facial muscle activity starts (e.g., smiling and frowning). Additionally, the sympathetic and parasympathetic nervous systems evoke a series of physiological changes that are independent of self-report and indicate arousal. For instance, heartrate measurably changes after two seconds (fig 2.2) and pulse amplitude after approximately eight seconds (Dan-Glauser & Gross, 2015). Skin conductance

⁹higher N100 amplitude

¹⁰higher P1 amplitude

¹¹N170 amplitude

¹²Cz at fronto-temporal site

¹³early posterior negativity at PO10

¹⁴late positive event related potential components at Pz

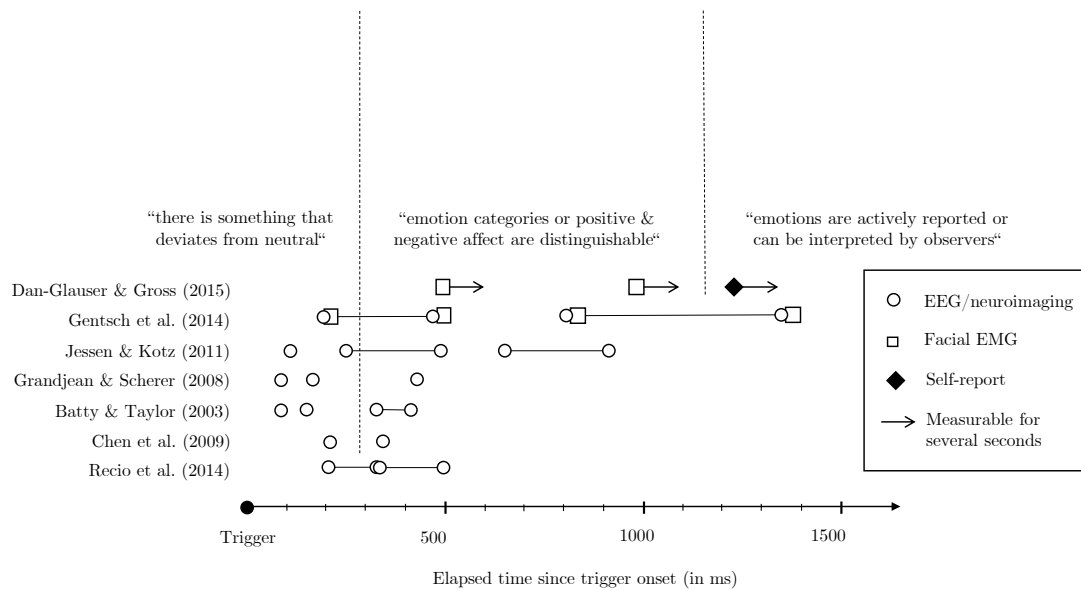


Figure 2.2: The timeline gives an overview of the temporal dynamics of measurable emotions. Note that most studies examined user reactions on still pictures in a lab environment and likely underestimate latency. Observability of facial emotions is theoretically not possible before the detectability through electromyography and likely begins later, varying between emotions (Losonczy & Brandt, 2003).

For comparison, measurements of facial muscle activity are faster with differences from neutral stimuli after approximately 200-500 ms for negative affect (*corrugator supercilii*) and approximately 500-1000ms for positive affect (*zygomaticus major*, Dan-Glauser & Gross, 2015; Gentsch et al., 2014). Facial muscle activity increases and can be detected through facial electromyography (facial EMG) before, for example, a smile is identifiable by human observers. However, the time from trigger onset to observable reaction varies greatly between emotions (Losonczy & Brandt, 2003). For instance, it took over 2 seconds on average for reportedly amused subjects to produce an observable smile (Keltner, 1995). Once presented, it takes observers at least 3 seconds for happiness and up to 7 seconds for all other *basic emotions* to document their judgment (Kirouac & Doré, 1983). A more recent experiment with healthy adults showed that the reaction time can be brought down to 620-660 ms when only distinguishing between angry and happy faces (Conte et al., 2018). Lab-experiments with positive and negative stimulus material show how participants begin with reporting their own emotional experience approximately 1250 ms after the onset of emotional stimuli (Dan-Glauser & Gross, 2015). Note that all above listed experiments presented stimulus material with only binary differences in valence or at most Ekman and colleagues' six basic emotions. Yet, in a similar study with binary judgments (true-false) about the correctness of labels to emotional pictures from ten categories, response latencies varied between 540-840 ms depending on emotion category and cognitive load (Tracy & Robins, 2008). As an exemption to the intervals reported so far, a person's startle reaction when hearing a gunshot may be observable between 50-100ms already (Ekman et al., 1985). This is not included in figure

2.2 due to the low relevance for most HCI interactions.

In their study, Dan-Glauser and Gross (2015) showed that participants can influence their emotion expression and autonomic responses through voluntary suppression and acceptance. For instance, changes in the respiratory rate can be observed but are difficult to interpret as breathing can be steered both, autonomic and voluntary¹⁵, and may be attributable to the cognitive task of accepting the emotion. In a more applied HCI-study where users are more freely interacting with a system instead of consuming a linear stream of cues, it is therefore difficult to attribute physiological changes to voluntary activation, emotions or other cognitive processes. However, measured or observed reactions can still be timely linked to their triggers.

Our take on measuring emotions. Emotions are mental states (or processes) that modulate the nervous system and thus lead to bodily reactions. The idea to directly measure brain waves or physiological parameters early on and objectively is compelling yet inexpedient from our perspective due to the subjectivity of emotional experience. Even disregarding technical hurdles such as increasing facial EMG errors over time (Golland et al., 2018), we do not see physiological measurements suitable alone for determining user emotions in HCI, and we will list our arguments in the following paragraphs.

Most emotion-research is conducted in artificial lab settings with constructed unambiguous stimulus material and limited to basic emotions which are hence limited regarding their generalisability to real world expression and recognition of emotions. Measuring emotions directly in the brain or their unfolding over the facial muscles is possible. However, trackings of physiological parameters (EEG, facial EMG, EDA, pupillometry, heartrate) pick up events continuously and not each peak is necessarily caused by an emotion. We also cannot be sure that users are aware of their physiological changes in, for instance, arousal (Caruelle et al., 2019). Since we strive to measure users' experience, the awareness of reactions is crucial. Operationalising emotion through measurement of physiological arousal, not filtering by the experiential component, causes plenty of noise in the data and thus poses an infringement of construct validity.

For this reason, EEG and facial EMG are mostly deployed in studies where a stimulus is manipulated and a reaction within a certain timeframe expected. Our interest, however, lies in the opposite direction of this pipeline. We seek to identify emotional reactions and through timely proximity reversely determine the trigger. Affective computing may provide solutions for this issue in the future and has been getting lots of attention recently with its main journal ranking #3 in Google Scholar's category *Human Computer Interaction* (Google, 2021). Automatic facial expression recognition has long mastered Ekman's six emotions in perfectly illuminated lab scenarios but struggles with real human interaction in the field (Gunes & Hung, 2016). Machine

¹⁵In fact, breathing is one of the keys to voluntarily manipulate not only the same physiologic processes as the autonomic nervous system but the autonomic nervous system itself (Kox et al., 2014; Zwaag et al., 2020).

learning approaches will be particularly promising once they overcome issues of processing non-frontal head poses in varying illumination and contexts (S. Li & Deng, 2020). Multimodal approaches that retrieve emotions from how users rotate or apply pressure to input device are agnostic of the light situation. Pioneering works have so far achieved to distinguish four emotions during interaction with a tangible cube (Niewiadomski & Sciutti, 2021). In seamlessly adopting to varying contexts, however, lies a general problem.

Classifiers drawing from expression recognition, physiological parameters and brain wave data require to be trained on large datasets which have so far merely focussed on Ekman’s basic emotions. We appreciate the simplicity of emotion categories as they provide user researchers with emotional words and shared concepts for discussions with users. However, the variety of emotional events and experiences can hardly be pressed into universal categories, by drawing only from facial expressions (Barrett et al., 2019) and without accounting for the context of emotional experience. The complexity and richness of interaction situations outside the constructed lab-environment can hardly be captured by a set of physiological sensors. This lack of context awareness prevents computers from replacing the “uniquely human capacity providing a ‘richness’ [...] to our way of being in the world” (Dreyfus, 1992, p.53). Understanding an emotional event requires context knowledge which facial expression tracking does not offer (Ellis & Tucker, 2020). According to Ellis and Tucker (2020) even increasing the amount and variety of context sensors will not help because the issue lies within how today’s computers work in general: as long as computers are deterministic, they will never be able to perfectly read or imitate even neurologically primitive animals since even the fruit fly does not act in a programmed, deterministic way. How do we measure emotions then, if not through a technical apparatus?

Utilising human capacity to document emotion categories. Here we argue why computer-supported documentation of observed emotions as categories appears most appropriate for application. For now, we will just assume that humans are capable of recognising other humans’ emotions and dive deeper into empathic abilities in the next section.

As stated above, the categorical perspective represents humans’ mental model of emotions rather than appraisal processes. Common categories are rather broad, covering a variety of stages on the intensity dimension (e.g., Plutchik, 2001; Stamen, 2020). Dimensional models usually do not contradict the categorical approaches but rather extend them by organising the categories within predefined dimensions. The remaining challenge is the selection of the categories — and potentially dimensions — most appropriate for the research field or context under observation.

Whether the emotions are connected through their placement within dimensions or are considered as discreet categories is a question that may be addressed once the emotions are documented. During the observation, it is most critical to document emotions quickly, losing the least time possible while interacting with an interface. A short, oversimplified modelling of interaction reveals advantages of categorical over dimensional documentations:

- Homing, that means aligning the mobile device with the field of view, takes .95 seconds (Holleis et al., 2007).
- Tapping on the screen is estimated to take .1 seconds (Rice & Lartigue, 2014).
- According to rough estimates of the mental operator in keystroke level modelling (Card et al., 1980), mental operations add 1.35 seconds per decision (i.e., category or dimension) to each documentation — mental processing time that is consequentially lacking for focussed user observation. Making a decision regarding two dimensions (or more, e.g., four factors by Osgood et al., 1975) would consequentially take twice as long as (or four times longer than) a categorical decision.

In contrast to the categories which serve as buckets housing emotional notions that are roughly similar, points in the dimensional model impose precision. For instance, Yik et al. (2011) locate angles of affect, moods and emotion on their circumplex model with a precision to the degree. Freehandedly documenting points via touch (or any handheld input device) on a two-dimensional grid would be biased by inter-observer subjectivity as well as the touch screen typical “fat finger” problem (Perrault et al., 2013). Resulting dot clouds could mislead researchers to interpret minor distances between documented instances. Since for quickly and freehandedly documented emotions only the interpretation of larger clusters or regions makes sense those categories can be transparently communicated and superimposed right away on the documentation interface.

Once emotion categories relevant for a specific context are identified, labelling them is the next challenge. Finding one prototypical word describing the multiplicity of emotional notions might be hard. However, emoticons or emoji are a common way to convey snapshots of emotions quickly with reduced complexity (Ellis, 2018) which is also mirrored in the popularity of pictorial scales for quick formative evaluations (see chapter 1). A prerequisite for the documentation in whichever way is the recognisability of emotions in the first place.

2.3 Empathy - Feeling into Others

Humans’ ability to express and recognise a basic set of emotions in others has already been described by Darwin (1872). He observed the non-verbal, emotional communication in animals and humans and wondered whether it was innate or acquired early on. Studies from the last decades brought evidence for the existence of basic sets of emotions that are expressed and likely can be recognised across cultures (Ekman & Friesen, 1971) and contexts (Cowen et al., 2021) and thus support an innate component.

However, as argued in the section above, human emotion is highly context dependent and the spectrum of nameable emotional notions exceeds the basic set of 6-16 categories. Consequently, the recognition of culture- or context specific emotions must be learned by individuals. Eickers

and Prinz (2020) argue that emotion recognition is a social skill involving improvable, practical and flexible scripts. While from an evolutionary perspective, abilities to recognise emotions and further improve the skill in doing so likely fostered collaboration with other living beings and hence survival chances, today designers can harvest recognised emotions to improve UX with products, services and systems.

Prototypical studies on emotion recognition involve the review of picture series¹⁶ of faces expressing Ekman and colleagues' basic emotions at full intensity. Full intensity is a euphemism for grimacing expressions which are unlikely to be observed naturally. Still, in some studies those exaggerated expressions only achieve accuracy rates between 45% and 98% (e.g., Bandyopadhyay et al., 2020) which raises the question about how humans should be capable of interpreting more subtle emotional expressions.

One possible solution is familiarity. Following the theory of emotion recognition as an improvable skill, we are better at recognising emotions in faces of people whom we often socially interact with because we are better trained in reading their emotional expressions (Eickers & Prinz, 2020).

Another important aspect is context information. Naming the emotion expressed by an unfamiliar face may be less challenging when observers know what triggered the emotion and possibly even are aware of the cultural interpretation of the trigger (Eickers & Prinz, 2020).

However, humans capabilities with respect to sensing emotions in other people by far exceed the ability to name them. In fact, we are able to cognitively understand the emotion, simulate its affective component and socially react in appropriate ways — skills subsumed under the term *empathy* (Lawrence et al., 2004).

2.3.1 A Short History on Empathy

The concept of *empathy* was first brought up in the 19th century when two German philosophers independently described how we employ all our senses to feel the world around us. We refer to their work because some of their core ideas still serve as paramount examples for latest definitions of empathy in psychology.

Lotze (1858) invoked humans' appreciation for aesthetic notions in nature and speculated that this enhanced sensuality distinguishes humankind from animals. He described how we “extend our sensuality beyond the borders of our body in a sympathetic manner” (Lotze, 1858, p. 194) and elaborated his point with examples of feeling into persons, animals and even objects. To be fair, the provided examples include contexts we would not or hardly instantiate with the term empathy today, such as imagining the position of a stick in the room based on its perceived weight in the hand or “dreaming along with the narrow existence of a shellfish” [p.193].

However, Lotze raises two aspects that improve the ability to empathise in certain situations.

¹⁶e.g., the *Karolinska Directed Emotional Faces* (Goeleven et al., 2008)

The first aspect is prior experience, in a way that persons who experienced the effort of a particular physical exercise themselves understand what this exercise can evoke. This is an early description of what we label as *cognitive empathy* today (Lawrence et al., 2004).

The second aspect covers the beneficial potential of immersion in the emotional situation — presented in a sadistic example of how a cruel person can only enjoy the victim’s pain if they are able to tangibly feel the impact of the weapon on the victim’s body. Interestingly, this violent example serves the full spectrum of the etymological roots of empathy from the Greek *empathia* (passion, state of emotion) assimilated from *pathos* (feeling, suffering; Harper, 2002). Furthermore, the simulation of other persons’ bodily experienced emotion resembles the first stage of *emotional empathy* (Lawrence et al., 2004).

Fifteen years later and reportedly unaware of Lotze’s book, Vischer (1873) published his thoughts on visual perception and aesthetics in which he introduces the term *Einführung*¹⁷ [from German: feeling into]. Comparable to Lotze, Vischer presumes the ability of humans to feel into living creatures as well as inorganic objects. Most notable for the topic of this work is his description of feeling into ones neighbour and empathising with them — an act for which Vischer (1873, p. 23) uses the term “self-duplication”. Today, psychologists label these sub-processes *emotional simulation* and *perspective taking* and found correlates in brain activity (Elliott et al., 2011). The third sub-process comprised in empathy concerns *emotion regulation* (Elliott et al., 2011) which already the philosopher and psychologist Lipps (1903, pp. 106–107) qualitatively described along with the other two sub-processes:

“We express all kinds of affects, emotions, types of inner excitement, such as fright, joy, astonishment, directly in sounds. [...] And if I now hear a sound similar to the one in which I myself announced my affect, then I find - not connected with it, but directly in it - this affect again. This “finding” seems at first a mere immediate co-imagining. In fact, it is more. I not only gain the idea that the sound is based on the affect, but I also learn it. I make it inwardly, the more surely and fully, the more I am inwardly completely turned to the sound. I am inclined to rejoice with the rejoicing person, that is, to inwardly join in his rejoicing. And I actually do this, if nothing prevents me, to be completely devoted to what I hear.”

2.3.2 Relevance of Empathy for User Research

Similar to Lipp’s quote cited above, user researchers may sometimes feel like rejoicing or steaming of anger along with their users. Being well-trained, researchers then hopefully are able to regulate their emotions and instead react in more appropriate ways, that is not emphatically joining the user emotions but rather validating them. The required ability to convey an understanding of

¹⁷According to Ewald (1908), Titchener (1909) was the first who translated *Einführung* with empathy — who again claims to have adapted his views from the Würzburg School.

the communication partner's emotions mostly cited in the context of therapeutic counselling is *behavioural empathy* (Bayne & Hankey, 2020). While emotional and behavioural empathy are certainly of advantage for moderating user tests, cognitive empathy is most relevant for understanding a person's emotions and, hence, for our attempt to capture the users' emotions and fathom their experience. In particular, the ability or sub-process of *perspective taking* has long been associated with the cognitive part of empathy (Davis, 1983; Elliott et al., 2011) and recent brain imaging evidence supports cognitive empathy¹⁸ to predict everyday perspective taking (Hildebrandt et al., 2021).

In section 2.2.4 we cited research on temporal dynamics of emotions that hints towards explicitly observable emotions in persons' facial expressions beginning after approximately half a second. Through a study setup where two conversation partners wore facial EMG electrodes, Riehle et al. (2017) showed that participants mirrored their dialogue partner's smile after less than 200 ms already. This extremely short synchronisation time hints towards an anticipated response. Mirroring emotional expressions is considered as part of emotional empathy. Neurologic interdependencies between cognitive empathy and emotional empathy are not yet fully understood but evidence suggests that emotional empathy does not predict everyday perspective taking (Hildebrandt et al., 2021).

While anticipation is well researched in rational perspective taking (Zhang et al., 2012), we are not aware of research on response times for anticipated emotions. Brain wave data from participants asked to classify dynamic facial expressions indicates neurological responses¹⁹ around 180 ms with stimuli only being presented for 600 ms (Recio et al., 2017). We would therefore expect anticipation intervals for emotions to be closer to the 200 ms for mirrored expressions (Riehle et al., 2017) than the seven seconds for anticipation of rational decisions (Zhang et al., 2012).

2.3.3 Determining Cognitive Empathy in Observers

A person's ability to empathise in a specific situation depends on several external and internal factors. External factors subsume all events and circumstances that distract observers from their task and consequentially reduce observation quality. When conducting lab experiments, we aim to control or randomise external factors. This is only restrictedly desired during user tests in the field which is why internal factors are decisive for observation quality.

As argued in the last sections, the skill titled "cognitive empathy" by social psychologists is the most important internal factor for a person's ability to empathise in a specific situation. Cognitive empathy is considered a personality trait and has shown to be stable over time (Quince

¹⁸Cognitive Empathy is referred to as *Theory of Mind* in the paper. We stick to the term *cognitive empathy*, because Theory of Mind is also used in more general for perspective taking and anticipation, representation and distinction of others' mental states and actions (Quesque & Rossetti, 2020) within the rational research field of Game Theory (Zhang et al., 2012).

¹⁹N170 latency

et al., 2011) and be affected by disinterest rather than due to increased age (Richter & Kunzmann, 2011).

Several questionnaires exist to measure or approximate persons' ability to judge observed emotions in oneself and others. Some scales focus on fictitious situations (Leibetseder et al., 2007), have unclear dimensions (Hogan, 1969) or are optimised for psychopathological diagnostics, such that their sensitivity is optimised for the lower end of the spectrum (e.g., Lawrence et al., 2004). Others are more suitable for observer selection through providing subscales to measure cognitive empathy (Carré et al., 2013), recognition of emotion in face-to-face discussion (Cassé-Perrot et al., 2007) or taking the perspective of others (Davis, 1983). However, all of them are self-report scales and thus potentially affected by biases that occur when people are asked to judge their own abilities. In fact, a meta-analysis on 85 studies indicates that self-report tools only explain 1% of the behavioural cognitive empathy and hence may not be suitable for assessing cognitive empathy (Murphy & Lilienfeld, 2019). One explanation may be that self-report rather examines the motivation to empathise and thus resembles another construct (Dang et al., 2020).

An established behavioural operationalisation of cognitive empathy is the *Eyes Test* (Baron-Cohen et al., 2001) which also served as behavioural baseline in Murphy and Lilienfeld's (2019) review. The Eyes Test directly assesses a person's ability to recognise emotions from another person's facial expression rather than their impression of this ability. Baron-Cohen et al. originally introduced the Eyes Test as a "mentalising" test that overlaps with empathy. The overlap is so convincingly apparent that Lawrence et al. (2004) assumed the term "mind reading" as a synonym for cognitive empathy. The Eyes Test owes its name to participants' task of judging the displayed emotions based on clipped pictures only revealing the pictured persons' area around the eyes.

Determining observers' general ability to empathise is one way to select observers. Further selection criteria include their suitability for the respective context. For instance, an observer's familiarity with culture, context or a specific person are considered beneficial for recognising emotions (Eickers & Prinz, 2020). Judging emotions of cognitively impaired people from another generation poses particular challenges.

2.3.4 Emotions and Empathy in Dementia

People with dementia undergo changes of personality and get worse at recognising emotions which is attributable to a loss of white matter integrity (Multani et al., 2017). Emotional symptoms vary by the type and severity of dementia (Balconi et al., 2015; P. Wang et al., 2021). Caregivers may not be aware of these deficits which increases their own stress — also known as caregiver burden (Martinez et al., 2018). In general, emotional expressivity declines with the progression of dementia but functional emotions such as anger are maintained until late stages and facilitate the communication of needs, wants or goals (Magai et al., 1996). For some causes of dementia,

details are already known about the association of neurological changes and emotions. Patients with early stage frontotemporal dementia can not feel emotions any more, even though they are still capable of explicit emotion appraisal (Balconi et al., 2015). Progressed atrophy in emotion critical brain regions (cortical and subcortical regions) results in weaker or incongruous responses even on a basic physiological level. That means arousal measurements via skin conductance or valence measurements via facial EMG produce abnormal results compared to healthy controls (Kumfor et al., 2019).

In sum, restricted emotion expressivity of people with dementia “[makes] it difficult for others to interpret and receive these cues” (Lazar et al., 2017a, p. 2177). Designers or researchers who have no prior experience with dementia can hardly imagine what dementia feels like for affected persons and how it changes their perception and experience. To tackle this issue and improve compassion of designers and researchers, sophisticated dementia simulations have been suggested (Smeenk et al., 2018). Compassion interventions have shown to improve empathy for and communication with patients (Brown et al., 2020). From our own observation experience we admit that increased exposure to residential groups increases understanding of their needs and emotions. And who would have more insight into people with dementia’s life than their caregivers? We argue that it is vital for researchers and designers to immerse themselves into the field and enhance compassion and understanding. However, when it comes to design critical interpretation of ambiguous user expressions, it is no shame to modestly seek advice of those who have been in close contact with the user group for years and silently trained their empathising skills.

2.4 Wrapping up Theoretical Perspectives

We conclude that UX is a buzzword with a variety of interpretations and operationalisations, most of which have user emotions as a core component. Emotions alone do not suffice to explain UX, however, using emotions as an anchor point may aid to reveal underlying needs, motives and experiences. When self-report is impractical, emotions are best captured through the observation of users’ expressive behaviour as well as subtle vocal or visual cues. Documenting observed reactions in predefined categories promises to be most efficient in formative evaluations. User researchers with high trait levels in cognitive empathy may be at advantage, but perspective taking and interpretation of users’ emotional reactions can be trained and improved for specific cultures, contexts and even individuals. Before diving into practice by applying our insinuated concepts to the field, in the next chapter we will provide reviews of existing formative evaluation methods in the domains of ATC and dementia.

Chapter 3

Review on UX evaluation methods

The application domains considered in this work are the contexts of dementia and ATC. The evaluation approaches occurring in the two domains are as diverse as the tasks, technology and users who are involved. To answer the explorative research question, how suitable the existing methodology is for formative UX evaluations with users who have no spare cognitive resources, in this chapter we review the UX evaluation methodology reported in the literature of both domains¹. Because the two fields are united by users' lack of spare cognitive resources, we seek to identify methods that require little to none mental resources on the users' side. Formative evaluation methods shall provide insights to inform the design of a product's or service's next iteration. Therefore, a relevant criterium for the methods is to deliver detailed information about what aspect of the interaction caused which experience. A separate review for each domain stands to reason since there is no thematic overlap and literature is mostly published in distinct journals and conferences. Our utmost goal for this review is to find a systematic method that meets the requirements for any of the two contexts. In pursuit of the perfect formative evaluation method we encompass and structure the variety of evaluation methods that have been applied in the two fields of dementia and ATC. More precise definitions of the exact context in which UX evaluations shall be conducted will be given in the according section.

The purpose of formative evaluation methods is the involvement of diverse yet plausible users to capture as many issues and experiential aspects of the design under evaluation as possible. Similarly, in this review, we seek to uncover the diversity of reported methods. Hence, we follow a systematical approach in searching for different methods but report our findings in a narrative

¹A preliminary review covering part of the dementia literature described in this chapter together with the criteria for formative evaluations of reminiscence sessions has been published in Huber, Preßler, Tung et al. (2017)

manner without explicitly quantifying the occurrence frequency of each method. We first report literature on dementia and ATC respectively, before discussing the strengths and limitations of identified UX methods.

3.1 Evaluation Approaches in Dementia Literature

There are numerous examples of usability evaluations and acceptance of assistive technology for community-dwelling people in their early stages of dementia (e.g., Hattink et al., 2016; F. J. M. Meiland et al., 2012). According to a review of literature in medical databases such research is optimistic regarding assistive technology and rarely considers quality of life (Holthe et al., 2018) which is similar to the concept of UX (see chapter 2). Whereas assistive technology supports people in achieving certain goals towards autonomous living, this focus on tasks — and hence the purpose of usability — loses relevance for people living in dementia care facilities. People with (advanced) dementia in care facilities rarely have a clear task on which performance needs to be measured. Hence, providing them a good experience in whatever they pursue is the highest priority. Days are mostly structured by meal-times and common activities in between include singing, performing art, playing, physical activity or reminiscing, with each other, caregivers or family members, respectively.

In the project *Interactive Memories* (<http://intermem.org>) we explored in an iterative user centred design process how reminiscing in people with dementia can be enriched through technology. We investigated both scheduled reminiscence group activities (Bejan et al., 2018; Huber, Berner, Uhlig et al., 2019) and short snaps of the past throughout the day (Gall et al., 2020). But how could their UX be formatively evaluated? How can we determine in a structured manner what part of a prototypical interface triggered positive thoughts of the past and which aspects of the reminiscing experience needed to be iterated? A deep view into literature revealed that the UX method we sought did not exist yet. Even for usability evaluations, standard methods are inapplicable in the dementia context (Gibson et al., 2016) and appropriate methods for people with severe dementia do not exist (F. Meiland et al., 2017). Popular qualitative methods such as think aloud techniques do not even work with healthy elderly persons (Franz et al., 2019). In the following we present a catalogue of criteria that need to be considered when formatively evaluating UX for people in all stages of dementia and show the limits of existing methods.

Based on literature and the contextual design process described in more detail in Huber et al. (2016), we derived the following requirements for the evaluation of prototypes with persons with moderate to severe dementia in care facilities:

- R1 Avoid overexerting people with dementia by keeping cognitive load to a minimum. Particularly people in advanced stages of dementia struggle to follow instructions or maintain focus (Kashimoto et al., 2016).

R2 Plan for residents with disabilities in speech.

R3 Embed evaluation into the daily routine.

R4 When facing restrictions in communication (Critchley, 1964) and self-reflection, do not use self-report methods. Even people in an early stage of dementia struggle to handle an interface with only three options (Rasquin et al., 2007).

Additionally, for optimising prototypes we need to identify which interactions are good and which need to be adapted. This need led to two further requirements:

R5 Map reactions to specific interactions.

R6 Emotions need to be documented instantly because some reactions are only interpretable in the context.

3.1.1 Method

In this review, our focus lay on formative UX evaluation methods, but we included reports of summative UX evaluations as well. Reasons for this fusion are that, on the one hand, not all authors explicitly state the purpose of the evaluation they conducted and on the other hand, some methods can be used for both evaluation purposes. Our initial review of reported methodology on UX in the dementia literature was conducted in December 2016. We updated our findings with literature that had been published since then in June 2018 and September 2021 in the ACM digital library², IEEE Xplore³, or PubMed⁴ — the most relevant databases for HCI and healthcare. The final search queries (table 3.1) comprising literature published until September 2021 resulted in a total of 448 papers. An examination of the titles reduced the literature to 154 papers. After reading the abstracts, 79 papers remained. For further inspection we applied the inclusion criteria that papers either had to present a novel UX evaluation method or describe the application of existing techniques that are — in accordance with the above stated requirements — suitable for the dementia context and produce outcomes from which design decisions can be informed. Here, we broadly defined *UX* as the report of some user reactions or emotions that exceed mere task fulfilment.

A priori, we excluded reviews without own empiric contributions as well as our own preliminary publications which will be described in more detail in this work. In our filtering process we excluded publications due to either of the following reasons:

²<https://dl.acm.org/>

³<https://ieeexplore.ieee.org/>

⁴<https://pubmed.ncbi.nlm.nih.gov/>

Table 3.1: Search queries by literature database, last updated on September 16th 2021.

Database	Search query syntax
ACM Digital Library	[[Abstract: dementia] AND [[Abstract: evaluation] OR [Abstract: design]] AND [[Abstract: usability] OR [Abstract: ux] OR [Abstract: "user experience"] OR [Abstract: usability] OR [Abstract: emotion]]] OR [[Publication Title: dementia] AND [[Publication Title: evaluation] OR [Publication Title: design]] AND [[Publication Title: usability] OR [Publication Title: ux] OR [Publication Title: "user experience"] OR [Publication Title: usability] OR [Publication Title: emotion]]]
IEEE Xplore	(dementia AND (design OR evaluation) AND (ux OR user experience OR usability OR emotion))
PubMed	((((dementia[Title/Abstract]) AND (design[Title/Abstract] OR evaluation[Title/Abstract])) AND (ux[Title/Abstract] OR user experience[Title/Abstract] OR usability[Title/Abstract] OR emotion[Title/Abstract])) NOT (review[Title]))

- Users were still in very early stages of dementia and, for instance, self-organised in online forums or were able to live autonomously with only minor support from assistive technology. This does imply that the same standard methods are applicable as with healthy people.
- People with dementia played a role but were not the users of, for instance, apps that predicted disease progression or supported the organisation among caregivers.
- Healthy people participated in preliminary studies with technology that was designed for people with dementia.
- Other stakeholders or experts were surveyed about the assumed needs of people with dementia without their immediate involvement.
- Exclusively quantitative measures are used such as summative questionnaires that serve no formative purpose.
- Results of an evaluation are reported but the description of deployed methods is missing or insufficient.

Furthermore, we excluded work describing research methods that rely on verbal communication or even questionnaires because they require communicative abilities that restrict the deployment to persons with only mild-cognitive impairment or early stages of dementia. From the references of our remaining papers we complemented the literature list with appropriate papers that had not occurred in our original search but did match our criteria. We summarise the methods of the final 43 references in the findings.

3.1.2 Findings

Qualitative observation notes as a standalone. The most common method of documentation in formative evaluations are manual notes. Some authors create standardised forms which they then use for multiple studies. An example for this is Bejan et al. (2017) who conducted three studies with different prototypes of multimedia systems. They took note about people with dementia's interactions and reactions which then informed the iteration of prototypes and the derivation of general guidelines. In another study, Gündogdu et al. (2017) use the self created forms to evaluate how 16 people with dementia interact with a digital fishtank and derived general insights and implications for future design of virtual experiences. Jönsson et al. (2019) created a protocol as well to facilitate the systematic documentation of caregivers' observations. Through a quantitative analysis of those observation protocols Jönsson et al. learned how often two residents noted meal-time notifications on a reminder system and that residents reacted happy or curious and followed the invitation to the dining room.

Other authors do not report how exactly the observed events were documented and analysed. For example, observations during user tests with a virtual planting platform led to an evolution of the interaction concept (Siriaraya & Ang, 2014). During another study within the context of art therapy, observers took notes during art therapy sessions and iterated the prototype of an interactive art frame together with the therapist in between sessions (Lazar et al., 2017b). Finally, Bouvier et al. (2016) collected qualitative data through observation and video recordings in a small study on user acceptance of a virtual training coach. They report participant's comments verbatim and identified a confusingly unnatural instruction gesture of the virtual coach when touching her nose as optimisable. Yet it is unclear whether the critical insights were gained in context or from reviewing the video recordings.

Another role of observational notes can be to serve as memory aid that supports extensive recall shortly after. Morrissey and McCarthy (2015) took field notes in three different environments during music workshops with persons with dementia and extended these notes to fair copies of field texts later the same day. They then proceeded with an analytic Grounded Theory approach (Charmaz & Mitchell, 2001) to gain insights on processes, actions and meanings in the data. The authors later report the same methodology in a publication on experience centred approaches in dementia (Morrissey et al., 2017). When evaluating a music emitting pillow with people in advanced stages of dementia, Houben, Brankaert et al. (2020) also developed field notes to field texts and additionally captured photographic artefacts. Similarly, (Stoeckle & Freund, 2016) used a combination of direct observations and context information via screen captures and audio recordings to identify experiential patterns of people with age related memory loss who used a prototypical music player.

Rich observational data can compensate the lack of video recordings. In an observational study, Chang et al. (2014) placed a specimen of the popular robotic baby harp seal Paro in two

different semi-public areas of a dementia care facility. Observers developed an online form to document information about interactors as well as their interactions with Paro and other people. They entered data in-situ using a laptop and took additional field notes describing the context since they had no permission for video recordings. Descriptive frequencies of interactions give an overview of who interacted how with Paro. Most valuable insights into what kept residents from interacting with Paro or how the interaction was initiated are based on the rich context information of the field notes.

Qualitative observation notes as a foundation. A further popular option is to use observation notes directly as a communication basis for debriefings and then extend them with other qualitative data. For the iterative design and development of a virtual worlds experience, Siriaraya and Ang (2014) made detailed observation notes and validated their observations in the aftermath with caregivers. Together with insights from focus groups, those observations informed the design choices of the next version.

Unbehaun et al. (2018) explore how exergames affect the social life of people with dementia over eight months. Notes from regular observations are interwoven with transcribed interview data acquired in parallel and all data is thematically analysed together. Insights are reported but are not directly applied to the design of the exergame platforms design in the same study. Kok et al. (2018) taught a Pepper robot to play reminiscence stimulating music. They evaluated Pepper's ease of use and effect on reminiscence and mood through a combination of direct observations and proxy ratings by caregivers. Based on their results, they give concrete recommendations for future designs.

Huber, Berner, Uhlig et al. (2019) rely primarily on observational field notes for their formative evaluation and iterative design of three tangible prototypes. Experienced dementia care evaluators took note of moments of reminiscence and identified meaningful experiences in people with advanced dementia. Huber, Berner, Uhlig et al. back up their interpretations through video analysis and interviews with caregivers.

Tabbaa et al. (2019) invited people with dementia to virtual environments that allowed them a view outside their locked facility. In their study, a researcher observed participants' interactions with the head mounted display and their facial expressions while interviews with caregivers contributed a professional perspective on the observed experience. They report instances where emotions could only be interpreted through the caregivers' mediation.

Similarly, Thoolen et al. (2020) designed the inclusive, personalisable multi-media experience AmbientEcho and evaluated it in a semi-public space of a dementia care unit. The researcher took note of residents' behaviour together with interpretations of facial expressions as positive or negative. She collected family members' and caregivers' comments during focus groups to enhance the observational data.

Muñoz et al. (2021) offered eight cooperative and competitive tablet-gaming experiences to

people with moderate dementia and their visitors. During the gaming sessions, they took notes whenever an activity of the gaming app triggered an interaction between the dyad (touching, laughing, talking, eye contact). They enriched their observational data with automatic interaction logs and interviews. From quantitative data, Muñoz et al. learned which games dyads preferably played and how those preferences and interaction patterns developed over time. From the qualitative data they drew explanations for said developments.

Systematic, quantitative observation notes. Analysing observed instances of emotional interaction permits a view on quantitative long-term developments. We still include these studies in our review of formative methods because in both cases the emotion tagged video data would also allow for a qualitative analysis.

A five-year longitudinal study on social interaction with robots in a nursing home (Chu et al., 2017) used observational notes that were cross-checked and validated with video data. Researchers coded the frequency of *approaching robots*, *interacting with robots*, *interacting with others* and instances of observable *pleasure*. They used the frequency data together with quality of life tools to demonstrate an improvement over the years but did not qualitatively analyse individual instances.

Researchers of the Digital Timelines Project (Colibaba et al., 2015) recommend producing short personalised video snippets consisting of material gathered with the person with dementia, their families and caregivers. The resulting memory medleys, so-called *personal digital memories*, should be played back to persons with dementia on a regular basis and trained observers should take notes in-situ or record the sessions on video. Prepared coding sheets invite the observer to take note on valence and intensity of emotions and their triggers. When changes in people with dementia's reactions over time become obvious, those shall be discussed with family members and guide decision-making regarding the adaptation of the personal digital memories.

Video analysis for mapped reactions. The first advantage of video analysis is transparency. Though being very time-costly, video analysis is a popular research method because in contrast to mere observations the whole chain from raw data to design decision is preserved. A second major advantage particularly for formative evaluations is that emotions and behaviour can be tagged and precisely linked to critical interactions. During the early user tests of CIRCA — a research product for multimedia supported reminiscence — Gowans et al. (2004) captured all reminisce sessions on video to be transcribed and analysed. In addition, they used coding sheets to capture over 30 people with dementia's positive responses to the programme (laughter, smiling, singing) and interaction with the moderating caregiver (talking, eye-contact). As a result from the analysis, insights on the experience and unexpectedly reawakened abilities of people with dementia were identified next to usability issues which were considered in further developments of the system. A slightly different purpose has Kiro, an anthropomorphic social robot who

performs exercises adhering to the instructions of a therapist and is supposed to increase residents' engagement. In a preliminary feasibility study, Cruz-Sandoval et al. (2018) recorded videos of two therapy sessions and analysed how seven people with dementia interacted with the robot during therapy. A systematic scheme is not presented but insights on interaction, engagement and adoption are derived. The most direct mapping of emotions to their triggers was carried out by Iwamoto et al. (2015). They presented digital photographs to people with dementia and video recorded the interaction. Subsequently, they coded the users' facial expression ("degree of smile") to learn which topics evoked the most joy and happiness in people with dementia. An even more holistic approach with respect to data consolidation has been carried out to thematically analyse people with dementia's reactions and associations when interacting with a soundboard prototype. In addition to analysing transcribed audio recordings and field notes from moderators and observers, Houben et al. (2019) synchronised the video recordings with the interaction logs (i.e. which sound files were played at what time) and annotated observable user reactions (e.g., yawning, acting surprised, gestures). From this rich, consolidated data they could conclude which sound experience evoked which emotions and associations and derive design considerations for future interventions. While their workshops were restricted to participants with mild dementia, the methodological approach seems suitable for all stages of the syndrome. Finally, to evaluate the effect of a music therapy intervention, Solé et al. (2014) conducted a systematic video analysis, coding verbalisations, physical and visual contact, active participation and emotions. The emotion categories were coded based on facial affect and body expressions and included *happiness*, *sadness*, *relaxation*, *anger* and *agitation*. Even though only the descriptive frequency of emotions is reported, emotion tagged video files would support speculations on what triggered the emotions.

Video analysis for (shared) off-site rating. Alternatively, video recordings can be used to take the pressure out of analyses which could have been conducted on-site. This is particularly relevant when many nuances of a short experience shall be captured or a lot is going on. When the density of information is high, analysing recorded videos has the third advantage of practically unlimited chances to re-observe the critical moments. In the evaluation on how a virtual forest experience affects the mood of people with dementia, Moyle et al. (2018) recorded videos and then had a research assistant complete the Observed Emotion Rating Scale (OERS, Lawton et al., 1999a) while watching the videos. The OERS allows documenting how long an observed emotion category was prevalent in the predefined time intervals never, < 16s, 16 – 59s, 1 – 5min and > 5min. Hence, they could find that the forest experience had no impact on the OERS dimensions *sadness* or *anger* but increased *pleasure*, *alertness/interest* and *anxiety/fear*. Since they did not take field notes or mapped emotions to specific events in the video, no qualitative statements can be made about what exactly brought residents joy or frightened them.

Distributing the effort of video analysis on multiple heads is an approach to alleviate its high

time costs. To evaluate a therapeutic card game, G. D. Cohen et al. (2008) videotaped the interaction sessions and had two research assistants assess emotion prevalence of 33 people with dementia based on observable reactions in the video recordings. G. D. Cohen et al. found their game to reduce sadness and increase pleasure but do not qualitatively link emotional reactions to specific events in the game.

Video analysis for reliability (measurements). Rewatchability presents the fourth advantage of video recordings: the exact same perspective can be repeatedly relived by the same person or different persons and an agreement of observations can be determined. Data on inter-rater reliability is mostly reported when video analysis are utilised for summative evaluations. Pérez-Sáez et al. (2020) video recorded sessions of three participants with moderate to severe dementia over four weeks while they participated in dog-assisted therapy or the same activities without dogs. Amongst other measures, they used the OERS to show how the dog’s presence increased positive emotions. A critical, unwanted outcome was that the two raters who had been jointly trained to code the video material — both psychologists with experience in dementia assessments — only achieved surprisingly low intra-class correlation coefficients for sadness (.34), pleasure (.28) and interest/alertness (-.05). Anger and fear could not be calculated because one of the raters did not use these categories.

Alternatively to redundant raters, data can also be rated twice by the same rater to increase robustness. Hammar et al. (2011) showed that morning care is more pleasant for residents with dementia who additionally resist less when the caregiver is singing. In their study, they video recorded the morning routine of caregivers with people with dementia and coded the OERS first live and then based on video recordings once again after 10 days which resulted in a high test-retest reliability of .97⁵. The authors do not discuss the reasons behind events with disagreement. Discrepancies of a second coder’s ratings who scored all the videos were resolved by consensus but the amount of critical events is not reported. Hence, we do not know to which amount either coder matches with the finally agreed upon “ground truth”. This information would have been helpful to judge whether coding in-situ is beneficial for data quality.

A very conservative approach to data analysis and interpretation was taken by van Rijen et al. (2020). The group of researchers iteratively designed RelivRing, a device for people with dementia to auditory relive the prior visits of their family members. In the second and third iteration, the authors split up data sources and individually coded non-verbal behaviour from observational notes, audio and video recordings and only followed interpretations that had been identified in at least two sources (that is, by at least two researchers). This included the emotional responses to interactions.

In an attempt to train sensors for automatic recognition of “challenging behaviour”⁶, Krüger

⁵Calculated as $\frac{\text{agreement} - \text{disagreement}}{\text{total observations}}$

⁶“Challenging behaviour” is a controversial concept as the term does not implicitly relate to the lived experience

et al. (2017) used six categories (*apathy, general restlessness, mannerisms, pacing, aggressive behaviour* and *trying to get to a different place*) for live annotations with video recordings in two dementia care facilities. Subsequently, Krüger et al. complemented their documentation with offline annotations for more fine-grained data of 1 ms instead of 5 min. Offline annotation increased the reliability of their data drastically from Cohen's $\kappa = .38$ to $.56$. Another interesting finding from their study was that when observing up to eight residents at a time, there were instances where one person walks out of sight. It was upon the researcher to decide whether to follow the individual or remain with the group. Krüger et al. (2017) estimate that even when all people were in the room, about one tenth of the annotated behaviour was not comprehensible from the video recording.

Physiological data. Living the dream of ongoing waves of industrial revolution, researchers seek to delegate time-costly tasks to machines. When it comes to observations of behaviour and emotions, an automation imminent promise is that physiological measurements may be more objective than human judgment. Alarcão (2017) describe in their futuristic vision how electroencephalograms (EEG) automatically recognise emotions in people with dementia and accordingly adapt the presented reminiscence triggers in real-time. Independent of the ethical considerations associated with making people with dementia wear a net of up to 64 electrodes, the achieved accuracy of emotion detection today ranges somewhere between 36 – 100% (Alarcão & Fonseca, 2019). A currently more feasible approach is to implement a music recommendation system based on residents' heart rate. Hsu et al. (2019) gave people with dementia a commercially available heart rate sensor (FitBit Charge 3) and had a system play music in three time slots per day based on residents detected activity. While the continuous logging of physiological reactions to music is very promising, the researchers so far shared only the caregiver's opinion of the system's feasibility. Steinert et al. (2020) had people with dementia wear a wristband that can capture EDA, temperature and blood volume pulse but did not report any of the physiological data in their study.

Optical and acoustical detection. Machine learning approaches to extract emotions from users' voice, gestures, body posture or facial expression through external recordings are slightly less intrusive than body worn physiological sensors. While it is not explicitly necessary, all studies we found on these approaches directly use sensors that are part of the prototype and thus avoid external recording hardware.

The companion robot Ryan, for instance, was designed to remind users about their schedule and interact with them socially (Abdollahi et al., 2017). For successful social interaction, Ryan was given the capacity to decipher emotions from facial expressions and speech. Conveniently, all subject interactions, facial emotions, speech sentiment and conversations with Ryan can be

of the person with dementia but rather summarises behaviour patterns considered challenging by the caregiver.

directly logged. Unfortunately, it is not reported how high the accuracy of emotion detection is. However, three patients with early stage dementia frequently conversing with the robot over a 4-6 week period and enjoying the interaction indicate that the embedded emotion detection has potential. Matilda, a social robot for home-based care comes with a similar set of skills (Khosla et al., 2014). Matilda "sees" the user through a camera, can extract facial features and interpret emotional valence (negative, neutral or positive) through Learning Vector Quantisation with accuracies between 80 – 100%. The most practical aspect is that the video stream together with the emotion log and interaction data can be saved directly by the robot. No additional hardware is required.

Similarly, Steinert et al. (2020) used the built-in camera of a tablet to capture people with dementia's facial expression during the use of user-specific activation apps. The purpose of their study was to show how users' emotional valence (negative, neutral or positive) can be automatically detected through machine learning, and they report agreement of Cohen's $\kappa = .45$ and $.49$ between their system and human raters. The emotion annotated videos which were in this case a mere by-product of research can help to identify critical content or interactions and ultimately contribute to improve the UX.

Combining it all, Parekh et al. (2018) propose a video-based system that automatically recognises and labels users' input, gaze, emotion and behaviour. The emotion detection module adopts the six categories from Ekman and Friesen (1971) and was trained on image data of elderly people. Apparently, recognising emotions in people with dementia was so hard that the researchers turned to use the facial emotion detection results of the healthy person sitting next to them as proxy data for the approximation of resident's mood.

Live annotation tools. Quality of life tools are common in documenting observed behaviour of residents in dementia care facilities and often are deployed to evaluate the care facility. Their primary aim is quantifying the concept of quality of life and the sheer amount of existing scales is overwhelming: two decades ago Thorgrimsen et al. (2003) claimed the existence of over 1000 methods and Gill and Feinstein (1994) found that of 159 reviewed instruments, 136 had been used only once. Since then, Quality of life methods evolved and today one of the most comprehensive and complex instruments that requires trained expert evaluators is the *Dementia Care Mapping* (DCM, Kitwood & Bredin, 1992). It consists of 23 behaviour categories (e.g. articulation, handicraft) and quantifiers (+5 very positive to -5 very negative) for mood and engagement shown by the observed person. Combined, the weighted behaviours are used as representative labels describing observation periods of five minutes. Observations are thought to cover five to eight people at once and usually take place in the context of care evaluations independent of technology. A meta-evaluation of the DCM over different research teams revealed questionable psychometric properties such as low variability of scores and an inter-rater reliability for the behaviour categories of $\kappa = .54$ (Sloane et al., 2007). However, Sloane et al. believe that a

shortening of the observation period could improve validity, reliability and practicality of the DCM. A less complex tool for evaluating quality of life is the aforementioned OERS (Lawton et al., 1999a).

Hamada et al. (2016) introduced a therapeutic robot into recreational group sessions with the aim to activate people with dementia. As evaluation measure they assigned observed behaviour and reactions within five minute intervals to one of nine categories. The categories could be grouped into active/passive behaviour or positive/negative reactions. Positive reactions were *accompanying* the robot, *laughing*, *touching*, *paying attention* or *talking* with the moderator. *Sleeping*, *not reacting* or *disliking* were considered negative. Hamada et al. descriptively analysed the gathered data on category frequency. The categorical way data were documented without context prevents a formative evaluation of individual human-robot interactions but allowed the researchers to identify an increasing activity when the person moderating the robot interaction was present.

To systematically evaluate humour therapy, Casey et al. (2014) developed BEAM, a touchpad for the documentation of Behaviour, Engagement and Affect Measures for up to four residents at a time. Regarding affect, they differentiated the dimensions *angry*, *anxious*, *happy*, *neutral* and *sad* — with *agitation* as a separate binary category (*low* — *high*). Observed durations of affect and agitation were logged, recorded and could be analysed to gain insight on pattern changes over time. For instance, Casey et al. could identify descriptive differences in affect between mealtime, free time and scheduled activities. The achieved inter-rater reliability of the entire BEAM form ranged from Spearman's rho of -.04 to 1.0, with the affect measures at the higher end of the range ($\rho_{angry} = .56$; $\rho_{neutral} = .94$).

During the iterative design of LiveNature, an interactive installation in a care home, (Feng et al., 2019) rely on caregiver feedback and observations which are not described in more detail. However, they conclude their project with a summative comparison of LiveNature versus another interactive installation regarding their abilities to evoke engagement and positive affect. Affect was operationalised by having an observer and the moderator complete the OERS after each interaction. They merely received data for the three negative emotion categories (anger, sadness, anxiety/fear) and report Cohen's kappa scores of .68 (interest/alertness) and .74 (pleasure) for the positive categories.

Interaction logs. How are interactive art installations in hospitals used by patients with dementia? Through interaction logs and interviews with staff, Wallace et al. (2012) derived design implications for persons with dementia's interactions with an art piece where short films could be played through placing globe props on a TV set. Insights are mainly based on the recalled observations of proxies, the interaction log indicates the most popular film-topics but gives no reason for their popularity.

Ethnography. If details of the interaction are of no importance but the researchers' focus is broader and lies, for instance, on how the presence of and interaction with a novel device in the care home influences the social role of residents, more descriptive, ethnographic methods are appropriate. Foley, Welsh et al. (2019) deployed a small receipt printer that inspires group communication in a care home and studied the evolution of residents' participation and agency over two years.

Participatory and Co-Design. Speaking of user experience, we mostly refer to a concept that is formatively evaluated with a small group of people who shall represent a larger user group. Thus, the expectation is that improving the UX for a representative group of persons, will lead to an improvement of the UX for a broader public — even though the context and prior knowledge may vary between individual users. However, if users are extremely diverse and designers have the resources to create an individual experience for each person, participatory design or co-design may be the appropriate approach. HCI and Dementia literature hold a number of examples where reminiscence artefacts (Czech et al., 2020; Wallace et al., 2013), or art therapy (e.g., Lazar et al., 2017b) were iteratively optimised for or with individual persons.

3.2 Evaluation Approaches Reported in ATC Literature

In contrast to the dementia context, safety critical domains are dominated by a Human Factors mindset where UX has so far not been the highest priority “and might even be considered too ambiguous, irrelevant or even risky” (Mentler & Herczeg, 2016, p. 5). Leading thinkers of the IEA (International Ergonomics Association) insist in a strategy paper that performance and well-being are equally important outcomes (Dul et al., 2012) which sounds similar to pragmatic and hedonic qualities of UX (Hassenzahl et al., 2003). Yet, the strategy paper's first author Jan Dul emphasized in a panel nine years later that performance is the highest priority we should design for and well-being can be achieved via good performance (Dul, 2021). From a human factors point of view, good UX means people are productive: “HFEs [human factors engineers] working in these domains strive to provide a total system solution that is useful, usable and efficient and a user experience that is productive, satisfying and engaging” (Savage-Knepshield et al., 2016, p.2053).

UX in a sense of positive emotions instead of mere satisfaction or productivity has long been treated as a secondary outcome in safety critical domains. However, Grundgeiger et al. (2020) argue that primarily designing for UX can inspire users to live up to their potential and ultimately increase performance. And user-centred design is not entirely new to safety critical domains. On the contrary, ATC has a long history of ethnographically inspired designs and involving end-users in cooperative refinements of prototypes (Twidale et al., 1994). User-centred methods are necessary to support user researchers and engineers in gathering and fine-tuning

requirements that inform the design of interfaces for domains as complex as ATC (see Vaccaro and Duca (2011) for an exemplary project). On the following pages, we review the approaches that have been made so far with the aim of learning from air traffic controllers' experience during formative evaluations. In the course of this, we are inclusive of all air traffic control positions. Even though air traffic controllers' specific set of tasks and the layout of their workstations may vary greatly between positions, controllers' management work comprises a common set of abstract tasks that is mainly communicating with actors on multiple positions and documenting clearances while maintaining situation awareness regarding all events in their respective sectors.

The most critical requirement for evaluations of all working positions in air traffic control is to avoid deteriorating controllers from their tasks. In situations with real traffic this is self-evident in order to maintain safety. In simulations, methods with the least distraction minimise bias and uphold data quality. Apart from this restriction, we adapt the requirements for formative evaluations from section 3.1 that are mapping reactions to specific interactions (R5) and documenting emotions instantly in context for optimal interpretation results (R6).

3.2.1 Method

Analogue to the previous review in dementia care, we searched for publications that propose or apply formative UX methods in the domain of air traffic control. Due to the stark focus on performance, studies that look beyond safety and efficiency are scarce. This forced us to broaden the focus and include methods which have the potential to capture experiential aspects of use even though in the corresponding publication the methods had been merely utilised to detect effectivity issues. Again, we used the most popular databases for HCI research, ACM digital library⁷ and IEEE Xplore⁸. Since air traffic control is one of the domains that brought up the discipline of Human Factors, we additionally searched on SAGE⁹ in all journals and conference proceedings published by the Human Factors and Ergonomics Society (HFES). The final search queries (table 3.2) included papers published until September 2021 and revealed 166 papers. A total of 55 papers withstood an examination of the titles and 23 papers remained after reading the abstracts. To further compensate the small amount of publications that cover real-time evaluations of novel systems, we additionally included ethnographic works on existing systems. As inclusion criteria we defined whether the method or a combination of methods could be applied for formative UX evaluations. The main reasons for the exclusion of papers were any or a combination of the following:

- The article describes a preliminary study with participants who did not have any prior training in air traffic control. While involvement of “real users” is advisable in any field,

⁷<https://dl.acm.org/>

⁸<https://ieeexplore.ieee.org/>

⁹<https://journals.sagepub.com/action/doSearch?>

Table 3.2: Search queries by literature database, last updated on September 27th 2021.

Database	Search query syntax
ACM Digital Library	[[Abstract: "air traffic control"] OR [Abstract: "atc"]] AND [[Abstract: "user experience"] OR [Abstract: or "usability" or "ux" or "emotion"]] AND [Abstract: "design" or "evaluation"]
IEEE Xplore	((Air traffic control AND (design OR evaluation) AND (ux OR user experience OR usability OR emotion)))
SAGE journals	[All "air traffic control"] AND [All "user experience"] *Filters were applied to search within <i>Human Factors</i> , <i>Ergonomics in Design</i> , <i>Proceedings of the Human Factors and Ergonomics Society Annual Meeting</i> , <i>Journal of Cognitive Engineering and Decision Making</i> and <i>Transportation Research Record</i>

it is imperative to conduct all user research with the target group in safety critical domains. Studies with students repeatedly indicated how a collision of aircraft resulted in no observable reaction (Brunett et al., 2018) or even a relieved reaction (Truschzinski, 2017).

- The study consists of focus groups on icon design, first contact with non-functional prototypes or cooperative prototyping without in-context use of the technology. Evaluating a system outside its context of use or omitting the task typical time constraints inhibits authentic user experiences.
- The interface design processes or evaluations were exclusively grounded in cognitive models or task models of air traffic controllers instead of empirically measured performance or experience.

Additionally, we had to exclude two papers that were only available in French or Turkish. We enriched the findings with one recent publication that was not listed in any of the databases but contributes a crucial summary of today's state of the art regarding physiological measurements. After our revision of the literature, 14 papers remained whose findings were summarised in the following section.

3.2.2 Findings

Logfiles. The simplest, most direct benefit of logfiles was drawn by Ahlstrom and Arend (2005), who asked air traffic controllers during a training day to take the time and set their colour preferences on weather information overlays. From the saved settings, Ahlstrom and Arend (2005) learned how diverse controllers' preferences were and suggest that limiting the colour spectrum up for choice could avoid configurations that are hardly legible.

Observations and interviews. In a series of workshops, Conversy et al. (2011) asked teams of two or three air traffic controllers to engage with a table-top prototype and solve predefined tasks. They observed how the usage of certain features affected controllers' communication, collaboration and situation awareness and validated their observations in debriefing interviews. One of the design implications emerging from the study was the necessity to increase the interface size when three controllers needed to use it simultaneously. For the evaluation of a 3D interface for approach control, Rozzi et al. (2007) placed the prototype between conventional displays on a workstation during human-in-the-loop simulations with four controllers. Post-simulation, controllers were interviewed and given large paper sheets to draw the traffic situations they were describing. This process provided researchers without ATC training with a deeper understanding of the controllers' strategies and the information controllers required to resolve conflicts. A similar way was chosen by Traoré and Hurter (2016) who asked controllers after the completion of various tasks with a novel interface to note their preferences and remarks in a feedback document by themselves.

Seeking to understand strategies, Malakis et al. (2014) observed air traffic controllers during shifts ensued by interviews with the observees and so-called on-the-job training instructors. The systematically reported strategies that controllers follow during critical incidents include design-critical descriptions of how they interact with the radar and which cues they are specifically looking for. Huber et al. (2020) followed a similar approach as part of a contextual design process. They deployed teams of 2-3 observers during ATC shifts and validated their observations in subsequent interviews. In a more focused, ethnographical study Mackay (1999) observed over several months how controllers handle paper flight strips. From her learnings on the role of the paper flight strip she concluded it was time to augment the artefact but keep its spirit.

Analysing recorded verbal exchanges among air crews and ground controllers can be considered an asynchronous observation. Joyekurun (2007) found through conversation analysis that work is drastically redistributed during weather induced critical events. Analysing the exact redistribution of tasks could inform possible more robust adaptive interfaces.

Performance/Usability. Performance outcomes are most critical for air traffic control and therefore serve as the last resort when deciding about the appropriateness of new interfaces. Depending on the way performance is operationalised, it may inform detailed design choices. Doble and Hansman (2004) asked departure planners in a within-design experiment to sequence ten aircraft and clear them for departure using five different designs of paper and electronic flight progress strips. The primary performance measurements were efficiency operationalised in runway occupancy time as well as effectivity and efficiency of detecting aircraft that mistakenly turned towards the runway during taxiing. Detecting critical situations on the radar quickly indicated a low head-down-time and thus an efficient operability of the flight strips. Additionally, participating air traffic controllers rated the perceived difficulty and their preferences regarding

the various experienced flight strip designs. Interestingly, the electronic flight strip configuration that gave the best sequencing performance was not the one controllers preferred most. Qualitative data revealed that in spite of rating the electronic flight strips experience higher than the paper flight strips, controllers found the early hardware holding the electronic flight strips too clumsy. In sum, the differentiation between five different flight strips in details of the configuration allowed Doble and Hansman to draw conclusions for future design even though the study setting resembled a summative usability test.

Physiological data. Determining workload through physiological parameters such as heart rate has applied in human factors research on air traffic since the 1960s (Fowler, 1969). However, recent years saw a rise in novel techniques, further variables and ideas of direct coupling between physiological parameters and adaptations of the task or interface. For example, Hill and Bohil (2016) provide an introduction to functional near-infrared spectroscopy (fNRI), a light reflection based neuroimaging technique, and point out its potential for measuring mental workload and emotional valence in domains such as ATC. This sounds promising but has not been tested so far.

Typically, eye-tracking has been used to either measure workload via pupil dilation or mapping gaze data on a picture or video of the environment and learn about users' current focus of visual attention or their attention distribution over time. A preliminary study with only one user shows that taking a closer look at the scan-path permits differentiating whether the air traffic controller was monitoring, planning or controlling traffic (Imants & Greef, 2011).

Y. Liu et al. (2019) calibrated a 14-channel EEG to distinguish three states of emotional valence (positive, neutral, negative) and four levels of workload in 12 subjects. They during a two-hour training, air traffic controllers had on average neutral emotions and the lowest levels of stress and workload were detected after 50 minutes. While in theory peaks of either measure could be mapped to specific situations in formative evaluations, in this study only 5 minute averages were reported.

A preliminary publication by Reisman and Kaliouby (2007) promised that via non-invasive optical real-time analysis of facial expressions emotional states and cognitive states (agreement, disagreement, confusion, disinterest) could be measured in pilots and controllers more conveniently. One decade later, experiences of a larger research project indicate that the promise of autonomous, reliable state monitoring cannot easily be fulfilled yet:

The project StayCentered attempted to physiologically determine controllers' current stress level and emotional state with the long-term aims of overload prediction or automatic system adaptation (Buxbaum, 2019). As means of exploration which measurement captured emotions best, the research consortium applied all that were available and feasible for long-term application. This included multiple video streams for the extraction and analysis of a) interaction between controllers, b) facial features, c) body posture and gestures as well as measuring d) eye-tracking

and pupil dilation, e) skin conduction and heart rate, f) running a sentiment-analysis on the audio recordings and finally g) subjective measures of workload and situation awareness. Their results disillusion the hopes in automation for routine application in every shift. First, variations in the light situation and head rotation were beyond the algorithms' error tolerance in optical measurements. Second, the "professional coolness" of selected and well-trained controllers resulted in few changes of mimic and gestures during their shift. Third, the sentiment-analysis is challenged by bilingual communication (here English and German) between controllers that additionally includes meaningful non-verbal units. Fourth, the identified emotion categories of the sentiment-analysis contradicted the results of the Facial Action Coding System (Ekman & Rosenberg, 2005) run on facial video recordings. In sum, the traditional physiological measures of galvanic skin response (EDA) and pupil dilation served as reliable indicators for workload whereas none of the data- and processing expensive methods sufficiently detected emotions (Buxbaum, 2019)¹⁰.

3.3 Need for a New Formative UX Evaluation Method

So far, we revisited methods from the domains of dementia and ATC that have either already been deployed to capture UX or are promising to do so. In this section, we proceed with summarising their strengths and weaknesses and conclude by suggesting a new formative UX method that combines the strengths of several methodological families.

Summary of UX evaluation methods in the Dementia Context. People with dementia are a diverse user group. HCI researchers utilise a broad repertory of techniques to better understand their needs and alleviate the deterioration of quality of live. The methods applied during formative evaluations range from very individualised participatory approaches to potentially scalable facial recognition and interaction logs. In spite of all technical progress, the predominant methods that inform design choices are variations of observation notes or manual annotation.

Handwritten notes appear to be the most frequently deployed method today, either alone or in combination with other methods. Their huge advantages are on the one hand the inexpensive setup and execution, on the other hand the freedom to write within or outside predefined codes. Manual notes allow descriptively mapping observed emotions and reactions to interactions. However, often no systematic approach is followed or described in the publication; thinking about what and how to write on manual notes as well as the writing itself capture a great amount of the observer's attention which then is unavailable for the actual observation.

Across systematic observation tools, distinguishing many categories of behaviour, interactions and emotions in-situ results in a decrease in psychometric quality, mostly reported as inter-rater reliability. The highest inter-rater reliability was achieved in studies with the OERS, a systematic observation tool with only five categories. Additionally, most quality of life tools are designed to

¹⁰More technical details are given in the project's final report (Brunett et al., 2018).

inform summative decisions. Their low timely resolution during which predominant emotions can be documented (e.g., down to 16 seconds for the OERS or 5 minutes for the DCM) hardly suffice to capture the high frequency of triggers and reactions for interactive technology. Both effects can be alleviated with thorough video annotations which allow precise linking of any number of emotions to triggers and make the results accessible to everyone. However, video analysis is very time costly that is, it can take more than 10 hours of coding for one hour of video in the dementia context (e.g., W. Liu et al., 2020). A limitation in terms of thoroughness is that documenting observed emotions in-situ provides richer context information and some emotions may not be detected accurately in the aftermath with the video as a sole source. Most promising appeared the tool BEAM (Casey et al., 2014) which allows observers to document affect via timestamps. A combination with video recordings could establish the bridge between each observed emotion and the situation that triggered it.

Collecting and analysing physiological data is very complex and methods that rely on body worn sensors render it hardly feasible and ethical questionable for the dementia context. Optical feature tracking and facial interpretation appear to be interesting but mostly distinguish only valence when more complex information could be extracted by human observers.

Interaction logs were merely queried for quantitative insights about use which does not allow for causal inferences. However, one could utilise the precision of interaction logs in combination with timestamped observations to directly map user reactions to their triggers.

Summary of UX evaluation methods used in ATC. Ultimately, an interface will only be approved if it meets performance criteria and supports air traffic controllers in maintaining safe, efficient and structured management of air traffic. Therefore, in summative evaluations, the performance that teams can achieve with an interface is the final frontier in ATC.

Summative evaluations of minimal-difference prototypes are one possible way to compare the influence of single features on UX and performance. However, due to the associated acquisition of enough participants for identifying inferential differences this is not practical in early stages of design.

The umbrella term physiological measures summarises a pool of instruments that are established in ATC and other fields as indicators for workload. So far, ambitious projects aiming at ample assessment of controllers' cognitive and emotional states through physiological parameters have not succeeded. For instance, recent studies show that physiological measures cannot reliably distinguish the subtle notions of emotion in controllers yet.

Taking notes of observations is feasible in the isolated atmosphere of workshops with fully functional prototypes and reduced traffic where the environment is controlled and hence does not need to be described in-situ. Even then, the high frequency of interactions is challenging for the evaluator. However, an evaluator can barely fathom all crucial events in the high complexity and fast pace of real shifts while writing manual notes in parallel. Multiple observers may be

required to note and capture the events in detail.

Logfiles and videos can contribute to accurately record the context, reduce the amount of information the evaluators need to manually describe in-situ, increase the evaluators' attention on the observation and ultimately increase data quality. Another potential advantage of video recordings is that controllers can make use of sequences or screenshots during debriefing interviews instead of drawing pictures that convey their perspective.

Proposing Proxemo as a novel formative UX method. Formative UX evaluations are thought to establish links between certain interactions with a prototype and triggered user reactions in order to inform design decisions in the next iteration. Self-report methods generally disrupt the user experience and in our application domains are not appropriate due to a) the communicative abilities of persons with severe dementia, or b) their introduction of unnecessary risk into safety-critical procedures. Note-taking through observers during formative evaluations is established in both fields and can produce rich descriptions of critical moments. Unfortunately, the process of writing distracts evaluators from the observation and may result in missed or misunderstood situations. Thorough analyses of video recordings allow for re-observations of missed reactions but are very time costly and still may contain situations that remain opaque in the recording.

Therefore, our suggestion is to combine the precision of timestamps with the rich context from video analysis. In the *Proxemo pipeline* (figure 3.1), we fuse the computer's ability to precisely capture and process extensive amounts of data with skills of human observers in recognising emotions and critical events through context cues. In detail, evaluators shall (a) observe user reactions such as emotions and (b) document them by proxy as timestamps in an application. Here, the required evaluator-application interaction should be minimal to maintain the evaluator's focus on the observed situation. Additionally, the user interaction shall be recorded on video. Synchronising evaluator's timestamps of observed emotions with the video recordings results in a pre-annotated video document where users' emotions are mapped to triggers. The role of users' emotions is not identical to UX, but a critical component of it and as argued in chapter 1, emotions are the observable key to gaining an understanding of a user's experience. In formative UX evaluations involving people with dementia, (c1) this file of thoroughly documented critical events can serve as the foundation for a more detailed video analysis where in-context documented emotions support the interpretation of opaque situations. In ATC, where users' media-supported recall of experiences and their verbal communication in detail can be presumed, (c2) annotated files facilitate methodological combinations of observations and subsequent interviews. For instance, when the thoroughly documented reactions and video material are quickly synchronised, the resulting file can be utilised in debriefings to efficiently navigate between critical observations and systematically discuss design choices.

We envisioned Proxemo to deliver direct mappings of emotional user reactions to their triggers

which provide the thorough foundation for systematic video analysis or discussions with users. Proxemo's main target group are practitioners whose focus is the formative optimisation of prototypical products or services. Qualitative researchers who seek to gain insights beyond the iterative design may choose to complement Proxemo with field notes or texts on key insights or crucial observations written up after the interaction session.

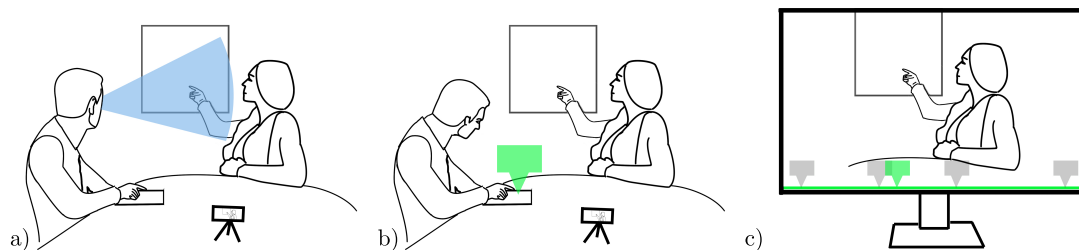


Figure 3.1: Schematic illustration of the *Proxemo pipeline* where interactions are captured on video and evaluators a) observe and b) document emotional reactions by proxy. Subsequently, c) emotional situations can be efficiently recovered via timestamp and analysed in more detail.

3.3.1 Discussion

Our intention was to find a suitable UX evaluation method for our situation in either domain and describe the methods we discover along our path. As none of the methods perfectly matches our needs, we identified the need for a new method. Following the terminology proposed by Baumeister and Leary (1997) our review therefore represents a mixture of the categories *state of knowledge* and *problem identification*.

We were aware from prior experience that human observers' ability to process a variety of context cues when assessing a person's emotional state cannot easily be matched by a machine. One insight that still came as a surprise during the review process is how weak recognition of emotions in air traffic controllers through machine learning is today. Controllers are by engagement mostly young healthy persons who already work in a very controlled environment. Those should be the perfect conditions for facial recognition which is highly dependent on the light situation (S. Li & Deng, 2020) and works best with young adults with a frontal pose (Bhattacharya & Gupta, 2019).

Limitations and future research. As in any review, we could only include the literature we were aware of. We briefly mentioned that there are hundreds of tools for quality of life assessment in the dementia context. We decided to include the two most popular in care settings (DCM) and most appropriate for technology evaluations (OERS) in our review but did not extend our search queries with the term *quality of life* because this added thousands of articles with a mostly medical focus to our matches. Whereas it was difficult to limit the search query in dementia, we

struggled finding enough relevant literature in ATC construable as relevant for UX. One reason for this is that in favour of performance the role of UX or emotions of controllers has barely been considered so far. Another issue is that even civil ATC research often takes place on a national level and not all reports are accessible to the international public. This safety concern may particularly affect preliminary formative evaluations which potentially reveal safety critical issues of the system or employees' emotional state.

In the dementia context, we explicitly designed Proxemo to support considering needs and experiences of persons with advanced dementia in formative evaluations who are not capable of communicating their needs. When able, persons with dementia should speak for themselves and be actively involved in the UX design process (Gilfoyle et al., 2021; Span et al., 2018).

One common issue of narrative literature reviews is their confirmation bias which can be avoided by a thorough search for counterexamples (Baumeister & Leary, 1997). On first sight, the suspicion of a bias towards finding the research gap that legitimates Proxemo seems natural. However, arguments countering this view are woven into the development of our work which is why we will share a short chronology here. Our initial dementia literature research in 2016 followed the aim to identify appropriate methods and best practices we could simply adopt during the formative evaluations of prototypes in the reminiscence context (project Interactive Memories, <https://www.intermem.org>). Employing a method that serves our purpose out-of-the-box would have cost the least effort by far. Yet, in the diversity of existing methods, none was suitable for the requirements in our project, and so we started to develop Proxemo. After two years of user centred iterations of the Proxemo method and an app that implements the method (we will describe in chapter 4) from 2018 on, we searched for a method that could serve as a fair opponent to Proxemo in summative evaluations. Surprisingly, among the variety of different measures and sensors, handwritten notes appeared to be the standard form of documentation in formative evaluations in either context. In the application domain of ATC, the consideration of user experiences beyond mere performance and workload is such a novel concept that it would have been hard to not find the research gap.

Having envisioned the concept of Proxemo as a method we will proceed in the next two chapters to check its feasibility in practice. Chapter 4 encompasses the iterative development of Proxemo as a method and the implementation of an app that facilitates its deployment in the dementia context. Chapter 5 describes the adaptation of the method for ATC and proposes another format for the app.

Chapter 4

Proxemo in the Dementia Context

In dementia care facilities, scheduled activation sessions are part of the daily routine and stimulate creativity, social interaction and reminiscence. Research artefacts deployed in this context can be optimised for individuals (Wallace et al., 2012) or groups (Huber, Berner, Uhlig et al., 2019) and range from musical instruments (Houben, Lehn et al., 2020), social robots (Cruz-Sandoval et al., 2020) and personalised interactive media (Hodge et al., 2019) to ambient multisensory setups (Feng et al., 2019) or virtual reality experiences (Bejan et al., 2018). In short, research focuses on technology which improves persons' quality of life through positive, technology mediated reminiscence experiences which we would simply refer to as a part of positive UX. When designing positive UX with and for people with dementia, we should make sure that the methods we intend to use are appropriate for this potential user group.

We argued in the last chapter that conventional usability methods that require self-reflection such as thinking aloud protocols and questionnaires do not produce reliable data in the context of dementia (Gibson et al., 2016) whereas observation methods are more promising. Neither assessment techniques for quality of care nor the versatile set of user experience evaluation techniques offer any structured observation methods so far that are optimised for UX evaluations in the dementia context. Following the need for a new method in the last chapter, we proposed *Proxemo* — a novel formative proxy UX evaluation method.

In this chapter, we report the user-centred development of Proxemo to meet the requirements constituted by evaluating interactive technology in the dementia context. To probe the feasibility of the method Proxemo we conceptualised and iteratively developed the Proxemo App as a tool¹ that allows for an efficient and discreet application of the method in technology mediated

¹In this chapter we focus on the *method* Proxemo and only summarise the design iterations of the *tool*. Studies 1 & 2 of this chapter have been priorly published in (Huber, Bejan, Radzey & Hurtienne, 2019) — there both

remembrance sessions. The design focused research question overarching all four studies is “How can we enable evaluators to document observed emotions by proxy in the context of reminiscence sessions for people with dementia?” Over the course of four qualitative field studies we explore the feasibility of using Proxemo as primary or secondary task, the ease of use and suitability of the Proxemo App, and the utility of the generated data for teams designing interactive systems. Insights gained along that formative path directly contributed to an iteration of the Proxemo App.

4.1 A Structured Observation Method for the Dementia Context

In formative evaluations, users’ reactions to a (prototypical) interface are captured and inform the iterated design. As shown in (Huber, Berner, Uhlig et al., 2019), user reactions are a result of users interacting with both, the interface and the presented content. While often hard to distinguish, both need to be considered in formative evaluations and Proxemo shall allow for documenting emotions triggered by either factor. However, since the exact nature of a trigger is not important for meta evaluations of Proxemo, for the sake of reading-ease, in the following we restrict to the term *interactions* without distinguishing in each instance whether the content or the interface were central for triggering a user’s interaction. To improve the system design based on users’ reactions a *mapping of reactions to specific interactions* is required. As highlighted in section 3.3 some reactions are only interpretable in context, and it is helpful to instantly document these (figure 3.1). However, when documenting the UX during actual use (ISO, 2018) by proxy there are three major challenges that need to be met in observational UX assessments before users’ experiences can inform design improvements.

Observability. As a first challenge, users need to display an observable reaction. This comprises the premises that interactive technology triggers emotions which are reflected in users’ facial expressions (Thüring & Mahlke, 2007) as well as observers’ ability to recognise those emotions (see section 2.3). Reminiscence technology is designed to evoke autobiographic memories and satisfy emotional needs of people with dementia (Lazar et al., 2014). A reminiscence intervention that does not trigger any emotional reaction is extremely unlikely and would indicate a conceptually wrong approach altogether. When the premise of emotionally rich interaction is met, how are emotions displayed? In our experience, people with dementia communicate their thoughts directly and are not known to intentionally hold back their emotions (opposed to some users in business context). A reason for this may be that masking their actual experience would require higher cognitive capacities (Proske, 2021). However, cognitive decline also limits the

formative studies are summarised as ‘study 1’ and contain more detailed findings on the tool’s form-factor.

communicative abilities and, for example, a progressing Parkinson’s Disease (one possible illness eliciting symptoms of dementia) hampers the interpretability of facial emotions (Rinn, 1984).

For an improvement in observability of facial emotions, this leaves the observers’ ability to recognise even subtle emotions as the most promising influence factor. In other words: it is important to involve observers who are highly trained in interpreting users’ emotional expressions. There are two groups of candidates for capable observers of emotions in a person with dementia. One group consists of *formal and informal* (e.g., family members) *caregivers* who have known the person with dementia for years and are highly trained to read a person’s emotions through daily contact. Through their intimate relationship, personal caregivers can differentiate individual ticks from meaningful expressions and best interpret the current emotional state of their patients. A second group is made up of *general dementia experts* who regularly evaluate care settings to consult care facilities and hence have contact with large numbers of patients with a variety of symptoms. In contrast to the personally trusted caregivers and for optimal results, these general experts need to shortly familiarise themselves with the person with dementia before starting the evaluation session. On the plus side, general experts are most familiar with evaluations and take the “outside view” on the interaction situation more naturally.

Documentation. A need for unobtrusive documentation of observed events poses the second challenge. In the best case, users’ experience forms a continuous stream uninterrupted by the evaluator. Unfortunately, evaluators’ documentation of observed experience shifts their attention away from the user. During that time evaluators either miss periods of users’ ongoing experience or ask users to pause the interaction while they finish their documentation — thus interrupting the users’ experiential stream. *Unobtrusive* means in this context that the documentation itself is very efficient and causes merely minimal distraction from the observation such that no subtle, yet important cues on user experience are missed. For evaluations where manual notes on few outstanding observations are desired, the new method shall keep redundant documentation to a minimum. However, we would advise against handwritten notes during overt observations as in a prior evaluation observers’ open notebooks gave the impression of a workplace to people with dementia and discouraged them from interactions assumed to be distracting (Huber, Berner, Ly-Tung et al., 2017).

A way to achieve an unobtrusive efficient documentation is offering *predefined categories* of expected and relevant emotions in the context of use. Representations of the emotion categories can be presented on an interface and simply be clicked or tapped to log the emotion. In contrast to handwritten notes, clickable categories potentially reduce the attention shift towards documentation and allow observers to keep their focus on the user.

Communicable through computer support. As a third challenge, once documented, the notes or codes shall be readable by other team members. Other than loose notes in the observer’s

handwriting, logfiles are in principle already human-readable by other members. However, the most important advantage of computer-supported documentation is that each documented event can be saved with a timestamp that allows a synchronisation with other data sources such as interaction logs or a video recording of the situation. The resulting emotion annotated video file enables a rich documentation capturing the users' experience in the context of their interactions with the system and other actors. This piece of consolidated information facilitates subsequent detailed analysis by the evaluators themselves as well as communication to others.

4.2 Design Solution: the Proxemo App

Technical requirements for a tool implementing a documentation aid for the Proxemo method are low. It needs to enable evaluators to set precise timestamps for logged emotional reactions and record those timestamped events in a file readable by video analysing software. Furthermore, as described above the tool shall be unobtrusive so that the documentation does not keep evaluators from observing or interacting with residents. Last, the tool needs to be always-on and at hand to log observed emotions whenever they occur. For example, a touch screen presenting an intuitive pictorial interface would satisfy the latter requirement. As our first prototype, we decided upon an application running on a smartwatch, which fulfils all above stated requirements and can be worn discreetly to attract the least possible attention from the persons being observed. A smartwatch allows evaluators to spontaneously log an observed emotional event while performing two-handed activities a second before and after its use.

Regarding the emotion categories for the dementia context, we took inspiration from literature. Lawton et al. (1999a) identified in their studies on quality-of-life in dementia the five frequently occurring emotions *pleasure*, *sadness*, *anxiety/fear*, *anger* and *general alertness* which we adopted for our first version of the Proxemo App. Taking inspiration from existing pictorial evaluation tools (Huisman et al., 2013; Laurans & Desmet, 2012) we decided upon a set of five emoji² representing each of the five emotion categories (see figure 4.1 and table 4.1).

Smartwatch interfaces generally underlie tight limitations of space. However, the screen size of most models suffices to display five emoji-buttons. For the interface layout we strived to avoid the impression of a hierarchy among emotion categories. Additionally, we aimed to establish a maximal spacing between emoji to prevent the so-called fat finger problem on the small display (Perrault et al., 2013). After initial scribbles on paper we set up a wireframe with the dynamic Prototyping software Axure (<http://www.axure.com>) and iterated the design with two UX evaluators and two experienced dementia care evaluators. This resulted in the additional requirement that the Proxemo App shall enable evaluators to document observed

²Intermediate versions of the Proxemo App have been presented as posters or demonstrations at various conferences to invoke discussions on the latest results at the time with experts in dementia (Huber, Preßler & Hurtienne, 2017) and user experience (Huber et al., 2018; Huber, Bejan, Radzey & Hurtienne, 2019; Huber, Preßler, Tung et al., 2017). These publications contain more details on individual design decisions.

emotions for multiple users at the same time and save them distinguishably. The seven cumulated requirements for the Proxemo App described hitherto are contained in the following list:

- R1 The interface shall always be available for the evaluator to document emotions with precise timestamps.
- R2 The form factor shall allow for an efficient documentation, causing the least distraction for the evaluator.
- R3 The form factor shall allow for a discreet documentation of emotions.
- R4 The predefined emotion categories need to be adapted to the evaluated domain (here: reminiscence sessions for people with dementia).
- R5 The interface shall provide sufficient space to accommodate all emotion categories.
- R6 All emotion categories shall be equally accessible.
- R7 The app shall facilitate a distinguishable documentation for multiple simultaneously observed users.

We could meet all those requirements by designing the Proxemo App for a round watch face. As displayed in figure 4.1, emoji were arranged with equal spacing in a circle. We implemented a first version of the Proxemo App on TizenTM(The Linux Foundation) for the smartwatches Samsung Gear S2 and Gear S3 (Samsung Electronics, Seoul, South Korea). Saving emotions distinguishable for multiple observed users (R7) was not realised in the first version of the Proxemo App.



Figure 4.1: The two screenshots show the first (left) and iterated (right) version of the Proxemo App with the feedback provided directly after an emotion was logged. Starting clockwise from the top, the emoji in the first version represent the categories *pleasure*, *sadness*, *anxiety/fear*, *anger* and *general alertness*. In the iterated version, the emoji represent the categories *pleasure*, *wistfulness*, *pride*, *general negative emotions* and *general alertness*. Additionally, pressing the centre button in the iterated version, documents an observed instance of *sense of agency* in the currently observed user. Rotating the bezel switches between users. For reasons of anonymity, the current user is here depicted in German as "links" [leftmost user]. However, portraits and names of participating users could be loaded into the app in preparation for a session.

Table 4.1: List of emotion categories with description and example as presented to evaluators. In studies 1...3 the categories *pride*, *wistfulness* and *sense of agency* were missing, and the three emotions pooled in *negative emotion* were still listed separately. Descriptions for *pleasure*, *anger*, *anxiety/fear*, *depression/sadness* and *general alertness* are adopted verbatim from Lawton et al. (1999a). Emoji are provided for free by emojihone.com.

Emotion	Description	Instantiations from early observations
 pleasure	laughing, singing, smiling, kissing, stroking or gently touching other, reaching out warmly to other, responding to music, statements of pleasure	A list of folk songs is displayed. The moderator wants to talk about the next title, but a resident spontaneously starts talking about past trips: “We always used to sing there. That was so beautiful! It made me think ‘now I feel alive!’”
 general alertness	participating in a task, maintaining eye contact, eyes following object or person, looking around room, responding by moving or saying something, turning body or moving toward person or object	When a caregiver directly addresses a resident with moderate dementia she only gives one-syllable answers but stays attentive almost for the entire session.
 pride	similar to pleasure; additionally: autobiographic relation, e.g. being proud of home town, special event/trip, grandchildren/children, own skill	“We always had two cows. [...] because, like I said, we had no money and then we just improvised. I like to reminisce about that.”
 wistfulness	resident is delighted by/ tells about a beautiful event from the past while being conscious about this time having passed; looking back with mixed feelings	Resident recognises the installation as aquarium/sea and exclaims: “Oh you could take a dive into this”. He then continues to tell that he was able to swim.
 negative emotion	<i>depression/sadness</i> : crying, frowning, eyes drooping, moaning, sighing, head in hand, eyes/head turned down and face expressionless (only counts as sadness if paired with another sign), statements of sadness <i>anxiety/fear</i> : shrieking, repetitive calling out, restlessness, wincing/ grimacing, repeated or agitated movement, line between eyebrows, lines across forehead, hand wringing, tremor, leg jiggling, rapid breathing, eyes wide, tight facial muscles, statements of anxiety/fear <i>anger</i> : physical aggression, yelling, cursing, berating, shaking fist, drawing eyebrows together, clenching teeth, pursing lips, narrowing eyes, making distancing gesture, statements of anger	Resident knows the [German folk] song <i>Hoch auf dem gelben Wagen</i> , remembers how she used to sit on a yellow chariot herself but does not want to listen to the song. She points at her head: “Memories are always there.” Resident leans back as far as possible on the sofa, away from the dog [avatar] displayed on the monitor wall. She declines the offer to pet it. Resident originating from Palatinate who is invited to comment on pictures from the Black Forest: “I have never been there and I do not want to see that.”
 sense of agency	special skill that is not taken for granted, targeted action, remembering and telling, recognising, ability to recite texts by heart, reading, singing	Reading a text aloud; recognising and naming a person/song/animal; imitating movements for playing the piano; remembering that a song used to be well known

4.3 Feasibility of Proxemo in the Dementia Context

People with dementia have unique skills and behaviours. Consequently, the emerging situations in reminiscence sessions are unpredictable and hardly comparable. Quantitative studies, let alone random-control-trials in the dementia context are therefore very rare³. To better understand the feasibility of Proxemo in formative evaluations of reminiscence technology and iteratively optimise the design of the Proxemo App, we conducted a series of four small qualitative studies looking at a variety of plausible roles and settings.

Context: Interactive Memories on a wall sized screen. All evaluations of reminiscence technology were part of the project *InterMem* (Interactive Memories). They took place in dementia care facilities in the Black Forest area in southern Germany. Residents with various stages of dementia participated voluntarily in the reminiscence sessions and agreed to the presence of an observer. During the ongoing session, caregivers continuously monitored the residents' mood and willingness to continue which is referred to as *process consent* (Dewing, 2007). Residents' legal representatives had provided written informed proxy consent in advance. In each study, participants taking the evaluator's role were introduced to the purpose of Proxemo and functionality of the Proxemo App, had time to explore its categories (table 4.1) and features (figure 4.1) and signed informed consent.

Unless indicated otherwise, all reminiscence sessions involved the *interactive wall*: A wall mounted cluster of screens that worked as one large touch screen⁴ with approximately 1.5 × 2.5 meters. Optionally, periphery devices for gesture recognition and remote control of the content could be integrated. The interactive wall allowed residents to experience multimedia presentations about personally meaningful topics (e.g., farm animals, hometown, current season) in group sessions or to explore virtual environments in single sessions (see figure 4.2). Both formats were moderated by a caregiver and lasted about 30 minutes.

4.3.1 Researcher

The researcher had conceptualised Proxemo and the Proxemo App and had two years of experience of user-centred reminiscence research in the dementia context. He disclosed his role as owner of Proxemo to all participants and highlighted its incompleteness, the need for further development and hence the relevance of honest participant feedback. The researcher and three of the observers were acquainted, having worked as partners in the research project InterMem. The researcher's role and level of participation varied slightly between studies and included the setup and introduction of the Proxemo App at the beginning, overt observations during the use of Proxemo and short interviews and debriefings at the end of each session. Interview questions (see appendix A.1) focussed on specific experiences during the studies rather than participants' general willingness to use or adopt Proxemo. While we cannot say with certainty that the researcher's role and acquaintance had no impact on the results, feedback from participants

³Astell et al. (2018) recruited an exceptionally large sample to show how cognitive stimulation improves cognition and quality of life over time. Even though their study lacks a control group, it is the closest call to quantified evidence for the effect of technological interventions we are aware of. For therapeutic approaches that cherish persons' individuality, small qualitative studies or even participatory design are more common (e.g., Hendriks et al., 2014; Houben, Lehn et al., 2020; Morrissey et al., 2016).

⁴Technically, gestures were recognised via infrared and could have been executed in a distance of 10 cm from the screen. However, in our observations, moderators mostly touched the screen directly.

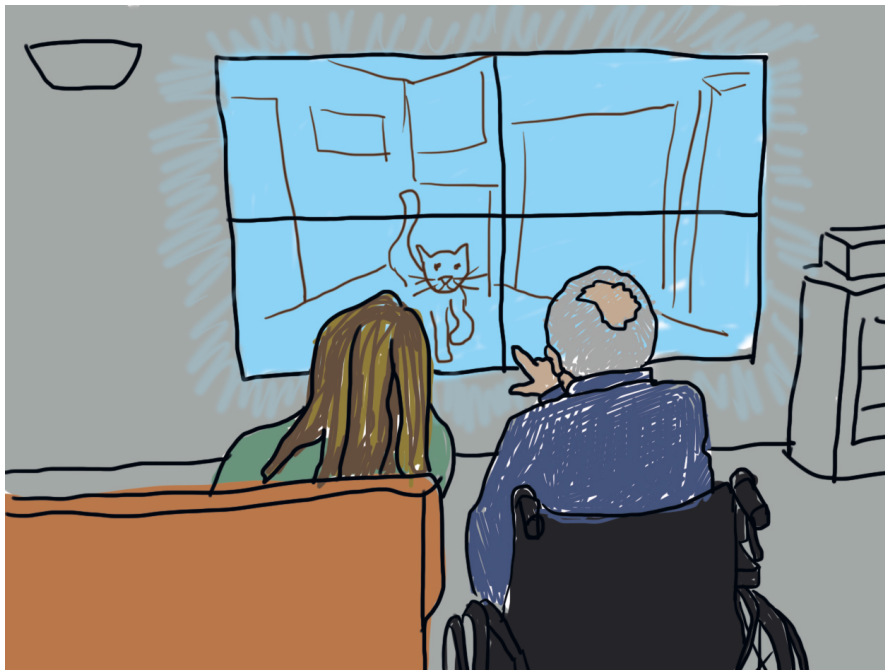


Figure 4.2: The drawing visualises a reminiscence session, where a resident (right) explores together with a caregiver (left) different topics on the interactive wall. Illustration printed with anonymous artist's permission.

appeared honest and covered negative aspects.

4.4 Study 1: Evaluating Reminiscence Interventions With Proxemo

Our first study aimed to explore the feasibility of Proxemo as an evaluation method in the least stressful scenario. That means, sessions were guided by a separate moderator, so evaluators could lay their sole focus on the observation.

Method. In nine reminiscence sessions of 11-38 minutes ($Mdn = 27.5$), a moderator guided varying groups of three residents through a multimedia presentation about meaningful topics or supported individual residents exploring a virtual Black Forest house together with a pet avatar (for details see Bejan et al., 2018). Evaluators consisted of three undergraduate students of health sciences and one caregiver with administrative responsibilities taking turns in observing. All four participating evaluators were female and between 21 and 27 years old ($Mdn = 21.5$). They were familiar with reminiscence technology in the dementia context but had few prior experience in observations and no previous contact to the residents in this study. Evaluators sat orthogonally

to the residents and the displays so that they could see both, the users' facial expressions and the content on the interactive wall triggering the emotions. We equipped them with the Proxemo App on a smartwatch and a clipboard for optional note-taking. A wide angled camera recorded videos of residents' profiles and the interactive wall. After the session, we interviewed evaluators on their experiences with Proxemo and asked them to validate or clarify our notes. Additionally, we asked them to complete the questionnaires QUESI on their perceived consequences of intuitive use (Naumann & Hurtienne, 2010) of the Proxemo App (scales ranging from 1 to 5 with 5 being best) and RAW TLX to measure subjective mental workload (Byers, 1989) (scales ranging from 0 to 10, with 0 indicating the lowest workload). Questionnaire data were collected to descriptively complement the qualitative statements.

Analysis. Handwritten notes of observations and participants' responses during the interviews were digitised. For this and all three subsequent studies, we openly coded and inductively categorised the textual data, methodologically guided by Qualitative Content Analysis (Mayring & Fenzl, 2019). The researcher coded qualitative digital data within days after the collection of a study was completed.

Findings. Evaluators' first impression was that the logging of emotions took them only 1-2 seconds which they subjectively rated hardly distracting and more efficient than taking notes on paper. Evaluator (E) 1 found it "easily manageable to take notes on the side" when residents displayed only few emotions. Others considered writing "not necessary in this situation" (E4) and considered "the watch [serving] better as a standalone unit since taking notes next to it needs too many resources" (E3) so "[they] would miss important emotions" (E4). E3 came to the conclusion that "the watch was really good in the situation and afterwards — meaning directly after the intervention [I did] the writing. Using paper [during the intervention] would seem like [being] in an exam [for the residents]. The watch is more discreet. Wearing the watch is more practical than holding it in ones hand." Evaluators' ratings on the QUESI scale ($Mdn = 4.46$, $range = 3.93 - 4.64$) and the RAW TLX ($Mdn = 1.75$, $range = .83 - 4.42$) indicate that the Proxemo App was perceived as intuitive to use and did not cause a high workload. This intuitive interaction and low workload is supported by a statement from E1: "one simply has to push the right button".

During group sessions, evaluators used Proxemo to document the emotions of all three participating residents. Regarding the pre-defined emotion categories, E2 noted that Proxemo "contains all the important emotions". E1 reported that she missed emotion categories for situations when residents were very surprised/astonished by either the presented content on the interactive wall or the comments of the moderating caregiver. Evaluators also pointed out the lack of emoji for "residents falling asleep" (E1), being "disinterested" (E3), or "distracted — as opposite to general alertness" (E2). E4 reported that she "could never observe anxiety [and] that anger

and sadness were not always clearly distinguishable in people with dementia”. Examining the Proxemo logfiles, the most frequently documented emotions over the course of nine sessions (221 minutes in total) were pleasure (242) and general alertness (91), followed by sadness (20), fear (13) and anger (4). An explanation for the high prevalence of fear is that E1 misused the category to log instances of “surprise about content or explanations from caregivers”, making up for 3 instances alone.

In sum, evaluators found Proxemo to cause little obtrusion for themselves and the residents. We learned that Proxemo changes the character of an evaluation. Through Proxemo, observers gain more time to actually observe because they do not need to spend their cognitive resources on note-taking. The predefined set of emotion categories is beneficial for novices. Evaluators instantiated the positive categories frequently while they rarely observed negative emotions which they found hard to distinguish. Instead, evaluators expressed the need for additional categories without offering coinciding suggestions apart from general disinterest in the activity.

4.5 Study 2: Video Analysis With and Without Proxemo Data

In the second study, we explore the utility of Proxemo. We investigate specifically whether the Proxemo data generated in study 1 is beneficial for video analysis by team members not being present during the data collection.

Method. We synchronised video recordings from study 1 with the Proxemo timestamps logged during the sessions using the video annotation tool ELAN (Max Planck Institute for Psycholinguistics, The Language Archive, Nimwegen, Netherlands). The resulting file was made available to the student teams who had created the multimedia presentation and the virtual house and avatar experience in study 1 (*Proxemo Team, PT*). The team members analysing the videos (undergraduate students of information sciences in healthcare) had not taken part in the evaluation. Their task was to thoroughly evaluate the video with respect to people with dementia’s emotional reactions to the prototypes.

As a contrast, a second team with backgrounds in nursing sciences analysed video material from interventions with the interactive wall where Proxemo had not been used. Those evaluators also extracted emotions from the video material without the supporting timestamps set in situ but, therefore, also without being biased by or limited to the set of five emotions of the Observed Emotion Rating Scale (Lawton et al., 1999a) used in Proxemo (*Video-Only Team, VOT*). Evaluators who performed the video analysis shared their experiences with us during a short interview which we inductively coded.

Findings. Since emotions were not saved distinguishably for multiple observed users in the early version, the Proxemo Team had to manually assign timestamps from group sessions to respective users. Assigning the documented emotions to one of the residents in the video caused some extra effort but was almost always unambiguous. “Mostly [the emotion timestamp] belongs to the person who is either laughing or with whom the caregiver is talking, hence the observer’s focus is on them as well” (PT1).

The data from Proxemo gave “additional assurance [and] definitely made the analyses easier” (PT2). Proxemo data became particularly relevant “when residents’ faces [were] not visible in the video because residents moved beyond the captured area, turned away from the camera or caregivers stood between the camera and the resident. Sound alone was sometimes not enough to recognise the emotions so those would have remained ambiguous without the Proxemo data” (PT1).

Just the descriptive frequency of events documented in Proxemo already offers some insight. PT1, who evaluated several sessions, appreciated the aspect of how “the tables with raw data — without having to read or even write the whole transcript — gave a quick overview on which sessions worked for which resident.” Analyst PT2 used the Proxemo timestamps that were imported in ELAN with a standard duration of one second as a starting point to mark the observable duration of an emotion in the video recordings. He found it “difficult to say ‘I experienced pleasure 50 times’ [but] better to say ‘I had pleasure for 12 minutes’.

Regarding the selection of emotion categories, the analysts agreed with the observers in study 1 upon the lack of a category for “absentmindedness because residents often appear distracted and one resident even dozed off for two or three minutes” (PT1). For the categories of positive emotions, PT2 found a need to „maybe add something between general alertness and pleasure”.

The Video-Only Team analysed the videos of sessions without Proxemo and proceeded differently which inspired our further development of Proxemo. The VOT did not code each emotional expression but only emotions that occurred in association with moments interpreted as autobiographically informed meaningful for the person with dementia. In total, the VOT coded fewer emotions but identified moments of pride or wistfulness — emotions that go beyond the set initially derived from the Observed Emotion Rating Scale (Lawton et al., 1999a). For example, residents proudly talked about their hometown and the amount of cattle they once owned, or they wistfully reminisced over a chapter in their life that had been good but was clearly over. The evaluators also tagged residents’ skills (e.g. singing, reading, remembering something) that became apparent during the intervention and that gave the persons with dementia a feeling of pride and pleasure.

From this study we learned that the predefined set of emotion categories alleviates video analysis but limits the richness of data noted in context. Consequently, the more detailed interpretation of observed events is shifted towards the video analysis. Proxemo data supported

the navigation in video files and facilitated the interpretation of situations where relevant information was not visible or audible in the video. As one observer stated in study 1, the most comprehensive data is generated when using Proxemo during the observation and subsequently writing down insights in addition.

4.6 Study 3: Proxemo Usage on Top of Moderation

The focus of this diary study lay upon the feasibility of Proxemo as a secondary task next to moderation. Reminiscence sessions were moderated by a caregiver who additionally logged the observed emotions.

Method. We conducted an event-contingent diary study (Hyers, 2018) with one caregiver as sole diarist whose demographic data can not be shared in more detail without violating anonymity due to the small research group. The diarist was recruited after her participation in a prior study because she regularly moderated reminiscence sessions with people with dementia. We provided her with written instructions on how to start, charge and operate the smartwatch and how to start the Proxemo App after a reboot so she could use the smartwatch for a longer period. Furthermore, we instructed the immediate logging of observed emotions and documentation of each trigger by a separate timestamp. Finally, the instructions contained a list of explanations for the emotion categories adopted from Lawton et al. (1999a), see table 4.1. As the focus lay on the feasibility of using Proxemo next to moderation and not the analysis of emerging data, we followed the principle of data economy and spared video recording the sessions.

The diarist decided to make notes during or immediately after the session and type her formulations subsequently at the end of her shift. In each entry, the diarist recalled the general context of each session, described the events in which she used Proxemo and reflected upon her experience. We first read the diary, finding that all entries were fit for further analysis and then inductively coded the data.

Findings. The diarist had access to the Proxemo App for about four months. Due to her administrative responsibilities, illness and staff shortage, she only managed to use Proxemo three times. All three usages happened within the first month.

In the first documented session, the diarist used a tablet application (Mediademencia, Media4Care, Berlin, Germany) to explore an illustrated book about spring flowers with a 78-year-old resident in the early stage of dementia who absolutely enjoyed the pictures and associated memories of her own garden. During the session, the diarist tried to use the watch and additionally take notes next to moderation which she found stressful. She reflects in her diary that she forgot to log separate events for distinct triggers, i.e. “pleasure about the pictures’ content, then [pleasure about the] self-efficacy when the resident was capable of reading a text [on the same page] by

herself and was happy afterwards”. In addition, the diarist was confused, whether she should also document *general alertness* as in her understanding attention and interest pose a precondition to situations that lead to pleasure or other emotions.

For the second documented session, the diarist watched a dog video on the tablet application together with an 81-year-old resident with middle stage dementia. The resident was upset and cried before the session started but laughed on first sight of the dog video. Behaviour of the video dog triggered further laughter throughout the session and reminded her of her daughter’s dog. The diarist noted that “in this form of session, the emotions could be linked to video sequences in which the trigger is identifiable”. However, she admits that in between she had trouble to decide whether an emotion was triggered anew or still continuing from the previous trigger. Overall, we agree with the diarist that “this irritation may result in inaccuracies in documentation”.

In the third documented session, the diarist gathered a group of four residents to reminisce about the topic *garden in springtime* and subsequently play a short ball game. Three residents had mid-stage to advanced dementia coupled with aphasia and one resident was in an early stage of dementia. Pictures of roses and lilac spread joy in the group and reminded the fittest resident of own gardening experiences and springtime songs. Catching and throwing the ball triggered self-efficacy and pleasure. The diarist noted that for people with “dementia and aphasia [it is] difficult to distinguish whether pleasure is triggered by a memory or the picture itself”. On top of moderating and using Proxemo, the diarist tried to take qualitative notes during the session. The reason for this is that she wanted to capture the context in such detail that she could tell afterwards which resident experienced which emotion in which situation. Thus, she re-enacted the effortful manual note-taking that inspired Proxemo in the first place — except handwritten notes in-situ are usually taken by non-participating evaluators, not the moderators. Of course, we had not stipulated this approach but now know for certain that moderating a group session while using Proxemo and *additionally* taking handwritten notes is too much. Interestingly, the diarist was not deterred by her experience and concluded that “Proxemo is well applicable in group sessions in combination with either notes or video recordings”.

From this short diary study we learned that using Proxemo next to moderating the session is feasible. However, a way to distinguish plain pleasure from pleasure due to own achievements (self-efficacy) or memories of former achievements (pride) is required. Furthermore, a possibility to document emotions for distinct residents in the app might diminish the urge to take notes about who showed what reaction. Finally, we must elaborate our instructions to clarify that each identified trigger requires the separate documentation of an emotion even if the previously triggered emotion was the same. A clarification of instructions should also suffice to inform evaluators that *general alertness* — or interest (as the category is labelled by Lawton et al. (1999b) in some versions of their scale) — represents the most positive reaction displayed by people in advanced stages of dementia. For those still capable of expressing more extreme emotions than a shift of attention, *general alertness* serves only as basis for all other positive

emotion categories and hence only needs to be documented if there is no emotion observable that goes beyond general alertness.

Redesign of the Proxemo App. The subsequent novel features of the iterated version of the Proxemo App addressed several of the former shortcomings regarding emotion set and user distinction. In detail, evaluators could switch the user for whom an observed emotion should be logged by rotating the bezel of the smartwatch. On the interface, this rotation in either direction iterated through the list of predefined user portraits or placeholders displayed in the centre of the watch face. When documenting an event, the title of the currently present user was then written in the logfile with emotion and timestamp. Additionally, we had learned that the emotion set of the observed emotion rating scale for generic evaluations in care settings (Lawton et al., 1999a) was not optimised for evaluations of reminiscence technology. Study 1 and 2 taught us that the negative emotions anger, sadness and fear were neither very frequent nor easily distinguishable. As the documentation of negative experience is extremely important for formative evaluations, we decided upon keeping the definitions of all three emotions but merging them into one generic *negative emotion* category. In return and as suggested by the VOT in study 2, we added *pride* and *wistfulness* as distinct emotions to the set of categories. Finally, we turned the centre picture of the currently observed user into a clickable button. Pushing the centre button wrote a timestamp with *sense of agency* in the logfile — as implicated in study 2 and 3.

Thus, with the novel version of the Proxemo App we enabled evaluators to distinctly document emotions for multiple users. Particularly relevant for the reminiscence context was the fact that evaluators could now distinguish in their documentation mere *pleasure* from pleasure triggered by own accomplishments (*sense of agency*), pleasure about past achievements (*pride*) and consciousness that beautiful, joyful events lay now in the past (*wistfulness*). See table 4.1 for descriptions.

4.7 Study 4: Expert Evaluation in Reminiscence Sessions

The fourth and final study in the context of dementia has two aims. First, we tested whether the iterated version of the Proxemo App was easily understandable and applicable by caregivers and evaluators. Second and more importantly, the suitability of Proxemo was explored and judged by an evaluation expert for dementia care settings.

Method. We tested how applicable the final version of Proxemo is for logging observed emotions of people with dementia over the course of four scheduled reminiscence sessions with 1, 2, 3 and 4 participants. All sessions took place in an urban care facility within a display wall setting similar to that from the formative evaluation in study 1. Four different caregivers with no prior knowledge of Proxemo and up to two years of experience with smartwatches (all female, aged

51 – 57, $Mdn = 57.5$ years) conducted the interventions and decided upon content and media they used depending on the residents' mood during the sessions. Proxemo was primarily used by a trained evaluator with over 15 years of experience in evaluating dementia settings using the Dementia Care Mapping Method (Innes & Surr, 2001) and who was also familiar with the Observed Emotion Rating Scale (Lawton et al., 1999b) but did not personally know the residents in advance. She received a short introduction to the Proxemo method as well as the features and functionality of the Proxemo App. She was asked to document all observed emotions and moments of agency so a person without much experience in the dementia context could understand the users based on the video and the accompanying timestamps. For two sessions each, she tested Proxemo running on the smartwatches Gear S2 and Gear S3. The caregiver moderating the session received the same introduction to Proxemo and was given the other smartwatch (Gear S2 or Gear S3). However, we instructed the caregivers to keep their focus on moderating the session and only document emotions if they had spare capacity. Each session lasted for about 30 minutes. Caregivers summarised their impressions in brief statements upon returning the smartwatch but we spared interviews due to their tight schedule. Analogue to study 1, we asked the caregivers to complete the questionnaires QUESI (Naumann & Hurtienne, 2010) and RAW TLX (Byers, 1989) after they finished moderating “their” reminiscence session. The expert evaluator filled in the questionnaires after her fourth reminiscence session to capture a more thorough experience with Proxemo. Questionnaire data were collected to descriptively quantify the evaluators' experience with Proxemo.

Findings. After using Proxemo for four sessions with 1..4 participants, the evaluator rated Proxemo as easy to use (QUESI score = 4.43, all subscales ≥ 4), causing low effort (RAW TLX score = 1.75, all subscales ≤ 2.5) and described it as appropriate for the context. Her only concern was the amount of users observed at a time. When residents with high levels of activity were observed, she found four residents to be the upper limit. She preferred the Gear S3 over the S2 for evaluations due to the larger display and more gentle haptic feedback during bezel rotations. Smaller size allowing for a more discreet interaction was identified as sole advantage of the S2. Over the course of all four sessions (150 minutes in total) the expert descriptively documented more ($n = 200$) observed emotional events than caregivers ($n = 97$): the highest number of registered timestamps pertained to instances of agency (expert evaluator: 95 | caregivers: 39) and pleasure (81 | 31), followed by general alertness (16 | 13), wistfulness (5 | 5), negative emotions (2 | 8) and pride (1 | 1).

Caregivers' statements indicate that the main reason for fewer timestamps set by them is that they were primarily engaged with moderating the session and paid less attention to documenting its effects. Our observation confirmed this — particularly the caregivers who interacted with three and four persons barely used Proxemo. From caregivers' feedback we learned that mere tapping on emoji buttons is a manageable interaction on top of the moderation. However, rotating the

bezel steals too much attention from the main task of moderating the session. Caregivers reported descriptively lower scores for intuitive use and higher subjective workload ratings than the expert with QUESI scores averaging below the centre of the scale ($Mdn = 2.46$, $range = 1.64 - 3.5$) and RAW TLX ratings above the centre of the scale ($Mdn = 6.38$, $range = 2.67 - 9.0$). Interestingly, *negative emotions* is the only category documented more frequently by caregivers than by the expert evaluator. We only spotted this difference after the study when retrieving the logfiles and cannot determine post-hoc whether caregivers were generally or situationally more sensitive to negativity or had a better angle to perceive these emotions. Caregivers did not report having glitched exceptionally often on the emoji button representing the *negative emotions*.

From our fourth study we learned that an expert considered Proxemo as suitable for formative technology evaluations in the dementia context with up to four residents. For caregivers who document emotions in Proxemo while moderating a reminiscence session observing a single user is the limit. In contrast to caregivers who document emotions next to moderating the session, the expert documented about twice as many emotions. A higher amount of documented emotions does not necessarily implicate a higher thoroughness (sensitivity) in documentation as it could also point towards large amounts of nonsense-data and a decreased validity (specificity). However, since caregivers admitted they barely had time to document emotions, we assume the expert's documentations were more thorough.

4.8 Discussion

In this chapter, we first introduced the novel structured observation method Proxemo and iteratively developed the Proxemo App – a tool that facilitates the deployment of the method and enables evaluators to document observed emotions. Moreover, we conducted a set of four studies to qualitatively evaluate the feasibility of Proxemo and the Proxemo App in reminiscence sessions in the dementia context as well as the utility of generated data. From various staff constellations we learned that as a method, Proxemo is generally considered suitable and leads to best results when used by an evaluator who can fully focus on the observation of up to four residents. Since the expert who used Proxemo in study 4 is more experienced and possibly more capable than the average evaluator in the dementia context, we recommend restricting the number of users observed at a time to three. Evaluators developed the best practice to thoroughly document observations with the Proxemo App during the session and take detailed handwritten notes of particularly meaningful events in the aftermath.

The generated Proxemo data from observations showed to be useful for video analysis as it sped up navigation in the video files and allowed interpreting ambiguous situations. However, the pre-defined set of emotions limits the richness of documented emotions. In order to limit this adverse effect, careful consideration needs to be given to the exact set of predefined emotion categories.

Regarding the set of emotions, we started off with adopting a pre-existing scale from literature (Lawton et al., 1999a) and adapted it based on qualitative data from initial applications. In detail, we consolidated the three negative categories (fear, anger, sadness) and extended the list with categories particularly relevant in the reminiscence process (pride, wistfulness, sense of agency). Apparently, other researchers who recently used the observed emotion rating scale (Lawton et al., 1999a) found only few instances of negative emotions as well and consequently collapsed them into one generic category from the beginning (Steinert et al., 2020), post-hoc (Cohen-Mansfield et al., 2012; Feng et al., 2020) or ignored them entirely during analysis (Feng et al., 2019). The self-conscious emotion of pride is part of most models and emotion word lists (Petta et al., 2011; Remington et al., 2000; Scherer, 2005; Yik et al., 2011) and is accurately recognised by observers (Tracy & Robins, 2008). Wistfulness is our label for a category of mixed emotional reactions that has been titled *nostalgia* by other researchers (Cowen & Keltner, 2017; Watson & Stanton, 2017).

Regarding the Proxemo App, a descriptive examination of the QUESI scores and RAW TLX ratings reveals that the expert evaluator's ratings in study 4 are similar to the evaluators' ratings in study 1. This indicates that extending the set of predefined categories and adding multi-user documentation did not drastically boost the workload or make the Proxemo App appear less intuitive to use.

4.8.1 Limitations

For earliest explorations in study 1 and 2, all observers but one had little prior knowledge of dementia. While this did not result in higher subjective workload it may have led to fewer correctly identified emotions. Of course, our four qualitative studies with small samples cannot suffice to fully fathom the method's performance in all possible technological interventions. However, the constellations are representative of many reminiscence technology settings and may give impressions that help assessing the appropriateness of Proxemo for future projects. In all four formative tests, evaluators were only shortly introduced into the purpose of the method and the functionality of the Proxemo App. To ensure a high representation of observed users' emotions in the data set, evaluators should be selected or trained a) regarding their general empathy, b) their understanding of the context of study, and c) instructed to pay less attention to their own emotions while coding the inferred emotions of others.

In the findings of our fourth study, we speculated about thoroughness and validity of Proxemo. So far we found Proxemo to be generally feasible in the context of dementia but due to the small sample size and qualitative nature of our studies we are not able to make a final judgment regarding the quality criteria of Proxemo. Thus, the values for validity and thoroughness of the Proxemo method are still unknown. To determine these, a proper experimental evaluation of the method regarding its quality criteria will be reported in chapters 6-8. Before examining

Proxemo's quality criteria in lab studies we tackle another limitation of generalisability and probe Proxemo in a different field. So far, we applied Proxemo only in the context of dementia which limits our knowledge of the structured observation method's generalisability to other contexts where users can barely speak for themselves during interactions. We will address this limitation in the next chapter, introducing Proxemo in the domain of air traffic control where a selection process assures the general ability of observed users, but their available cognitive resources are restricted by the highly demanding tasks.

Chapter 5

Adapting Proxemo for Air Traffic Control

As described in chapter 1, the safety-critical task of air traffic controllers binds most of their cognitive resources and additionally requires ambidextrous interaction with the workstation. Therefore, depending on current traffic load, the reflection of perceived emotions or even a documentation of situations is not continuously feasible by the users themselves during real operational shifts or high fidelity simulations. Proxemo may serve as a suitable alternative to self-report in formative UX evaluations in the context of air traffic control, even in periods of complex traffic. Hereby, the main difference to the dementia context is that the limitation of air traffic controllers' cognitive resources is only temporary. By selection, air traffic controllers possess a high cognitive capability which they can use to reflect upon their experiences and communicate them – once their shift is over. For the Proxemo pipeline introduced in chapter 3 this means that after (1) observation and (2) documentation of emotions by proxies, the (3) analysis of relevant situations may take place in participatory workshops with the users themselves, for example as debriefing interview. In this chapter we first outline briefly how we adapted the set of relevant emotions and the form factor of the documentation aid referred to as *Proxemo App* to ATC and then conduct a case study to answer the research question whether Proxemo is feasible for formative evaluations in the context of simulated approach control.

5.1 Design Solution for the ATC Context

Task and context of ATC. In approach control the roles of *pickup* and *feeder* work closely together. The pickup controller picks up aircraft from previous lower airspace sectors, then decreases their altitude and speed before handing them over to the feeder at a predefined area and in a previously agreed upon state. The feeder controllers manage the continuous stream

of aircraft onto the downwind leg and final approach which they feed into the runways. To efficiently harvest runway capacity, the feeder's task involves optimizing the separation between aircraft before handing them over to the tower controller. For a close collaboration between both positions, the workstations are typically located side by side, in some instances in a mirrored arrangement of the respective interfaces (see figure 5.1). We can use this mapping for the Proxemo App in order to allow intuitive user allocation side by side on the interface.

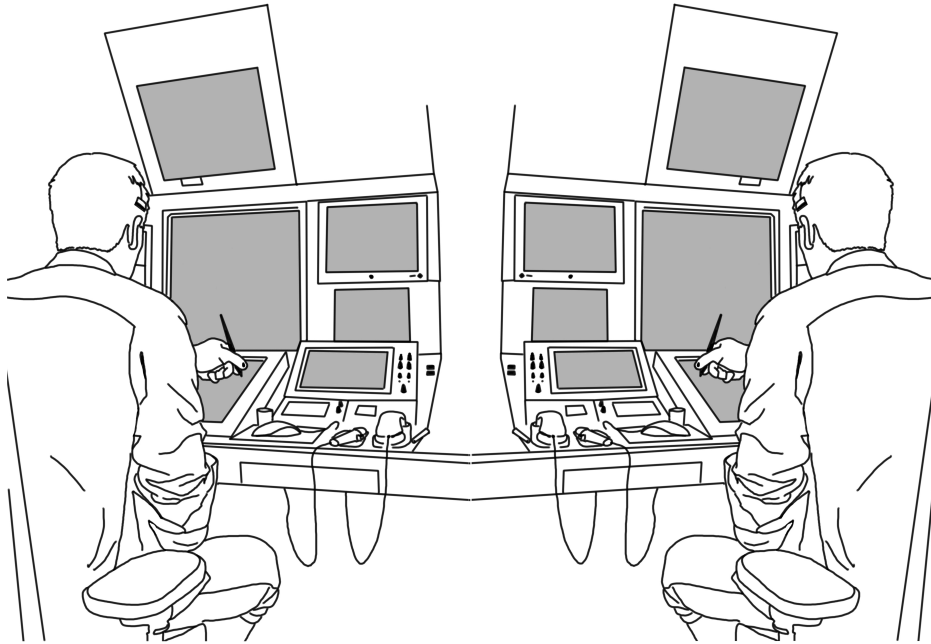


Figure 5.1: The sketch schematically represents the mirrored workplace of two collaborating controllers on the pickup (left) and feeder (right) position. The figure is based on an illustration by Cordula Baur.

Task and context of observer. In the context of ATC, colleagues and supervisors make the perfect observers because they can empathise best with task and context and are also familiar with the users (fellow controllers). To optimise direct communication among controllers of different sectors, the workstations in the operation room of a control centre are arranged in rows with colleagues from adjacent sectors being positioned next to each other. Therefore, the selection of possible positions for the observer using Proxemo is limited. We took inspiration from training situations where the coaches typically sit or stand behind the controller. From there, they have the second-best view on the screens after the controller, only a small chance of distracting them by their presence and are even able to gain a better view of the adjacent workstations. A possible restriction for the use of Proxemo in this setting is the restricted view of the observed controller's facially displayed emotions. The case study detailed later in this chapter will assess

the importance of this restriction.






Identifying emotion categories. Naturally, emotion categories observable in air traffic control are not identical to those in reminiscence sessions. To identify appropriate emotion categories for air traffic control, we followed a bottom-up approach. We conducted a directed content analysis (Hsieh & Shannon, 2005) on transcribed interviews with approach controllers. The interviews had already been conducted as part of an ethnographic study in approach control that is described in more detail in Huber et al. (2020). Interactions with the workstation and other actors as well as the complexity of traffic triggered the emotional experiences highlighted in the following.

Controllers' responsibility for sectors is partly dynamic and designed in a way that they are usually required to handle average amounts of traffic. They describe this default state as a *"relaxed shift"*. Periods of low traffic, during night shifts for instance, in combination with an interface that requires only few inputs evoke *boredom* in controllers. Periods with a high amount of complex traffic or exciting situations keep controllers *"busy"*, set them *"under pressure"* and cause *"positive stress"*. If they resolve these situations through *"competence [and] professional collaboration"*, they experience *"self-efficacy"* and *"pride"*. Only in rare occasions, stressful situations cause *"overextension"*. *"Anger about colleagues, the system [or] a situation badly dealt with"* is rather rare. Similar to other domains, teamwork induces experiences of *"joy and fun"*, *"conflict"* and *"solidarity"*. Since controllers are trained to expect everything and are prepared for all kind of situations, *surprises* occur mostly when *"the interface reaction differs from my expectation"* or when — despite all their technostress — controllers *"get along surprisingly well with the [new] system"*. In sum, we decided upon the emotion categories *pride*, *surprise*, *stress*, *anger* and *boredom* with *relaxation* as the default state (Table 5.1).

Arrangement of emotion categories on the interface. According to Nilsen (1996), the time users require selecting items (here: emotion categories) from a list depends not only on the length of the list but also on its structure. In his experiments, Nilsen used natural numbers (1 – 9) as items which can easily be arranged in a natural order. Even though some emotion models suggest a natural order of emotions in two-dimensional space (Plutchik, 2001; Scherer, 2005; Yik et al., 2011), there is no natural conclusive arrangement known today presenting the output in a simple ordered list. Attempts to create accurate mappings of emotions resulted in high-dimensional clusters (Cowen & Keltner, 2017). Alternatively, one could order emotion categories by the expected frequency of their occurrence. While this might speed up the selection process on the interface, it would likely increase the observer bias — especially if observers were made aware of the order criterium. Therefore, we arranged the emotion categories represented by labelled emoji (Table 5.1) on the interface to be equally accessible.

Together with an undo-feature that had been missing in the watch face implementation of

Table 5.1: List of emotion categories relevant in air traffic control where the default state is considered *relaxation*. Emoji are provided for free by emoji.com.

Emotion	Description from ethnographic data	Instantiation in ATC
 anger	negative emotion; being frustrated, annoyed or upset about something that went wrong or was not achieved; feeling like ranting or swearing	e.g. being frustrated by miscommunication or the interface
 boredom	being unchallenged and impatient, because nothing interesting is happening and one is condemned to idleness	e.g. low traffic
 stress	emotional or mental tension caused by e.g. imminent loss of control	e.g. high amount of traffic with an additional emergency, challenging communication or interaction concepts
 surprise	being confused by an unexpected event	e.g. being surprised or irritated by behaviour of colleagues, pilots or the interface
 pride	joy, caused by an achievement; success through one's own skills; the cause can be an event or experience in which self-efficacy was experienced	e.g. being proud of a mastered situation, one's own performance

Proxemo, we implemented the App for Android (Google LLC, Mountain View, CA; see figure 5.2) complying with the following list of requirements:

- R1 The user allocation on the interface of the Proxemo App shall map to user allocation on the work station.
- R2 The form factor shall allow for efficient and least distracting documentation of emotions.
- R3 The emotion categories need to be adapted to the work domain of air traffic controllers.
- R4 All emotion categories shall be equally accessible.

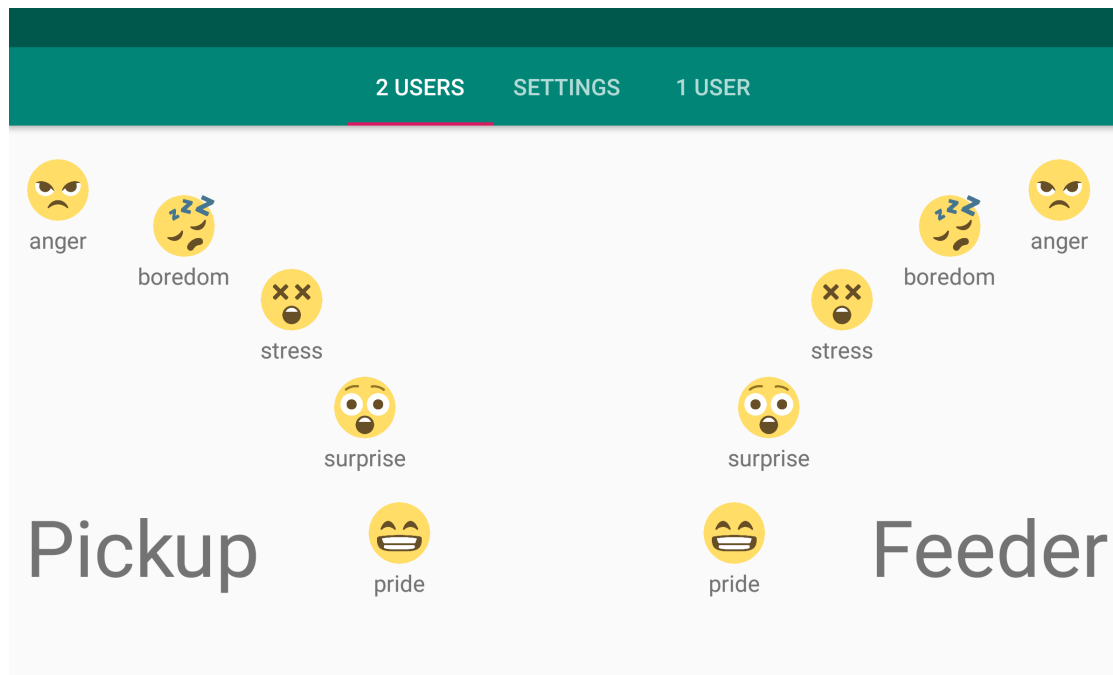


Figure 5.2: The screenshot displays the documentation screen for two users of the Proxemo App running on a 6” Android phone in landscape mode. Emoji are arranged to be accessible quickly by the twitch of a thumb. The observer sets the usernames to approach positions.

Analogue to the documentation aid implemented for the dementia context, the Proxemo App is kept as simple as possible. Tapping an emoji sets a timestamp with the respective emotion category for the respective user. After this documentation event, a confirmation message (*snackbar*) is displayed for three seconds at the bottom of the screen. It informs which emotion was logged and offers an “undo” button allowing the user to delete the latest timestamp.

In contrast to evaluations of reminiscence sessions in dementia care facilities where Proxemo logfiles and video recordings are thoroughly analysed days or weeks after the observation, Proxemo files in air traffic control need to be immediately synchronised with video recordings and analysed with controllers in-situ. Due to the typically tight schedule of simulation runs or shifts, a dense workflow starting with the file transfer needs to be supported.

Most requirements listed up to this point have been focusing on optimizing the documentation tool for the very specific environment of approach control. Thinking beyond those two working positions and even beyond air traffic control, the app needs to be more flexible. Regarding the settings of the app we, therefore, extended the set of requirements of which R5-R7 have already been implemented today (figure 5.3) since they benefit the context of air traffic control as well:

R5 Logs shall be easily transferable to video annotation programs.

R6 The observer shall be able to easily rename users.

R7 The app shall support the observation of only one user.

R8 Emotion categories shall be easily exchangeable.

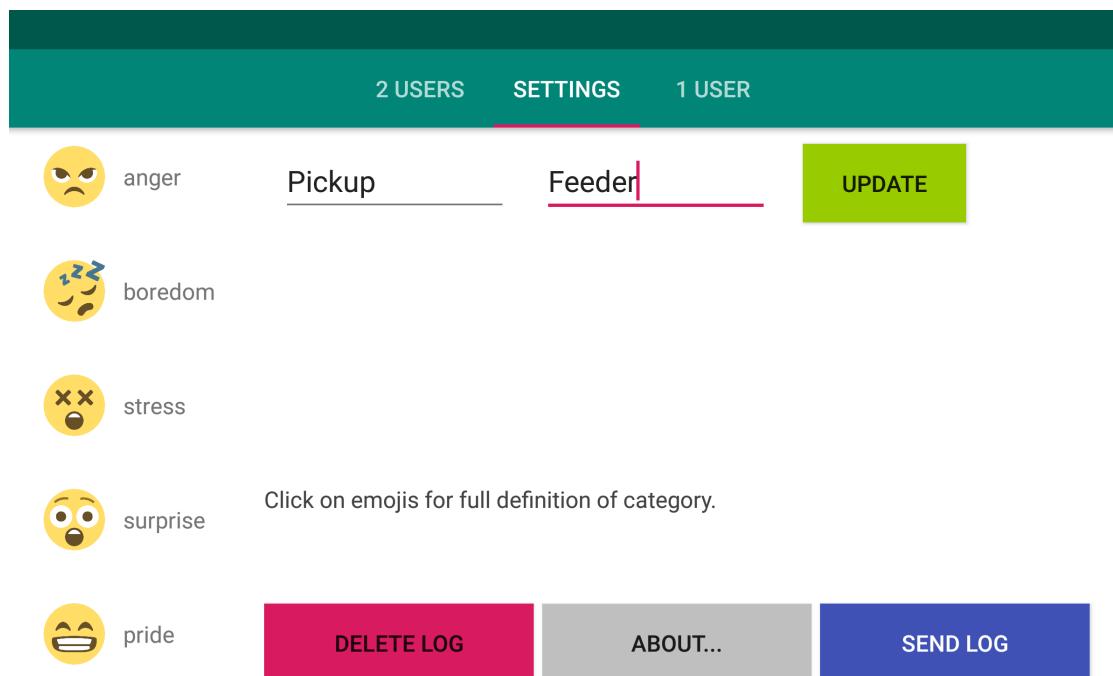


Figure 5.3: The screenshot displays the settings tab of the Proximo App where users can be renamed, logfiles sent or deleted. In the future, emotion categories will be exchangeable. In the current version, tapping on an emoji category on the left triggers a pop-up with a description of the emotion category instead.

5.2 Method

In order to test whether Proximo and the above described implementation of the Proximo App add value to formative evaluations in air traffic control, we conducted a case study. In

short, we joined a scheduled evaluation study of a novel support tool for air traffic control and offered Proxemo as an additional evaluation method. During the study, we examined Proxemo's utility, appropriateness and UX for controllers and observers as in-situ documentation tool during simulated ATC shifts. Additionally, we explored the same parameters for controllers, observers, ATC researchers and developers in the subsequent debriefings using Proxemo annotated videos. Finally, we gathered descriptive data providing first impressions of downstream utility. The researcher triangulated data from different sources and clustered notes in an affinity diagram before retrieving insights. The study was approved by the air traffic controllers' work council.

5.2.1 Context of Simulation

We conducted this case study during a simulation event taking place within a larger research project. The objective of the simulation was to formatively evaluate a prototypical spacing assistant for air traffic controllers in the final approach which is described in more detail in Haugg and Konopka (2022). The prototype under evaluation followed a novel interaction concept that deviated from the controllers' operative workstations. Briefly summarised, the novel system allowed the documentation of clearances directly on the radar screen via mouse interaction whereas the current operative system uses a separate screen with stylus interaction for the documentation of clearances. Testing Proxemo during a novel tool's formative test is appropriate as it authentically represents the context in which a formative evaluation method would be deployed in the future.

5.2.2 Participants

Several groups of stakeholders for the prototype under evaluation participated in this case study and hence had contact with Proxemo at different levels of the pipeline. Two supervisors and four air traffic controllers from the approach control of two major German airports had volunteered to participate in simulation runs with the prototypical spacing assistant. They were organised as two operative teams each consisting of two experienced approach controllers and one supervisor. We had given them advance information on the Proxemo evaluation method and asked for their consent to include Proxemo in the upcoming simulation. Their willingness to use Proxemo was independent of their participation in the simulation study. All participants were curious to test the new method and signed an informed consent document.

In addition to the approach controllers, a team of ATC researchers and developers was present during the simulation study. They operated the simulator-backend, observed the interactions and held the debriefings. They are neither primary users of Proxemo nor were they observed but were present at all times and interacted with the participants. Most appropriately, they can be denominated as tertiary users of Proxemo since they took note of the controllers' statements induced by Proxemo annotations in the videos during the debriefing. We did not ask participants for their demographic data, because more information than their working position would have

removed anonymity due to the small teams in an already small population.

5.2.3 Researcher

The researcher had conceptualised Proxemo as a method as well as the Proxemo App and gathered experience with Proxemo over three years in field evaluations and lab studies. He disclosed his role as owner of Proxemo to all participants but highlighted its incompleteness, the need for further development and hence the relevance of honest participant feedback. The researcher had met four of the six participants in prior prototype tests. Due to the advanced error-culture in air traffic control and judging by the broad spectrum of feedback received, we assume that neither the researcher's relation to Proxemo nor his acquaintance with some participants had relevant impact on the outcome.

The researcher's role and level of participation varied between simulation runs and debriefings. After handing over the smartphone with the pre-configured Proxemo App to the supervisor, the researcher merged into the research and development team and conducted overt observations during the simulation runs. During the debriefing, the researcher took the role of a participant observer as he moderated the debriefing and controlled the Proxemo annotated video recordings.

5.2.4 Data Collection and Analysis

Simulation

During the simulation, the pickup controller and the feeder controller sat at adjacent workstations. The prototype under evaluation ran on the radar display and required mouse interaction. The supervisor took the role of the observer and sat behind the controllers in order to gain a good view on the workstations, the controllers' body posture, the facial expression from a steep angle and most importantly hear their utterances (figure 5.4). In order to get a better view on the facial expression of the feeder, one supervisor tried to observe the scenario from the workstation next to the feeder during one run, that means watching the feeder from the side rather than looking over their shoulder. However, this position did not allow sufficient oversight of the pickup and their workstation. Therefore, the supervisor went back to the original observation position.

The supervisor used the Proxemo App on a Oneplus 5T (OnePlus Technology Co., Ltd., Shenzhen, China) with a 6" screen running OxygenOS [Android 7.1.1]. Resulting logfiles from the Proxemo App were saved as comma-separated values on the smartphone. To capture video recordings of the simulation we used a Samsung S20+ (Samsung Electronics Co., Ltd., Seoul, Korea) mounted on a tripod. Video files were written in 4K UHD (3840 x 2160 pixels) with a 30 Hz sampling rate. The camera was oriented in such a way it could capture the feeder's face as well as the radar screen of their workstation. Its resolution was sufficient to allow for the

identification of callsigns and the built-in microphone captured voices from both workstations. The pickup's radar screen was visible in the background but callsigns on it were illegible. The smartphone's camera sensor would have been able to capture in 8K but we did not have access to a monitor supporting that resolution. Additionally, the file transfer time would have quadrupled unnecessarily extending the break between simulation and debriefing. In the second run we experimented with recordings in Full HD (1920 x 1080 pixels) resolution only to reduce file transfer time and because the callsigns had not been evoked during the first debriefing. However, during the second debriefing, attendees complained about the now illegible callsigns which is why we went back to 4K videos for the remaining runs.



Figure 5.4: The sketch visualises the setup of the Proxemo deployment in the simulation. The approach controllers are sitting at their workstations on the positions pickup (left) and feeder (right). They use a mouse to document clearances and interact with traffic displayed on the large radar screen and communicate with each other directly via voice (blue highlights). The supervisor is sitting behind them and documents observed emotions in the Proxemo App whose logfile will later be synchronised with the video recording of the feeders' workstation (green highlights).

Debriefing

Between the simulation and the Proxemo guided debriefing, the researcher transferred the Proxemo logfiles via Bluetooth and the video recordings via USB-C to a computer. The Proxemo logfiles required preprocessing in Excel (Version 2016, Microsoft Corporation, Redmond, Washington, USA) before being synchronised with the video file in ELAN The Language Archive (Version 6.0, Max Planck Institute for Psycholinguistics, Nijmegen, NL). During that time a conventional, open debriefing took place which was characterised by free recollection of critical situations¹ and verbal description of recalled parameters.

The debriefing was held in a conference-like setting with 11 to 16 attendees (always two supervisors, four controllers and one researcher plus a varying constellation of the research and development team) facing a large 4k monitor. To maximise the ratio of the video on the screen, we downscaled the space ELAN controls required by reducing the font size in the system settings. This led to tiny, illegible buttons and necessitated controlling the ELAN software with shortcuts. Similarly, since emotion categories were not legible for all participants, the researcher announced them before replaying a situation from video. Audio was presented on external sound boxes. See figure 5.5 for a schematic drawing of the setup and appendix A.3 for an exemplary screenshot of the ELAN software.

5.2.5 Procedure

Data was collected during six simulation runs scattered over three consecutive days. To be exact, both teams conducted a run during each session, making it theoretically twelve individual runs. However, there was only time for one extended debriefing so in each run we set up camera equipment behind the feeder of one of the teams and supervisors took turns in using Proxemo.

On the first day, a general briefing took place where all attendees were introduced to the prototypical spacing assistant and the purpose and features of Proxemo as a formative evaluation method. We explained the set of emotions, their descriptions together with instantiations and what data they were derived from. Supervisors were instructed to document observed emotions and if in doubt rather set one timestamp too many than too few since timestamps can easily be skipped in the video but searching for a situation without markers is time-consuming. Because the spacing assistant was designed to support the feeder, the focus of the evaluation also lay on the feeder position.

During the following six simulation runs, each team was assigned workstations in the simulation room where they used the prototypical interface to control simulated peak traffic approaching their familiar airport while considering the varying weather conditions and departures. Approach

¹In this work, we use the term *critical situation* for instances that triggered emotions (regardless of valence) and, hence, were critical for the overall user experience. We refrain from the term critical “incident” (Flanagan, 1954) which is more common in HCI but rings alarm bells regarding the affected “safety of operation” in aviation (International Civil Aviation Organization, 2020, p10).



Figure 5.5: The sketch visualises the setup of the debriefing. All attendees had a clear view of the monitor where the Proxemo annotated video was replayed. There was space in front of the monitor allowing attendees to walk up to the screen and discuss details of specific scenes. The researcher controlling the video was sitting next to the monitor. Green highlights indicate how the researcher could control the ELAN software either via shortcuts to jump between timestamps in the video (vertical list on the right) or via mouse to fine-tune the starting point in the timeline (horizontal, below the video). Blue highlights indicate 1) the feeder’s radar display and 2) audible utterances from the video in combination with 3) memory based explanations by the approach controllers or supervisors as the three main sources of information in the debriefing.

controllers took turns on the pickup and feeder position between runs. Each of the six runs consisted of approximately 50 approaching and 15 departing aircraft and lasted 43 to 64 minutes ($Mdn = 53min$).

Just before a new run began, the researcher started the video recording, synchronised it with the Proxemo App and handed the smartphone to the supervisor. The supervisors positioned themselves behind (or in one run next to) the approach controllers and documented the controllers’ emotions during the simulation run. As soon as the run was finished, the researcher stopped the video recording and reclaimed the smartphone from the supervisor. Logfiles from the Proxemo App were transferred to a computer, preprocessed and then synchronised with the video recordings in ELAN. During this procedure which took approximately 10-15 minutes ATC researchers and developers initiated an *open debriefing* where approach controllers reflected about the latest run and reported the issues they remembered.

In the debriefing with Proxemo annotated videos, the researcher told all attendees the number of timestamps set and asked the supervisor whether any among them should be omitted or

were particularly relevant. He then started to play the video recording on a large monitor, starting from the first relevant timestamp and pausing the video as soon as approach controllers or supervisors indicated memories of the situation and started to explain what had happened there. All timestamps had been synchronised to -5 or -10 seconds in the timeline before their actual occurrence in the video file, so that based on the replayed video insight into how the situation evolved leading to the emotion could be gained more easily. When participants or other attendees indicated that the situation did not require further discussion, we proceeded to the next instance of documented emotions. The researcher took observation notes during the debriefing and subsequently conducted short individual semistructured interviews with participants to complement the notes. After the last simulation run, a closing meeting took place where findings were summarised and requests for prototype changes prioritised. All participants jointly prioritised the requirements ATC researchers and developers had already derived during the debriefings. We took that chance to ask ATC researchers and developers whether they were under the impression that Proxemo contributed relevant insights to promote the conceptualisation and development of the prototype. Subsequently, all video recordings and raw Proxemo logfiles were deleted.

5.2.6 Hygienic Measures

The ongoing corona pandemic necessitated special procedures during the data collection. The total number of people in the room was limited and medical masks or FFP2 masks had to be worn at all times with one exception: when sitting at their workstation facing the screens and separated by shields of acrylic glass, approach controllers were allowed to take off their masks. Since we observed no instances where audibility or communication were impacted, we will not further discuss these measures. Additionally, devices such as the smartphone running the Proxemo App were disinfected prior to each change of user.

5.2.7 Data Collection

During the runs, we made few observations and had Proxemo logfiles and video recordings generated. Most observational data and statements were collected during the debriefings lasting 15 to 20 minutes as well as during the subsequent short structured interviews with individual participants. Video recordings were deleted after the debriefings and all remaining data are in written form. Questions and follow-up questions to the participants are presented in appendix A.2.

The importance of questions shifted over the course of the six runs. After the fourth run, each approach controller had been in the feeder position twice and had once been observed with Proxemo while being there. The two teams in the last runs had used Proxemo before which resulted in the emergence of a saturation effect after the fourth run regarding the participants' feedback to Proxemo. As a countermovement to the saturation in controllers' and supervisors' feedback, the development team needed some time to get comfortable with the new method.

As a result, their involvement in the debriefings and active use of the Proxemo annotated video material began to increase after the third run.

In the closing meeting the researcher thanked all attendees for participating in testing the novel method Proxemo and its emoji based, playful interface untypical for the serious domain of ATC. He emphasised the helpfulness of the collected data and critically posed whether the 10-20 minutes extra effort Proxemo added to each debriefing had been worthwhile. With this check, participants validated the most important prior observations and statements. The researcher also wanted to know if the participants had stayed in the extended debriefings because they felt obliged to do so or whether they saw a real advantage in the annotated videos that justifies the extra time for future evaluations. Additionally, ATC researchers and developers wrote up a protocol of insights and prioritised requirements which was made available to all attendees afterwards.

5.2.8 Data Preparation and Analysis

Observational notes and statements in interviews were recorded with pen and paper in-situ and digitalised within 24 hours. In order to structure the qualitative data and capture insights regarding our questions we transferred statements and observations onto digital sticky notes on Miro (RealtimeBoard Inc., San Francisco, California, USA) and clustered them into a small affinity diagram (Holtzblatt & Beyer, 2016; Kawakita, 1991) resulting in seven categories on the highest level and 32 groups on the lowest level. The written protocols from the closing meeting were already filtered by requirements addressing interaction elements or system behaviour and clustered top-down by priority. For the interpretative step of extracting insights we triangulated between notes of statements and observation during the simulation, written protocols from the debriefing as well as validations from the semistructured interviews.

5.3 Results

5.3.1 Descriptive Statistics

We deployed Proxemo during the six simulation runs incorporating a new spacing assistant for more precise separation in final approach. Supervisors' focus lay more on the feeder position. Across all sessions they descriptively documented more emotions for the feeder ($Mdn = 17.5$, $range = 5 - 23$) than the pickup ($Mdn = 1.5$, $range = 0 - 14$). Supervisors observed 117 emotional situations across all runs and documented them in the categories surprise ($n = 64$), stress ($n = 23$), boredom ($n = 16$), pride ($n = 9$) and anger ($n = 3$). Due to the focus of the formative evaluation we decided not to review documented instances of boredom in the video during the debriefing. In two debriefings we refrained from redundantly revisiting the few timestamps set for the pickup because the supervisor insisted that they marked situations that had been emotional

for both controllers and hence had been already discussed when reviewing the video on the basis of timestamps set for the feeder. During the debriefing none of the approach controllers noted any emotional situations not represented in the Proxemo timestamps. However, as the main focus of this study lay on qualitative data, we did not systematically ask for and quantitatively determine “missed emotions”. Therefore, a computation of Proxemo’s sensitivity or specificity is not possible in this study.

5.3.2 Qualitative Insights

In this subsection we report observations and statements from the debriefing and the interviews. The insights presented in this section follow the structure of the affinity diagram.

Controllers do not feel disturbed by video recordings and observations

Approach controllers unanimously stated that the camera and observation did not disturb them — “*not at all*” (Controller 2,3) — and they ignored the context of the study. For instance, they claimed they “*had intermittently forgotten that [the recording] was running*” (C4) or explained how “*the observer is the first thing you forget as soon as something is happening [on the radar]*” (C1). This statement is validated by our observation of one instance where the simulator crashed, the controller turned around, rediscovered the running camera and cheered and grimaced for the camera.

The currently implemented set of emotions is suitable and sufficient

As a first reaction to the question on the appropriateness of implemented emotion categories, all controllers agreed that they did not miss any further emotion categories. Yet, on second thought, they came up with ideas on how to extend the set of emotions for other contexts than the formative evaluation of a novel artefact. C4 noted that in scope of the formative evaluation of a novel interface, such as the study at hand, “*surprise is most frequent*”. If the observations were conducted in the operation room and “*live, surprise would be rarer and at the same time associated with stress*” (C4). A reason for this are the varying triggers for surprise. In the formative evaluation of a novel interface, surprise was elicited by unexpected system behaviour. During a live session in the operation room, all controllers should be sufficiently trained with the interface and part of their control task is to expect variations in traffic behaviour. Thus, the frequency of surprises is deliberately reduced and unexpected events demand for immediate action causing stress. With trainees, C2 expected to see more instances of “*confusion — negative, but with varying degrees*” and suggested that “*surprise should remain but occur combined with confusion in an overarching category*”. Furthermore, “*sudden enlightenment — when trainees get it — could form a category of its own [as well as] frustration*” (C4). Supervisor 1 told that judging by their experience despair could form a category of its own. The suggestion of

sudden enlightenment is a novel idea, *frustration* and *despair* are already considered part of the general negative category *anger*, and *confusion* is an element of *surprise*. The categories had been instructed accordingly and, therefore, we interpreted the statements as support for our predefined set of emotions.

Since the consideration of UX is new to the domain of air traffic control (chapter 3), talking about emotions on the job and especially during a formative evaluation of prototypes was still new to controllers. Even though we were jumping from one emotion tag to the next during the debriefings, controllers mainly talked about how the situation evolved and rarely mentioned the emotions. C2 exclaimed, “*that’s something new, I entirely suppressed the pride!*” and explained “*controllers never learned to feel pride, we receive little positive feedback*”.

In the closing meeting, S2 stated that “*one generic timestamp would be sufficient — emotions do not need to be differentiated*”. Using just one generic *something-interesting-happened*-button is a worthwhile thought as it would potentially make the interaction with the Proxemo App even more efficient. However, if the researcher moderating the debriefing had missed announcing the emotional category of the currently reviewed situation, attendees always asked for the documented category to support recall and better make sense of the situation in the video snippet. Interestingly, in the fifth run, S2 even remembered a situation through their memory of the associated emotion they had documented: “*and once I logged stress for pickup — this was because of [...]*”. In conclusion, we observed several instances where the variety of emotion categories in the method was beneficial for the evaluation process and we, therefore, intend to uphold the distinction of observed emotions.

The framing of emotion categories needs to match controllers’ perception of their job

Controllers did not feel quite comfortable with the term *pride* and suggested “*joy*”(C2) or “*content, affirmation*”(C2) as better alternatives. However, controllers described an exemplary source of perceived joy as having “*sovereign control despite high amount of complex traffic*”. This is in line with observed behaviour of C2, who — while having a good run in a high traffic scenario — clapped their hands and exclaimed “*send me more aircraft!*”. Even though controllers avoid the term *pride* and prefer to talk of *joy* when recalling their experiences, the source of joy lies in their recent achievements. Thus, the joy resulting from self-efficacy can be interpreted as pride.

Similarly, C3 suggested relabelling *surprise* as “*the discovery of something unexpected*” and S2 explained their desire to change the category *stress* to “*very busy*” because bad presets resulted in “*an unnecessary amount of clicking [...] but no emotional stress*”. The other supervisor stated that “*there is no such thing as boredom — when controllers are about to get bored they start to chat with each other*”. In sum, participants never declared single categories as obsolete but requested more precise definitions. In future research, we recommend keeping the short labels of categories and complement the already existing instructions with concrete examples

that illustrate how the categories can occur in the context of the pending evaluation.

The Proxemo App fosters supervisors' concentration and shifts their focus

Supervisors took turns in using the Proxemo App over the course of the simulation study because only one combination of camera and smartphone with the Proxemo App was available. Both supervisors volunteered to begin and after the first eventful run, the supervisor who did not use Proxemo during this run noted *"we would definitely have had something to click: [so many] emotions!"*. Both supervisors stated that the Proxemo App did not distract them but the method changed their view on the simulation giving them a novel view on controllers' emotional expressions. S2 considered Proxemo to have *"even contributed to my concentration because it gave me a task"*. After the last run S2 stated *"I attentively waited for something to log again"*. S1 saw a focus shift that came with the method as well: *"when I watch without Proxemo I am more involved in the traffic because I do not need to pay attention to the mimic. Here I noted that the controller went with their hand through their face, that means stress."* Paying close attention appears to be necessary, since *"controllers hide their emotions. That makes it difficult, even though I have been involved in situations like these myself"* (S1). When controllers had the time to reflect on their emotions, they sometimes supported the supervisor in documentation, for example thinking aloud about what confused them (thus indicating surprise) or in one occasion even jokingly exclaiming *"surprise!"*(C2) as the system behaved in an unexpected fashion. There was no instance during the debriefing where controllers objected a documented emotion or situation as non-relevant or misinterpreted, indicating the high capability of supervisors to understand their teams behaviour and emotions.

Timestamps and the video recording help to review the simulation experience

Proxemo timestamps and the respective video sequence served as memory triggers. Even though, C4 claimed that remembering the situations *"is easy directly in the aftermath"*, in the first run, a situation that was tagged with an emotion in the Proxemo App had not been mentioned in the open debriefing.

We observed that during the joint video review the team made up of pickup, feeder and supervisor always remembered the situation causing the emotion. However, in contrast to the initial open debriefing, the explanations became richer in detail when reviewing the recording. Beyond the study at hand, controllers saw potential *"for trainees with long debriefings and plenty of action where it is important that the trainees have exactly the same situation in mind and before their eyes"* (C4).

Sound and screen capture are most important for reflecting the experience

The most important features of the video recording were sound and the section of the video capturing the radar screen. While C3 “[does] not have to be in the frame of the video”, C1 found it helpful “to see and especially hear myself.” In the closing meeting they highlighted that sound and radar are important whereas the view of ones back of the head is not. C1 found “*sound and screen capture should suffice and sound is more important than picture*”. However, the complete picture is crucial for the purpose of sharing the experience.

Video recordings form a communication bridge between controllers and developers

Developers especially tend to ask questions about details that are far easier explained on the screen than verbally from memory, which poses the challenge for controllers to convey their mental model of the situation. The video review of selected scenes helps with this: “*My memory is fresh in the debriefing, yet the video helps to explain the situations to others — non air traffic controllers — because everyone has their own vocabulary and mind-set and explanations often lead to frowning. But the video establishes a common clearly understandable platform*” (C3). This is required since “*despite close collaboration, there are two different perspectives and linguistic worlds*” (C3) between controllers and developers. ATC researchers and developers labelled the method as a “*top translator between controllers and developers*” because the video recording assisted them in comprehending the controller’s depictions of experienced situations even when the callsigns were not recognisable.

In several occasions, the feeder pictured in the video recording and a developer walked up to the 4K monitor, taking a closer look at the recorded radar picture and discussed how exactly the tagged situation evolved, what the expectations had been and what should have happened instead. On the basis of video recordings, issues were comprehensible for developers that would have been costly to reproduce due to complex dependent parameters such as wind and traffic constellations. For instance, the video-mediated discussion revealed how a simulation artefact affected the system behaviour (second run) or how system behaviour worked as designed but needs to be briefed differently in the future (fourth run). Developers summarised the method as “*cool thing*” because it “*helps me to better visualise [the issue]*”.

Video recordings offer evidence in-situ and as take-away

Jokingly, the controllers compared the Proxemo debriefing with larger sport events where video replay supports the judgement of the referees. In fact, this came true after the last run when most attendees had already experienced five debriefings with Proxemo. The supervisor ended the open debriefing by declaring that he had “*one more aspect to discuss, but it’s better to directly have a look at this in the video*”. Developers accepted jointly discussed situations together with the video as sufficient evidence to base decisions upon. After one situation in the sixth run where

the controller noted a malfunction of the spacing assistant, the developers shortly discussed the configuration with the controllers and consequentially adjusted the threshold of parameters in the spacing assistant. From the third debriefing on, one of the developers repeatedly asked for screenshots of specific situations on the radar screen in order to get a starting point for checking how that situation looked in their logfiles. Screenshots were particularly popular in situations where the algorithms performed as designed whereas the air traffic controllers considered the system behaviour as inappropriate.

Proxemo changes the workflow and saves developers' time

In the closing meeting, controllers summarised how timestamps enormously helped by directly accessing specific situations. They emphasised that pausing situations in live-simulation is not possible. Therefore, Proxemo would be particularly interesting for debriefings in apprenticeship and further qualification. ATC researchers and developers noted how an exact reproduction of the situation after simulations without Proxemo often is not possible or very complex. The extra-effort Proxemo causes during the debriefing is therefore well invested because it reduces developers' efforts in the aftermath and presumably saves time. Research and development expressed interest in applying the method in future simulation runs.

5.3.3 Downstream Utility

We analysed the written closing meeting protocols pertaining to the resulting requirements. Of the 13 requirements specifically addressing interaction elements or system behaviour, attendees classified one requirement as *priority 1*, ten requirements as *priority 3*, two requirements without priority and one as optional. Five requirements (two *priority 2*, three without priority) were specified that addressed issues solely occurring in the simulation environment but still being relevant for an authentic experience of air traffic control such as appropriate and timely speed reduction of simulated aircraft. Finally, three requirements were specified about how the system's behaviour shall be instructed and trained.

While we can only report a descriptive list for comparison with future studies, we would like to emphasise that ATC researchers generated this protocol right after the last simulation run. Since some issues occurred redundantly over the course of the six simulation runs we can not clearly distinguish which of the issues were detected during the open debriefing and which were only discussed due to Proxemo. However, Proxemo's contribution was to increase the developers' understanding of the underlying problem for unexpected system behaviour during the debriefings already and to give them a better idea of which requests were realistic to be addressed within the next iteration. By the time the protocol was sent around, developers had already advanced the implementation of the prototypical system to meet two requirements including the requirement with *priority 1*.

5.4 Discussion

In this chapter we adapted the design of the Proxemo App to the context of air traffic control. We then conducted a case study to examine the feasibility of the Proxemo method in a high fidelity simulation. The purpose of the simulation was to formatively evaluate a prototypical spacing assistance for approach controllers in the feeder position during peak traffic. We synchronised logfiles resulting from the documentations in the Proxemo App with high resolving video recordings and subsequently reviewed emotion-tagged scenes with controllers, supervisors, researchers and developers during the debriefing.

5.4.1 Contextual Fit of Proxemo for Air Traffic Control

None of the supervisors using the Proxemo App and controllers being observed and recorded on video considered the method a disturbance. According to supervisors, the Proxemo App even fostered their concentration. Supervisors reported how the Proxemo App shifted their focus from traffic events towards paying attention to the controllers' emotions which was not easy because controllers are used to hide their emotions. Considering that the replayed video sequence together with the timestamp always resulted in controllers or the supervisor recalling what this situation had been about supervisors did a great job in detecting and documenting relevant emotions. In few instances, controllers verbalised their experienced emotion to make the supervisor aware of the relevance of documenting the current situation. This reminds of Sanderson et al.'s (2007) study where air traffic controllers had to regularly announce aloud their workload for researchers to document it. Whereas an occasional self-disclosure about currently experienced emotions may contribute to higher data quality, we do not recommend expecting, instructing or relying on that behaviour since it adds a prospective-memory task to the air traffic controllers job (as argued in chapter 1). For highly complex tasks van den Haak et al. (2003) suggest retrospective think aloud protocols based on videotaped interaction — similar to the approach used in our debriefing — which in their study lead to similar results as concurrent think aloud without the negative effect on task performance.

The information channels that supported participants best in recalling situations were the recorded sound and the segment of the video recording showing the radar screen. Even though one supervisor suggested collapsing all emotion categories into one generic button, knowing the documented emotion associated with the currently replayed video sequence supported the memory of controllers and supervisors. They requested the associated emotion category for each situation if the researcher had not announced it already. Controllers were interested in the emotion categories but never objected to the documentation. We interpret that as agreement with the supervisor's interpretation since we explicitly had invited to veto irrelevant or incomprehensible timestamps. Skipping snippets would have saved time and was welcomed by all attendees when controllers and supervisors stated the situation was a duplicate.

Only few controllers considered seeing their own silhouette in video recordings beneficial. Therefore, we can adjust the focus on the radar screen in future work and leave out filming the controllers. For this purpose, a sole screen capture taken directly from the simulation system may be even more advantageous due to increased legibility of details on the screen such as aircraft labels.

5.4.2 Suitability of the Predefined Emotional Set

Surprise. Observations of *surprise* alone made up more than half of the documented emotions. The main reason for this imbalance is that the simulation held lots of potential for surprises, confusion and unexpected behaviour. First, the novel assistance system under evaluation redefined their task in suggesting separation between aircraft. Second, the system ran on a simulation workstation that followed a more advanced interaction concept (mouse only) than the workstations currently deployed in operative control centres (mouse, touch and pen interaction). Third, there were few simulation based artefacts such as unrealistic speed reduction of aircraft on their final approach.

Boredom. The focus of the formative prototype evaluation — assistance in stressful situations involving complex traffic — allowed omitting documented *boredom* during debriefings. Additionally, the simulated traffic was based on busy hours before the onset of the corona pandemic resulting in fewer causes for boredom than stress situations. However, in spite of the simulation’s focus and scenario design to cause more stress and challenge controllers, in over 6 hours of simulated traffic 16 instances of boredom were recorded. This means, boredom is not an overall negligible experience for air traffic control. On the contrary, the effects of boredom caused by increased automation in air traffic control or other safety-critical surveillance tasks have been discussed and studied for decades [e.g. Thackray (1980) and Westgate and Steidle (2020)]. Boredom gains importance when controllers need to uphold vigilance over longer periods of low traffic, such as night shifts, where boredom is likely predominant. Recently, interaction researchers are suggesting more involving, playful interfaces to tackle boredom in air traffic control [e.g. Badea (2021) and Gramlich et al. (2022)]. With respect to emotion theory, boredom is mentioned in most models and emotion word lists across theories (Petta et al., 2011; Plutchik, 2001; Remington et al., 2000; Russell, 1980; Scherer, 2005; Watt-Smith, 2015). Westgate and Steidle (2020, p. 4) argue for boredom as an emotion because, “like other emotions, [boredom] is reliably elicited by specific situational appraisals.”

Stress. Compared to boredom, *stress* is more questionable and listed across models either as emotion (Petta et al., 2011), used as synonym for the affective dimension *tension* (Scherer, 2005) or used in a statement describing the circumplex segment of *activated displeasure* (Yik et al., 2011). Watt-Smith (2015, p. 294) lists the more specific term *technostress* as a title for the stress

and anger elicited by ill-designed technology that hinders rather than supports humans' tasks: "They are supposed to be making our lives easier, these wilful electronic slaves of ours. But mostly it feels as if they're in charge, forcing us to negotiate with them, cooperate, read their manuals. . .". Whereas the subcategory of technostress perfectly suits formative evaluations, we construe the category stress as broader, including stressful reactions that were not elicited by technology (see the description in table 5.1) but require technology to respond adequately.

Anger. In air traffic control, *anger* is an already scarce emotion that is most frequently triggered by behaviour of other actors. To reduce the risk of infection, the number of participating actors were artificially reduced for this simulation, thus further decreasing the possibility of anger prone situations to occur. In future evaluation studies, simulation pilots who control the aircraft parameters and participate in radio communication will be invited again. This may increase socially induced emotions such as anger, stress and joy in successful or entertaining teamwork.

Pride. Finally, *pride* is an emotion that controllers hesitate to show. This is a pity due to its importance for user researchers and designers who could build upon interactions that triggered self-efficacy and pride towards increasing opportunities for positive UX in a system (Huber et al., 2022).

Since supervisors observed the whole spectrum of emotion categories during the simulation and the high prevalence of surprise is accounted for by the nature of the simulation setup, we consider the currently implemented set of emotions as suitable. The variety of alternative titles for emotional nuances already covered by our pre-defined categories indicates that the existing set of emotions is sufficient as well. Supervisors and controllers urged to rename categories in order to sound less bold and invite the inclusion of more subtle nuances of emotion categories. However, crowding the interface of the Proxemo App with sub-categories such as "*not-yet-stress*" does not seem expedient when observees have the chance during the debriefing to explain the trigger, emotion nuance and intensity of an experience. In order to improve clarity in future deployments of Proxemo about the purpose of bold sounding umbrella terms representative of multifaceted emotion categories, we recommend investing more time during the introduction and to more thoroughly instruct evaluators on the broad set of notions counting into the seemingly simple categories.

5.4.3 Benefits for Developers

Proxemo changed the nature of the debriefing process from the recollection of critical situations and verbal description of recalled parameters to a structured and chronological debriefing with the audio-visual data in its centre. Thereby, Proxemo brought a number of changes into the debriefing process that attendees considered beneficial. Most importantly, discussing situations

based on Proxemo annotated video recording facilitated communication between ATC researchers, developers and air traffic controllers. While the audio channel helped controllers and supervisors recall the situation, the video recording of the radar screen mediated communication. Proxemo annotated video recordings promoted a common understanding among all attendees of the assistive system's state and operator's strategies in a specific situation during the debriefing. Without Proxemo, it is common practice for controllers and developers to compliment explanations with flip-chart drawings of traffic situations from memory. Additionally, the video recording served as evidential material for ambiguous situations. ATC developers and researchers could discuss with air traffic controllers based on the imagery material how the system was configured at the moment and how it should be better implemented for future responses. Screenshots² of such critical situations were popular among developers as they could be fed into issue tracking systems and either proved system errors or served as a reminder to review the implementation in regard to a specific situation. The timestamps and aircraft constellation visible on screenshots served developers as guidance to find the related situation in their logfiles and efficiently reproduce the scenario. This changed the developers' workflow and saved time used to spend reading logfiles or reproducing scenarios according to controllers descriptions. How much time was actually saved on the developers' side and whether this will result in more cost-effective iterations — considering the cumulated extra time of air traffic controllers during debriefings — is an economic question to be addressed in future work.

Law (2006) argues in her evaluation of downstream utility how important it is to consider the *developer effect*. The extent of this developer effect is presented in a quantitative study where Law shows how severity and frequency of issues but also the length of the issue description and the collaborative relation between stakeholders have an impact on the effectivity of fixes submitted by the development team. We strongly agree with Law's demand for tracking discovered issues and their fixes through multiple iterations. While this case study accompanied the first deployment of Proxemo in ATC and focused on feasibility, future work should pursue a quantitative approach on Proxemo's downstream utility. The collaborative nature of debriefings in our study reported in this chapter where users directly conveyed the validity and relevance of emotions and associated triggers and the annotated video recording together with the combined knowledge of all participants contributed to a rich shared knowledge base. From this shared knowledge base, all stakeholders involved could form consensual decisions about a prioritisation of issues to be addressed in the upcoming iteration. These observations are in line with case studies reported by Borneo and Stage (2017) who report a reduction of the developer effect through actively involving developers in usability testing and redesign workshops. Due to this promising constellation, we see high potential in Proxemo to alleviate the developer effect as long as developers are involved in the debriefing or at least granted access to the annotated video

²Screenshots of the annotated video were captured and saved only with participants' consent, as they survived the deletion of video material.

material. We want to encourage the tracking of downstream utility in future applications of Proxemo in projects with multiple iterations.

5.4.4 Limitations

Through the open debriefing and Proxemo in this formative evaluation, we learned *what* aspects and *how* the prototypical interface of the workstation should be adapted to meet air traffic controllers' expectations. However, we gained only few insights about *why* a certain scenario triggered an emotion on the psychological needs level. While attendees were interested in the emotion category linked to reviewed video snippets, we did not spend time discussing the controllers' emotions or underlying needs. During the debriefings documented emotions served as link to critical situations. Air traffic controllers were quick in suggesting design solutions that would, for instance, improve their trust in the system. In contrast to the valence method (Burmester et al., 2010) where the *laddering technique* guides user researchers question by question from the users' experience to the need, we omitted that drill down. Since the controllers never actively vetoed a documented emotion, we consider Proxemo timestamps to be set at appropriate points marking relevant experiences during the interaction. Hence, the annotated video recordings could serve as useful material to conduct retrospective interviews with UX laddering in future work.

A limitation of internal validity is that we did ask controllers (and supervisors) only about the set of emotions but not whether the controllers were content with the supervisors' judgement of their emotions. Since supervisors stated how using Proxemo entertained them during the observation it is possible that they tended to document more positive emotions. What speaks against supervisors being affected by such positive bias is that they also reported how it was hard to interpret controllers' emotions despite having experienced similar situations before. This means they did not just infer the emotion categories from their own impression of the respective situation but attempted to empathise with the controllers under observance. Additionally, controllers never objected a named category during the debriefing, and we did not observe notions of disagreement with emotion categories. However, the content or disapproval of proxy-ratings should be included as additional question in future work. Additionally, we strive to ascertain quality criteria including validity and reliability in studies that are reported in later chapters of this work.

The set of emotion categories used in this study and the layout of the Proxemo App were optimised for teams of two in air traffic control. In our study, teams consisted of the roles pickup and feeder in the approach position. The findings regarding the applicability of the Proxemo App could generalise to teams in other sectors in air traffic control where teams typically consist of an executive controller and a planning controller. Participating controllers and supervisors saw potential in Proxemo for its use in training. However, the emotion categories used in this

study were derived from qualitative data collected during routine shifts from experienced operatives. When training with air traffic control apprentices, the set of emotion categories may require adaptations. Since Proxemo is a method for formative evaluations typically taking place in simulation, we are not sure whether our findings need to be transferable to the control centre operation room. However, the high fidelity simulator is set up to be a realistic reproduction of workstations used in the near future. Therefore, Proxemo may be used to gather controllers' emotional experiences during regular shifts. The general concept of video recording and documenting observed emotions could generalise to other safety critical surveillance and control domains such as train control, power plant control or anaesthesia if the set of expected emotions is adapted accordingly.

We faced the trade-off between video resolution and transaction speed resulting from file size as technical limitation. Future hardware availability may allow transfer and synchronisation of multiple high resolving videos in such a short time that the debriefing is not delayed and still allows drawing on replays of the whole scene including both feeder's and pickup's screens.

To speed up the clearance from the work council we did not collect any quantitative demographic data such as work experience, age or gender which could facilitate the identification of individuals in our small sample. However, to get the right perspective on the experience of air traffic controllers who contributed the data, we would like to emphasise that it is in the supervisors' interest to bring seasoned and highly motivated team members to simulation runs on novel interfaces that may shape the workstation for decades.

Without disturbing the interaction experience of operatives, Proxemo provides a useful method for formative evaluations in the context of air traffic control. Controllers remembered most situations and even callsigns but learned fast to utilise the annotated videos to convey their experience to ATC researchers and developers. Developers found value in the stills extracted from highlighted video snippets for an efficient comprehension of the controllers' experience and possible changes in the associated algorithms. Therefore, bridging the different "languages" and perspectives formerly impeding communication during debriefings stood out as the greatest advantage of Proxemo. In sum, we answer this chapter's research question with yes, Proxemo is not only feasible but showed to be very helpful during formative evaluations of novel interfaces for air traffic control. Future studies in ATC or other safety critical domains should investigate systematically, whether the retrospective debriefing structured by timestamps constitutes the most important aspect of the Proxemo pipeline – for instance by deploying a thorough debriefing without the Proxemo categories as a baseline. A thorough evaluation of Proxemo regarding its quality criteria in controlled experiments remains to be done and will be addressed in the following chapters.

Chapter 6

Inter-Observer Reliability

Observation methods such as Proxemo are prone to subjective bias because absolute neutrality of the observer is not possible (Beveridge, 2002). When designing and deploying psychometric methods, measures can be implemented to improve objectivity. However, it is difficult to directly evaluate the emerging objectivity. On the other hand, the two quality criteria reliability and validity which are based on objectivity can be measured. The interrelation of the three quality criteria is that objectivity is the foundation for both, validity and reliability and “unreliability limits the chance of validity” (Krippendorff, 2004, p.212). How the deployment of Proxemo can be improved during observations to contribute to objectivity has been discussed in chapter 4. Therefore, we will now examine the reliability of Proxemo.¹

6.1 Selecting the Appropriate Quality Criteria

So far, we have introduced Proxemo as a novel UX-method for formative evaluations and have demonstrated its usefulness and feasibility in two specific scenarios. A thorough evaluation of the method is still pending but crucial before Proxemo can be recommended for formative evaluations in a broader variety of contexts. Only if novel methods are thoroughly evaluated and the resulting quality criteria are published, practitioners gain proper guidance when choosing appropriate methods for their projects. Evaluation methods are held up to a large variety of criteria in meta-evaluations with main emphases varying between authors. Authors from the fields of HCI, human factors and psychology agree *that* evaluation methods need to be evaluated but have different opinions on *which* quality criterion is the most important in such meta-evaluations (e.g. Hartson et al., 2001; Law, 2006; Salmon et al., 2020).

Classical test theory lists objectivity, reliability and validity as quality criteria which build upon each other as described above (Krippendorff, 2004). Other disciplines adopted those criteria

¹A brief summary of this study has been published in Huber, Bejan, Radzey and Hurtienne (2019).

and sometimes coined different names for them or even established novel criteria.

Researchers from human factors and ergonomics — a domain rooted in psychology and engineering — analyse, predict and influence human behaviour in safety critical or complex sociotechnical systems. They require their analysis and evaluation methods to be reliable and valid in order to promote safety but also to establish credibility during collaborations with other engineering disciplines (Salmon et al., 2020).

In usability research the occasional lack of objectivity has been discussed as *evaluator effect* (e.g. Hertzum et al., 2014). Reliability between researchers is defined as one part of *consistency* (Hartson et al., 2001), with the other part being *repeatability*, that is consistency of results across different methods. Usability evaluation methods mainly focus on the identification and prioritisation of issues in interfaces. They are required to reveal real (valid) problems only but as many of these as possible. Thus, Hartson et al. (2001) promote *effectiveness*, the product of validity and thoroughness, as the ultimate criterion. However, there is also the practitioners' perspective: even the most effective way to identify relevant instances does only improve the user's experience if identified issues are actually fixed and found to benefit the user in the next test. This is why the tracking of issues throughout multiple design iterations of the development process should be considered (John & Marks, 1997). This quality criterion is referred to as *downstream utility* (Hartson et al., 2001) and is heavily understudied due to the effort it entails. Downstream utility has been elaborated in detail by Law (2006) and shortly discussed in relation to Proxemo in chapter 5. As an important aspect for practitioners, the UX of the method and its implementation need to be taken into account (Hartson et al., 2001). As Stanton (2016) speculates, the popularity of methods may be related to the ease of acquiring the proficiency to deploy them. When designing Proxemo and the Proxemo App we took great care of the observers' experience and reported descriptive statistics (chapter 4) and qualitative data (chapter 5) on how observers perceived the method and the app during their studies.

There is one final quality criterion, often invoked and almost impossible to determine: *cost-effectiveness*. Hartson et al. (2001) point out how the cost-part of cost-effectiveness is challenging to precisely sum up across all variables involved in learning and applying a method. We believe that the effectiveness-part of cost-effectiveness is even harder to capture when thinking beyond the mere quantity of detected issues or emotions. This aspect is not even satisfyingly assessable when the downstream utility is tracked from the first kick-off meeting to the launch of a product. The more persons are involved in designing, developing and testing a product, the harder it becomes to estimate the exact cost or revenue of design decisions. Will the deployment of a method that involves multiple disciplines tie teams closer together resulting in fruitful long-term collaboration? What impact will user involvement in testing with a particular method or the public's knowledge about those tests have on the overall customer- and brand experience? These questions show how the impact of a method and the resulting money-trail cannot be followed up conclusively. Therefore, in this work we focus on criteria that are calculable. As an alternative

to the high goal of cost-effectiveness, more immediate measures of efficiency are realistically measurable such as observers' workload caused by Proxemo and their resulting spare cognitive resource to focus on the users' emotions. Further indicators for cost-effectiveness are Proxemo's high learnability and its contribution to more efficient navigation in video files and more efficient interpretation of video data [chapter 4].

The above listed criteria may not be comprehensive but include the most prevalent quality criteria formative evaluation methods are held up to in meta-evaluations. Examining and publishing all appropriate quality criteria of a novel method is particularly relevant for practitioners as they rarely have the resources to determine the quality criteria of each method they intend to use by themselves (Lindgaard, 2006). Hence, we aim to evaluate Proxemo as a method regarding its quality criteria. In former chapters we already addressed observers' experience when using Proxemo and discussed Proxemo's potential with respect to downstream utility. In this chapter and the two subsequent chapters, we will discuss or evaluate Proxemo — thus making it comparable to other methods of formative UX evaluation or observation — regarding the criteria objectivity, reliability, validity, thoroughness, effectiveness, efficiency with associated consequences and again the observers' user experience.

6.1.1 Objectivity

To assure the optimal degree of objectivity while conducting, evaluating and interpreting tests, the German Psychological Associations (Testkuratorium, 2018) highlight multiple criteria. Their guidelines were designed with a focus on self-report questionnaires and tests. However, some crucial aspects are applicable for structured observations as well. In particular, standardisation, exact and extensive instructions, descriptions of exemplary cases but also the prior knowledge are critical and hence relevant for Proxemo. In chapters 4 and 5 we reported pre-defined emotional categories reducing the degrees of freedom and thus standardising the observational outcome. We furthermore developed a detailed set of instructions for the observers using Proxemo and provided descriptions of the emotional categories' extent along with examples. As the Proxemo App is easy to learn and use, we did not define prerequisites regarding prior experience with technology. Neither did we provide clear thresholds for experience in the application domain but instead follow the principle “the more, the merrier”: Whereas generic empathic abilities facilitate the recognition of emotions in other humans, any further knowledge of the particular human and their abilities or an understanding of their context fosters the ability to recognise their emotions.

Despite all precautions, measurement errors such as subjective bias will always remain. In the following, we treat Proxemo as a psychometric instrument, select the most appropriate scale for its measurement error and conduct a study to determine the reliability of Proxemo.

6.2 Determining Measurement Error

Classical test theory acknowledges that scores measured with psychometric instruments do not directly represent *true scores* in subjects but are biased by measurement errors (e.g. Novick, 1966). Since only the resulting score of the psychometric instrument can be measured directly it is a challenge to determine the measured score's proportional composition from the true score and measurement errors.

To systematically determine or even reduce the impact of error, we first need to identify possible origins of error that influence reliability. Sources for errors in measurement which affect reliability, originate 1) from issues of internal consistency of the instrument, 2) instability of measurements over time or 3) measurement variances between observers (Hallgren, 2012). The first source for measurement errors in this list, internal consistency, is typically measured through correlating items within a scale. However, for Proxemo emotional categories are meant to be distinct from each other and used exclusively to describe a situation. Since the documentation of one single emotion suffices to tag a situation, there are no multiple items to be correlated within a scale and the requirement of internal consistency is not applicable for Proxemo.

The second source for measurement errors mentioned above, stability of measures over time, is typically determined through test-retest reliability when measuring the same subject over time (Hallgren, 2012). This approach is most feasible for psychometric instruments that measure traits which are by definition stable over time and, therefore, have a stable true score. The resulting variance between tests and retests can be interpreted as measurement error. However, user experience and resulting emotions are the consequence of multiple factors (e.g., Thüring & Mahlke, 2007). While deployed versions of software and hardware can be controlled over a series of tests, the context is variable. Especially the varying daily mood of users and the still not in-depth explored influence of the novelty effect (Rutten et al., 2021) and insights on the development of UX over time (Kujala et al., 2011) disallow for a retest with the expectation of identically re-experienced emotions. Mapped on the application domains referred to in this work, this leads to the following deductions: no pair of reminiscence sessions are the same and no air traffic control shift is identically replicable (especially if simulated on a novel prototype). To evaluate the reliability of Proxemo over time, one would not necessarily need to reconstruct the observable emotions live but could instead create a test-retest evaluation study based on video material, thus ensuring the stability of stimuli. However, rewatching videos likely results in noticing details that were missed the first time (Bentley & Murray, 2016). In fact, repeated replay of interesting video sequences is an essential part of multimodal sequential analysis (Luckmann, 2012). Therefore, our expectation is that a retest on the same stimulus material would result in observers gaining more detailed insights into the users' experience and hence generate more thorough Proxemo data. Replaying videos in order to complement Proxemo timestamps with annotations in a retrospective video analysis is optional but does not provide a meaningful

measure for the reliability of Proxemo.

Proxemo is a method to document emotions in a unique experience in-situ. Regarding reliability, we consider as most significant issue the extent of variation in documented data from different observers assessing the same situation. Therefore, we intend to measure the inter-rater reliability of Proxemo. For consistency across chapters of this work we will stick to the term “observers” and hence refer to the measure with the less commonly used term inter-observer reliability (IOR). Our expectation is that the agreement between observers is significantly above agreement by chance.

6.3 Method

The purpose of the IOR estimate is to assess the mean rating among multiple observers. Krippendorff (2004) recommends to measure reliability with clear instructions and equally capable observers who give truly independent ratings. In this section we describe the operationalisation of our study while adhering to these standards. From our two application domains we chose reminiscence sessions again. For ethical reasons and for standardising the procedure, we decided against deploying multiple observers in a live intervention. Instead, two observers reviewed video material of reminiscence interventions with single residents or groups of three residents and documented observed emotions with the Proxemo App.

6.3.1 Design

We used a fully crossed design in which both observers documented emotions in all subjects. Variables were the six emotional categories that could be documented in the Proxemo App and the number of residents participating in the video recorded reminiscence session as single participants or groups of three residents. We treated these as combined category on the nominal level with $emotional\ categories \times n_{single} = 6$ and $emotional\ categories \times n_{group} = 18$ possible emotional events to be theoretically set for any observed instance. We collected the perceived task load and effort as variables possibly confounding observers’ attention and thus their ability to judge the emotional situation.

6.3.2 Setup

The study was conducted in a controlled environment. In order to avoid issues of synchronisation, we presented the video material on a single large screen (165 × 93 cm) resolving FullHD (1920 × 1080 pixels). For clear sound we placed an external speaker on the table right in front of the observers. The two observers sat next to each other with equally good view of the screen, see figure 6.1. The Proxemo App ran on a Samsung Gear S2 and a Samsung Gear S3 (Samsung Electronics, Seoul, South Korea) which only differed marginally in size. Each observer chose to

wear the smartwatch on their wrist. To ensure independent ratings during the trials we put a visual barrier between the two observers that did not impact the view of the screen but prevented the observers from seeing each other interact with the Proxemo App. A list of emotion categories available in the Proxemo App was placed next to each observer at all times (table 4.1).

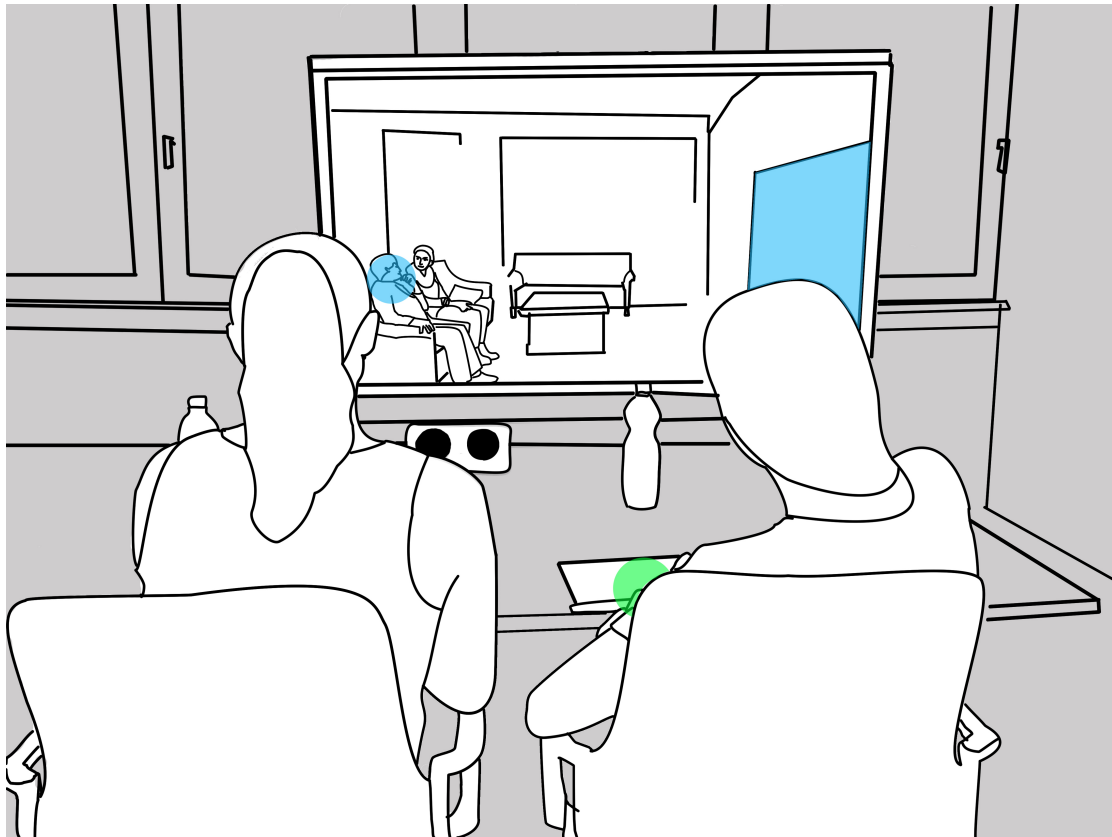


Figure 6.1: This schematic drawing shows the setup of the training session where both observers jointly documented emotions in a replayed video using Proxemo (green highlight). The subject material included a moderated reminiscence session with residents in front of an interactive wall. Here a session with a single resident (blue highlight) is depicted.

6.3.3 Participants

It is difficult to acquire experts with exactly the same amount of experience in dementia but more feasible to find persons with a comparable lack of experience. The two observers recruited for this study were students of Health Sciences or Human-Computer-Interaction, respectively, without prior knowledge of either Proxemo or reminiscing in dementia. Both participants were female, aged 21 and 26, and did not know each other prior to the study. Both participants consented to participating in our data collection. Including breaks, the study took 4 hours and

the participants were compensated with 8 €/hour.

6.3.4 Material

The video material showed reminiscence sessions in a residential care home setting in a rural area of Southern Germany. In each session, a moderator guided residents through interactive reminiscence content presented on a wall sized screen. Reminiscence topics included farm life, Black Forest sightseeing and an interactive tour through a virtual house with a cat avatar. For a detailed description of the scope of hardware, software and interactions see (Bejan et al., 2018). In the experimental run, observers assessed 80 minutes of video including two sessions with single residents and two sessions with groups of three residents. The video material used in the trainings was similar to the videos used in the experimental runs. We used the QUESI (Naumann & Hurtienne, 2010) and the raw version of the NASA TLX (NASA Task load index, Byers, 1989) as measures of perceived effort and task load caused by Proxemo to indicate potentially biasing effects.

6.3.5 Procedure

In the beginning, participants received a short briefing about the reminiscence project, an introduction to Proxemo and their role as observers. Both observers had the opportunity to familiarise themselves with the Proxemo App and the set of emotions used in the study. To make sure that observers shared a common understanding of the emotion categories, we conducted a two-staged training. First, observers documented emotions jointly and were allowed to pause the video and discuss ambiguous situations. Once they felt comfortable regarding their agreement we entered the second stage where observers watched and documented emotions in a full video of over 20 minutes without being able to pause it. For the experimental runs we put a visual barrier between the observers and disallowed communication (see figure 6.2). In total, the observers documented emotions throughout four video recorded reminiscence sessions and could take breaks between videos. After coding the last video session we asked observers to complete the questionnaires RAW TLX and QUESI.

6.3.6 Analysis

A perfect operationalisation for kappa calculations would require the rating of predefined units (e.g., time intervals). However, cutting emotional situations out of reminiscence sessions does not realistically reflect how Proxemo would be used in the wild. The videos of the reminiscence sessions were presented in their “natural” state as captured. This means they were not prepared, e.g. cut into pre-discretised snippets which would have allowed for a simpler calculation of IOR measures. Therefore, observers in our setting could not directly label predefined units with



Figure 6.2: During the experimental runs, we set up a visual barrier between observers. Both observers simultaneously documented emotions in a replayed video using Proximo (green highlight). In this figure participants view subject material of a group session with three residents (blue highlight).

emotions but faced two more degrees of freedom. First, observers had to identify emotional situations. Second — in the condition with three observed residents — observers additionally needed to judge who was affected by the emotional situation. Only then could they decide upon a predefined emotion category and set a timestamp for the corresponding resident in the app. Consequentially, the data includes instances where

- both observers documented the same emotion,
- both observers documented an emotion but not the same one and
- only one observer documented an emotion.

Handling instances of missing data where only one of the observers documented an emotion is not an easy decision. Semantically, we are not able to differentiate between instances in the data where both observers thoughtfully analysed the situation and came to different conclusions, so only one observer documented an emotion (i.e., an issue of reliability), and instances where one observer did not document an emotion because she missed the emotional notion (i.e., an issue of thoroughness). De Raadt et al. (2019) conducted mathematical simulations to compare three

possibilities of handling missing data. They concluded that the easiest and most unbiased way to handle missing data — whether at random or not — is listwise deletion of missing ratings. Therefore, in order to avoid an over- or underestimation of IOR, we only analyse instances with complete data pairs that is, time intervals where both observers documented an emotion. For the calculation of IOR we compared timestamps taken within 5 second intervals. This timeframe emerged from our exploration of how the amount of time may deviate between observers to process and classify an observed emotion, find it on the interface of the Proximo App and potentially navigate to the respective user with the bezel.

Inspired by an approach from Tscharn (2019) who faced similar challenges when calculating agreement between coders for unsegmented interview transcriptions, we report Krippendorff’s alpha in addition to the commonly used Cohen’s kappa. Cohen’s kappa (J. Cohen, 1960) is a score that relates the observed percentage of agreement $P(a)$ to the expected percentage of agreement $P(e)$:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

Krippendorff (2004) introduced a measure that compares disagreement rather than agreement between raters and allows for missing data-points. Additionally, Krippendorff provides formulae for nominal, ordinal, interval and ratio data allowing comparisons across metrics (A. F. Hayes & Krippendorff, 2007). In its most general form, Krippendorff’s alpha (Krippendorff, 2004) is defined as the ratio between D_o , a measure for the observed disagreement and D_e , a measure for the expected disagreement by chance:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Both agreement scores range from 1 (perfect agreement) to -1 (perfect disagreement). Hence, our hypothesis that agreement between observers is above chance agreement translates to the statistical hypothesis $\kappa, \alpha > 0$.

Statistical tests were run in RStudio (RStudio Inc., Boston, MA) using the Packages DescTools (version 0.99.41), vcd (version 1.4-8) and Stats (version 4.0.1). Agreement scores were redundantly calculated with the script provided by Freelon (2013) on <http://dfreelon.org/recal/recal3.php>. A significance level of $\alpha = .05$ was used for all statistical tests.

6.4 Results

Within 80 minutes of video recorded reminiscence sessions, observer 2 set 381 timestamps in the Proximo App and observer 1 set 447 timestamps, a descriptive difference of 15%. Among the documented emotions of both observers were 229 shared instances that fulfilled our criteria of co-occurrence within intervals of 5 seconds. Of these shared instances, 131 occurred in the single resident condition with timestamps between observers differing $M = 1.78$ seconds ($SD = 1.45$) on average. The distribution across emotional categories and observers is displayed in table 6.1. The modal value among agreed instances for single residents was in the category *pleasure* with 56 occurrences. A goodness-of-fit test revealed that the agreed instances (diagonal in table 6.1) are not equally distributed across emotional categories, $\chi^2(5, N = 112) = 181.04, p < .001$. This means that observers agreed in some categories more frequently than in others.

Table 6.1: Contingency table of observed emotions in single residents across two observers.

observer 1 \ 2	self- efficacy	pleasure	wistfulness	pride	negative emotion	interest	Σ_2
self-efficacy	48	0	1	1	0	2	52
pleasure	2	56	1	1	1	0	61
wistfulness	3	0	5	0	0	0	8
pride	1	0	0	1	0	0	2
negative emotion	0	1	0	0	2	0	3
interest	4	1	0	0	0	0	5
Σ_1	58	58	7	3	3	0	131

To detect potentially confounding variables in group sessions, we first tested for a dependency between observers and residents. A chi-square test indicated that there was no significant association between the observers and the frequency of emotional instances documented for any of the three residents, $\chi^2(2, N = 359) = 3.0, p = .223$. In group sessions, observers' documentation matched in 98 emotional instances each within a 5 seconds interval with timestamps between observers differing $M = 2.03$ seconds ($SD = 1.62$). The distribution of observed emotions in three residents across observers is presented in table 6.2. The mode among agreed instances for three residents was in the category *self-efficacy* with 47 occurrences across three residents. Again, a goodness-of-fit test revealed that the agreed instances (diagonal in table 6.2) are not equally distributed across emotional categories and residents, $\chi^2(11, N = 73) = 96.15, p < .001$.

Observer 1 reported a descriptively higher task load (RAW-TLX = 5) and a lower QUESI-score (3.71) than observer 2 (RAW-TLX = 0.5, QUESI-score = 4.93). While observer 1's medium RAW-TLX score resulted from high ratings on all scales except physical demand and frustration, the QUESI score rating particularly decreased through the subscale learning effort.

Table 6.2: Contingency table of observed emotions in groups of three residents across two observers. Residents are distinguished by their seating order on the sofa into left (l), middle (m) and right (r). Missing columns and rows indicate that none of the observers documented the corresponding combination of emotion and resident. Note that the *generic negative emotion* was not documented at all during group sessions.

observer 1 \ 2		self-efficacy			pleasure			wistfulness		pride	interest			Σ_2
		l	m	r	l	m	r	l	r	r	l	m	r	
self-efficacy	l	13	1	0	1	0	0	0	0	0	0	0	0	15
	m	0	6	0	0	0	0	0	0	0	1	0	0	7
	r	1	0	17	0	0	3	0	1	0	0	0	0	22
pleasure	l	2	1	1	5	1	0	0	0	0	0	0	0	10
	m	1	0	0	0	22	0	0	0	0	2	0	0	25
	r	0	0	0	0	0	4	0	0	0	0	0	0	4
wistfulness	l	0	0	0	0	0	0	1	0	0	0	0	0	1
	r	0	0	1	0	0	0	0	0	0	0	0	0	1
pride	r	0	0	0	0	0	0	0	0	1	0	0	0	1
interest	l	0	1	1	1	0	0	0	0	0	2	0	0	5
	m	1	0	0	0	2	0	0	0	0	0	0	0	3
	r	0	0	0	1	0	1	0	0	0	0	0	2	4
Σ_1		18	9	20	8	25	8	1	1	1	5	0	2	98

6.4.1 Agreement Scores

We calculated Cohen's kappa and Krippendorff's alpha to determine agreement² between two observers' judgement of residents' emotions. When independently coding the emotions of one resident only, the agreement between observers revealed a Cohen's κ of .764, (95% *CI*, .672 to .855), $p < .001$ (observed agreement: .855, expected agreement: .387) and a Krippendorff's α for nominal data of .764 (131 cases, 136 missing values, observed disagreement: .145, expected disagreement: .713).

When independently coding the emotions of three residents, the agreement between observers revealed a Cohen's κ of .696, (95% *CI*, .596 to .796), $p < .001$ (observed agreement: .745, expected agreement: .162) and a Krippendorff's α for nominal data of .697 (98 cases, 163 missing values, observed disagreement: .255, expected disagreement: .838).

According to the commonly used benchmarks, the kappa agreement scores can be interpreted as *substantial* (Landis & Koch, 1977) or *very good* (Regier et al., 2013). Similarly, the α scores are above the smallest acceptable level (Krippendorff, 2004).

²We report three decimal places here to highlight that kappa and alpha scores result in similar yet not identical values.

6.5 Discussion

In this study we aimed at determining the reliability of Proximo. We operationalised IOR with two observers who assessed residents' emotions in video recorded reminiscence sessions that were replayed in a controlled environment. The two calculated agreement scores Cohen's kappa and Krippendorff's alpha are similar and can be interpreted as substantial agreement. The inferential statistics for the kappa-score support the hypothesis that agreement between observers is clearly above chance.

There are several factors that contributed to a potential underestimation of the agreement scores. First, we conservatively chose observers without prior experience in the dementia context and only trained them together on video data. If observers had had more experience with people with dementia they might have encountered more emotional nuances and thus would have scored more instances consistently and raised kappa. Second, the experimental setup bore small differences between the observers. The smartwatch worn by observer 1 was slightly larger. While the observers' view on the screen was equal, the arrangement of interactive wall and residents presented in the video was not balanced. All video material showed a similar scenery with the residents sitting on the left side (better viewing angle for observer 1) and the interactive wall visible on the right side of the screen (better viewing angle for observer 2). This is a trade-off that comes with the single-screen setup which we chose in order to avoid issues of synchronisation or different replay devices. In a perfect operationalisation, agreement would not be influenced by different perception of the video. Third, the subject material consisted of reminiscence sessions from a rural area of Southern Germany. After the study, observer 2 noted that she had occasionally had trouble understanding utterances of residents with broad dialect. Fourth, frequencies of the emotional categories detected by both observers are not equally distributed. *Self-efficacy* and *pleasure* were more prevalent than other categories. Byrt et al. (1993) demonstrate how data with skewed distributions of prevalences leads to an underestimation of Cohen's kappa. They present an adjustment which is, however, only applicable for 2×2 contingency tables.

The distribution of emotions deserves to be discussed beyond its potential influence on kappa. Hence we will discuss potential reasons for the unequal prevalence of observed emotion across categories. Naturally, not every resident showed all emotions during the short reminiscence sessions. Whereas observers detected emotions of all six categories during sessions with single residents, only 12 of 18 possible emotion-resident combinations occurred in group sessions. Luckily, the atmosphere in the group sessions was so positive that there was no reason for observers to make use of the generic negative emotion category. The main reasons for this may be the carefully crafted reminiscence material and well moderated sessions which primarily stimulated positive autobiographic memories. Ironically, the high prevalence of agreed positive emotions in the replayed sessions which is a good testimony for the person-centred reminiscence sessions impedes comparison to former research. In the care homes we collaborated with, residents looked forward

to technological interventions and asked caregivers for the next opportunity, for example to use the interactive wall. Two decades ago, Lawton et al. (1999a) reported from their observation studies that few residents who were mobile enough “actively avoided organised activity settings” (p.73). This hints at the character of those activity sessions and why observations of pleasure (*mean affect rating* = 1.4) were closely followed by the three negative categories anxiety (1.3), sadness (1.1) and anger (1.1) — that is for residents who could not physically evade the activity. To be fair, those observations took place throughout the day and activity sessions had slightly higher pleasure-estimates than morning care, meal time and down-time, yet negative emotions were similarly prevalent across all four time slots. More recent research deploying the original scale from Lawton et al. (1999a) supports our observations of low prevalence for negative emotions in technological interventions (Feng et al., 2019; Steinert et al., 2020) or when engaging residents with non-technical artefacts (Cohen-Mansfield et al., 2012).

With respect to inter-rater reliability of the observed emotion rating scale in interventions, Feng et al. (2019) report scores ranging from .68 to .74 [which is Cohen’s kappa, as confirmed in personal communication with Feng (June 5, 2021)] and (Feng et al., 2020) report scores of .64 and .78 which appear similar to the scores we found for Proxemo. Steinert et al. (2020) report Cohen’s kappa of .8 but reduced the emotion categories to only differentiate valence between positive, neutral and negative. All these scores are higher than the inter-rater reliability between neural networks and human raters which in prior studies has reached levels of up to $\kappa = .49$ for people with dementia (Steinert et al., 2020) and $\kappa = .38$ for healthy elders (Ma et al., 2019).

It may come as a surprise that the group sessions with three residents did not produce three times as many emotions as the sessions with single residents. However, an explanation for this lies within the important role of the moderator. Emotional moments are often not triggered by technology alone but facilitated through the guidance of the moderator. The crucial role of moderation in the reminiscence context is discussed in more detail in Huber, Berner, Uhlig et al. (2019).

Despite the scores by Cohen and Krippendorff indicating substantial agreement, instances where observers did not agree are still worth a closer look. In group sessions, oftentimes only the resident directly addressed by the moderator displayed any emotion. The contingency table 6.2 indicates that during group sessions one observer sometimes documented *pleasure* or *self-efficacy* for one resident while the other observer documented *interest* — the most passive emotion category — for another resident (mostly the one sitting on the left). Unfortunately, our data in this study does not allow for an answer to definitely distinguish as to whether a) residents were not displaying any emotions beyond *interest* unless directly addressed by the moderator or b) the observers’ focus lay on currently addressed, hence activated residents. However, *interest* was not only documented in passive residents while pleasure or self-efficacy was documented for others. Judging by the descriptive frequencies in the contingency table 6.1, *interest* was difficult to tell apart from the more active category *self-efficacy* during observations of single residents

as well. Additionally, observers only agreed half of the times or less upon *wistfulness* or *pride* and confused them on other occasions with *self-efficacy* and *pleasure*. A reason for the indistinct classification of wistfulness can be that it was difficult to capture the resident's awareness of a reminisced event already concluded in the past.

We have no explanation yet for why observers agreed upon the category *generic negative emotion* in two instances and why in two more instances one observer classified the situation as *pleasure*. This is surprising, considering the different descriptions and valence of the categories. In this study, Proxemo's generic negative emotion category was created by merging all three negative categories of the observed emotion rating scale (Lawton et al., 1999a) into one. Insofar our descriptive results are in line with studies by Lawton et al. (1999a) who found questionable psychometric qualities for all the three negative categories Proxemo's category *generic negative emotion* is merged from.

6.5.1 Limitations & Future Work

The kappa scores reported here slightly deviate from those reported on the same study in Huber, Bejan, Radzey and Hurtienne (2019), yet without effect on the scores' room for interpretation. This is due to a change in preprocessing and that kappa scores in the prior publication had been calculated in SPSS (Version 24, IBM, Armonk, NY) which does not allow for calculations of the "original" κ_{co} by J. Cohen (1960) but according to Hallgren (2012) automatically computes the bias-adjusted κ_{sc} suggested by Siegel and Castellan (1981). In this work we followed the recommendation by Eugenio and Glass (2004) to use the original κ_{co} statistic because the assumption of equal distribution across categories could not be upheld.

Observer 1 reported a medium task load and learning effort. The increased task load may be associated with an increased amount of timestamps set during the study. However, we have no explanation for the difference in perceived learning effort since both observers had exactly the same amount of training and the Proxemo App is particularly easy to use. As mentioned above, the viewing angle on aspects of the video material was not identical for both observers. While there was no significant dependency between the observers' and residents' position in this study, we strive to eliminate this potential bias in future studies. In field studies, observers should be allowed to choose and take a position in the setting where they feel that they are able to capture best what is going on.

To additionally decrease the measurement error between observers it is best selecting observers who are familiar with the individual person whose emotions are assessed or at least with their cultural context. Validity and reliability might further increase by choosing observers who have extended prior knowledge with respect to the subject population beyond what can be conveyed in a ten minutes study introduction. The predefined set of emotions we chose limits the generalisation of our findings. Choosing a set of more extreme emotion categories may increase

the reliability whereas emotion categories with only subtle distinction may decrease reliability. In this study we did not intend to select categories for a particularly high or low reliability but instead stuck to the set of emotions empirically grounded in the case studies from chapter 4.

Substantial reliability does not guarantee high validity of a method. Even when observers' documentations have a large shared variance it is still possible that the instrument does not measure the intended construct. While we let observers train together, we did not measure their individual empathic abilities neither did we define a ground truth for emotions in the subject material. The fact that in this study none of the observers was familiar with dementia in general or the residents occurring in the videos in particular may have increased this issue of validity. In the next study we will, therefore, address the validity of Proxemo. Lastly, we are aware that the listwise deletion of missing ratings raised issues regarding thoroughness. Therefore, measures of Proxemo's thoroughness will be part of the next study as well.

Chapter 7

Effectiveness, Efficiency and Observer Experience

Before deploying UX methods in evaluation scenarios with users, their appropriateness for the intended use needs to be examined. In the last chapter we highlighted how important results of so-called meta-evaluations are for practitioners because they facilitate the comparison of different methods regarding their advantages and limitations (Hartson et al., 2001; Koutsabasis et al., 2007). We found the reliability of Proxemo to be substantial yet faced limitations regarding the lack of knowledge about observers' empathy and thoroughness in documentation. In this chapter, we will take observers' empathy into account and complement our knowledge about Proxemo's quality criteria.

According to Hartson et al. (2001), the ultimate criterion to compare methods is their effectiveness which they define as a product of validity and thoroughness. Therefore, we conduct a larger lab-study with one pre-study and one post-study to measure Proxemo's appropriateness for the intended use. We impose the usability standard criteria in accordance with the ISO standard 9241:11 (ISO, 2018) and complement the measurement of Proxemo's effectiveness with its efficiency and observer experience. In short, we firstly gather information about meaningful artefacts for students from that generation and host two reminisce sessions which we videotape. Secondly, we present the video recordings to observers and evaluate the quality criteria listed above for Proxemo and handwritten notes which are the current documentation standard. Thirdly, we test our assumption on whether students from our participant pool possess a sufficient level of empathy to serve as observers in a validation study.

For their calculation of effectiveness from validity and thoroughness, Hartson et al. build on equations originally postulated by Sears (1997). They contrast real instances with instances detected via a method. In our case, those instances are emotional responses — Hartson et al.'s (2001) research originally addressed the identification of usability problems. In Sears's definition,

thoroughness describes the ratio of real instances identified by a method to the number of real instances existing. Sears’s thoroughness is also referred to as *sensitivity* in psychology or *recall* in data sciences:

$$\textit{Thoroughness} = \frac{\textit{number of real emotions detected}}{\textit{number of real emotions that exist}}$$

A method is valid when it allows evaluators to focus on relevant instances (Sears, 1997). Thus, validity describes the proportion of detected instances that really exist. The outcome of this formula is referred to as *positive predictive value* by psychologists or *precision* by data scientists:

$$\textit{Validity} = \frac{\textit{number of real emotions detected}}{\textit{number of instances classified as emotions}}$$

For better comparability of methods Hartson et al. suggest effectiveness as a single indicator ranging from 0 to 1 (perfect) that indicates when either thoroughness or validity are low. They define effectiveness as the product of thoroughness and validity:

$$\textit{Effectiveness} = \textit{Thoroughness} \times \textit{Validity}$$

However, to calculate these factors one needs to contrast instances detected by the method with “real” instances. Hartson et al.’s first suggestion of identifying real usability problems by comparing them to a problem inventory does not translate easily to interaction-triggered emotions. Here, it is more appropriate to identify realness through judgment by users or experts. With respect to validity, the relation between detected instances and manifest external criteria corresponds to criterion-related validity in psychology (Döring & Bortz, 2016, p.447). *Specificity*, a value that is typically reported in clinical studies (Bortz & Schuster, 2011, p.56, 176), cannot be calculated in our studies because there is no finite number of non-emotional instances in episodes of human experience. Without knowing the fixed amount of ‘negatives’ the specificity – or true-negative-rate – cannot be computed.

Usability as defined by the ISO 9241-11:2018 (ISO, 2018) comprises effectiveness, efficiency and satisfaction. Effectiveness in terms of usability is defined as the accuracy and completeness

with which users achieve their goals and Hartson et al.’s equation for effectiveness is one possible way to measure it. Efficiency in terms of usability, however, extends beyond the objective dimension of time and cost saving and includes human effort which is often measured as subjective efficiency or mental workload, for instance with the NASA Task Load Index or its shorter variation RAW TLX (Byers, 1989). Satisfaction as the third factor “includes the extent to which the user experience that results from actual use meets the user’s needs and expectations” (ISO, 2018, p. 3.1.14) and can, therefore, be measured appropriately with single item questions on satisfaction and fun or validated UX tools such as the User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018; Schrepp et al., 2017). A summative usability study examining the effectiveness, efficiency or UX of Proxemo has not been conducted yet and a benchmark study against handwritten notes during user observations is lacking. Proxemo was designed to quickly document emotions but spare observers the taking of detailed notes in a dynamic situation. With the increased documentation efficiency observers should be able to focus more on user interactions and capture emotional responses more thoroughly using Proxemo. Additionally, we expect that reducing documentation demands reduces observers’ perceived mental workload. Both the reduced workload and the awareness of a better performance increase observers’ satisfaction and improve their overall experience. Because Proxemo does only support the documentation but not recognition of emotions we have no reason to expect a difference in validity. In the following, we present an empirical meta-evaluation to test the appropriateness of Proxemo as a documentation method for observational evaluation studies. In a controlled environment that replicates aspects of an in-the-wild observation, we compare Proxemo with handwritten notes as the currently prevalent practice in observational studies. We refer to the control condition as Pen&Paper. Our main contributions are (a) to report quality measurements for Proxemo and (b) to benchmark Proxemo against Pen&Paper. Based on our assumptions, we tested the following hypotheses:

Effectiveness

- H1.1 Observers achieve higher thoroughness with Proxemo than with Pen&Paper.
- H1.2 Validity does not differ between the conditions.
- H1.3 As a result of H1.1 and H1.2, observers achieve higher effectiveness with Proxemo than with Pen&Paper.

Efficiency

- H2.1 Values of the RAW TLX are lower in the Proxemo condition than in the Pen&Paper condition.
- H2.2 As a consequence of efficiency, after using Proxemo observers can answer more questions on observed details compared to Pen&Paper.

H2.3 Observers are aware that they remember more details of their observation after using Proxemo rather than Pen&Paper.

Observer Experience

H3.1 Values of the UEQ-S are higher in the Proxemo condition than in the Pen&Paper condition.

H3.2 Observers report higher values for satisfaction after using Proxemo than after using Pen&Paper.

H3.3 Observers report higher values for fun after using Proxemo than after using Pen&Paper.

H3.4 Observers prefer Proxemo to Pen&Paper for future observations.

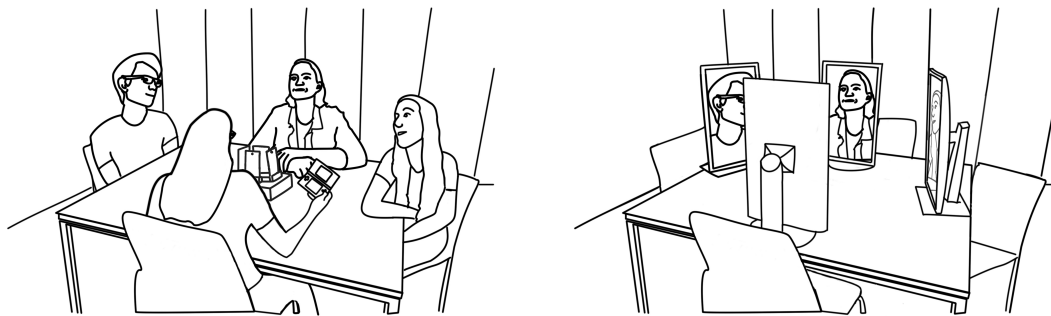
Prior to testing our hypothesis, in the following section we describe a pre-study with the aim of generating stimulus material of reminiscence sessions required in the main study.

7.1 Pre-study to Generate Stimulus Material

For the meta-evaluation of two evaluation methods, we needed to prepare an appropriate scenario. Its requirements were to (1) resemble properties from a reminiscence session from the dementia context, (2) be repeatable for multiple evaluators and (3) display emotional situations observable by untrained student evaluators. Correctly assessing the subjective state of persons with dementia is challenging (Lawton et al., 1996) and not the focus of this work. Emotions can be detected best when both the person expressing and rating the emotion origin from the same ethnicity, nation and region (Elfenbein & Ambady, 2002). Young raters are generally better at detecting facial emotions (G. S. Hayes et al., 2020) and perform best when observing people of a similar age (Riediger et al., 2011). To cater for best conditions concerning emotion recognition, we hosted sessions with groups of reminiscing students. We videotaped these group sessions and later displayed the recordings during the main experiment (figure 7.1) in order to have each evaluator code the same situations. Properties that we replicated from reminiscence sessions in the dementia context are a) triggers from formative years of one's life (between the age of 15 and 25; Martin et al., 2005) that are b) presented in a scheduled event where c) multiple participants attend who do not know each other too well and consequentially d) display emotions with varying frequency.

7.1.1 Procedure: Conceptually Replicating Reminiscence Group Sessions

First we identified culturally relevant reminiscence triggers for the age group under study in Germany by conducting an online survey. Respondents to the survey ($N = 19$, $M_{age} = 21.21$,



(a) Focus group of students reminiscing about a Nintendo DS™. (b) Video representation of the focus group in the experimental setup.

Figure 7.1: The illustrations show how we videotaped each participant of a reminiscence focus group with vertical smartphones in the table center and presented these videos to evaluators in the experimental setup.

$SD_{age} = 2.28$) ranked suggested reminiscence triggers and added items to the list. The resulting selection of appropriate reminiscence triggers used during the group session included:

- Objects: Diddl-Mouse merchandise, MP3 player, Nintendo DS, Window Color and an old cell phone
- Films and series: Harry Potter, H2O — Just add water, Takeshi's Castle, Tabaluga and Spongebob
- Music presented on the MP3 player with active speakers from the artists: Las Ketchup, Black Eyed Peas, No Angels, Avril Lavigne, Wheatus, Tokio Hotel, Atomic Kitten, Daft Punk, Gorillaz, Eminem and Train

We then hosted two group sessions with three invited participants each. The reminiscing students were between 20 and 25 years old, did not study in the same degree courses as the evaluators, consented to video recordings and received 25 € for their time. During the group session students sat around a table while a host presented them reminiscence triggers from their formative years in the 2000s. The host invited the group to interact with triggers whenever they wanted, reminisce and share with the others whatever came to their mind (figure 7.1a).

7.1.2 Analysis: Extracting and Annotating Meaningful Sequences

We videotaped each participant of the group session in portrait format using four smartphones mounted in a custom-made stand in the centre of the table (figure 7.1a). The sessions resulted in approximately two hours of video material from which we selected sequences of about 30 minutes

each that were rich and diverse in emotions from all reminiscing participants. We adopted the set of emotions from the latest iteration of Proxemo in the dementia context (chapter 4) but left out self-efficacy ¹ and relabelled general alertness as interest ² and *wistfulness* as *nostalgia* to better match the age group (figure 7.2).

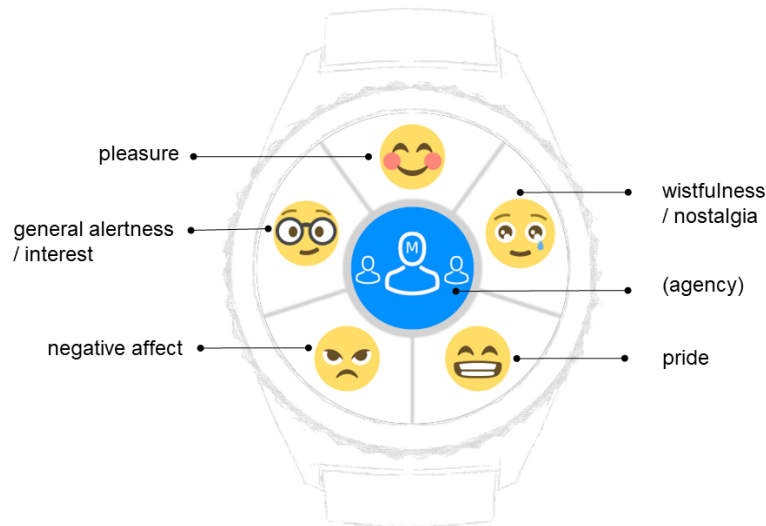


Figure 7.2: Smartwatch interface of Proxemo with emotion categories suitable for reminiscence. (Sense of) Agency was not used in the study at hand.

Calculating validity and thoroughness require the existence of a ground truth concerning emotions of each reminiscing person in the video data. Therefore, two experimenters who have 8 years of experience in observational studies and 2 – 4 years of experience in dementia and reminiscence coded the videos independently. The coders then merged and reviewed their ratings over three iterations and discussed critical situations until achieving consent for all instances. Because this so achieved expert rating defined the ground truth in this study, the validity to be measured in the main study can be considered external criterion validity.

¹Huber, Bejan, Radzey and Hurtienne (2019) and literature on dementia refer to self-efficacy often as “[a sense of] agency”

²In older versions of the OERS, the category was already titled *interest* but renamed to *general alertness* in the latest distributed version (Lawton et al., 1999b)

7.2 Method: Effectiveness, Efficiency and Observer Experience

After having resembled and videotaped reminiscence sessions as preparation we proceed to the main study. In this section, we describe the operationalisation of our hypotheses in a controlled experimental setup. In short, we used videotaped live sessions in order to provide each observer with the same stimuli (Hertzum et al., 2014). Then we presented those video recordings to observers and asked them to document observed emotions using the two different methods Proxemo and Pen&Paper. We conducted a meta-evaluation of both methods regarding their effectiveness, efficiency and observer experience. As a reminder, we hypothesised that Proxemo provided a better observer experience and was more efficient, thorough and effective than Pen&Paper but did not differ in terms of validity. To exclude biasing factors from affecting the performance of the two methods, we conducted our research in a controlled environment and optimised the detectability of emotions.

7.2.1 Setup

In the experimental setup, we reproduced the reminiscence group session by placing four pivot monitors in portrait format in the same way as reminiscing persons had been seated (figure 7.1b). On the monitors, we replayed the synchronised videos displaying one reminiscing person each plus the host. Sound was played from active speakers hidden under the table together with the laptop running the videos. Observers had the choice to either sit or stand behind the host-monitor or walk around the room. The room held a separate table for the experimenter and for the observers to fill in questionnaires.

7.2.2 Experimental Design

The experiment used a within-subjects design with the conditions Proxemo and Pen&Paper. In the Proxemo condition, observers used a smartwatch (Gear S3, Samsung Electronics, Seoul South Korea) running the Proxemo App to document observed emotions (figure 7.3a). In the Pen&Paper condition, observers used a clipboard with paper sheets and a pen (figure 7.3b). We randomised the order of the conditions and reminiscence sessions between observers.

Our primary dependent variables were thoroughness, validity and effectiveness of all documented emotional responses. We calculated those variables according to Sears (1997) and Hartson et al. (2001) as reported in the introduction of this chapter and with “real emotions that exist” corresponding to our ground truth extracted in section 7.1.2. We determined efficiency directly using the perceived task load questionnaire RAW TLX (Byers, 1989). In addition, we measured consequences of efficiency a) subjectively by asking participants in which condition they believe they caught more details of the observed situations and b) objectively by counting

correct responses to six questions about meaningful details from each session. One question on such meaningful details was for example what the person sitting in the middle was particularly proud of when talking about her Nintendo DS experience. We measured UX with the short version of the User Experience Questionnaire (UEQ-S, Schrepp et al., 2017) and single item questions with 5-point Likert scales on the fun and satisfaction the observers experienced when using each method. After the observers completed both conditions, we asked them in the final questionnaire to state their preferred method and justify their choice. In the final questionnaire, participants also had the opportunity to comment on positive and negative aspects of Proxemo or suggest improvements. As an exploratory variable, we measured mobility by noting whether participants were standing, walking or sitting during the observation.

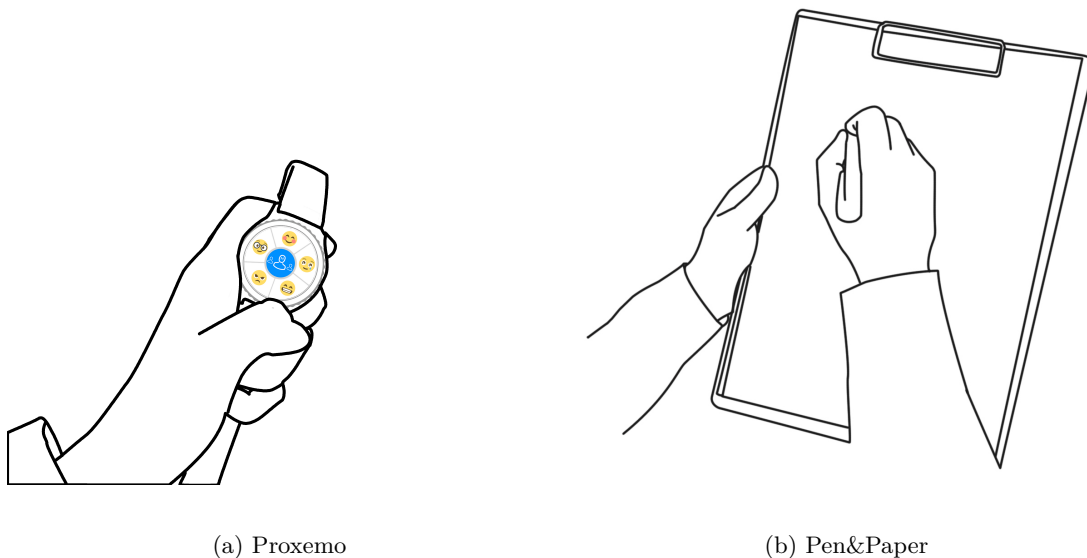


Figure 7.3: Illustrations of how observers documented the observed emotions during the experiment in the conditions (a) Proxemo and (b) Pen&Paper.

7.2.3 Participants

Without orientation from literature on the effect size of a documentation aid for emotions compared to handwritten notes, we expected a medium effect ($f = .25$). G*POWER (version 3.1.9.4, Faul et al., 2007) recommended a sample size of 54 for a two-tailed comparison between two dependent means with $\alpha = .05$ and $1 - \beta = .95$. We recruited $N = 52$ students from the institute's participant pool. To ensure that students participating in our study knew what the observations were about, we made it a precondition for participation that students had taken part and passed

one of two available courses on observation studies. All novice observers signed consent and participated in exchange for course credit. Prior to inferential analysis, we excluded one observer whose validity score in the Proxemo condition was 3.28 *SD* below the mean. The remaining $N = 51$ participants were aged 19 – 30 ($M = 22.0$, $SD = 2.49$) and included 41 female and 10 male students.

7.2.4 Procedure

After being welcomed and signing the informed consent, observers were asked to imagine their role in the experiment as follows: observers had the important task to capture emotions from a live user research session for a client. The captured information would be used towards the development of a novel system for specifically triggering positive emotions and avoiding negative emotions. The experiment contained the following steps:

1. Observers were asked to memorise a list with all relevant emotions along with a verbal description of indicators and representative emoji. Observers were allowed to consult the list during training but not during the two main trials.
2. The experimenter explained the documentation method of the first condition (Proxemo or Pen&Paper) followed by a training session with a 10 minute video sequence of the first reminiscence session. In the Pen&Paper condition the experimenter checked the comprehensibility of the handwritten protocol and if necessary repeated instructions on noting the emotions together with the trigger.
3. After the training the first observation trial with a 20 minute video sequence of the first reminiscence session took place. In the Pen&Paper condition observers received an unlabelled list with emoji to compensate for the presence of emoji on the interface in the Proxemo condition.
4. Subsequently, observers filled in intermediate questionnaires and responded to six questions of varying difficulty about the content of the previously viewed video sequence.
5. Observers then followed steps 2 – 4 with the second method and the video material of the other reminiscence session and then completed a final questionnaire.

7.2.5 Data Processing and Analysis

We used the video annotation software ELAN (<https://tla.mpi.nl/tools/tla-tools/elan/>) to synchronise Proxemo data with the videos and digitise Pen&Paper protocols by annotating documented instances in the related video sequence. We preprocessed the raw data with RStudio (RStudio Inc., Boston, MA) and considered Proxemo timestamps as matches when they occurred within an interval of 5 seconds with respective instances in the ground truth. Within

the Pen&Paper condition instances were identified based on scene descriptions and all emotional annotations generously allocated by the experimenter as timely matches in the related video sequence. Mismatches in the Pen&Paper condition, therefore, indicate manually documented emotions that do not match any of the emotional stamps within the described scene in the ground truth. Based on these matches we calculated thoroughness, validity and effectiveness for each participant in each condition. All statistical tests were computed with SPSS (Version 24, IBM, Armonk, NY). Graphs are based on export from Microsoft Excel if not stated otherwise.

7.3 Results

All tests were run against a Bonferroni corrected α level of .005. The order of conditions did not result in any effects that impede the interpretation of the main effect between Proxemo and Pen&Paper. For a better comparability across studies we report the mean and median for demographic data of our participants. There was no indication of a dependency between gender and degree courses ($p = .094$, two-tailed Fisher's exact test).

7.3.1 Effectiveness

Based on the emotions documented via Proxemo or Pen&Paper we measured effectiveness as a product of thoroughness and validity. The significant outcomes of paired t-tests indicate that compared to the Pen&Paper condition, observers achieved higher scores in thoroughness $t(50) = 8.25$, $p < .001$, $d_z = 1.16$ and effectiveness $t(50) = 3.59$, $p = .001$, $d_z = .51$ but lower scores in validity $t(50) = 7.68$, $p < .001$, $d_z = 1.07$ when using Proxemo. Descriptive data are presented in figure 7.4.

In the final questionnaire observers noted that they sometimes struggled with the bezel when switching between users ($n = 13$) or tapped the wrong emotion ($n = 4$). Some observers criticised that decisions could not be undone in the Proxemo condition ($n = 7$). This indicates that there are slips in the dataset. Unfortunately, there was no solution to clearly distinguish slips from erroneous decisions in the data. Post-hoc we calculated *documentation ratio* as the proportion of documented emotions to all occurring emotions in the session. The documentation ratio when using Proxemo ($M = .63$, $SD = .30$) is higher than when using Pen&Paper ($M = .28$, $SD = .10$), $t(50) = 9.06$, $p < .001$, $d_z = 1.27$. The Euler diagram in figure 7.6 visualises the proportions of emotions that were discovered with both methods and those that were documented exclusively in each condition.

Post-hoc we explored the thoroughness and validity of documented emotions by emotion category. A visual inspection of the descriptive data in figure 7.5 indicates that thoroughness was stable over all emotion categories when participants used Proxemo but varied greatly between

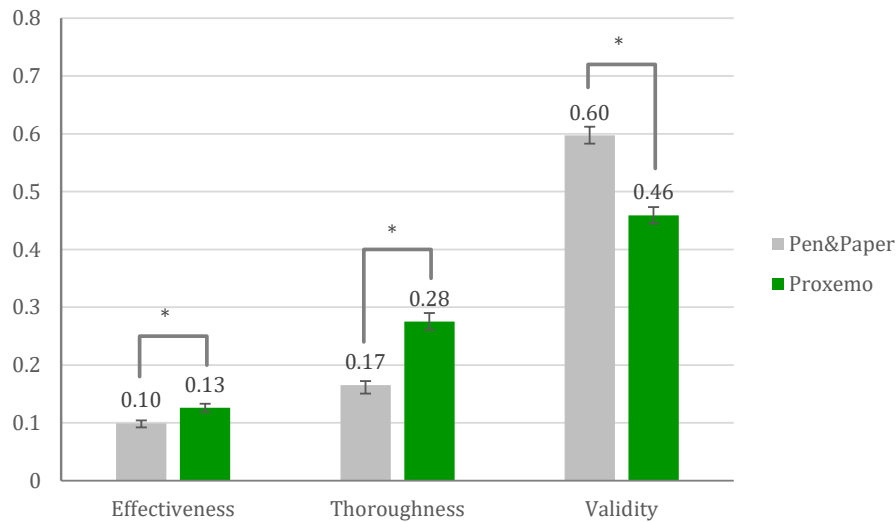


Figure 7.4: Bar graph displays results for effectiveness, thoroughness and validity. Error bars are SE_M .

emotions when using Pen&Paper. Significant correlations between thoroughness in both conditions were only detectable for the emotion categories nostalgia ($r = .60, p < .001$) and negative emotions ($r = .49, p < .001$). The descriptive difference in validity between conditions is particularly high for *pride* and smallest in *interest*. For validity, *pleasure* was the only emotion category that correlated significantly between conditions ($r = .41, p = .003$).

7.3.2 Efficiency

To measure subjective efficiency observers completed the RAW TLX questionnaire using Proxemo ($M = 41.73, SD = 15.134$) and Pen&Paper ($M = 59.69, SD = 16.09$). Paired t-tests showed that the subjective workload is significantly lower in the Proxemo condition, $t(50) = 7.78, p < .001, d_z = 1.09$. Descriptive data of all subscales is presented in 7.7. There were outliers on subscales of the dataset. Excluding them from this test had no impact on the significance of the outcome which is why the reported data still includes the outliers. As a consequence of documentation efficiency we inquired how much observers remembered from the observed reminiscence sessions. From the overall 306 responses in each condition observers could descriptively answer more ($n = 242$) when using Proxemo compared to Pen&Paper ($n = 225$). However, this difference was not statistically significant, $\chi^2(1) = 2.61, p = .13, V = .07$. Finally, we asked participants with which method they caught more of what was going on in the reminiscence sessions. A binomial test indicated that the proportion of participants that believe they observed

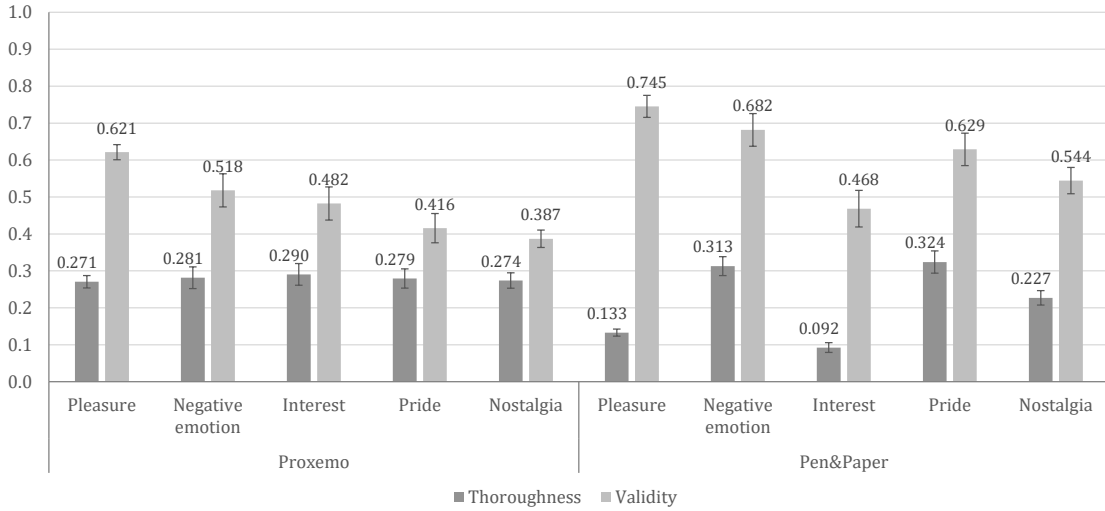


Figure 7.5: Bar graph displays results for thoroughness and validity by condition and emotion. Error bars are SE_M .

more when using Proxemo ($n = 40$) is higher than the expected value, $p < .001$, $g = .29$. Qualitative data indicates that even when using Proxemo observers found it hard to capture everything during emotional situations between multiple persons ($n = 4$).

7.3.3 Observer Experience

Observers documented their experience by filling in the UEQ-S after using Proxemo ($M = 1.59$, $SD = .68$) and Pen&Paper ($M = -.95$, $SD = .91$). Paired t-tests show that this difference is statistically significant, $t(50) = 14.97$, $p < .001$, $d_z = 2.1$. Descriptive values are presented in figure 7.8 next to international benchmarks (Hinderks et al., 2018). Furthermore, we asked participants directly how much fun and satisfaction they experienced (0 = none; 4 = a lot) in the current condition. Participants experienced more fun ($t(50) = 7.62$, $p < .001$, $d_z = 4.0$) in the Proxemo condition ($M = 3.16$, $SD = .81$) than in the Pen&Paper condition ($M = 1.41$, $SD = 1.25$). The score for satisfaction was also higher ($t(50) = 5.56$, $p < .001$, $d_z = 4.71$) in the Proxemo condition ($M = 3.22$, $SD = .90$) than in the Pen&Paper condition ($M = 1.94$, $SD = 1.17$).

After completing both conditions, we asked observers which method they would prefer for similar tasks in the future. Of $N = 51$ observers, $n = 44$ observers named Proxemo as their preference and $n = 7$ would prefer to work with Pen&Paper. According to a binomial test those values differ significantly from an equal distribution between conditions ($p < .001$, $g = .36$) and comply with our hypothesis. Reasons for this choice can be found in the qualitative data. Observers who preferred Proxemo argued that it was simpler, more efficient and less cognitively

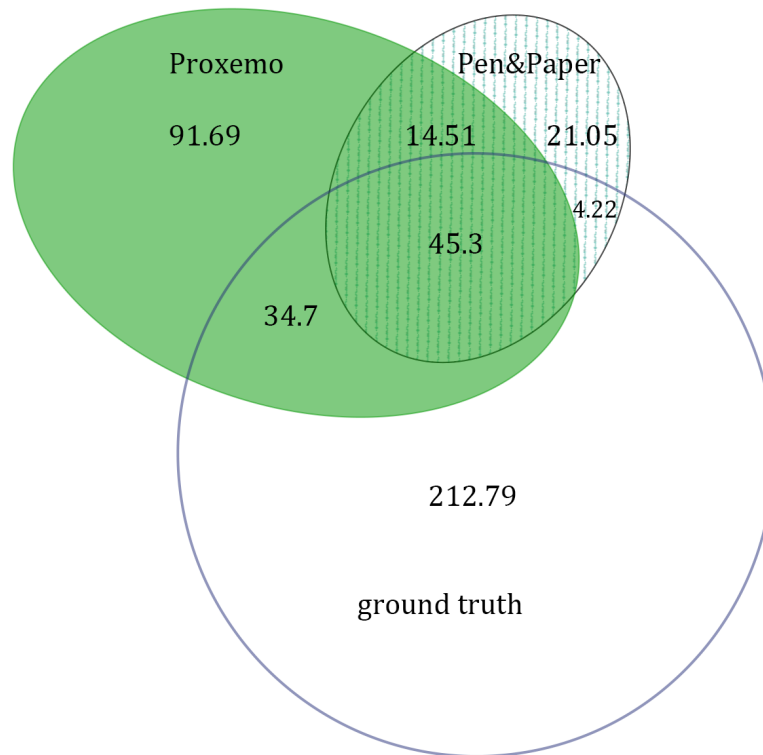


Figure 7.6: The Euler diagram displays overlaps between the ground truth and emotions documented with Proxemo and Pen&Paper. It visualises the proportion of shared and exclusively documented instances. Provided values are arithmetic means per participant. The graphic is based on *eulerAPE* (Micallef & Rodgers, 2014, software distributed on <http://www.eulerdiagrams.org/eulerAPE/>).

demanding than Pen&Paper ($n = 39$). A second argument was that Proxemo left them more capacity to follow the conversation and capture details, thus allowing for a more precise or effective documentation of emotions ($n = 33$). Few observers stated explicitly that Proxemo caused them less pressure ($n = 5$) and the documentation style was innovative and more fun ($n = 3$). Finally, we explored observers' mobility. During the study the experimenter noted whether observers sat down during the evaluation or chose to stand or walk. As reported in table 7.1 observers were more likely to stand or walk when using Proxemo, $\chi^2(1) = 19.63$, $p < .001$, $V = .62$. However, the varying mobility did not correlate with effectivity or its factors, all $r < .15$. Additionally, when using Proxemo, only $n = 5$ observers put the smartwatch on their wrist, the other $n = 46$ held it in their hand. This behaviour did not affect the effectiveness ($U = 114$, $p = .913$).

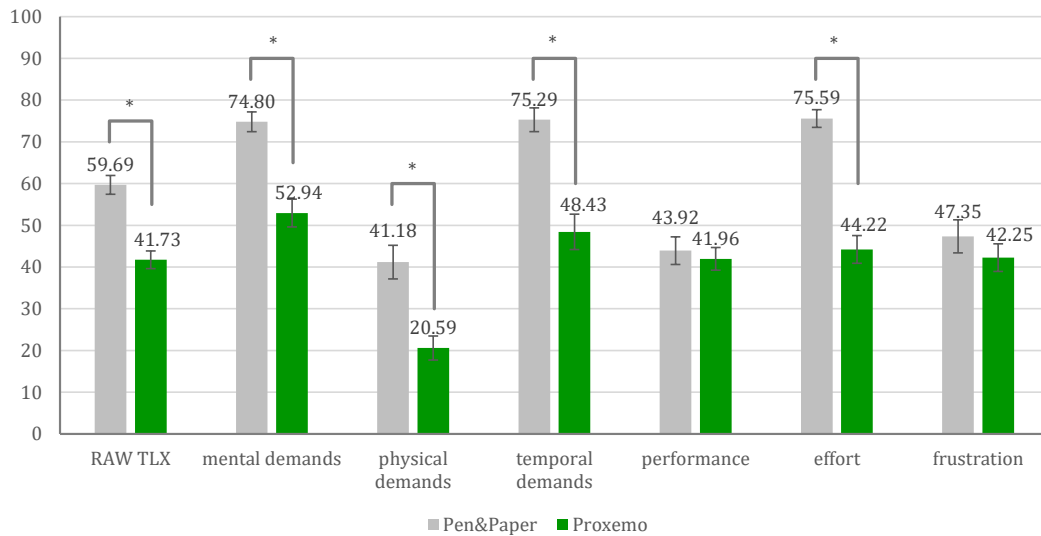


Figure 7.7: Bar graph displays results of RAW TLX and all subscales. Error bars are SE_M , (*) indicates significance with $p < .001$.

7.4 Discussion

In this randomised controlled study we tested quality criteria of Proxemo and compared it to handwritten documentation of observed emotions. Our hypotheses were in short that Proxemo offers higher effectiveness, efficiency and observer experience than documentation with Pen&Paper.

The higher values for effectiveness and its factor thoroughness in the Proxemo condition support our hypotheses H1.1 and H1.3. A closer inspection of the documented instances regarding their overlap suggests that most emotions were consistently discovered with both methods. There are only few valid emotions documented singularly with Pen&Paper. Without Proxemo over 2/5 emotions would have remained undetected. Qualitative data indicates a reason for the high thoroughness score of Proxemo. The simple interface of the Proxemo App allowed observers to remain focused on the conversations and emotions of reminiscing persons. This impression is consistent with the lower workload measures (H2.1) when using Proxemo. We expected that

Table 7.1: The cross table lists the frequencies of participants who chose to sit or stand/walk during both conditions they experienced.

Proxemo	Pen&Paper		Total
	Sitting	Standing/Walking	
Sitting	27	0	27
Standing/Walking	11	13	24
Total	38	13	51

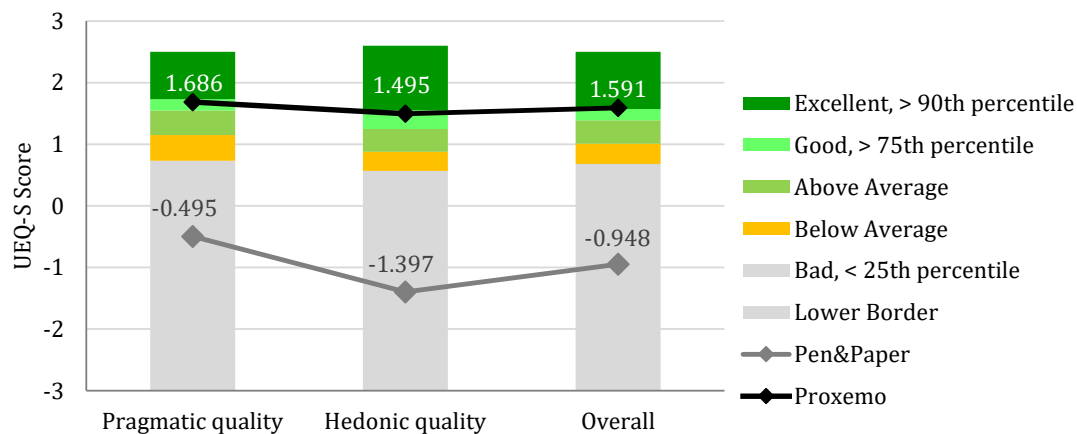


Figure 7.8: Data points represent mean values of UEQ-S in both conditions. Stacked bar graphs indicate actual meaning behind the scores in context of other studies (Hinderks et al., 2018).

observers' reduced workload in the Proxemo condition results in a deeper understanding of the reminiscence situation. According to our observers' comments and responses (H2.3) they share this belief. However, more objective data does not confirm the feeling since observers were not able to answer more questions on details after using Proxemo (H2.2). Apparently, the advantage of a cognitively less demanding documentation method resulted in capturing more details and documenting these significantly more thoroughly at first. Yet, details on reminiscence situations did not remain in the observers' memories to the same extent until the end of the session. To be fair, we only instructed observers to document emotions not to memorise details. During an evaluation in the field the details could easily be retrieved from the video timestamped with the Proxemo App, especially if the documentation data is sufficiently thorough.

In terms of observer experience all results were consistently in favour of Proxemo with our hypotheses H3.1-4. Observers reported to have experienced more satisfaction and fun after using Proxemo and preferred it to the documentation with Pen&Paper. The score of the UEQ-S indicated that Proxemo resulted in a higher observer experience than Pen&Paper. UEQ-S was developed for interactive products and the validity of applying it to Pen&Paper is questionable. Therefore, we additionally compared Proxemo to a benchmark containing data of over 240 interactive products (Hinderks et al., 2018) where Proxemo is in the range of the 10% best results. When using Proxemo observers embraced the chance for increased mobility and stood or walked during the observation. They also preferred single-handed interaction to putting the watch on their wrist. By holding the watch in one hand, a twitch of the thumb sets a timestamp. Neither of these classified variables affected effectiveness.

Finally, we need to discuss the unexpected outcome for validity. We hypothesised (H1.2)

that validity does not differ between conditions because neither of the methods actively supports the recognition of valid emotions. Due to the large effect of validity, we can no longer assume that it does not differ between the conditions Proxemo and Pen&Paper. Descriptively, values for validity were higher in the Pen&Paper condition. One reason for this may be that documenting an emotion in Proxemo causes a lower effort – particularly lower mental, physical and temporal demands. Hence, Proxemo is so fast and convenient to use that observers may tend to click before they think. Thoughtless documentation despite uncertainty about the observation can reduce validity. In contrast, the anticipation of the upcoming effort associated with manually noting down an observation possibly makes observers think twice, and once decided upon taking a note, they still have sufficient time to reflect their observations while they write. In support of this explanation, observers documented a higher ratio of emotions when using Proxemo. Additionally, qualitative data indicates that the resulting dataset includes slips that could not be undone in the app. Incidentally documenting an emotion and forgetting to cross it out is an unlikely scenario for the Pen&Paper condition which none of the observers reported. Consequentially, slips and rashly documented observations only diminished validity in the Proxemo condition rendering it a conservative measure for validity. The validity score of Proxemo could be increased in future work by adding an undo function to the tool or artificially decreasing its efficiency, which contradicts the main purpose of the tool. While an undo feature can simply be added to the app, reflected documentation rather is better tackled with training observers.

7.4.1 Implications

Above we mainly discussed the direct comparison of Proxemo with Pen&Paper as our baseline. When interpreting the meaning of quality measures it is hard to tell which scores for effectiveness would be a “good” score. In the literature, the quality of observation methods is mainly judged by their reliability which has already been shown to be substantial for Proxemo in chapter 6. Benchmarks for effectiveness values exist in the domain of usability evaluation methods, from where the ultimate criterion of effectiveness originates (Hartson et al., 2001).

In an article about quality criteria of ergonomic methods, Stanton (2016) lists the validity of observed errors as .47. Compared to this validity benchmark, Proxemo ($M = .46$) performed similarly and Pen&Paper ($M = .60$) even achieved a high validity score. Stanton, in his discussion of the reliability and validity of ergonomic methods (Stanton, 2016), does not report benchmarks for thoroughness, and we are not aware of any further meta-evaluations of observations. When broadening the scope to include expert ratings, there are benchmarks for usability errors on websites. We are aware that these domains do not match as perfect comparison for observed emotions during formative evaluations. Due to a lack of closer related work we still contrast their results with our findings. Koutsabasis et al. (2007) asked students with comparable background to our study to conduct expert evaluations of a website and found thoroughness scores ranging

from .2 to .41, validity scores ranging from .71 to 1 and resulting effectiveness scores between .15 and .3. Compared to these benchmarks, thoroughness of Proxemo ($M = .28$) is placed in midst of the range whereas validity and resulting effectiveness ($M = .13$) are below this range. Again, a possible explanation is that the observation of emotions in group settings is more complex and not comparable to expert usability ratings of websites. Future work should address the lack of publicly available quality criteria for observational methods.

Irrespective of complexity or other factors that influence our results we must emphasise that the absolute scores for validity and thoroughness achieved in our study do not occur high for either condition. We constructed a setup where we conceptually replicated reminiscence sessions as they take place in dementia care homes but controlled all variables to allow the best results possible. In detail, no other persons such as fellow residents distracted the session which should improve thoroughness. Voices and faces of young persons catered for best understanding of natural emotions, thus improving validity. Furthermore, we preselected the screened scenes to display a variety of clearly distinguishable emotions and recruited observers from the same age group as the reminiscing persons. This should increase validity compared to settings in dementia care homes where typically the observers are younger than the residents. Nevertheless, about every second documented emotion in the Proxemo condition and 4/10 of emotions in the Pen&Paper condition were invalid. What is the resulting implication for field observations? User researchers validate their observations and interpretations directly with the users (Holtzblatt et al., 2004). However, reduced communicative abilities render this impossible for persons in later stages of dementia. The only indicators to the validity of observations is the experience of the observer and their familiarity with the observed persons. Therefore, the scores for validity that resulted from the controlled conditions in our study are alarmingly low and question the validity of observed emotions without confirmation by observers in general and reported in dementia research in particular. To boost thoroughness of detected emotions, observers could add to their in-situ ratings in the aftermath when analysing the video.

Regarding thoroughness, Molich and Dumas (2008) found values of .056 – .2 for usability issues identified by nine test teams. Our Pen&Paper condition ($M = .17$) can be found within that range — however, participants were more thorough in our study when using Proxemo to document emotions ($M = .28$). In another study on the evaluator effect in usability tests Hertzum et al. (2014) presented videos of single interacting users to evaluators who were allowed to pause and replay the video, achieving levels of thoroughness between .32 and .33. Their participants took on average 8 hours to analyse a 33-minute video. From this perspective, the achieved thoroughness of .275 with Proxemo when documenting emotions in real time for three users at a time is impressively efficient. But again, meta-evaluation from the domain of usability does not serve as perfect benchmark for the logging of observed emotions.

When using Proxemo thoroughness was not only higher but additionally appeared to be more stable across emotions. Our post-hoc exploration of thoroughness and validity by emotion

revealed that thoroughness correlated in only two emotions among conditions and validity even merely in one emotion category. Since the use of documentation methods was manipulated within participants, these variances in emotion documentation can be attributed to the method itself. While Elfenbein and Ambady (2002) show how the recognition of specific emotions varies cross-culturally and Mill et al. (2009) report an age-related decline to recognise negative emotions, our data hint towards a difference in documentation method. For practitioners this implies that the choice of method for the structured documentation of observed emotions is critical and may boost or impede the factors of effectively detecting different emotions. For instance, the detection of *pride* was descriptively more effective in the Pen&Paper condition while *interest* was detected descriptively more effectively when using Proxemo. Of course, our descriptive data does not allow for conclusive recommendations. Yet, the great variance of validity between emotions and conditions is conspicuous and calls for replication studies. Future research should try to replicate the differing effectiveness between emotions, inferentially determine how large the effects are and contribute to an understanding of why the detected emotions depend on the used method.

We calculated effectiveness as the simple product of thoroughness and validity because we had no reason to assume that one of the two measures was more important than the other. Hartson et al. (2001) adapt a formula that originates in natural language processing (Manning et al., 1999) to calculate a weighted product of thoroughness and validity. They argue it is often more important for evaluators to “[find] what they are looking for, even at the cost of having to sort through some irrelevant items retrieved.” Proxemo timestamps usually provide the basis for subsequent video analysis (Bejan et al., 2018) that already involves sorting through data. Therefore, calculating effectiveness as an unweighted product of thoroughness may be considered conservative in this context and Proxemo’s true benefit over Pen&Paper can turn out higher when applying weights to the measure of thoroughness.

The most important question of the study was whether Proxemo is an appropriate method for the documentation of observed emotions. When relying on effectiveness as the “ultimate criterion” (Hartson et al., 2001), Proxemo is better than Pen&Paper. Additionally, our subjective results indicate that Proxemo offers higher efficiency and UX for observers than Pen&Paper. Using Proxemo is particularly worthwhile, when a more thorough documentation of emotions is important. With the implementation of the Proxemo App we used in this study, the price for fast documentation of emotions and a reduced interface was erroneous data and consequentially reduced validity. Validity may improve with training of observers towards reflected documentations and the simple implementation of an interaction to undo the last timestamp. Until then, Pen&Paper should be the method of choice for projects requiring higher validity than thoroughness. Of course, in the end every researcher or practitioner needs to decide on a suitable method for their current question by themselves.

7.4.2 Limitations & Future Work

In contrast to the practice in usability studies we did not distinguish the severity between detected instances (Koutsabasis et al., 2007). Therefore, we cannot say whether the emotions detected in either condition were the most or least important occurrences. Additionally, there is a limitation of internal validity to our study. We implemented an iterated expert rating to establish the ground truth in the video sequences. A better solution would be to let the persons in the video validate their own emotions (Hartson et al., 2001). In the study reported in the next chapter, we will ask users to directly annotate their own emotions after the session and potentially distinguish the instances by intensity. A limitation of external validity is the choice of young observers for the reminiscence sessions. On the one hand, people with dementia tend to display emotions less frequently which may reduce the large difference in thoroughness between conditions. On the other hand, observers in the context of dementia often face ten or more people with dementia per session and sometimes take the role of steward observers (Huber, Berner, Uhlig et al., 2019). Therefore, the importance of documenting emotions quickly to keep focus on the situation or re-engage with persons was possibly underrated in our study.

We used a publicly available implementation of the Proxemo App³. Its lack of an undo feature resulted in observers not being able to delete slips in-situ. This caused an advantage for the Pen&Paper condition where observers were able to cross out wrong statements or just wait and think longer before they write since handwritten notes are not time critical. According to observers' comments, the missing undo function of the app descriptively reduced the validity of the app and ultimately may have biased the inferential statistics between conditions. For future implementations of Proxemo, we recommend adding an undo function or instruct observers that — in the light of 40 – 50% invalid data — single slips do not matter. When observers are aware of the missing feature, they may keep focused on the situation under observance and ignore slips to sort them out later. Embedding further features or more emotion categories in the app can result in longer interaction time, distract observers from the situation and reduce thoroughness. A further limitation in terms of generalisability is the predefined categorisation used in our study. Other contexts may require a different set of emotions and may be easier or harder to recognise.

We decided upon Pen&Paper as control condition without providing pre-printed tables of emotions because blanc sheets of paper best represent the method currently used in the field, and hence implemented our striving for high ecological validity (Barrett et al., 2019). Alternative operationalisations in future work may contrast Proxemo with the Valence Method (Burmester et al., 2010) or the Facial Action Coding System (Ekman & Rosenberg, 2005). Replications in virtual reality may be an option as well that come with a higher effort in time and required technology for producing and presenting stimulus material but potentially offer a higher level of control. For instance, full body avatars could take the place of our four monitors.

³<https://github.com/bja-engineering/Proxemo>

Finally, we followed the definition of UX as [observer] experiences that result from actual use of Proxemo as defined in the revised ISO 9241-11 (ISO, 2018). A broader definition of UX extends the experience from the situation to include anticipation of the situation or lasting consequences on quality of life (Hassenzahl, 2010; ISO, 2019). During the meta-evaluation observers used two different methods in quick succession. This consecutive use makes it difficult to separate the experience from the products in a timely manner and cannot measure longer-lasting experiential qualities after having used either method. We, therefore, must assume that observers reported their experience during actual use when filling in the questionnaire.

7.5 Post-study to Determine Cognitive Empathy

A critical precondition for observers to document emotions is their ability to recognise another person's emotions in the first place. In chapter 2 we argued that a person's ability to empathise in a specific situation depends on their familiarity with the context and persons as well as their trait cognitive empathy.

To maximise Proxemo's potential regarding validity and effectiveness in controlled studies we would need to select participants with high empathic abilities. However, this would likely produce euphemistic outcomes that impede the generalisability of our findings. In the preceding field studies we relied on observers who were highly trained in their context. We assumed that supervisors in air traffic control who could empathise with their colleagues or care professionals are generally capable of recognising emotions in persons being in their care. However, there is evidence suggesting empathy in caregivers is not necessarily higher than in the general population (Jütten et al., 2019). Since we strive towards benchmarking the Proxemo method for the average observer and have no intentions of limiting the observer group to highly trained care-professionals, our interest is in a representative sample regarding empathic abilities. In this chapter's main study we had recruited student participants naively assuming their empathic abilities being representative of the public. Empathy has been shown to correlate with verbal intelligence (Pfaltz et al., 2013) but we had never tested the cognitive empathetic abilities of our participants.

Consequentially, we now run a post-study to check whether we can uphold the assumption that empathic abilities in our participant pool do not exceptionally vary from the general population. In particular, our interest lay on the ability to attribute the emotional state to another person through observation. In this short survey we measure our participants' ability to recognise emotions in others and discuss whether it is just to assume an average level of empathic abilities in our participants. Due to Murphy and Lilienfeld's (2019) critical review on self-report scales, we chose to utilise the Eyes Test (Baron-Cohen et al., 2001) which instead captures participants' performance.

7.5.1 Method

The Eyes Test (Baron-Cohen et al., 2001) consists of a set of 36 photos of human faces which are clipped to include only the area around the eyes. The dominating element in the pictures are the eyes — hence the test’s name. Participants only have the eye area as a cue to judge the pictured persons’ emotion and select the correct answer among four suggested items (one correct, four foils, see figure 7.9). In our study, participants did not get immediate feedback but received their overall score “ $X/36$ ” after completing the online study. We randomised both the order of items and answers between participants. The resulting scores will give us a descriptive impression of whether empathic abilities from students in our participant pool are comparable to published benchmarks. Additionally, we collected demographic data on participants’ age, gender and semester of their degree course.

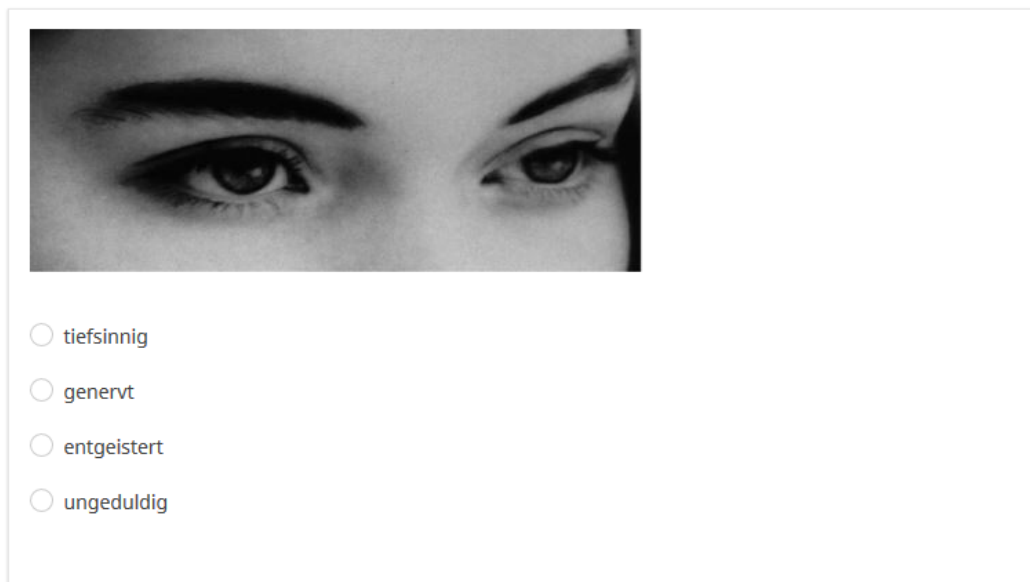


Figure 7.9: Screenshot of our online survey, displaying item #29-reflective, [foils: irritated, aghast, impatient] of the Eyes Test. Translations are taken from Pfaltz et al. (2013).

For this purpose we invited all participants of the effectiveness study to complete the German version (Pfaltz et al., 2013) of the Eyes Test. This post-study took part in the last week of September 2020 that is 14 – 15 months after their participation in the effectiveness study (from mid-June until mid-July 2019). We conducted the Eyes Test in form of an online-study that was locally hosted on the university’s LimeSurvey server (LimeSurvey GmbH, Hamburg, Germany) and participation was compensated with course credit.

Due to functioning anonymisation of data in the main study and the unscheduled nature of this subsequent check of trait empathy as a potential confounder no allocation of the participants

between the studies was possible – thus we could not calculate a moderation analysis but were restricted to mostly descriptive examinations. The majority of the former participants we addressed had completed their quota of study participation or graduated already. A systematic error in the participation pool software offered the declined study slots to students who had not participated in our effectiveness study and whom we had not intended to invite for this reason. Before we noticed this error and could close the study, $n = 14$ invited and $n = 24$ uninvited participants had already completed the study. Not wanting to waste collected data, this bug allows us an exploratory comparison between the groups of former participants of the effectiveness study with other students from the participant pool. All statistical analysis tests have been run in JASP Team (2020).

7.5.2 Results

Of the 38 participants who completed our study we identified one as an outlier whose Eye Test score lay 3.06 SD below the mean. Since the participant's score was two points below 13 and thus indistinguishable from chance (Baron-Cohen et al., 2001) we excluded their data. The remaining $N = 37$ participants were descriptively slightly older $M = 22.378$ ($SD = 2.802$, $Mdn = 22$, $IQR = 21 - 23$) than the $N = 51$ participants in the effectiveness study described above ($M = 22.0$, $SD = 2.49$, $Mdn = 21$, $IQR = 20 - 22$). Visually examining the data in figure 7.10, the age distribution in both studies looks similar.

Considering that the Eyes Test was conducted 14 – 15 months after the effectiveness study, older participants are not surprising. Participants who had previously taken part in the effectiveness study were aged by over one year (now $Mdn = 22$, $IQR = 21 - 22.75$) and now in a higher semester ($Mdn = 5.5$, $IQR = 5 - 6$). Participants who had not taken part in the effectiveness study were two semesters below them ($Mdn = 4$, $IQR = 3 - 5$) and slightly younger ($Mdn = 21.5$, $IQR = 20 - 23.5$) which is partly attributable to the Eyes Test taking part three months later in the academic year, explaining an age difference of .25 years alone. According to a Mann-Whitney test, the age difference between subgroups is not statistically significant, $U = 186$, $p = .34$.

The sample consisted of 3 male and 34 female participants who pursued their bachelor degree course in either human-computer interaction (2) or media communication (35). The skewed distribution in gender can be attributed to the high popularity of the larger degree course on media communication among female students generally leading to female predominance in samples (e.g., Breves & Schramm, 2021; Brill & Schwab, 2020). Yet, the biased distribution of gender across degree courses was not statistically significant ($p = .158$, two-tailed Fisher's exact test). Neither the distribution of genders ($\chi^2(1) = 2.252$, $p = .133$, $V = .16$) nor the distribution of attended degree courses ($p = 1.0$, two-tailed Fisher's exact test) was dependent on the study. For detailed frequency data of gender and attended degree courses from both studies see table

7.2.

Table 7.2: Frequencies for degree course and gender in both studies. The degree courses are abbreviated in the table as MCS (Human-Computer-Systems) and MK (Media Communication).

Study	Gender	Degree course		Total
		MCS	MK	
Eyes Test	male	1	2	3
	female	1	33	34
	total	2	35	37
Effectiveness study	male	2	8	10
	female	1	40	41
	total	3	48	51
Total	male	3	10	13
	female	2	73	75
	Total	5	83	88

7.5.3 Eyes Test Score

The Eyes Test score of all $n = 37$ participants averaged on $M = 24.95$ ($SD = 3.94$). There was no correlation between scores in the Eyes Test and age ($r = -.09, p = .59$) or semester ($r = -.05, p = .75$).

The Eyes Test score of participants who had taken part in the effectiveness study ($Mdn = 25, IQR = 22.25 - 28.75, n = 13$) was descriptively comparable to other participants from the pool who had not taken part in the effectiveness study ($Mdn = 25, IQR = 22 - 27.25, n = 24$).

7.5.4 Item Difficulty

Table 7.3 lists all items and the portion of participants who correctly chose the target emotion. On six items (1, 10, 13, 23, 27 and 29) less than 50% of participants selected the target word. On nine cases, more than 25% participants selected a single foil (1, 2, 10, 13, 17, 23, 25, 27 and 29). A foil of item 29 was selected more often than the target. All difficult items are summarised in Table 7.4.

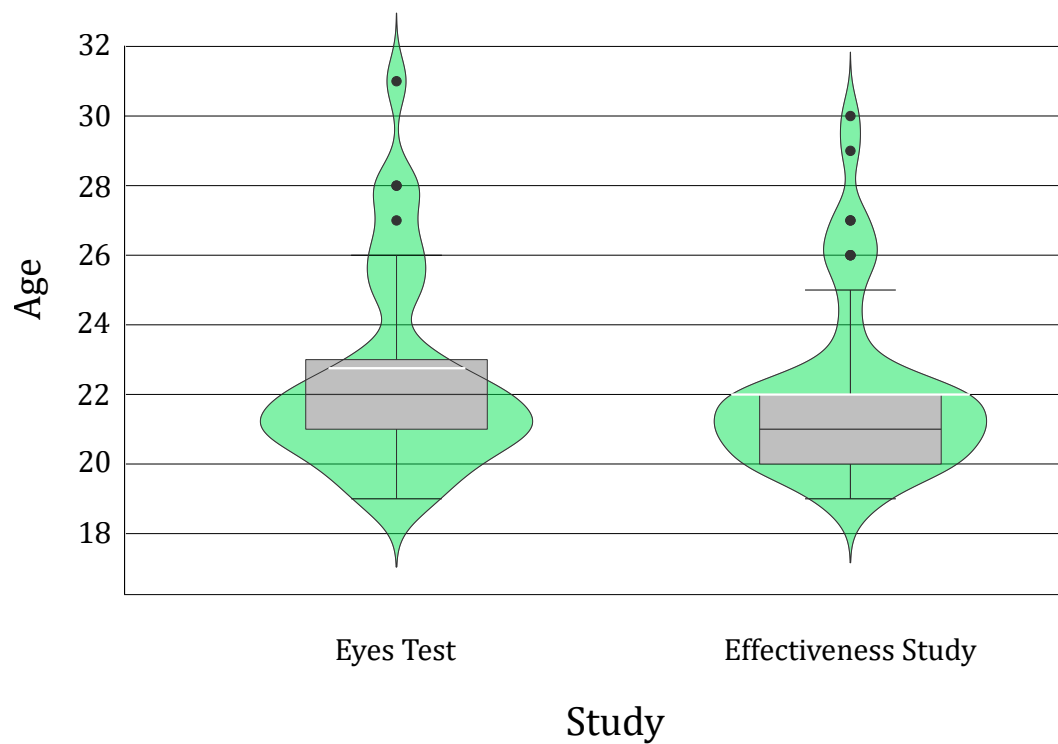


Figure 7.10: These violin graphs display the age distribution in both studies. For each study, the boxplot's centre line represents the median, the boxes cover the interquartile range and dots are outliers. The green violin element represents a smoothed distribution plot of the data with the mean indicated as a white horizontal line. Graph is based on export from JASP Team (2020).

Table 7.3: List of German items of the Eyes Test (Pfaltz et al., 2013) together with English originals from Baron-Cohen et al. The numbers depict the percentage of participants who chose the target item in our online study and in comparison to (Pfaltz et al., 2013) and Baron-Cohen et al. Item “#0 - panisch [DE] - panicked [EN]” served as a test item to explain the study and does not count towards the score.

#	German	Our study	Pfaltz et al.	English	Baron-Cohen et al.
0	panisch	86.8	–	panicked	–
1	lustig	36.8 ^a	65.8	playful	70.9
2	bestürzt	60.5	49.4 ^b	upset	85.4
3	begehrnd	84.2	85.1	desire	83.5
4	darauf bestehend	89.5	74.2	insisting	87.4
5	besorgt	71.1	64.5	worried	82.5
6	tagträumend	57.9	72.9	fantasising	77.7
7	unruhig	57.9	49.0 ^b	uneasy	78.6
8	verzweifelt	89.5	77.4	despondent	83.5
9	geistesabwesend	86.8	78.6	preoccupied	91.3
10	vorsichtig	47.4 ^a	76.0	cautious	63.1
11	bedauernd	63.2	74.3	regretful	80.6
12	skeptisch	78.9	87.7	sceptical	83.5
13	vorausahnend	47.4 ^a	55.8	anticipating	76.7
14	beschuldigend	78.9	73.4	accusing	94.2
15	besinnlich	78.9	84.5	contemplative	83.5
16	nachdenklich	78.9	76	thoughtful	82.5
17	bezweifelnd	60.5	50.3	doubtful	60.2
18	entschieden	84.2	81.9	decisive	79.6
19	zögerlich	60.5	57.4	tentative	58.3
20	freundlich	86.8	81.3	friendly	87.4
21	tagträumend	55.3	39.4 ^b	fantasising	81.6
22	geistesabwesend	81.6	72.9	preoccupied	91.3
23	aufsässig	44.7 ^a	61.7	defiant	84.5
24	nachsinnend	78.9	57.4	pensive	77.7
25	interessiert	50.0	42.6 ^b	interested	57.3
26	feindselig	63.2	78.1	hostile	81.6
27	vorsichtig	47.4 ^a	67.1	cautious	63.1
28	interessiert	76.3	63.9	interested	65
29	tiefsinnig	31.6 ^a	69.0	reflective	64.1

Continue on the next page

Table 7.3 — continued from previous page

#	German	Our study	Pfaltz et al.	English	Baron-Cohen et al.
30	kokett	65.8	86.5	flirtatious	89.3
31	zuversichtlich	50.0	32.3 ^b	confident	52.4
32	ernst	92.1	66.5	serious	72.8
33	beunruhigt	76.3	77.4	concerned	74.8
34	misstrauisch	78.9	71.0	distrustful	81.6
35	nervös	73.7	60.6	nervous	82.5
36	argwöhnisch	92.1	85.0	suspicious	87.4

^a Very difficult item - selected by less than 50% in our study

^b Very difficult item - selected by less than 50% in Pfaltz et al.

Table 7.4: List of items with high difficulty according to the percentage with which the target and the most prevalent foil were selected. The German translation of the Eyes Test is taken from Pfaltz et al. (2013).

#	Target			Critical Foil		
	German	English	%	German	English	%
1	lustig	playful	36.8	beruhigend	comforting	31.6
2	bestürzt	upset	60.5	verängstigt	terrified	31.6
10	vorsichtig	cautious	47.4	darauf bestehend	insisting	26.3
13	vorausahnend	anticipating	47.4	schüchtern	shy	28.9
17	bezweifelnd	doubtful	60.5	zärtlich	affectionate	26.3
23	aufsässig	defiant	44.7	neugierig	curious	39.5
25	interessiert	interested	50.0	ungläubig	incredulous	36.8
27	vorsichtig	cautious	47.4	arrogant	arrogant	39.5
29	tiefsinnig	reflective	31.6	genervt	irritated	47.4

7.5.5 Discussion of Results for the Eyes Test

We conducted an online study with the German version of the Eyes Test (Baron-Cohen et al., 2001; Pfaltz et al., 2013) in order to receive an estimate of cognitive empathy in the participant pool of the institute of Human-Computer-Media, University of Würzburg. In the following we will compare the resulting scores to those reported in the literature.

The study by Pfaltz et al. from which we retrieved the German Eyes Test reports a test with $n = 155$ Swiss participants who scored on average $Mdn = 25.0$ — exactly the same as participants in our online study. This supports the representativity of our participant pool in general regarding their empathy. Researchers using the original English version generally report higher scores. Baron-Cohen et al. (2001) reports scores for $n = 122$ general population controls

($M = 26.2$, $SD = 3.6$) and $n = 103$ undergrad students from Cambridge ($M = 28$, $SD = 3.5$). Lawrence et al. (2004) reports that $n = 53$ volunteers recruited from health clinical staff and general population scored $M = 27.6$ ($SD = 4$). One reason for this difference between the English and German versions may be that the German translation is not perfect. In fact, based on the questionable psychometric properties Pfaltz et al. (2013) discovered in their study, they recommend to only use a reduced list of 24 items. As marked in table 7.3, the items found to be problematic in their study are not identical to the items that turned out to be difficult in ours. A translation that is closer to the English original version exists (Bölte, 2005) but had not been validated at the time of our data collections. Recently, Kynast et al. (2021) published age- and sex-specific standard scores. For the age norm group 20-29, in which most of our participants were, Kynast et al. report scores of $M = 26$ ($SD = 3.2$), which are higher than the German Pfaltz translation but still slightly lower than the English original. It is possible that we underestimated the true abilities of our participants due to our lack of control over the setup and conditions in which they completed the Eyes Test.

The gender distribution in our study is certainly not representative of the general population in Germany. Unfortunately, we cannot say whether the gender distribution, age, or advancement in degree course are representative of the participant pool because none of the data required for self-registration (first name, last name, user ID, email address and phone number) allows to derive the age and an explicit full census would be inappropriate. Other studies involving the Eyes Test with psychology students had a similar gender distribution (Fernández-Abascal et al., 2013).

Considering that the large samples reported by Baron-Cohen et al. (2001) and Pfaltz et al. (2013) did not suffice to detect a gender difference in the Eyes Test score, the skewed gender distribution in our study is not a grave problem.

We found no descriptive difference between students who had previously participated in the effectiveness study and those who had not. Regarding age or degree course advancement there were no correlations with the Eye Test score. This adds further support to the argument of cognitive empathy being a personality trait which is stable over time. Our finding on independency of participants' age is in line with literature that reported few differences between age groups (Richter & Kunzmann, 2011) and satisfactory test-retest reliability (Fernández-Abascal et al., 2013).

Post hoc we examine those items that are specifically relevant for the set of emotions expected and pre-selected when using Proxemo in the context of reminiscence. *Interest* is easiest to match because it appeared even twice in the Eyes Test. Item #28-interest had acceptable agreement in our study while item #25-interest appeared to be a difficult item in our study as well as in Baron-Cohen et al. (2001) and Pfaltz et al. (2013). Other emotion categories used in Proxemo in the context of reminiscence do not directly occur in the Eyes Test but aspects of them are represented. We proceed with the *generic negative emotion* which combines the emotions anger,

anxiety and sadness. *Anxiety* is best described by the items #5-worried, #33-concerned, #35-nervous and #0-panicked which all achieved a participant agreement of 70% or greater. The only item approximating *sadness* is #8-despondent with an agreement of almost 90%. *Anger* is a possible aspect of being upset. However, the picture of item #2-upset does not show an angry but rather a worried expression. *Pride* and *self-efficacy* have no appropriate matches in the Eyes Test either. The remaining two categories, pleasure and wistfulness, are affected by suboptimal translations in the German version of the Eyes Test we used. Our description of *pleasure* (table 4.1) contains laughing and smiling. In the German translation by Pfaltz et al. (2013) item #1-playful is translated with the German word “lustig” which is most commonly associated with being funny or hilarious (Hemetsberger, 2021). Bölte (2005) uses “verspielt” as a more verbatim translation of playful. While playfulness may lead to pleasure, the facial expression does not exactly appear hilarious which made the item difficult for our participants who frequently misinterpreted it as a comforting expression (table 7.4). The last emotional category that found no direct match in the Eyes Test is *wistfulness*. Wistfulness is made up of good memories of an event from the past together with the awareness that this time has passed. Admittedly, the interpretation of this complex emotional category requires more context than a facial expression and usually was documented based on people with dementia’s utterances. Since one’s own reminiscence of a beautiful event and the awareness of its lying in the past require actively thinking about the past, the items #24-pensive and #16-thoughtful may serve as appropriate proxies for one aspect of wistfulness. Both items had achieved a satisfactory agreement in our study but fail to capture the nostalgic joy that emerges from retrieved memories of the past. We can summarise that participants agreed satisfactorily on items that have a good match to emotional categories used in Proxemo. When it is important to determine not only the general level of empathy in participants but measure their ability to recognise specific emotions, it might prove advantageous in future work to compliment the set of facial expressions provided in the Eyes Test (Baron-Cohen et al., 2001) with the subscales anger, fear, happiness and sadness from a similar test suggested by Allen-Walker and Beaton (2015).

In this post-study, our goal was to estimate whether the empathy of students in our participant pool deviates from values reported in literature. Our descriptive findings match reported medians of other German samples and lie below scores reported for the original English version. While our sample had a gender bias, this is also true for our effectiveness study and likely represents the participant pool. We conclude that the assumption of a representative level of empathy in our participant pool can be upheld. For future studies on quality criteria of Proxemo or other evaluation methods based on observation we recommend measuring each observer’s empathic abilities in advance.

7.6 Preliminary Conclusion on Effectiveness

In this study we contributed a meta-evaluation of the method Proxemo in contrast to handwritten notes as the current documentation standard. Our focus lay on main quality criteria as defined by Hartson et al. (2001) because they appeared to be the most relevant and appropriate for the context of reminiscence sessions. We extended the catalogue by measuring efficiency and observer experience of the evaluation methods. Proxemo outperformed handwritten notes in the control condition on all criteria except validity. Compared to scores from usability observations (Stanton, 2016) the validity of Proxemo is already on eye-level and has the potential to gain validity through the simple implementation of an undo function and deployment of trained observers. Our results showed that Proxemo is a thorough and effective method appropriate for collecting observational data in group sessions.

As listed in chapter 6 other criteria exist and can be relevant to practitioners (e.g., learnability, downstream utility (Law, 2006)) or for the application in safety critical domains (e.g., intrusiveness (Eggemeier et al., 1991)). The final study will complement our knowledge about Proxemo's quality criteria with measurements of intrusiveness and an alternative measurement of validity. Future research should additionally address the surprising variance we discovered in the effective detection of emotion categories between conditions.

Chapter 8

Effectiveness and Intrusiveness

The main motivation behind our optimisation of Proxemo for observational studies in air traffic control was the assumed distraction self-report methods would cause. Therefore, in this final study we let participants play computer games and examine the spared intrusiveness through the use of Proxemo in direct comparison to self-documentation. Additionally, we re-evaluate validity, thoroughness and effectiveness. This time we use annotated emotions from the users themselves as outside criterion or “ground-truth”. A critical precondition for the validity of this approach to ground-truth is that users are capable of recognising their own emotions.

Self-report methods are not uncommon in air traffic control. Sanderson et al. (2007) for example regularly asked controllers to state their currently perceived workload. SASHA-L, a real-time assessment technique for situation awareness requires controllers to read a statement on a specific situation and judge its content as true or false (Jeannot et al., 2003). Hagemann et al. (2020) use a similar approach by prompting single item questions after predefined intervals on a touch input device to query situation awareness and workload. Such a regularity is not comparable to Proxemo where in a self-report scenario, controllers should not only reveal their experience when prompted but either constantly share their state of mind in a think aloud method (van Someren et al., 1994) or communicate their emotions together with triggers whenever they are aware of any. Since controllers are trusted with using touch-input-devices next to their task and participants in former studies of this work emphasised the low effort Proxemo caused, we considered self-documentation with the Proxemo App a fair comparison to the observation by proxy in terms of intrusiveness. Due to the scarcity of the user group of air traffic controllers we conceptually replicated aspects of their task for an experimental study with student participants.

8.1 Intrusiveness

Following the philosophy of popular user research approaches (e.g., Holtzblatt & Beyer, 2016), observations of user behaviour and emotions are best validated by the users themselves. Arguable is, however, whether users are able to recognise emotions in themselves without prompt. Assuming that users are capable, we suspect communicating these emotions concurrently while experiencing them has a measurable impact on the user experience itself — in particular on workload, performance and affective aspects.

8.1.1 Intrusive Effect on Workload

Asking users to self-report their emotions as they play a game poses a secondary task. Following the argumentation by Kahneman (1973), even when pursuing two distinct tasks some interference will arise. The factors that cause said interference were refined by Wickens (2008) in his multiple resource model over decades. The model structures stages and codes of processing in four dimensions with the idea that tasks can be better adhered to in parallel when they consist of different levels of the dimensions. For example, one can more easily sing a song (vocal/verbal resources) while riding a bike (manual/spatial resources) than sing a song while reading a book (both vocal/verbal resources).

In ATC, the context of our focus and a paragon for multitasking, operatives follow already both, spatial activities (i.e., localising and controlling aircraft on a radar screen) and verbal activities (i.e., communicating clearances to pilots and written documentation of those clearances). Therefore, there are no generally free capacities a self-report task could fill without impact on the performance of either task. Even the simplistic one-click documentation of own emotions that result from the use of the Proximo App as a self-report tool require processing perceptual cognitive activity and manual responses which are required for the primary task already. As a consequence, the self-report of emotions competes with the primary task leading to increased workload.

Matthews et al. (2015) compared different workload metrics including eye-tracking, subjective self-report (NASA TLX) and EEG. They found measures to be sensitive but not equal. To cover this divergence, we use several instruments to measure workload, including subjective self-report (NASA TLX, Byers, 1989), pupillometry and skin conductance.

8.1.2 Intrusive Effect on Performance

The quality criterion describing the amount to which an assessment method deteriorates the primary task performance is called *intrusiveness* (O'Donnel & Eggemeier, 1986). In human factors, intrusiveness has been studied in detail by researchers who raised workload measures and needed to know whether their deployed methods increase said workload, e.g., Eggemeier et

al. (1991). In fact, a considerable amount of the research on intrusiveness of measurements has been done in the safety critical domain of air traffic. Resulting disadvantages of intrusive methods include “possible compromises in system safety associated with primary-task intrusion, the lack of operator acceptance of tasks that are perceived to be artificial or bothersome and associated problems with the failure of the operator to perform such tasks” (Eggemeier et al., 1991, p.233). When measuring workload, researchers can avoid intrusion by choosing less intrusive techniques such as logging the performance in the primary task, physiological measurements or subjective self-assessments that are handed out after task completion (O’Donnel & Eggemeier, 1986). For the intrusiveness of secondary tasks, there are mixed outcomes reported in literature concluding that intrusion is not a huge problem. For instance, in a series of studies, Wierwille and colleagues examined 16-20 mental workload measures including two secondary task measures with pilots on simulated flights. They found a *time estimation task* to intrude cognitive task performance (Wierwille et al., 1985) but no psychomotor tasks (Wierwille & Connor, 1983) or perceptual tasks (Casali & Wierwille, 1984). In all studies, performance in the secondary tasks deteriorated with increasing workload. This makes perfect sense because a performance loss in the secondary task was meant to indicate a sensitivity to increased workload and the outcome corroborates the tasks’ suitability to do exactly that. In a similar study, Casali and Wierwille (1983) even observed how subjects entirely disregarded the secondary task when the load in the primary task was high. For self-report UX methods this would mean omitting emotional events and thus lead directly to a loss in documentation quality. The literature cited above indicates that self-report of emotions as a reflective (cognitive) secondary task competes with the primary task and participants cannot succeed at both. Hence, we assume that participants individually set their focus on either the primary or the secondary task and on average perform worse at both compared to participants who have no secondary task.

8.1.3 Intrusive Effect on Experience

Experiences during gaming are rich and reducing them to sheer performance is not fair. IJsselsteijn et al. (2007) argue that gaming experience is multi-dimensional and its approximation requires a set of scales including *flow* and *immersion*. In their later developed Game Experience Questionnaire (GEQ), they add the dimensions *competence*, *challenge*, *tension* as well as *positive* and *negative affect* (IJsselsteijn et al., 2013). Interruptions of the game can diminish the game experience even when they occur only between levels of the game (Santos et al., 2019). Thoroughly self-reporting emotions throughout the game causes an interruption for each emotional event which we would expect to cause a drastic reduction of game experience in all dimensions. We quantitatively assess gaming experience with the GEQ (IJsselsteijn et al., 2013) and qualitatively note individual statements. Law et al. (2018) point out the questionable psychometric properties of the GEQ. However, we decided to still use it as we are not aware of another tool

with comparable multi-dimensionality and translations.

Flow is a special case due to its clear distinction from simultaneous reflection of the experience. Csikszentmihalyi and Larson (2014) defined flow as a period of clear focus and dedication to the activity at hand — without awareness of the flow condition itself. To describe this condition, Csikszentmihalyi and Larson borrow the term “loss of self-consciousness” from Maslow (1971, p.63). Regaining the awareness for one’s experiences constitutes an exit condition for the flow. Deploying a self-report method for user experience therefore counteracts the emergence of a flow experience because in order to maintain concentration, “potentially intruding stimuli must be kept out of attention” (Csikszentmihalyi & Larson, 2014, p. 139). Each reflection about the current experience forces users into an outside perspective, draws attention away from their current task and thus breaks the flow. We, therefore, expect flow to be decisively lower in the self-report condition.

8.2 Effectiveness

To determine the “ground truth” for computations of thoroughness, validity and effectiveness (Hartson et al., 2001), in the last study we relied on consolidated expert ratings as a criterion. In this study, we operationalise criterion validity differently and compare documented emotions from both, the participant and the observer, with emotions communicated directly by the participant.

Peute et al. (2015) found effectiveness of concurrent think aloud to be higher (.8) than retrospective think aloud (.62) when seeking usability problems. However, retrospectively thinking aloud while reviewing videos of the prior interaction did produce unique insights that were not found through concurrent think aloud (and vice versa). In situations such as air traffic control (or gaming) where concurrent think aloud is not feasible for reasons of performance (and thus potentially safety), we consider retrospective think aloud as an acceptable alternative. We operationalise the ground truth in this study as the emotions communicated by participants when reviewing their experience on a video recording (“cued recall” Bruun et al., n.d.).

Following the argumentation on intrusiveness and task switching we assume that participants will be less thorough in documenting their own emotions next to gaming than a focussed observer. Results from the previous study on effectiveness in chapter 7 indicated that emotions which are costly to document (i.e., with Pen&Paper opposed to Proxemo) were more valid. In the study described here, emotions are costly to document through self-report as well in a sense that their documentation requires an attention shift to the secondary task. For the observer, documenting emotions is the only task they have, thus requiring no extra effort. We therefore assume that participants will only document emotions when they are really sure about them and hence achieve higher validity than the observer.

As introduced in chapter 2, emotional reactions may activate the sympathetic nervous system which again stimulates measurable changes in organs such as the heart, skin, muscles or eyes.

Initially, we intended to gather a mixture of physiological data that offers a further criterion of external validation for emotional situations. Namely, we sought to approximate emotions by measuring indicators for arousal and valence. As for valence, facial EMG has shown up as indicator for positive (*musculus zygomaticus major*) and negative (*musculus corrugator superciliaris*) experiences (Golland et al., 2018). EDA is most commonly known as an indicator for arousal (Caruelle et al., 2019). It must be noted, however, that recent feature extraction approaches suggest it might even offer hints for valence (Jainendra et al., 2019). In EDA data, there is an expected latency of 1 – 4 seconds between the stimulus/trigger and the resulting phasic change in skin conductance response that indicates emotion (Caruelle et al., 2019). This latency lies well within the ± 5 seconds interval we granted for emotions documented with Proxemo across studies and thus renders phasic changes in EDA a worthwhile external criterion. Heart rate variability on the other hand is typically analysed over five minutes or more and even ultra-short term heart rate variability which is most promising for emotion classification requires intervals of 15 – 30 seconds (Schaaff & Adam, 2013). Finally, there is pupil size which is commonly known to change primarily due to variations in luminance. However, the arousal level does affect fluctuations in pupil diameter and was found to correlate with the aforementioned physiological measurements of skin conductance and heart rate (C.-A. Wang et al., 2018).

Consolidating physiological parameters would have allowed us to at least compare peaks in physiological data as “emotional reactions” to documented instances in Proxemo data. In that way, physiological parameters would have provided an additional indicator for participants’ emotions even though we are aware and also have argued before (see chapter 2) that the two-dimensional valence&arousal model does not allow for a distinction between emotion categories as detailed as those definable in Proxemo. Unfortunately, we learned in piloting and from self-experience that we cannot expect the ability of participants to appropriately place facial EMG electrodes on first trial — and assisting them would have required physical contact which opposed our hygienic concept for limitation of infection risk in midst of the COVID-19 pandemic. Losing our sole indicator for valence to pandemic precautions leaves us with a combination of indicators for peaks in arousal and the opportunity to examine valence in future work. Heart rate does represent arousal, but its adjustment needs too long for our intended identification of densely succeeding emotional triggers which is why we dropped it for this study. Pupil diameter is mainly affected by light, and we have no feasible way to subtract the illuminating influence of the stimulus screen. This means we could not distinguish whether phasic changes of pupillometry are due to dynamic luminance of the screen or a result of variance in workload. Therefore, we restricted ourselves to explore the overlap between EDA peaks and human labelled emotions.

The aim of this final study was to examine the validity and intrusiveness of Proxemo. Additionally, we strived to show the independence of the Proxemo method from the Proxemo App’s form factor and emotion set through a replication of findings from former studies. From the argumentations given above we derived the following hypotheses. Hypotheses addressing a

between-subjects comparison relate to participants whose sole task is to play a game (*gaming-only* condition) and participants who additionally are asked to concurrently self-report their emotions (*self-documentation* condition).

Effectiveness

H1.1 Compared to a ground truth defined through retrospective self-reports, observers using Proxemo achieve higher thoroughness than users do through concurrent self-reports.

H1.2 Compared to a ground truth defined through retrospective self-reports, users achieve higher validity through concurrent self-report than observers with Proxemo.

H1.3 As a result of H1.1 and despite H1.2, observers achieve higher effectiveness with Proxemo than users with concurrent self-report.

Intrusiveness

H2.1 Values of the RAW TLX are lower in the gaming-only condition than in the self-documentation condition.

H2.2 Pupil dilation is higher in the self-documentation condition than in the gaming-only condition.

H2.3 Skin conductance is higher in the self-documentation condition than in the gaming-only condition.

H2.4 Participants in the gaming-only condition score higher in the games than participants in the self-documentation condition, i.e. higher ranks, shorter time to finish and better streaks.

H2.5 Participants in the gaming-only condition report higher values in the GEQ than participants in the self-documentation condition.

H2.6 Participants in the self-documentation condition are aware that using Proxemo during the games affects their game experience.

Relation to physiological data

H3 Compared to a ground truth defined through retrospective self-reports, observers using Proxemo achieve higher effectiveness than skin conductance peaks.

8.3 Method

We set up a lab study to determine the intrusiveness of self-documented emotions and whether observed emotions can be documented effectively. In short, we conceptually replicated fundamental aspects of air traffic controllers' workstation and task including a desktop environment and different input methods. The tasks comprised a constant demand for attention and readiness for interaction, required the understanding of (game-)physical constraints and led to success when a safe, structured and efficient approach was followed. We deployed a variety of measures to capture the influence the additional task of in-situ self-documentation had on the participants' main task. Finally, we compared the emotional instances documented by an observer during the game with those reported by participants in a retrospective interview.

An even more realistic replication of the air traffic control domain such as cooperation between actors and more complex interactions had to be omitted due to the ongoing pandemic during data collection. The institute's hygienic concept and common sense required to limit the infection risk to a minimum thus disallowing two participants working closely to each other in the same room or disproportionately extending their presence by extensive training for a complex environment.

8.3.1 Setup

During the experiment, participants interacted in a classical desktop setup with mouse, keyboard and a 24" monitor in front of them. They wore a Pupil Core eye-tracker (Pupil Labs GmbH, Berlin, Germany) and electrodes that measured conductivity (EDA) on the palm of their non-dominant hand. Next to the keyboard and mouse we placed an Android phone (Oneplus 5T, OnePlus Technology Co., Ltd., Shenzhen, China) with a 6" screen running OxygenOS [Android 7.1.1] that ran the Proxemo App in the self-documentation condition and had a deactivated screen in the gaming-only (control) condition. As Proxemo App we used the single user mode from the app also deployed in the qualitative study with air traffic controllers – and due to the task similarity between the chosen games and air traffic control we also adopted the set of emotions from chapter 5. Analogue to the dual user mode, upon emotion documentation it displays a confirmatory *snackbar* including an “undo” button to delete the latest timestamp (figure 8.1). The same app was used by the observer. An illustration of the entire setup is provided in figure 8.4.

To minimise the risk of infection with COVID-19, the experimenter and observer were in an *operation room* (figure 8.2) separated from the *experiment room* with the participant during the experiment and the retrospective analysis of video recordings (figure 8.3). An external webcam with integrated microphone (figure 8.4) streamed visual and audio impressions of the participant as well as a screen capture to the operation room via local network. Participants perceived game sounds and instructions via active speakers. The bidirectional stream was facilitated by TeamViewer (TeamViewer AG, Göppingen, Germany) and locally saved with OBS Studio (OBS

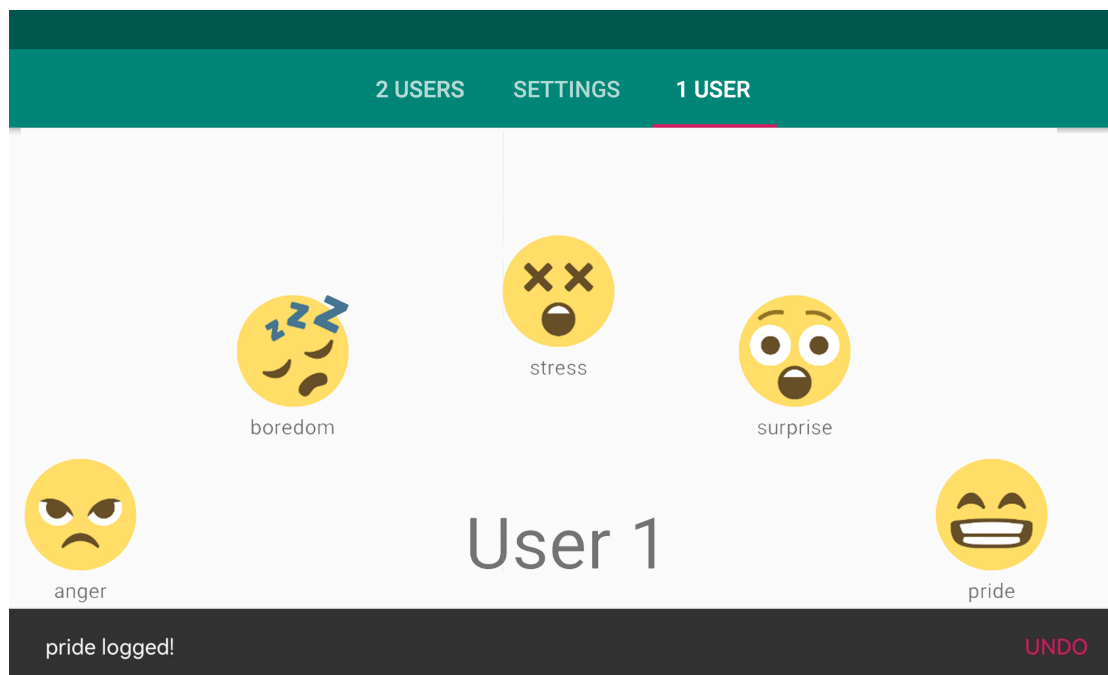


Figure 8.1: The screenshot displays the documentation screen for one user of the Proxemo App running on a 6" Android phone immediately after *pride* was documented. Emoji are arranged to be accessible quickly for one-handed use when the smartphone is lying on a table.

Project, available on <https://obsproject.com>). Participants completed all questionnaires manually on paper.

Physiological data were captured using the Biopac MP150 with an EDA module for electrodermal response and the AcqKnowledge software (Biopac Systems Inc., Goleta, CA). Eye-tracking data was processed with the software Pupil Capture (Pupil Labs GmbH, Berlin, Germany). We used the video annotation software ELAN (version 6.0, Max Planck Institute for Psycholinguistics, The Language Archive, Nimwegen, Netherlands, available on <https://tla.mpi.nl/tools/tla-tools/elan/>) to annotate the emotions participants retrospectively reported during the interviews and subsequently synchronised the emotions with Proxemo data.

Games

Appropriate games for this study with student participants needed to be learnable in the short period of time for novices and still pose a challenge for more experienced gamers. Hence, we chose games that followed the old concept *easy to learn — difficult to master* in that they offered simple controls and a clear goal but enough variability to both, pose a challenge (Malone, 1982) and evoke emotions (Kosiński et al., 2018).

With *Flight Control HD* (Electronic Arts Inc., Redwood City, CA, available on <https://store>.



Figure 8.2: During the game session, the experimenter (right) supervised and communicated with the participant (who was in the experiment room) via TeamViewer and the observer (left) followed the OBS stream showing the shared game screen and a webcam image and annotated observed emotions in Proxemo (green highlight).

steampowered.com/app/62000/Flight_Control_HD/) as our first choice, we selected a game of skill that simulated the air traffic controller’s task on the approach positions in a very simplified manner. On the map of Flight Control HD (appendix A.4), players are responsible for the approach sector of an airport with two runways and one heliport. The player’s responsibility for an aircraft begins as soon as it enters the fixed screen. Aircraft vary in size and speed and are assigned a runway already when entering the screen/sector which is conveyed through their colour. The player’s task is to guide aircraft to their runway while maintaining separation between them. Aircraft are controlled by clicking on the aircraft and holding down the mouse-button while drawing the desired flight path. Vertical separation is not possible and aircraft “descend and land” automatically when they cross the runway threshold. There are no restrictions on the intersection angle of this threshold. The system emits an audio-visual alarm to help the player in avoiding a collision whenever aircraft get too close to each other. There are no departures to be dealt with. Players receive one point per landed aircraft and their current score is displayed next to their high-score on top of the screen. Over time, amount and complexity of traffic increases. The game ends immediately when a collision occurs. This fatal consequence for the game should avoid the effect observed by Truschzinski (2017) with participants benefitting from collisions in terms of a decreased workload. The game is titled *flight control game* in the following.

In *SuperTuxKart* (SuperTuxKart Team, available on <https://supertuxkart.net/Download>),

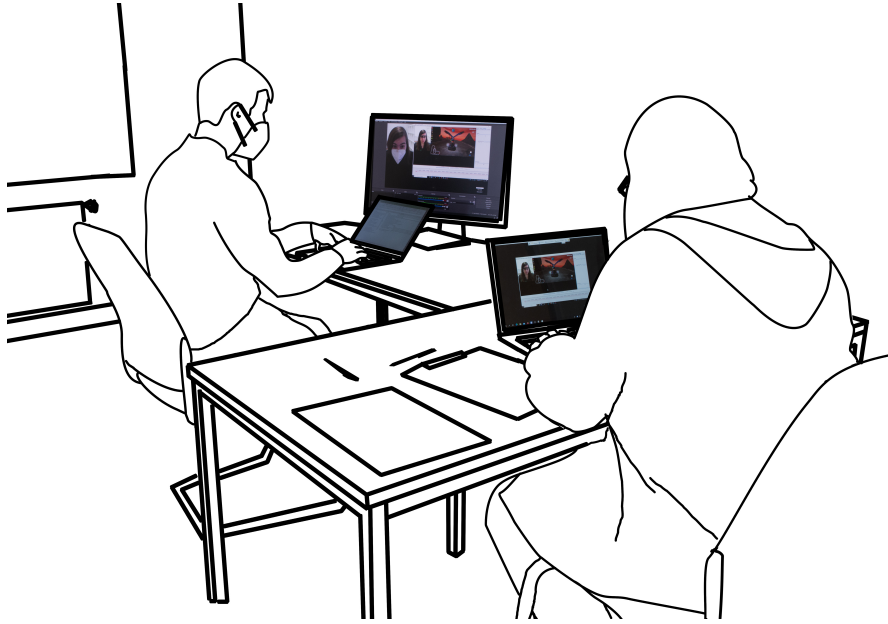


Figure 8.3: During the retrospective analysis, the experimenter (right) communicated with the participant (who was in the experiment room) via TeamViewer to jointly annotate emotion categories in the video recordings. Meanwhile, the observer (left) logged participants' detailed descriptions about their emotions including triggers.

an open source racing game, a social component is simulated through competitive behaviour of non-player characters (appendix A.4). The racing-track and number of rounds can be configured as well as the number and ability of non-player characters. We chose a rather difficult map for the gaming session where players could fall off the track and needed to avoid obstacles falling onto the track. As a compromise, we set the difficulty to easy (i.e., low ability of all seven non-player characters) in both training and experimental gaming sessions. In order to win, players needed to complete multiple rounds on the track and be the first to cross the finish line. On their way, they could collect items to gain speed or harm competitors. Players lost time when they got hit by competitors' items and/or fell off the track. Right-handed participants manoeuvred the kart, activated boost and fired items with marked keys on the right side of the keyboard, left-handed participants with marked keys on the left side of the keyboard. Status information on players' current position was provided through a ranking on the left and a mini-map in the bottom-left corner. The current round and the accumulated time on track were displayed in the top-right corner. The game is titled *racing game* in the following.

8.3.2 Experimental Design

The experiment used a between-subjects design with the conditions *self-documentation* and *gaming-only*. In both conditions an observer documented participants' emotions with Proxemo

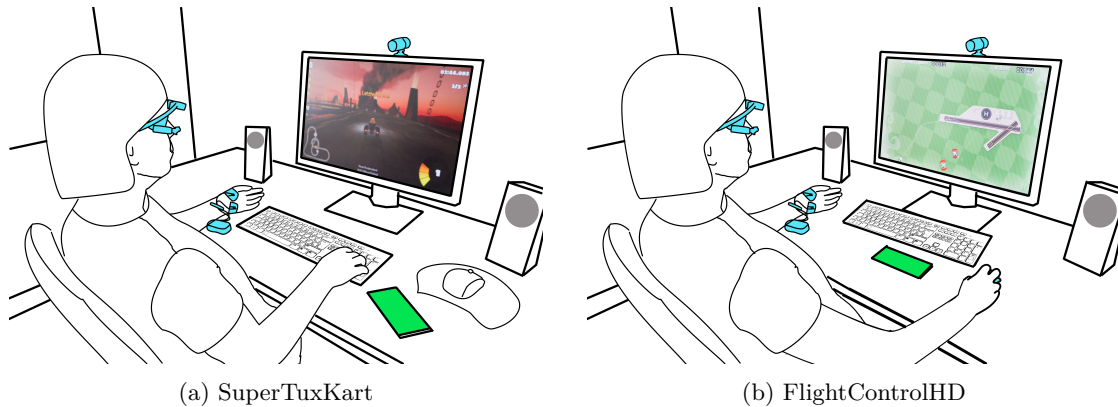


Figure 8.4: Schematic illustrations of the setup in the experiment room during both games. In the SuperTuxKart game (a), participants controlled the kart with the keyboard, in the FlightControlHD game (b), participants controlled aircraft with the mouse. During both games, participants were filmed, wore an eye-tracker and EDA-electrodes on their non-dominant hand (blue highlight). Participants in the self-documentation condition used the smartphone (green highlight) during both games to log their own emotions.

while they played a racing game and a flight control game. In the self-documentation condition, participants were additionally given the Proximo App to document their own emotions in-situ, see figure 8.4. All participants played both video games in the same order. We randomised the order of the conditions self-documentation and gaming-only between participants. The study was approved by the Ethics Committee of the Institute Human-Computer-Media.

Our first group of dependent variables were thoroughness, validity and effectiveness of all documented emotional responses. We calculated those variables according to Sears (1997) and Hartson et al. (2001) as reported in the introduction of chapter 7. This time, we regarded participants' self-reported emotions in the retrospective interviews as "real emotions that exist". For a more robust ground truth we asked participants to not only retrospectively name their emotions but also explain what caused them. These prompts mainly aimed at improving participants' self-reflection, and the resulting qualitative explanations will not be analysed in this work.

As second group of dependent variables we examined intrusiveness which we measured indirectly through its effects on workload, operationalised threefold through physiological parameters, subjective impressions and objective performance. With respect to physiology we measured participants' average pupil diameter and their average skin conductance level. In the racing game we operationalised performance as participants' time needed to finish the race and the position they were on when crossing the finish line. In the flight control game we operationalised performance as the ratio of aircraft landed to aircraft crashed as well as the longest streak of landed aircraft before a crash occurred. Regarding participants' subjective impressions we measured their perceived task load with the RAW TLX questionnaire (Byers, 1989) and their game experience with selected scales of the GEQ (Nacke, 2009). From the GEQ we left out the subscale *immersion*

and item 35 of the subscale *negative affect* due to their focus on the story, explorative character and impressive aesthetics which were not given in either game.

Finally, we explored the relation of Proxemo data and physiological data. We computed the effectiveness of physiological data based on peaks in phasic skin conductance responses. As a prerequisite of participants' ability of experiential self-report, we collected participants' demographic data and measured their perceived ability to recognise their own emotions — an aspect of the controversial (Schuler, 2002) construct of emotional intelligence — using the scales *attention to emotions* and *clarity of emotions* from the German version (Otto et al., 2001) of the Trait Meta-Mood Scale (Salovey et al., 1995). Strong deviation on the perceived emotional intelligence scale served as criterion for post-collection exclusion.

8.3.3 Participants

For an estimation of the sample size we oriented ourselves by the large effects of Proxemo in the previous study and expected Proxemo to have a large effect opposed to self-documentation ($d = .8$). We parametrised G*POWER (version 3.1.9.4, Faul et al., 2007) with $\alpha = .05$ and $1 - \beta = .95$ which resulted in a sample size recommendation of 2×35 participants for one-tailed independent t-tests. Between November 2020 and March 2021 we recruited 70 students from the institute's participant pool who were aged 19 – 28 ($M = 21.57$, $SD = 1.7$). There was no dependency between the distribution of genders across the gaming-only condition (10 males, 25 females) and the self-documentation condition (13 males, 22 females), $\chi^2(1) = .58$, $p = .45$, $V = .01$. All participants signed informed consent and participated in exchange for course credit.

Two participants in the gaming-only condition scored over three standard deviations below the mean of the attention subscale for emotional intelligence. Since this indicates the participants' restricted ability to pay attention to their own emotions, we excluded them from all tests that involve the self-report of experience. One participant did not complete the questionnaires after the racing game, but we include the data of both in all other tests.

Regarding our EDA setup we struggled with issues of connection and adhesion of the electrodes which may be due to temperatures ranging from $18.1 - 27.3^\circ C$ as a result of frequent ventilation and reheating in the winter months and possibly participants' dried out skin after excessive disinfectant usage. During a visual inspection we noticed unrealistic spike patterns in 35 participants and conservatively chose to exclude them. EDA data of $n = 20$ participants from the gaming-only condition and $n = 15$ participants from the self-documentation condition was suitable for further calculation. Additionally, we had to exclude the eye tracking data from nine participants as the format and size of their glasses were incompatible with our head worn eye-tracker, or we found during data analysis that participants' glasses rendered the pupillometry unsuitable for further analysis.

8.3.4 Procedure

After welcoming participants, we told them that the study's aim was to capture emotions occurring in video games. Participants had time to ask questions and then signed informed consent. After this, the experiment contained the following steps:

1. Participants completed the pre-questionnaires for emotional intelligence and demographic data.
2. Participants put on the eye-tracker and aligned the cameras — when necessary with the experimenter's assistance.
3. Participants followed a visual instruction to place the EDA electrodes on the *thenar eminence* and the *hypothenar eminence* of their non-dominant hand and mount the associated radio module on their lower arm (see figure 8.4). While waiting for the electrolyte gel to bind with the sweat gland ducts for optimised hydration and signal quality (Boucsein et al., 2012) participants proceeded with the training session.
4. The experimenter started the video recording and explained the controls and aims of the first video game. The participants then received a short training consisting of 3 rounds on a simple test track (“Gran-Paradiso-Island”) in the racing game. Subsequently, the experimenter explained the controls of the second video game and the participants trained the flight control game for about 2-3 minutes. During the training, the experimenter clarified questions and assisted when it was apparent that participants did not understand an aspect or constraint of either game.
5. Between completion of the training and commencement of the experimental block we started the EDA with a baseline measurement that required the participants to sit still and upright with their feet on the floor and relax for 2 minutes. Participants in the self-documentation condition were additionally familiarised with the list of emotions (table 5.1) and instructed to document their own emotions on the Proxemo App while playing the upcoming games. The observer in the operation room got ready to document observed emotions during both games in both conditions.
6. In the gaming session, participants first played 5 rounds of the racing game (on the “Fort Magma” track) and then at least 6 minutes in the flight control game. When the 6 minutes were up, we let the participants finish the current game (i.e., waited for a plane crash) before interrupting them. After each game, participants completed the RAW TLX and the GEQ.
7. We offered participants a short break and instructed them to remove the EDA electrodes and the eye-tracker.

8. During the retrospective interview, we jointly watched the video recording of the gaming sessions via screen sharing. Participants in the gaming-only condition were familiarised with the list of expected emotions but invited to also talk about experiences that did not match with emotions on the list. The experimenter shared their screen showing the video playback in ELAN and asked the participants to name and explain each emotion they remembered. When participants named and described an emotion, the experimenter entered the label at the appropriate time and trigger in one general tier in ELAN and the observer typed the verbatim description of the experience (label, trigger and detailed emotional response) into a spreadsheet.
9. Before debriefing participants about the real purpose of the study we gave them the opportunity to talk about their experience during the gaming session more generally and posed the following questions:
 - Which game did you enjoy more? Why?
 - Which game triggered more emotions in you? Were these rather positive or negative?
 - Did reviewing the video support your memory of game experiences?
 - (Self-Documentation condition only:) Did you have the impression that using the documentation app distracted you from the game? Was there a difference between games?

Before and after the experimental block the experimenter measured the temperature ($Mdn = 23^{\circ}C$, $Range = 18.1 - 27.3$) and light conditions ($Mdn = 402\ lx$, $Range = 363 - 509$) in the experiment room.

A short note on pandemic-specific precautions. We followed the institute's hygienic protocol with respect to frequent ventilation of both rooms, wearing face masks whenever possible and offered participants to wash and sanitise their hands prior to the experiment after they removed the electrodes and as they left. Measures for safe experimentation during the pandemic, including the schedule for disinfection and ventilation or additional information and documentation requirements, have not been reported in detail as they were constant across conditions, and we consider them negligible for a replication of the study. Participants continuously wore a face mask except during the gaming session.

8.3.5 Data Processing and Analysis

Similar to the prior study, we synchronised Proxemo data with the videos in ELAN where the retrospective annotations had been added already. Due to the different character of the two video games we analysed them separately. We preprocessed the raw data with RStudio (RStudio Inc.,

Boston, MA) and considered Proxemo timestamps as matches when they occurred within an interval of 5 seconds with respective instances in the ground truth. Based on these matches we calculated thoroughness, validity and effectiveness for each participant. Values missing a timely counterpart were not excluded but conservatively considered as mismatches. That means values existing in the retrospective annotations but missing in the Proxemo data decreased thoroughness, values existing in the retrospective annotations but missing in the Proxemo data decreased validity. All statistical tests were computed with JASP (version 0.16, JASP Team, University of Amsterdam, NL). Graphs are based on export from Microsoft Excel if not stated otherwise. Following the recommendations by Caruelle et al. (2019), we report preprocessing, detection and quantification of EDA as well as computation of EDA metrics. As recommended by Boucsein et al. (2012) and Biopac Systems (2015), we smoothed the signal with a 1 Hz FIR low pass filter and then visually detected and interpolated artefacts. Finally, we applied a 0.05Hz high pass filter to separate the phasic signal from the tonic signal. We treated the retrospectively annotated ground truth as stimulus events with a maximum of 10 seconds separation to phasic peaks in AcqKnowledge. To account for interindividual differences in the tonic signal we subtracted the tonic level measured during the games from the baseline measured during the rest period between the training and the trials. Tests were run exclusively on this delta.

8.4 Results

All tests were run against a Bonferroni corrected α level of .005. We report non-parametric data when assumptions for normality or equality of variances could not be met.

Ratings for emotional intelligence averaged $M = 4.17$ ($SD = .51$) on the attention subscale and $M = 3.58$ ($SD = .57$) on the clarity subscale. Equivalence tests following the two-sided tests procedure (TOST, Lakens et al., 2018) with the smallest effect size of interest set to $d = .5$ revealed that emotional intelligence differed slightly between conditions. On the clarity scale the difference between the gaming-only condition ($M = 3.61$, $SD = .51$) and the self-documentation condition ($M = 3.62$, $SD = .58$) was small $t_{lower}(65) = 2.0$, $p = .025$. On the attention scale ratings of participants in the gaming-only condition ($M = 4.14$, $SD = .45$) were lower than in the self-documentation condition ($M = 4.3$, $SD = .31$), $t_{lower}(65) = .39$, $p = .35$. This has no impact on the results regarding effectiveness as these are solely based on the self-documentation condition. There was a significant gender difference for the attention subscale with women reporting higher values ($M = 4.3$, $SD = .34$) than men, $M = 4.08$, $SD = .43$, $t(65) = 2.23$, $p = .029$, $d = .57$. In the clarity subscale men reported descriptively higher values ($M = 3.73$, $SD = .56$) than women ($M = 3.56$, $SD = .53$) without reaching statistical significance, $t(65) = 1.12$, $p = .24$, $d = .31$. Emotional intelligence for all participants was within the range of results reported in Otto et al. (2001) as is the higher value for female participants on the attention subscale, none posing an impediment to our hypothesis tests.

Participants rated their prior experience on 7-item scales from 0-*none* to 6-*very much*. The majorities' experience with racing games and casual games ranged from "0-*none*" to "4-*fairly much*" with only four participants reporting "5-*much*" or "6-*very much*" experience in racing games and one participant reporting "5-*much*" experience in casual games (both $Mdn = 2$, $IQR = 1 - 3$). Regarding specific experience with the games, the numbers indicate broad unfamiliarity: Except for two participants who stated "2-*little*" experience with SuperTuxKart, the sample rated their prior experience with this game as equal to "0-*none*". Similarly, all but one participant ("4-*fairly much*") rated their experience with the game FlightControlHD as equal to "0-*none*". Mann-Whitney tests indicate that prior experience with gaming did not differ between conditions for casual games (both $Mdn = 2$, $IQR = 1 - 3$) or racing games, $Mdn_{gaming-only} = 2$, $IQR = 1 - 3$; $Mdn_{self-documentation} = 3$, $IQR = 1 - 3.5$, $U = 505$, $n_1 = n_2 = 35$ $p = .2$.

8.4.1 Effectiveness

We used the same formulae as in the last chapter to calculate thoroughness and validity as well as effectiveness as a product of both. This time, the emotional instances reported by participants served as ground truth. The distribution of reported emotions across games is reported in table 8.1. Stress, pride and anger were the most prevalent emotions across both games while surprise and boredom occurred less often. Note that the frequency ranks are similar across games except for the two negative emotions of anger and stress which switch positions.

For hypothesis testing of effectiveness, only the data of the self-documentation condition can be used. Overall, those $n = 35$ participants documented 624 emotions concurrently during their game play, reported 1162 emotions retrospectively (ground truth), and the observer documented 1128 instances during both games.

The significant outcomes of paired t-tests indicate that the thoroughness of Proxemo timestamps documented by an observer were higher than emotions reported concurrently by participants themselves during the racing game ($t(34) = 4.89$, $p < .001$, $d = .83$) and the flight control game, $t(34) = 4.45$, $p < .001$, $d = .75$. Descriptive data is presented in figure 8.5 along with validity and effectiveness.

Validity was descriptively higher for emotions documented by observers, however, without statistic significance in one-tailed paired t-tests¹ in the racing game ($t(34) = 2.52$, $p = .992$, $d = .43$) and the flight control game, $t(34) = 1.92$, $p = .969$, $d = .33$.

A comparison of the computed variable effectiveness revealed statistical significance in asymptotic Wilcoxon-tests for both, the racing game ($z = 3.15$, $p < .001$, $r = .63$, $n = 35$) and the flight control game ($z = 2.65$, $p = .004$, $r = .54$, $n = 35$), indicating that the central tendencies for observed emotions are higher than for self-documented emotions.

¹p-values for two-tailed t-tests calculated post-hoc are $p_{racing\ game} = .017$ and $p_{flight\ control\ game} = .063$

Table 8.1: Frequency of emotional instances retrospectively reported by participants. Data is cumulated over all participants for both conditions and emotion categories are ordered descending by frequency.

Emotional category (frequency)	
Flight control game	Racing game
stress (460)	anger (400)
pride (227)	pride (251)
anger (211)	stress (187)
surprise (145)	surprise (178)
boredom (114)	boredom (28)

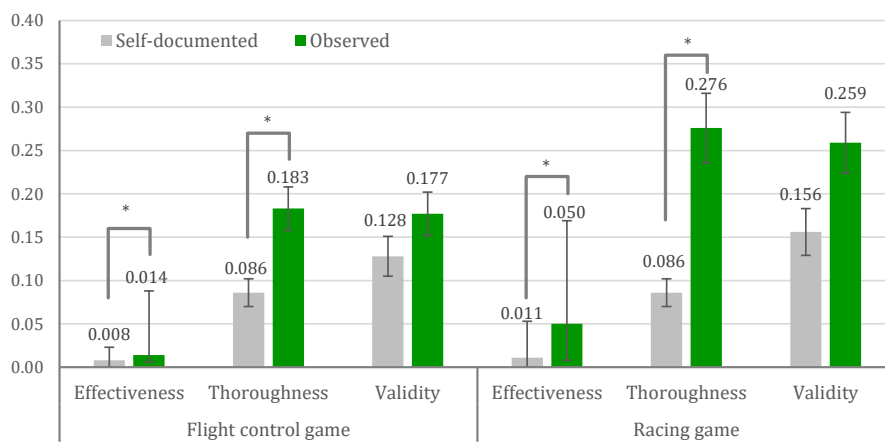


Figure 8.5: Bar graph displays median for effectiveness and means for thoroughness and validity. Due to the small values of effectiveness, we report three decimals. Error bars represent IQR and SE_M , respectively. Asterisks (*) indicate significance. Tables with descriptive data are provided in appendix A.5.1.

8.4.2 Intrusiveness

We expected self-documentation to be more intrusive for participants than being observed alone. We hypothesised that the intrusion affects the variables RAW TLX, pupil diameter, skin conductance, game scores and game experience as well as establishes an awareness of intrusion among affected participants.

RAW TLX. Against our hypothesis, workload measured with the RAW TLX was not higher for participants who were asked to self-document their emotions in the racing game ($t(65) = .94$, $p = .176$, $d = .23$) or the flight control game, $t(66) = .03$, $p = .49$, $d = .01$. Descriptive values are presented in table 8.2.

Table 8.2: Descriptive and inferential statistical results for questionnaire outcomes. Note that values of the GEQ depict agreement to statements from 0—*not at all* to 4—*extremely*. Consequently, higher values in the negative affect scale represent more negative affect. The RAW TLX ranges from 0 to 20 with higher values indicating more perceived workload.

	Condition	N	Mean (SD)	T-Test
RAW TLX — Racing game	gaming-only	32	8.50 (3.26)	$t(65) = .94$,
	self-documentation	35	9.31 (3.76)	$p = .35, d = .23$
RAW TLX — Flight control	gaming-only	33	11.01 (2.67)	$t(66) = .03$,
	self-documentation	35	11.03 (2.79)	$p = .98, d = .01$
GEQ_flow Racing game	gaming-only	32	2.64 (0.84)	$t(65) = .26$,
	self-documentation	35	2.59 (0.91)	$p = .4, d = .06$
GEQ_competence Racing game	gaming-only	32	2.23 (0.84)	$t(65) = .26$,
	self-documentation	35	2.29 (0.90)	$p = .6, d = .06$
GEQ_tension Racing game	gaming-only	32	1.72 (0.96)	$t(65) = .75$,
	self-documentation	35	1.91 (1.10)	$p = .77, d = .18$
GEQ_challenge Racing game	gaming-only	32	1.80 (0.71)	$t(65) = .43$,
	self-documentation	35	1.72 (0.69)	$p = .33, d = .11$
GEQ_positive affect Racing game	gaming-only	32	2.58 (0.74)	$t(65) = .98$,
	self-documentation	35	2.40 (0.77)	$p = .17, d = .24$
GEQ_negative affect Racing game	gaming-only	32	0.32 (0.42)	$t(65) = .55$,
	self-documentation	35	0.39 (0.44)	$p = .71, d = .13$
GEQ_flow Flight control	gaming-only	33	2.96 (0.66)	$t(66) = 2.19$,
	self-documentation	35	2.56 (0.84)	$p = .016, d = .53$
GEQ_competence Flight control	gaming-only	33	1.90 (0.81)	$t(66) = .68$,
	self-documentation	35	2.04 (0.85)	$p = .75, d = .16$
GEQ_tension Flight control	gaming-only	33	1.9 (0.71)	$t(66) = .15$,
	self-documentation	35	1.93 (0.83)	$p = .15, d = .04$
GEQ_challenge Flight control	gaming-only	33	2.3 (0.65)	$t(66) = 1.52$,
	self-documentation	35	2.06 (0.65)	$p = .07, d = .37$
GEQ_positive affect Flight control	gaming-only	33	2.26 (0.77)	$t(66) = .47$,
	self-documentation	35	2.17 (0.89)	$p = .32, d = .11$
GEQ_negative affect Flight control	gaming-only	33	0.29 (0.32)	$t(66) = 2.04$,
	self-documentation	35	0.50 (0.51)	$p = .023, d = .50$

Pupil diameter. As second indicator for workload we continuously measured participants' pupil diameter during the games. We consolidated the data by calculating the median diameter for each participant per game to smooth out the influence of single peaks of measurement error. Visual examination of all raw data revealed that in the racing game, pupil diameter levels were generally high and in the flight control game, pupil diameter increased over the course of the game. Opposing our hypothesis, there were no statistically significant differences of pupil diameter between conditions in either the racing game or the flight control game. Descriptively, pupil diameters were higher in participants who had to concurrently document their own emotions in addition to the game (see table 8.3).

Skin conductance. Compared to an EDA baseline measured prior to the gaming session tonic levels did not increase more in the self-documentation condition than in the gaming-only condition for either the racing game or the flight control game. Descriptively, the data points in the opposite direction (see table 8.3).

Table 8.3: Descriptive and inferential statistical results for the delta of tonic EDA levels ($n_{gaming\ only} = 21$, $n_{self\text{-}documentation} = 14$) and pupil dilation ($n_{gaming\ only} = 32$, $n_{self\text{-}documentation} = 29$). EDA levels are measured in micro Siemens (μS), with higher values indicating higher skin conductance.

	Condition	Median (IQR)	Statistic
EDA increase in the racing game in μS	gaming-only	2.19 (1.21 – 3.28)	$U(33) = 188$, $p = .899$, $r = .25$
	self-documentation	1.54 (0.91 – 2.51)	
EDA increase in the flight control game in μS	gaming-only	1.81 (1.36 – 3.10)	$U(33) = 165$, $p = .695$, $r = .1$
	self-documentation	1.61 (1.36 – 2.62)	
		Mean (SD)	
Pupil diameter in the racing game in mm	gaming-only	4.61 (.70)	$t(59) = 1.17$, $p = .12$, $d = .30$
	self-documentation	4.81 (.64)	
Pupil diameter in the flight control game in mm	gaming-only	4.02 (.58)	$t(59) = .40$, $p = .35$, $d = .10$
	self-documentation	4.08 (.43)	

Game scores. Descriptively, participants who had to document their own emotions needed more time to finish in the racing game and finished on lower ranks behind non-player-characters. In the flight control game participants were responsible for up to 14 aircraft at a time. Participants in the self-documentation condition landed descriptively fewer aircraft before a crash occurred, both in their best run and on average. However, none of those differences is statistically significant in independent samples t-tests or Mann-Whitney tests, see table 8.4.

Game Experience. In the GEQ, participants rated their agreement to statements on a 5-point scale from 0—*not at all* to 4—*extremely* (list with all items in appendix A.5.4). Independent samples t-tests of game experience between conditions did not produce significant results on any subscale. The two subscales which got closest to achieving statistical significance were *flow* and *negative affect* in the flight control game. The first trend indicates that participants in the gaming-only condition reported higher values for flow ($M = 2.96$, $SD = .66$) than participants in the self-documentation condition ($M = 2.56$, $SD = .84$), $t(66) = 2.19$, $p = .016$, $d = .53$. The second trend indicates that participants who had to document their own emotions reported higher negative affect ($M = .50$, $SD = .51$) than participants in the gaming-only condition ($M = .29$, $SD = .32$), $t(66) = 2.04$, $p = .023$, $d = .50$. For the racing game there were no observable trends regarding game experience between conditions. All descriptive data is presented in table 8.2.

The number of participants in our study was not sufficiently high for a component analysis. However, we computed Cronbach’s α for all components and found inconsistencies for the subscales *challenge* and *negative affect* (see appendix A.5.4) which are comparable to the scores reported by Nacke (2009). One reason may be that the items measure different dimensions of challenge and negative affect.

Awareness of intrusion and game preferences

Game preference. We clustered qualitative data from the short retrospective interviews in affinity diagrams and found a clear pattern of user needs (Desmet & Fokkinga, 2020) in the participants’ statements. Most participants agreed on having perceived the flight control game as triggering more intense emotions, mostly boredom in the beginning and then stress and frustration as more aircraft arrived. While some participants disliked those extremes, others found

Table 8.4: Descriptive and inferential statistical results for performance outcomes of all 70 participants.

	Condition	Mean (SD)	Statistic
average streak in flight control (the more aircraft the better)	gaming-only	28.24 (14.63)	$t(68) = 1.1$, $p = .14$, $d = .26$
	self-documentation	24.67 (12.50)	
best streak in flight control (the more aircraft the better)	gaming-only	37.60 (18.11)	$t(68) = .78$, $p = .22$, $d = .19$
	self-documentation	34.46 (15.51)	
time to finish in racing game (the fewer seconds the better)	gaming-only	238.82 (32.80)	$t(68) = -.75$, $p = .23$, $d = .18$
	self-documentation	244.05 (24.81)	
Median (IQR)			
rank in racing game (smaller ranks are better)	gaming-only	1 (1 – 2)	$U(68) = 559.5$, $p = .23$, $r = .09$
	self-documentation	1 (1 – 4)	

them motivating. They appreciated the challenge posed by an increasing amount of aircraft in the flight control game and by non player characters in the racing game. Other players simply named the game in which they were more successful as their favourite. Both statement categories seem to be motivated by a need for competence. Few participants explicitly stated how they liked the racing game better as it gave them more flexibility and a steeper learning curve. Both *environmental control* and *skill progression* are sub-needs of competence (Desmet & Fokkinga, 2020).

A further subset of participants based their preference on prior experience that means they reported to generally like or dislike racing games or games of skill (flight control). Some participants praised the joy, action and diversion the racing game caused as well as its larger variety of emotional triggers, thus pointing towards a need for stimulation. One main difference between the games was that even grave errors such as driving off track or crashing in the racing game only resulted in a respawn of the player along with a slight timely disadvantage, while errors in flight control caused an immediate game over scenario. Additionally, players perceived feedback in the racing game as more direct, providing a higher feeling of autonomy and control. Interestingly, neither of the games addressed to the participants' sense of aesthetics and beauty.

Video review as memory support. Participants widely agreed that the video aided their memory of gaming experiences in retrospective interviews. In detail, they found it helpful to review their own mimic reactions and events in the game. One participant even called their emotions out loud while playing in order to support their memory during the video review. This may be a clever strategy as two participants reported how they found it difficult to frequently switch their attention focus between the recording of their face and the game in an attempt to process both. Interestingly, reviewing the video of their game experience altered participants' perspective. They reported, for instance, how they perceived pride only in the retrospective when watching their own accomplishments, identified mistakes they made or even noticed some user interface elements from the video recording which they had not perceived during the game. Others thought that they would have remembered the most prominent emotional events also without the video or found it difficult to distinguish similar events despite the video. Critically, one participant reported memories of emotions which they could not identify by means of their own mimic. Another remembered emotions they had experienced during the game but no longer considered relevant in the aftermath. One participant who was in the concurrent self-documentation condition felt restricted by the categories and refrained from documenting their emotions as their experience did not perfectly match the category. This in fact is a strong argument for detailed retrospective analysis of experiential episodes which may not be elaborated during use.

The price of self-documentation. Regarding the concurrent self-documentation of emotions with the Proximo App participants in the self-documentation condition agreed that they a) consciously neglected this task and b) only documented their emotions when the demand of the game allowed for it. Participants prioritised achieving their personal goals in the games over the secondary task. To succeed, they focused on the game controls disallowing a concurrent use of the documentation app and resulting in overall fewer emotions logged or emotions documented with a delay when the game action ceased. The documentation of negative emotions after a crash in the flight control game was explicitly most simple because a crash was followed by the game restarting with no traffic at the beginning and, hence, sufficient resources for emotion documentation. Participants were aware that documenting their emotions in the app during the game influenced their game performance. In order to still log emotions two participants reportedly accepted the disadvantage of shifting attention and risking a crash or certainly losing velocity in the racing game (pushing an emoji in the app required lifting the finger off the keyboard and, consequently, interrupting acceleration). One participant reported that only the second game gave them enough time to reflect about emotions and several participants mentioned they were so concentrated on the game, they simply forgot about the emotion logging. Two participants even noted that the secondary task led to overextension which resulted in declining gaming performance and finally triggered anger.

Post-hoc we tested whether the subjective impressions on the performance trade-off between secondary and primary task or the flow experience were backed quantitatively across all participants. Regarding flow there was a weak yet not significant correlation between flow and thoroughness in the flight control game ($r = .3$, $p = .082$) and none in the racing game ($r = -.01$, $p = .94$). This means that across all participants there is no indication of a trade-off between concurrent self-report of emotions and subjective flow experience. There were also no significant correlations between thoroughness of concurrent self-report and game performance indicators in either game. However, thoroughness correlated between games ($r = .61$, $p < .001$) indicating that participants who chose to split their resources in one game did so in the other game as well. In addition to the negative affect scale reported above in table 8.2, we examined the retrospectively annotated emotions for differences between conditions. The distribution of emotions from the predefined categories did not differ between conditions ($\chi^2(4) = 4.18$, $p = .38$, $V = .04$) with overall 1151 emotional instances in the gaming-only condition and 1050 emotional instances in the self-documentation condition (see detailed descriptive data in appendix A.5.2).

8.4.3 Proximo and Physiological Data

We matched EDA peaks with retrospectively annotated emotions and due to interindividual differences in skin conductance response considered peaks as a match when they occurred up to ten seconds after an emotional event — thereby generously incorporating the 1 – 4s delay

of skin response (Caruelle et al., 2019). Since emotion categories cannot be distinguished in EDA, we adapted the Proxemo data and for this calculation considered a documented event as match when the timestamp occurred within five seconds of the retrospective annotation regardless of the category. Oposing our hypothesis, paired t-tests on consolidated data across both games showed no difference in effectiveness between Proxemo and EDA, $t(34) = -2.02$, $p = .974$, $d = -.34$. Proxemo achieved a higher validity than EDA, $t(34) = 3.49$, $p < .001$, $d = .59$. However, the EDA data were more thorough than observational data documented with Proxemo, $t(34) = -5.71$, $p > .999$, $d = -.97$. Descriptive results are depicted in figure 8.6 and reported in detail in appendix A.5.3. In theory, boredom should cause dips rather than peaks in skin conductance. However, subtracting matches between EDA peaks and instances of boredom from the retrospective annotations even slightly decreased all three quality criteria (see appendix A.5.3). Therefore, we conservatively ran above reported tests including the matches between EDA and boredom.

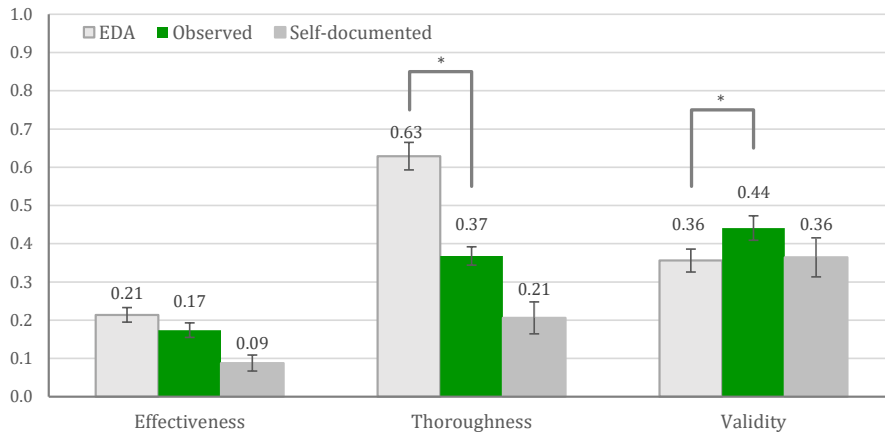


Figure 8.6: Bar graph displays means for effectiveness, thoroughness and validity for both EDA and observations with Proxemo. Error bars represent SE_M . Asterisks (*) indicate significance. Self-documented emotions, the third bar in each group, is based on the $n = 15$ participants only whose EDA data was usable and who were in the self-documentation condition.

8.4.4 Explorative Post-hoc Analysis of Quality Criteria Across Emotion Categories and Intervals

In section 8.4.1 above we reported the difference of validity and thoroughness scores between observed and self-documented emotions. Figure 8.8 complements the picture with a visualisation of emotional proportions that were discovered through self-documentation or observation and those that were documented exclusively with either method. In this paragraph we dip into a short post-hoc exploration of the descriptive differences between emotion categories. Due

to the small frequency of matches in each emotional category we calculated the validity and thoroughness scores only from consolidated data of all participants and cannot report measures of dispersion. A visual examination of the graphs in figure 8.7 clarifies that thoroughness suffering more under the self-documentation than validity is a pattern that spans across all emotion categories. Interestingly, among the observed emotions both criteria are approximately on level in the categories anger and pride and thoroughness is even higher than validity for boredom in the flight control game and for stress in the racing game. Between data sources validity scores for pride are approximately on level while observers' validity scores for surprise and anger exceed the self-documentation by far. A larger descriptive difference in validity of stress-ratings only manifests in the flight control game.

When ignoring the emotion categories entirely and only matching by timeliness, a descriptive comparison between figures 8.5 and 8.6 indicates that the thoroughness of observed emotions increases by approximately 60% while the validity increases by 100%. In the self-documentation condition, thoroughness even increased by 130% and validity increased by 150%. Note that these descriptive comparisons need to be taken with care as the latter calculation does not include all participants from the sample (the dispersion is similar though, see appendix A.5.3).

For increased timely precision when binding the timestamped emotion to the experience interaction captured on video, we decided early in this work to restrict the timely tolerance of documented emotions to 5 second intervals. Qualitative statements of participants in the self-documentation condition indicated that they sometimes delayed the documentation of emotions for longer until a suitable moment in the game action allowed them to. Speaking with the terminology of prospective memory research (e.g., Grundgeiger et al., 2014), participants self-determined the length of their interruption-lag and kept their attention on the primary gaming task before turning to the secondary documentation task. This strategy potentially resulted in longer delays until participants documented emotional events. For an exploration of this phenomenon we extended the interval from ± 5 to ± 10 and ± 15 seconds. Confirming qualitative statements, an extension of the time interval resulted in matches of the self-documented emotions tripling but matches in the observation condition only increasing by 70%, see table 8.5. Since the ground truth is unaffected by the extension of the tolerance interval and remains the same, all dependent quality criteria are directly proportional to the matches.

8.5 Discussion

In this lab-based gaming study we compared Proxemo with concurrent self-report of emotions and hypothesised that Proxemo achieved higher thoroughness and higher effectiveness despite lower validity. Furthermore, we expected the intrusion caused by self-documentation to negatively affect participants' workload, performance and aspects of gaming experience. As a bonus we explored the documentation effectiveness of EDA in comparison to Proxemo.

Table 8.5: Exploration of matches between documentation methods for emotions, when timely synchronicity is extended. Overall matches are computed between 1162 retrospectively reported emotions (ground truth) and 1128 observed instances or 624 concurrently documented emotions, respectively. Data is pooled over all participants from the self-documentation condition and merged over both games. Hence, scores for thoroughness (t), validity (v) and effectiveness (e) slightly deviate from former reports.

interval	Observed				Self-documented			
	matches	t	v	e	matches	t	v	e
± 5 seconds	277	.24	.25	.06	104	.09	.17	.01
± 10 seconds	377	.32	.33	.11	262	.23	.41	.09
± 15 seconds	406	.35	.36	.13	330	.28	.52	.15

Surprisingly, Proxemo not only delivered more thorough and effective results but also showed a statistical trend towards more valid results than concurrent self-documentation. This allows the conclusion that Proxemo is at least equal to concurrent self-documentation regarding validity and even more thorough and effective with effect size varying from small (validity) to medium (thoroughness) and large (effectiveness) according to J. Cohen (1992). For practitioners these results imply that in user scenarios requiring a constant demand for attention and readiness for interaction where interruptions would have adverse effects using Proxemo delivers better results than concurrent self-documentation of participants. An explorative extension of the time interval increased the quality criteria for both conditions. The extent to which this data is biased by timestamps mismatched with documented emotions from other situations occurring within a 30 seconds interval cannot be determined. Assuming at least 10 minutes of play time per participant, on average more than three emotions were retrospectively reported per minute, rendering their occurrence too dense for an extension of the tolerance interval. In less emotional environments, where only occasionally observed emotions are documented, longer periods may prove useful and maintain data quality.

Our results on intrusiveness are far less conclusive. Qualitative data points towards a clear awareness of the intrusive effects which caused annoyance and led to a negligence of the secondary task. Quantitatively, perceived workload, game performance and pupil diameter as a physiological indicator for workload hinted with small but statistical insignificant effects towards adverse effects of intrusion through concurrent self-documentation. The game experience aspects of decreased flow and increased negative affect support these tendencies though statistically insignificant with medium effects during a game imitating flight controllers' task. However, skin conductivity as an indicator for workload was descriptively lower in participants who had the secondary task to concurrently self-report their emotions. This small and statistically insignificant effect opposes the tendency of all other intrusion indicators and serves as a clear reminder on not drawing conclusions from descriptive data.

A simple theory to explain the advantage beyond expectations of Proxemo in terms of validity

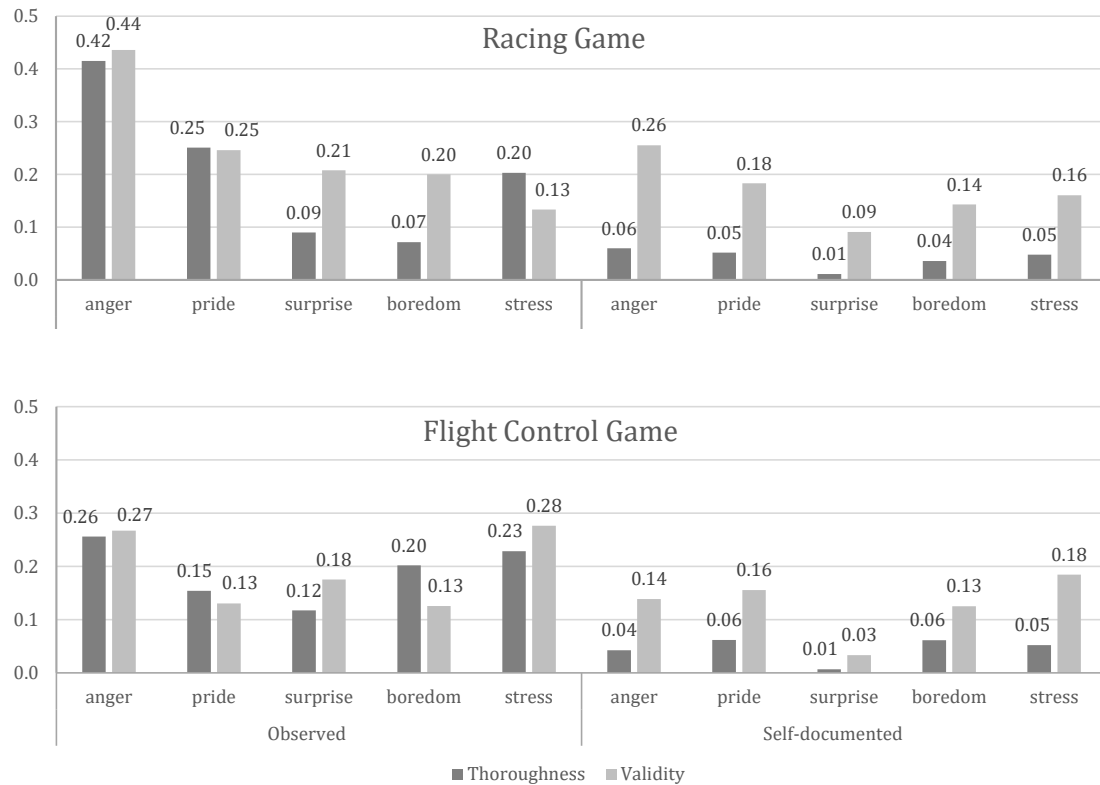


Figure 8.7: Bar graphs display thoroughness and validity by data source and emotion for the racing game (top) and the flight control game (bottom). Scores for thoroughness and validity are calculated from consolidated data over all participants.

and the small effects of intrusion would have been a trade-off between primary and secondary task performance in the self-documentation condition (Kahneman, 1973; Wickens, 2008). However, the lack of correlations between performance measures in both tasks does not support this perspective. Another possible explanation is that all participants' main focus lay on the games. Participants who still had spare resources attended to the secondary task with varying success. Consequentially, documenting emotions did not affect their game performance but resulted in a broad dispersion of documentation quality. Framing the observation in terms of prospective memory research, our participants deferred or blocked the documentation task. Eventually when their workload allowed for it and the smartphone with the Proxemo App or another cue reminded them of the documentation task they retrieved the emotion and documented it with great delay (Grundgeiger et al., 2014). For practitioners this leads to the implication that users can be bothered with an additional task such as concurrent self-report of emotions because a) users treat it as truly secondary and optional only turning their attention towards it if they have the time and b) in turn, the secondary task has little to no effect on their workload, performance and experience. This explanation is in line with literature as Casali and Wierwille (1983,

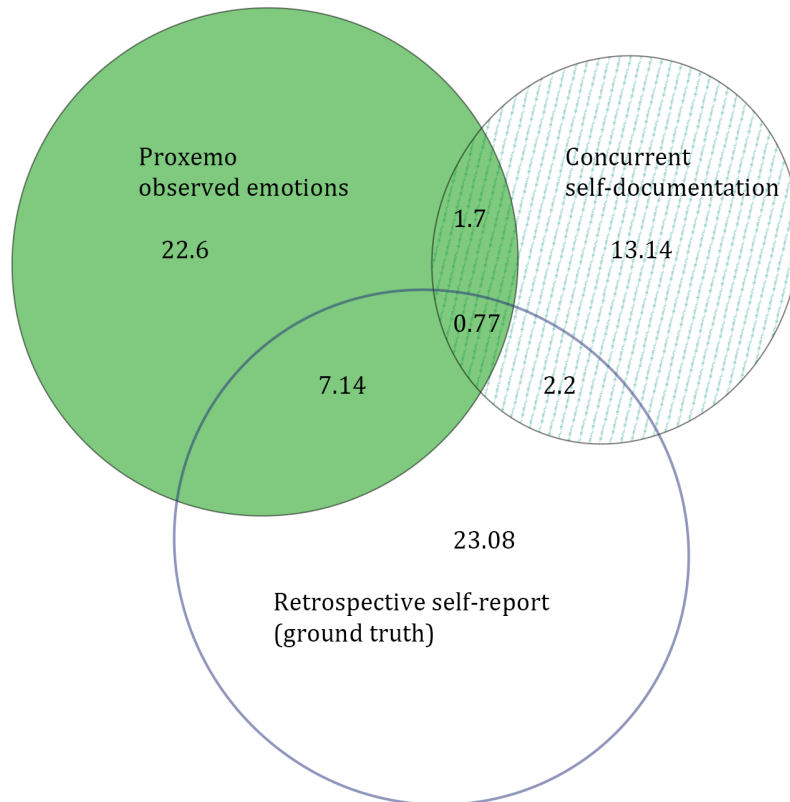


Figure 8.8: The Euler diagram displays overlaps between Proxemo documentations by the observer and the concurrent as well as retrospective (ground truth) self-documentation by participants. It visualises the proportion of shared and exclusively documented instances. Provided values are arithmetic means per participant. The graphic is based on *eulerAPE* (Micallef & Rodgers, 2014, software distributed on <http://www.eulerdiagrams.org/eulerAPE/>).

p. 640) found no intrusion differences between workload scales in studies with civilian pilots but report that “[s]ubjects were observed to disregard the secondary task at times when the communications burden was high.” In terms of generalisability, from our experience in observational studies in control centres we assume that air traffic controllers may likely decide more strictly than participants to ignore any secondary task if a potential performance loss in their primary control task is at stake. Yet, effects of intrusiveness may increase with workload in the main task (Casali & Wierwille, 1984), for instance when complexity of the environment and social interaction add to the attention required for the main task. Hence, we dare to recommend the generalised guideline for practitioners that if the effectiveness of experiential data is important to not make it a secondary self-report task for users.

In the racing game, anger was the most frequent emotion in the ground truth and also the most thoroughly and most validly documented emotion by observers. In the flight control game,

stress was the most frequent emotion in the ground truth and also the emotion documented most validly and second-most thoroughly by observers. A first implication for future research is to investigate the association between frequency and documentation quality of emotions. One reason for the varying frequency ranks of emotions across games may be that the racing game had non-player characters whose behaviour triggered anger, whereas the flight control game had less arbitrary events, leaving more control and thus stress to the player. Regarding the timely dimensions of emotions it would be interesting to re-analyse the data with different off-set tolerance values for emotions such as stress or boredom that may have multiple triggers (or none) and vary in intensity over time. Anger and surprise, in contrast, are rather short and intense emotions that capture the full attention of the participant in the game. The trigger that angered or surprised players afforded to be instantly dealt with in order to prevent further trouble (see also Stamen, 2020). Thus, no cognitive or timely resources were left for documentation. For practitioners, these instances are particularly interesting to be discussed in the retrospective as they may give insight into decision-making during crucial situations. The descriptive data in figure 8.7 indicates that evaluators using Proxemo capture distinctly more instances of surprise and anger than users through self-report which makes those categories promising for further studies.

8.5.1 Theoretical Implications

Thoroughness remained on a comparable level to the previous study (chapter 7). However, validity scores lie more than 50% under the validity scores from the previous study, in which we had listed slips as confounding factor potentially decreasing validity. After implementing a simple “undo”-feature, slips cannot serve as an explanation for low validity in this study any longer. This time, also the observer was trained in the use of Proxemo as a tool and method, resulting in better conditions on the observing side. Due to the similar amount of emotions both documented by the observer and the self-reported by participants retrospectively, we also cannot assume efficiency of the app a reason for over-reporting. Therefore, reasons for the decreasing validity in this study must be hidden in other aspects of the operationalisation, and we propose three contributing factors. First, the stimulus material in the previous study was preselected to display a variety of clearly distinguishable emotions – while in this chapter’s study no filter was applied to experiential episodes. Second, we created the stimulus material in the previous study during the social situation of people narrating their stories to other people as they reminisced with technology. The social and communicative character of those situations likely entailed more emotional expressivity than participants sitting in front of a screen and playing for themselves. Third, the variable serving as ground truth varied between the studies. As can be seen by the small overlap of concurrent self-documentation and retrospective self-report in figure 8.8, reviewing the game play along with facial, bodily and vocal expressions did not even

provide the participants themselves with enough information to cover all the emotions they found worth documenting in-situ. A reason for this low validity – even in between self-documented emotions – could be that participants reflected their emotions anew in the retrospective self-report, interweaving them with emotional appraisals of the course of the game. As a consequence, the three circles may as well measure three factors of the latent variable emotion. Following the components by Scherer (2005), the concurrent self-documentation is enabled through short *appraisal of emotional experience*, while the retrospective self-report is timely distinct from the emotional experience and requires a more thorough appraisal. Finally, observed emotions rely on bodily, *facial and vocal expression*. In the following paragraphs we will dip into three theoretical explanations of those factors visualised in figure 8.9.

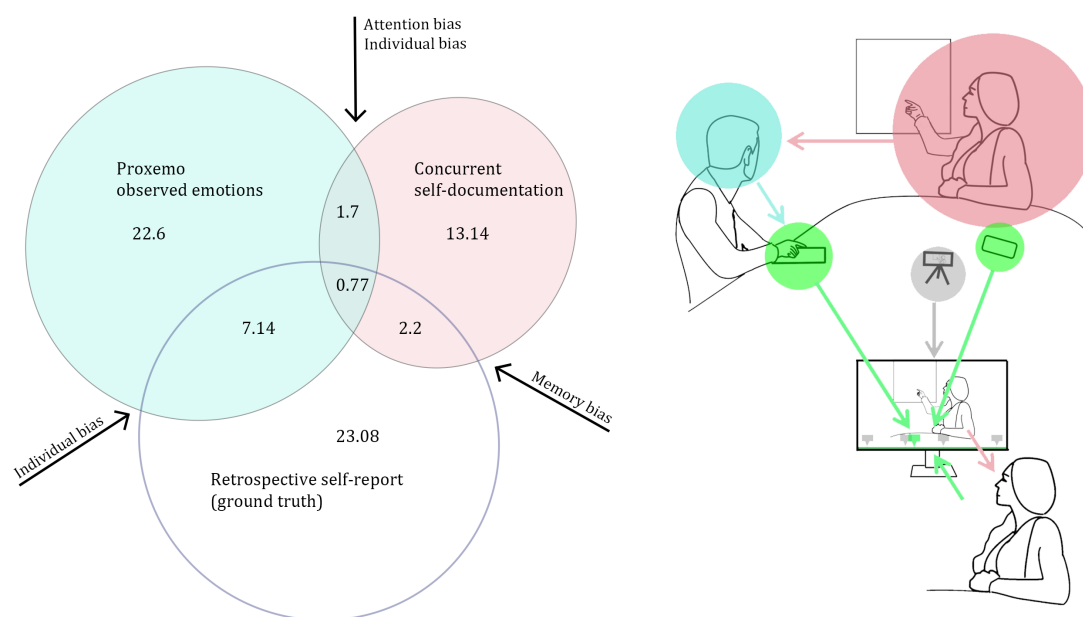


Figure 8.9: On the left, arrows added to the Euler diagram from figure 8.8 visualise where biasing factors affect differences in the operationalisations of emotion. The sketch on the right side spatially illustrates the respective context of those data sources: concurrent self-documentations with the Proxemo App by a participant (top right), Proxemo annotations from an empathic observer (top left), and retrospective annotations by the same participant based on video data (bottom). The green arrows illustrate how timestamps from all three sources were afterwards fed into the software for computational comparison.

Memory Bias. The retrospective self-report which built the ground truth can be seen as the emotional component of *remembered UX* (Wurhofer, 2018). The descriptive observation that the quality criteria for the self-documented emotions improved by more than double when the categories were neglected means that the retrospective evaluation of the emotional situations

by the participants differed not only in time but also in content. This alteration is in line with literature. For instance, Bruun et al. (n.d.) found only a medium correlation between dimensional report of emotions given at the end of a task and retrospective report of emotions given at the end of the experiment. L. Nielsen and Kaszniak (2007, p.366) even claim that “the shift from feeling to reporting can alter the experience fundamentally”. While self-report provides the only possible access to participants’ inner state, retrospective reports likely come with a bias caused by estimation strategies (Schwarz, 2007, p.11 et seq.). Another artefact of retrospective reporting is the increasing number of instances. Similarly to Petrie and Precious (2010), participants produced almost twice as many emotions in retrospective self-report than in concurrent self-documentation. As annotated in figure 8.9, memory bias affects the distinction between concurrent self-documentation and retrospective self-report.

Attention Bias. Attention bias is an influence working mostly on the observer, who has to interpret different signals from the sender (figure 8.9). Thereby observers’ ability to recognise emotions varies more between categories (Lewis et al., 2016) and less between channels (Connolly et al., 2020). However, if signals from multiple channels are contradictory, inferring the sender’s emotions becomes an issue even for machines, who are generally less affected by attention bias. This was the case in a study with air traffic controllers whose sentiment in spoken language was apparently not in line with their facial features (FACS), rendering the experimental data useless (Buxbaum, 2019).

Individual Bias. As argued above, we worked in this study towards consistency through keeping the observer constant. In spite of giving participants a list of emotions and their definitions for orientation it is very much possible that the comprehension and articulation of experienced emotions varied between individuals (Barrett, 2004). Therefore, with changing participants formerly unfamiliar to the observer, it is likely that an individual bias impacted the agreement between observed emotions and emotions stated by the participant – regardless whether concurrent or retrospective.

8.5.2 Physiological Data

Compared to phasic EDA data, Proxemo data were more valid and less thorough. Both effects are large (J. Cohen, 1992) with validity opposing our hypothesis. Effectiveness as product of both criteria was descriptively higher for EDA (insignificant, small effect). As a limitation of internal validity only affecting this measure it must be noted that the frequently occurring non-specific peaks lead to a systematic overestimation of sensitivity (thoroughness) and a systematic underestimation of the positive predictive value (validity). In our study, after subtracting matches from all peaks in the EDA data and dividing the remainder by a minimum of ten minutes playtime, the count of non-specific peaks left per participant during the game is approximately four per

minute. This is below the 5-10 non-specific reactions reported in other studies (Gertler et al., 2020; Zimmer, 2000) and hints towards ten or more non-specific peaks per participant that mistakenly went into our effectiveness calculations as match due to an incidental timely proximity to emotional instances in the retrospective ground truth. A further indication of the arbitrariness in the EDA data is the decrease of quality criteria when subtracting boredom for which no theoretical foundation is given. In contrast to Proxemo EDA did not allow for the detection of distinct emotion categories in this study which according to our case study with air traffic controllers (chapter 5) would be beneficial. Jainendra et al. (2019) propose a way to extract arousal and valence from a set of EDA features which theoretically may allow an approximation of emotion categories. However, they report validity issues with their training data — a strong correlation between arousal and valence — which reduces the generalisability of their approach. Another critical disadvantage of EDA is the extra huge effort in both material resources (BioPac Equipment, analysis software, disposable electrodes, gel) and timely resources (applying electrodes, waiting for gel to bind, measuring a baseline, filtering data). Finally, in our study only half of the EDA data sets were usable. According to Biopac Systems (2015) about 10% of participants may be non-responders. As we did not change the procedure between participants our only explanation for the remaining 40% of participants whose data sets were not usable are the variances in temperature and potential humidity of participants' skin. In the introduction to this chapter we argued that heart rate reactions are too slow, and pupil diameter is influenced by too many factors. In sum, we would discourage practitioners from using physiological arousal data as a basis for emotional event detection.

Physiological measurements for UX testing are on the rise especially in association with mixed reality (Lanius et al., 2021). Unfortunately, without valence indicators we could not harvest the full potential of physiological data. Pandemic regulations disallowed us to apply the facial EMG electrodes directly on participants' corrugator supercilii and zygomaticus major. Asking participants to apply the electrodes by themselves would have borne a risk of missed attempts followed by frustration which we deemed too high. Also, the electrode patches available to us did not stick to the pilot participants' skin and had to be fixated with additional tape. That acknowledged, we would like to take a more critical perspective on physiological data. For instance, in past UX studies the zygomaticus major has not shown to be a valid indicator for positive emotions (Thüring & Mahlke, 2007). Furthermore, Boehner et al. (2007) criticise the paradox treatment of subjective ratings in HCI where subjective ratings are used at first to validate or benchmark physiological measures but then disregarded as inferior. We agree and see the benefit of physiological data and specifically EDA in summative evaluations where no distinction between categories is required (or can be achieved through a combination with facial EMG). In formative evaluations where the users' subjectivity should be embraced, EDA merely provides a suitable baseline for discussions. From our experiences in this study we summarise that compared to EDA alone, Proxemo is better in situations, when 1) category distinction is relevant,

2) data shall be available quickly for debriefing, 3) participants are valuable (i.e., criticality of usable data) or 4) participant time is costly (e.g., EDA requires baseline and waiting time for electrolyte gel to bind with skin). Finally, 5) Proxemo is more flexible (EDA requires participant to sit still and not move their hand/foot where the electrodes are attached).

8.5.3 Limitations & Future Work

We based our sample size calculation on the expectancy of large effects in documentation quality. Consequentially, our study did not have the power to detect small and medium effects of intrusion with statistical significance. A power analysis indicates that a detection of the largest effect of intrusion in the GEQ ($d = .5$) would require 2×88 participants and for differences in the performance parameters, e.g., the average streak before a plane crashed ($d = .26$), even 2×321 participants or more. On the one hand, in a domain where decisions can cost lives it may be worthwhile to consider the impact of small effects beyond their statistical significance. As the negative affect scale of the GEQ has shown to be inconsistent, future research should look into which aspects exactly triggered the negative feelings. Then again, all our outcomes including the performance measures are based on a simplified gaming study with student participants and thus not directly generalisable to the tasks or abilities of air traffic controllers (see also Truschzinski, 2017).

While potentially uncomfortable for users, the permanent fixation of the non-dominant hand through EDA electrodes imitated the approach controllers' current workstation. There, air traffic controllers' interaction is restricted to one hand as well because they are constantly holding the radio control button in the other hand. What does limit the generalisability to air traffic control is the chosen set of emotions in this study. The games predominantly triggered negative emotions in participants, although positive emotions predominate in ATC. The reason for this is that the data collection for this lab study on effectiveness and intrusiveness started chronologically before the qualitative study in air traffic control (chapter 5). We selected the games according to their similar task character to air traffic control. At that time we knew already which emotion categories were to be expected from controllers, however, we were unaware of the high frequency of positive emotions in their work experience. Rather than switching the set of emotions, in a lab replication we would select games that offer more positive experiences to untrained participants. With respect to Proxemo, future work may additionally rearrange categories (left to right) on the Proxemo App based on their expected frequency or add more distinct categories for positive or negative emotions respectively.

A limitation of internal validity is our missing out of assuring participants' colour vision. In the flight control game, aircraft and runways were colour coded. We did not explicitly test or require participants' ability to distinguish colours but neither did we observe a colour confusion of aircraft to cause trouble for any participant during the training sessions. Furthermore, we

compared pupil dilation directly between participants. A cleaner approach would have been to measure a baseline and compare only the delta between conditions. Also, a lab environment with absolute control over light conditions instead of semi-transparent blinds would improve the reliability of pupillometry data.

Regarding external validity, we could have compared the intrusiveness of Proxemo with a no-observation control condition instead of self-report scales to describe the intrusiveness of Proxemo. Since the observer was not even in the same room, and we deployed the video recording anyway, we assume there would not have been any effect. Tuncer (2016) argues in her qualitative study that participants are aware of the recording, and it may change natural interactions. However, experimental studies have shown that camera presence has no effect on participants' decision-making regarding honesty in reports (Lohse & Qari, 2018) or pro-social behaviour (Jansen et al., 2018). A condition by Jansen et al. (2018) where participants were told their behaviour was evaluated by others who watched the video recording is closest to the situation in our study. Nevertheless, the knowledge of being observed and evaluated by others did not stop participants from cheating. We are not aware of any research on changes in participants' behaviour caused by the number of live-stream observers.

Neither of the games was particularly immersive (e.g., vivid illusion of body ownership and environment) by definition (Slater & Wilbur, 1997). Hence, we followed the principle of data economy and did not bother participants with the GEQ subscale. However, immersion may be a more relevant criterion for the UX in other potential application domains of Proxemo such as virtual reality and be evaluated there with more appropriate questionnaires (Wienrich et al., 2018).

There are a few limitations of generalisability that can be covered in future work. For instance, due to pandemic restrictions, we were forced to keep the study simple, leaving out the crucial aspects of cooperation with neighbouring sectors potentially combinable with multi monitoring. What further biases the generalisability of the scores is the observers' intended unfamiliarity with participants. The observer only joined the operation room towards the end of the training and participants wore medical face masks until the trial started. As a consequence of both, the observer had no opportunity of getting to know each participant's expressive behaviour in advance. The observer's validity scores can thus be considered as very conservative, not exploiting the possibilities. Potentially, in a post-pandemic setting, achieved validity can be increased simply through a short familiarisation with the participants or users at the beginning of the study, or even better through colleagues who are familiar with the domain and particular user already (see chapter 5).

Future work can attempt to distinguish the severity of emotions that means asking users to rank their self-reported emotions by importance. Our focus of intrusiveness lay on the global summative comparison between conditions. Eggemeier et al. (1991) speculate that intrusion is difficult to measure because the employed global workload measures are not sufficiently sensitive

to detect the effect of intrusion in those situations where peaks induced by secondary tasks occur. Further post-hoc explorative analyses on this study's pupillometry data may provide insight into the timely association between arousal peaks following an attendance to the secondary task.

Chapter 9

Conclusion

Emotions are a key to understanding users' experience. The main goal of this work was to introduce Proxemo as a structured observation method that facilitates the documentation of users' emotions during formative evaluations in contexts where concurrent self-report is not possible. In short, Proxemo comprises the documentation of emotions from pre-defined categories for multiple participants by setting timestamps in an application.

Revisiting the research questions posed in chapter 1, this chapter resumes what we learned so far about the feasibility and utility of Proxemo in the highly diverse scenarios of reminiscence activity in dementia care and high fidelity simulations in flight control. Furthermore, this chapter summarises the quality criteria tested for Proxemo comparing it to other methods where appropriate. We inspect the role of Proxemo in the light of HCI theories, discuss which UX definitions can be satisfied with the collected data and shortly contrast Proxemo with physiological approaches. For practitioners who seek to use Proxemo, we compile a short list of recommendations and further materials. Finally, we highlight possibilities for future work and give an outlook on future paths Proxemo might take.

9.1 Capturing User Emotions

In chapter 3 we searched literature for descriptions of a formative UX evaluation method suitable for users who have no spare cognitive resources (RQ1) and identified a lack of such methods. Consequentially, we strived to fill this gap by enabling evaluators to document observed emotions by proxy (RQ2). Hence, the resulting method Proxemo is not a universal answer for all user research situations but a precise method for formative evaluations in contexts where users' concurrent self-report is not feasible. Through Proxemo, qualified observers can facilitate users' emotions to flow into the iterative design process.

Utility and feasibility. Case studies in dementia care facilities (chapter 4) revealed Proxemo as a suitable way for conveying people with moderate to severe dementia's emotional reactions to design teams. The iterative implementation of the Proxemo App on a smartwatch provided observers with a discreet form factor to efficiently document emotions in the context of technology-supported reminiscing sessions. Where possible, the validation of observed emotions happened in the context of use as a natural part of person centred care concepts (Kitwood & Bredin, 1992). Annotated emotions in video snippets of interaction scenarios aided the interpretation of situations and the decision-making for design teams who were novices in the context of dementia.

In formative evaluations of novel ATC systems (chapter 5), Proxemo supported shift leaders to capture observed emotions while allowing air traffic controllers to uninterruptedly focus on the simulated traffic. The annotated emotions provided a thorough grounding for debriefings. Here, controllers validated their emotions in retrospective or explained in detail what had happened in the marked situation and whether it was relevant. Most important, Proxemo annotated video snippets served as communication bridge between controllers and developers. Not only did the annotated video support developers in critical situations with efficiently reproducing system behaviour. Short explanations of these video recorded situations gave developers and researchers a context-bound level of comprehension and insights which they did not remember having achieved in debriefings of former evaluation sessions with discussions based on handwritten notes. Qualitative data from both domains suggest that the tool and method Proxemo are useful and feasible for formative evaluations in the contexts of dementia and air traffic control, thus, answering RQ3 in the affirmative.

Since we varied the form factor between studies to match the context, we consider Proxemo to be agnostic of the documentation aid's implementation. Whereas the enthusiasm of controllers and developers indicates a satisfying level of quality in the Proxemo data, we systematically examined Proxemo with respect to quality criteria relevant for evaluation methods. Existing observation methods systematically lack published quality criteria. Therefore, the captured quality criteria for Proxemo may give an initial idea of the general dimensions quality criteria of structured observations of user emotions may reach. Our conclusions on Proxemo's performance in the quality criteria considered most relevant are reported in the following paragraphs and provide answers to RQ4.

Reliability. Towards a determination of Proxemo's quality criteria, we first conducted a study on inter-observer reliability (chapter 6). Proxemo has substantial (Landis & Koch, 1977) or even excellent (Cicchetti, 1994) reliability with values for Cohen's kappa and Krippendorff's alpha between .70 for observations of three reminiscing residents and .76 for sessions with individual residents. Compared to kappa values reported for paper versions of quality of life measurements in the context of technology-supported reminiscence by which our emotion set is inspired inter-rater reliability scores of Proxemo are on level with (Feng et al., 2020; Feng et al., 2019) or

above (Krüger et al., 2017; Sloane et al., 2007) other studies. We conservatively conducted the study with two student participants and expect that inter-observer reliability could be further improved if observers brought prior knowledge about the domain and user group and had more time to internalise the emotion categories.

Thoroughness. Two elaborate lab studies conceptually replicated live evaluations in reminiscence sessions (chapter 7) or aspects of air traffic controllers' work flow (chapter 8). Proxemo facilitated a more thorough documentation of emotions for observers than handwritten notes or participants' self-documentation respectively and achieved mean thoroughness values ranging between .18 and .28. Unfortunately, the lack of thoroughness values reported for emotional observations in literature disallows for direct comparisons to similar methods. Staying at least within the discipline of HCI, Proxemo's thoroughness values are within the upper range of thoroughness values for observed usability problems (Hertzum et al., 2014; Molich & Dumas, 2008). However, they can be achieved considerably more efficient because Proxemo annotations are based on real time observations and not detailed video analyses.

Validity and Effectiveness. In the same studies, Proxemo's validity ranged from .18 to .46. While observers achieved higher validity with handwritten notes, a trend indicates that Proxemo notes may be more valid than users' self-report when their main focus lies on another task. In both studies, effectiveness computed as the product of validity and thoroughness was higher for the Proxemo conditions implying that the advantage of thoroughness outweighed the deficit in validity. Similar to reliability, validity can be increased through the training or recruitment of observers who are familiar with the participants already. While we did not systematically collect descriptive validity data in the case studies, we have no documentation of any instance where air traffic controllers objected the documented emotion or the relevance of an associated situation. We argue that air traffic controllers would likely not have been as content with the Proxemo method if 50 – 80% of timestamps had been worthless or even misleading.

Efficiency. Efficiency is defined as ratio of effectiveness and effort. Both factors are most expressive when directly compared to other methods. Efficiency can be measured throughout the Proxemo pipeline at three major stages: (1) the observation of emotions, (2) the documentation of emotions and (3) the analysis of emotion annotated video files. Most clearly, observers rated their subjective effort as lower when using Proxemo to document emotions (stage 2) compared to handwritten notes (chapter 7). As a consequence of the more efficient documentation technique, observers had the impression of capturing more emotions (stage 1) that means achieving a higher effectiveness. This feeling was confirmed by objective data. When comparing Proxemo to users' self-documentation of emotions, there were only small, insignificant intrusion effects operationalised as subjective effort (chapter 8). The main reason for this is the users' unwillingness to waste

much thought on Proxemo which affected data quality. The overall diminishing effectiveness deteriorated efficiency. Finally, our field studies indicated that Proxemo time stamps speed up and support interpretative steps in video analyses (stage 3) together with users (chapter 5) or without them (chapter 4). Therefore, we conclude that observation, documentation and analysis gain in efficiency through the deployment of Proxemo.

Observer Experience. In this work, observer experience was mostly captured qualitatively with the exception for the effectiveness study in (chapter 7), covering the first two stages of the Proxemo pipeline. There, UEQ scores for Proxemo are high-ranking in an international benchmark and Proxemo's UX is clearly above the UX of handwritten documentation. The positive observer experience is mostly rooted in efficiency but also in increased mobility and hedonic aspects. Observers' feedback during case studies in the dementia context is similarly coined by the simplicity of the method and app which facilitates efficient and intuitive interactions. Observers experienced Proxemo as well applicable in group sessions with up to four residents. During one-on-one reminiscence sessions or when the availability of staff is restricted caregivers found it feasible to use Proxemo next to moderating the activity. As this reduced data quality we advise against making this double role the norm. For stage three, the analysis of video data, Proxemo annotations alleviated the effort through efficient navigation and, more importantly, conveying a feeling of certainty in ambiguous situation. In the ATC case study, supervisors would have observed their team members during the simulation anyway and found the documentation to cause no notable extra effort. The joint analysis of the Proxemo annotated video file evoked a great debriefing experience with improved cooperation and comprehension being the main advantages. Supervisors as observers and primary users of Proxemo but also observed controllers and developers as secondary users expressed their anticipation of using Proxemo in future evaluations. Independent of the context, Proxemo shifts the detailed examination and generation of insights to the video analysis and consequentially allows the observed users to have an interaction experience unbiased by the evaluators' questions.

9.2 Theoretical Considerations

Having based this entire work primarily on an empirical demand for a structured observation method, we would like to point out how some aspects of Proxemo are associated with theories of HCI, UX and emotion, thus completing the circle to the theoretical basis presented in chapter 2.

9.2.1 Proxemo Within the Three Paradigms of HCI

A chronological division of trends in HCI theory into three waves or paradigms (so far) is broadly agreed upon. As an oversimplified reminder, the three waves foci lay on (1) performance, (2)

more complex systems with multiple users and (3) emotional and meaningful experiences with ubiquitous technology.

The Proxemo pipeline has a clear focus on emotion and hence supports formative evaluations in the third paradigm that focuses on experience and meaning-making (Bødker, 2015). This is particularly important in our exemplary application domain of dementia care where users do not have a task based on which traditional performance measures (first wave) can be taken, but users instead may use technology for pure well-being and shaping their relationships with their caregivers (Houben, Lehn et al., 2020), visitors (Muñoz et al., 2021) or their own identity (Wallace et al., 2012). People with dementia who live in residential settings, in contrast to the caregivers who work there, rarely leave their environment. Consequently, any technology installed in this environment likely will affect residents' entire day including all actors and events. According to Bødker's (2015) view on HCI paradigms, this should be the ideal context for holistic third wave research as the dichotomy of work and leisure does not exist — except maybe in people with dementia's minds. To gain true experiential insight (third wave), the level of depth pursued during interviews or video analyses is decisive. Observers or interviewers need to go beyond a mere frequency analysis of emotional reactions and pursue an understanding of users' needs or experience with value-seeking techniques such as UX laddering (Abeele & Zaman, 2009) or phenomenological approaches (Prpa et al., 2020).

Regarding the researchers' perspective on experience, the documentation stage of the Proxemo pipeline is diminishing all user behaviour and expressions on emotion categories in a reductionist manner. In contrast, the retrospective interview can be led in a way to holistically understand the observed users' experience. Other than the variation of perspectives between stages of a design process (Blessing & Chakrabarti, 2009; Wurhofer, 2018) Proxemo alternates perspectives within one analytic method (similar to Burmester et al., 2010).

With our choice of both application domains we showcased how Proxemo creates a way of including user groups struggling to communicating their experiences (i.e., people with advanced dementia) or to efficiently include feedback from users whose experience is neglected so far in lieu of performance measures (i.e., air traffic controllers). Following the motivation of Wright and McCarthy (2010), Proxemo may become a facilitator for experience-centred design in these domains: “[T]he real excitement of experience-centred design is [...] to give people the chance to have a richer life, to include people who might otherwise feel excluded and to ensure that everybody has a chance to have their say, especially those who often feel voiceless” [p. 2].

In contrast to Proxemo's purpose which is situated in the third wave, our evaluation of the method draws on concepts and measurements originating from all three waves. The vision behind Proxemo as a coupled human-computer system with shared responsibilities (Fitts, 1951), joining their strengths in emotion recognition and interpretation (human) with precise capturing of large, transferable and synchronous data (computer) for optimal results is a typical picture from the second wave. Research questions regarding this harmonic interaction have been mostly

considered at the design stage where we worked out how observers feed information into the application and how the system then transforms and presents this information (Harrison et al., 2007).

Our user centred iterations of the application as well as descriptive and qualitative evaluations in the field took place within an experiential approach, bearing marks of the third paradigm (Bødker, 2015): we explored how evaluators in the dementia context wore or held the watch, we designed for the least distraction and gained insight on how Proxemo affected the evaluator’s role or changed the character of an evaluation.

Finally, performance measures which we applied when determining the quality criteria of Proxemo are a classical human-factors relict associated with the first wave (Harrison et al., 2007). Respective questions were ‘How efficient can evaluators use Proxemo?’, ‘How thorough and valid can evaluators document emotions?’.

9.2.2 The Relation Between Proxemo and UX

UX methods have long been used on consumer products (i.e., discretionary use) only and have flown under the radar of thorough tests regarding safety-critical aspects as intrusiveness. Recently the demand arose to consider the role of UX in safety-critical domains (Grundgeiger et al., 2020), or even deploy UX methods in design processes for the hospital (Klüber et al., 2020) or ATC (Gramlich et al., 2022). Consequently, UX methods will no longer primarily generate hardware and apps that offer welcome diversions for users of all generations but soon may revolutionise the way workstations and operation theatres are conceptualised. If well-being (hedonic aspects of UX) is as important as performance (pragmatic aspects of UX) in safety-critical domains (Dul et al., 2012) or performance outcomes shall even be increased via improved well-being (Grundgeiger et al., 2020), we need robust and flexible UX evaluation methods for these domains. Ensuring this makes proper meta-evaluations imperative. While summative UX questionnaires are satisfyingly validated (e.g., UEQ), structured observation methods or formative UX methods in general lack this level of testing. We wonder how it can be that user observation is a common practice, yet the methodology has so far barely been evaluated regarding its quality criteria. With Proxemo we present an evaluated method to document emotions in formative UX evaluations and call for meta-evaluations of other formative techniques in future work — especially if they are intended for use in safety-critical domains.

In this work we followed the definition of UX as experiences that result from actual use as defined, for instance, in the revised ISO 9241-11 (ISO, 2018) on two levels: the evaluation of prototypes and the evaluation of Proxemo. (1) Since emotions are by definition evoked through triggers, it is likely that events in timely proximity triggered this emotion. A broader definition of UX extends the experience from the situation to include anticipation of the situation or lasting consequences on quality of life (Hassenzahl, 2010; ISO, 2019; Kujala et al., 2011). When

persons with dementia are not able to communicate verbally, observers can only assume the reason for their expressed emotion. Applied in another context, interpretation is not necessary because workers in safety-critical domains like ATC have the cognitive abilities to answer questions concerning their expectation of a situation or in the aftermath of a critical situation. With Proxemo, emotional experience during actual use or “instant UX” (Wurhofer, 2018) is observed and inferred emotions are documented. In post-usage interview situations those emotions captured as expressed consequences of instant UX may be enriched with notions of “remembered UX” (Wurhofer, 2018). (2) During the meta-evaluation in chapter 7 observers used two different methods in quick succession. This consecutive use makes it difficult to separate the experience from the products in a timely manner and cannot measure longer-lasting experiential qualities after having used either method. We, therefore, must assume that observers reported their experience during actual use when filling in the questionnaires.

9.2.3 A Short Reflection About Emotion

Chapter 2 explained in detail the purpose, formation and bodily symptoms of emotions. Still, when it comes to data collection, it is arguable which aspects are strongest for the definition of an emotion. Are emotions only valid if they produce a human observable or instrumentally measurable physiological reaction? Are emotions only valid if the person experiencing and appraising the emotion is aware of it — most practically to a degree that allows self-report of the emotion?

With respect to content validity, in this work we varied between consistently observed emotions by experts (chapter 7) and self-reported emotions (chapter 8) as ground truth. Since cognitive appraisal and expressive reactions both are crucial parts of emotions (Scherer, 2005), there is no unmistakable theoretically grounded reason for a preference. In the light of both our studies resulting in a thoroughness of Proxemo of about one fourth, it appears to be up to future researchers to decide whether they choose consistently observable emotions or self-reported instances as ground truth. Interestingly, the low validity and thoroughness for concurrent self-report in chapter 8 leave the impression that instant emotional experience and remembered emotional experience are not as closely related with only one tenth of the retrospectively reported emotions overlapping with emotions documented concurrently (thoroughness). We argued how this *memory bias* already identified by other researchers alter instances of emotional self-report with time both in number (Petrie & Precious, 2010) and content (Schwarz, 2007). Future research should dissect whether instant experience and remembered experience (Wurhofer, 2018) are in fact almost distinct, or if this impression is an artefact of the concurrent experience self-report quality suffering under primary task load.

In addition to memory bias which only impacts the report of emotional instances within a user we suggested two further biases that affect the consistency of documented emotions between users

and observers. This non-exhaustive list of three biases between possible operationalisations of emotion documentation is visualised in figure 8.9. *Individual bias* affects the potentially varying comprehension and expression of experienced emotions between individuals, reducing agreement between users and observers (Barrett, 2004). A way to tackle it is either keeping individuals constant or increasing the familiarity between them and establishing a common understanding of emotions. *Attention bias* here describes how the human focus is drawn towards some signals of emotional expression more than others. This affects mostly the observer who has to interpret a user’s facial movements and bodily posture, interactions and utterances. While the observer in the study of chapter 8 only covered one user, the biasing effect likely increases if their attention is divided between signals from multiple users. One approach to reducing attention bias could be to increase the experimental control and, for instance, only make one channel accessible to observers. However, research currently indicates that persons’ ability to recognise emotions varies more between categories (Lewis et al., 2016) and less between channels (Connolly et al., 2020). In addition, observers perform best when all channels are present (Huber & Rathß, in press). Until this trade-off is studied in more detail, practitioners need to walk the fine line between granting observers unlimited access to all useful channels and coping with attentional bias. In our studies on Proxemo’s quality criteria, observers were given access to live-streams or recordings of in-situ audio (including system sounds and user utterances) and video of the interactions. Admittedly, the camera angle was optimised to show the users’ face not the entire body. Users in all studies’ stimulus material were sitting and at least the upper part of their torso was visible. This means, body postures which seem to be equally important for emotion recognition as facial cues (Connolly et al., 2020) could still be partially inferred. The influence of attentional bias on future studies could be reduced by maintaining a high quality signal of all available channels. Indeed, recent findings indicate that the emotion literature’s strong focus on facial expressions for emotion inference may be unreasonable (Huber & Rathß, in press).

Another way to increase validity would have been the deployment of formerly validated stimulus material. As elaborated in chapter 2, deploying established stimuli (e.g., Goeleven et al., 2008) may have increased Proxemo’s validity scores under lab conditions. However, our priority lay upon high ecological validity and outcomes of direct value for practitioners. Hence, ecologically valid cues such as live or video recorded user sessions were more important to us than building and validating a tool that only allows for the recognition of full intensity image stimuli under lab conditions.

It is arguable in how far the emotion categories themselves overlap or are related. Within the sets of emotion categories we chose, emotional reactions of pride and wistfulness overlap with pleasure. Some domain specific emotions are difficult to be classified according to categorical models (Ekman & Friesen, 1971). Wistfulness is particularly interesting as it contains both positive and negative valence — “good memory of a time now sadly over”. With categorical or dimensional models wistfulness would only be explainable through simultaneous coexistence of

opposing categories (Ekman & Friesen, 1971) or dimensions (Russell, 1980) that were elicited by the same trigger — a constellation unthinkable in the classical single-cue-single-reaction experiments. Within the wheel of emotions (Plutchik, 2001) that beyond its weaknesses listed in chapter 2 does allow for dyads of two complementary emotions, wistfulness could be explained as an amalgam of sadness and happiness. For wistfulness, the appraisal component is of great importance. Two events need to be evaluated at a time: (1) the event where the memory was created and (2) the event in which the memory is retrieved. A person’s appraisal process may come to the contradictory conclusion that the memory is pleasant in itself but in the process of reminiscing the awareness arises that this memory is in the past and unlikely to be relived — a sad insight. From the different appraisal outcomes emerge the idiomatic “mixed feelings” of joy and sadness labelled as cross-valence mixed emotions by Watson and Stanton (2017). Fokkinga and Desmet (2012) proposed a classification system solely for mixed emotions which allows the categorisation of wistfulness in even finer granularity. Wistfulness matches their cluster of *different stimuli* defined as emotions “evoked at the same time and by the same announcement but the actual stimuli [being] the different implications of the announcement” [p. 6].

So far, reminiscence has mainly been studied in the context of dementia (but see van Gennip et al., 2015). Ambiguous emotions such as wistfulness show how reminiscence offers an interesting domain to be examined in basic research as it may support the appraisal approach and foster our understanding of emotions. Experience sampling studies by Carstensen et al. (2000, 2011) showed correlations between age and co-occurring emotions of different valence, emphasising the suitability of older adults as user group for studies on mixed emotions.

With respect to the application domain of ATC, looking beyond the negative emotions evoked by high workload is still new ground. Similarly, basic research on emotion and physical responses is mostly dedicated to stressful situations where fear triggers either a fight or a flight response (e.g., Stemmler, 2004). This combination is difficult for ATC where running away is not an adequate option. Instead, we encourage future research to look into experiences of positive or mixed emotions in ATC.

Finally, we argued how cognitive empathy is a critical trait of qualified observers. The etymological roots of the term empathy (from Greek “suffering”) hint towards the negative spectrum of valence. However, in the context of UX, the emphasis of the observers’ task is headed to rejoicing with the users, validating their emotions in an effort to distinct, for example, between pride and pleasure.

9.2.4 A Glance on Automatic Emotion Detection Approaches

Artificial intelligence can already identify situations from video recordings in which users with dementia struggle with water faucets (Taati et al., 2011). However, nuanced emotions that give insights on users’ experience in more complex interaction scenarios cannot yet be automatically

extracted. So far, automatic recognition of emotions via facial recognition or EEG conceptualised for the dementia context cover either only one emotional expression (Rezaei et al., 2020) are not yet working reliably for people with dementia (Parekh et al., 2018; Taati et al., 2019) or have not been tested with this user group (Tseng et al., 2013; Wiratanaya et al., 2007). Approaches that continuously analyse utterances of people with dementia have so far only been used to extract vocal events of agitation (Salekin et al., 2020) or anxiety (Hernandez-Cruz et al., 2019).

Our explorations with EDA indicated that using skin conductance events as ground truth for future work would not be an option due to the high frequency of non-specific peaks which render the data unusable for our purposes. A fusion of different measurements may yield better results than individual data sources alone. For instance, D. Li et al. (2019) combined EEG and facial recognition and achieved concordance correlation coefficients between .62 – .63 which are still of good clinical significance (Cicchetti, 1994) but not as good as Proxemo¹.

Since Fitts’s (1951) enumeration of tasks at which machines are better, sensory functions have clearly outperformed humans. However, humans appear to be still superior in perception and particularly flexibility and judgement as computers’ performance without a human instance (see Proxemo) has not reached satisfactory levels yet and maybe never will (Ellis & Tucker, 2020). The question arises whether the constant comparison to computers is even necessary (Boehner et al., 2007), especially in a subjective domain where we still cannot agree with certainty what counts as valid emotion (see section 9.2.3). For future work we advise against replacing human judgement but rather supporting it. A human observer cued with suggestions based on automatically detected emotional activity could increase effectiveness of emotion detection. Yet, fusing multiple physiological measurements and image based calculations in real time might pose a challenge, especially as bodily reactions typically used in emotion detection (e.g., EDA) are only measurable with an offset of multiple seconds and consumer graded biofeedback equipment is not yet capable of distinguishing emotion categories (Schlör et al., 2020).

9.3 Recommendations for Practitioners

For practitioners who seek to use Proxemo, we aggregated implementation requirements (Matthews et al., 2015) of Proxemo, compiled a short list of recommendations and further materials and summarise the methods’ advantages and limitations. A condensed table of criteria and descriptions helpful for reviewing evaluation and design methods (Stanton et al., 2017) is presented in appendix A.7. A short summary and caveat in advance: Proxemo does not measure UX holistically but is specialised on facilitating the documentation of observed emotions. Users’ emotions are a key component of UX and in the Proxemo pipeline they serve as an anchor to discuss UX in more detail with users once they have the resources to do so (e.g., ATC) or as a clean

¹Assuming that scores of the concordance correlation coefficient and intra-class coefficients merely differ (Carol, 1997) and their scale is roughly comparable with kappa values (Cicchetti, 1994; Landis & Koch, 1977).

documentation baseline for more detailed interpretations when users are not able to verbally communicate (e.g., dementia). During collaborative debriefings, additional mental states as well as their underlying reasons can be inquired about that go beyond observable emotions.

Objectivity. Interpreting other persons' emotions is inherently subjective. Objectivity can be increased though through predefined fixed emotion categories as well as thorough training of the observers. To ensure a high representation of observed users' emotions in the data set, evaluators should be selected or trained (a) regarding their general empathy, (b) their understanding of the context of study, and (c) instructed to pay less attention to their own emotions while coding the inferred emotions of others.

Flexibility. Proxemo is particularly worthwhile when thorough documentation of emotions in context is important and validity can be optimised in the aftermath. Proxemo is interface agnostic as long as the observer can see the emotions. This makes Proxemo far more generalisable than, for example, the LEMtool (Huisman et al., 2013) that was bound to a web interface. When technology or clothing covers emotional expressions as it might be the case for large head mounted displays or surgical masks, observational validity of Proxemo is limited.

Conserving the valuable time of users and evaluators. In section 9.1 above we focused on efficiency from a user centred perspective and will now look consider an economical perspective on efficiency in formative evaluations with Proxemo.

In contrast to unsupervised self-report (e.g., diary studies), one extra observer is necessary. In most settings, however, a user researcher is on site anyway who can take the observer's role of documenting emotions if he or she is familiar with the domain and user group. Where this is not the case an extra observer needs to be scheduled for the observational slot. We showed how the extra effort of an additional observer increases effectiveness and, therefore, at least maintains efficiency. As always, in terms of cost efficiency, user research teams will need to consider individually whether the gain in effectiveness that is synonym to increased knowledge about the users' emotional experience justifies the personnel extra effort.

Reviewing the user experience together with the users on the basis of video recordings grants additional insights but comes at a price. J. Nielsen (1994b) recommends such retrospective testing but states: "The obvious downside is that each test takes at least two times as long, so the method is not suited if the users are highly paid or perform critical work from which they cannot be spared for long" [p. 199]. In relation to the entire session, not just the interaction phase, Peute et al. (2015) report that the retrospective has resulted in an average 72% increase in the length of sessions. When only selectively revisiting the Proxemo annotations in the debriefing, as we did in chapter 5, the additional time effort can be decreased to approximately 30% of the preceding interaction. The question remains whether user researchers' main goal is to minimise effort and

asymptotically strive towards 0 or whether an interdisciplinary review of critical situations with developers, users and researchers is time well spent and should be taken advantage of.

Acceptance in the dementia context. The Proxemo method relies on a technical implementation in order to get accurate time stamps. Recently, software tools for the evaluation of long-term care settings such as the BEAM emerged (Casey et al., 2014) and the well established Dementia Care Mapping method evolved from paper-based documentation to a digital form (Yamamoto et al., 2020). With a shift towards general software support in care setting evaluations, we also expect an increased acceptance for precise documentation tools such as the Proxemo App among staff. The discreet form factor of the Proxemo App running on a smart-watch as well as the placement of a camera far away from interface under evaluation should prevent people with dementia from confusing the prototypes' interface with recording equipment as reported by Gibson et al. (2016).

When to reach out for other methods. Proxemo increases efficiency compared to retrospective reviews of the entire video. When interaction is not time critical and load caused by a secondary task does not matter, concurrent think aloud may produce more insights than the selective debriefing which resembles retrospective think aloud (Peute et al., 2015). Furthermore, unvalidated use of Proxemo data (as we did with people in advanced stages of dementia) is only a choice when users themselves are generally unable to share self-reflection or communicate. Validating observed emotions with users is always preferable to the mere interpretation by (even skilled) observers. Finally, Proxemo's focus is the experience during actual use. The introduction of novel technology may be accompanied by future long-term consequences for the user or other actors. Estimating or measuring such broader impact lies beyond the scope of Proxemo. In ATC, specific modelling techniques are used to predict the results of transformative processes (Blom, 2019). Once a novel technology has been introduced in either domain, ethnographic approaches provide insights in how communications, interactions and strategies in the field changed (e.g., Huber et al., 2020; Mackay, 1999).

Preparing for Proxemo. Prototypical projects of the documentation app for Proxemo are available for watch interfaces on Tizen² and Wear OS³ as well as phones or tablets interfaces running Android⁴. A short implementation history together with perspectives of each version can be found in the appendix, section A.6. Before deploying Proxemo, there are several recommendations requiring consideration:

- When using Proxemo in other contexts than those described in this work, adapting the

²Tizen: <https://github.com/bja-engineering/Proxemo>

³Wear OS: <https://github.com/bja-engineering/EmoMem>

⁴Android: <https://github.com/bja-engineering/ProxemoTab>

categories or defining entirely new sets of emotions is required. Appropriate categories can be derived from literature, existing instruments and own ethnographic studies.

- Validate the set of predefined emotions and their meaning in the context with experts. This is an insight gained from the contrasting experiences of expert users in case studies and non-expert users in the lab-studies of this work. For instance, the emotion category *surprise* had been associated by air traffic controllers with unexpected interface behaviour (see case study in chapter 5). In the lab study with student participants (chapter 8), the interaction did not cater for any surprises and participants made use of the category surprise instead whenever the participants' vehicle was hit by items fired from non player characters in the racing game or in the flight control game, when aircraft entered the sector unexpectedly or got surprisingly close to other aircraft. Participants even suggested the category *relief* which only raised satirical comments when proposed to licensed controllers: "Whoops, I just dropped my pen — I'm really glad nothing severe happened!", or "The whole screen just had been red but a recalculation revealed that in fact everything is fine."
- Maintain a plausible set of emotions. Projects within the same user group may require totally different sets of emotions. For instance, within the dementia context, reminiscence triggers emotions different from those evoked by assistive technology. In ATC, emotions may strongly vary between positions or based on the expected traffic in shifts.
- If possible, always ask users to validate situations rated as critical by observers. Proxemo is not a free ticket to decision-making over persons' heads. Even in the dementia context, design processes should be preferably conducted *with* people rather than *for* people with dementia (Chopra et al., 2021). Residents should be actively engaged as mutual conversation partners (Foley, Pantidi et al., 2019) as long as their communicative abilities allow.
- As with any other method supporting formative UX evaluations, invite multiple and if possible diverse users to comprehend which emotions the interaction triggers. Lotze's (1858) prophecy is still valid today: "As for that qualitative content of sensation, it will always be impossible to decide whether the same sensation corresponds to the same stimuli in different souls."
- Carefully select observers. Towards a detection of valid emotions, chose familiarity with the person and context (users' colleagues or family) over general experience in user research.

9.4 Summary & Outlook

This work contributed a structured emotion observation method for situations that disallow users the concurrent self-report of their emotions. We implemented apps that facilitate the

documentation of emotions and validated both the methods' main quality criteria in lab studies and the feasibility and usefulness in context of two application domains. Proxemo's scores for quality criteria compare well to benchmarks where available. As structured observation methods for emotions are scarce and have not been thoroughly evaluated so far, Proxemo fills a gap in user researchers' method repertory.

It was surprisingly difficult to find meta-evaluations of observational methods conceptually on par as a benchmark for Proxemo. Conceptually closest to the structured observation approach of Proxemo are *behavioural observation scales* which usually require only pen & paper but lack validation studies (Stanton et al., 2017). Therefore, our report of quality criteria for manual note-taking of observed emotions appears to be a bigger contribution than intended. Future research should refocus on the evaluation of the methods practitioners strive to use in their research.

Looking beyond the quality criteria covered in this work, little research on downstream utility has been published since Law (2006). Even though the domain of UX has developed greatly in the last 15 years, practitioners' focus shifted towards the organisation and communication of business ideas (e.g., Osterwalder et al., 2014; Plattner et al., 2009), rather than seeking new ways to formatively comprehend users' experience (but see Holtzblatt & Beyer, 2016). While tool boxes typically promise a boost in knowledge about customers or users, we are not aware of thorough meta-evaluations supporting those claims. Future work could approach this by investigating the downstream utility of novel UX methods. For instance, the formula currently measuring a method's impact solely by the amount of fixed problems (Law, 2006) could be extended to cover positive aspects of UX.

Future paths for Proxemo. Proxemo was originally developed for formative evaluations of interactive technology in the dementia context and later adapted for formative evaluations in simulated ATC. The predefined categorisation embedded in our versions of the Proxemo App limits the generalisability of the meta-evaluations presented here and makes adaptations to the documentation aid a requisite before porting it to other contexts. The concept of documenting observed emotions for people who cannot communicate emotions is transferable. For instance, conventional UX methods are not applicable for children with autism spectrum disorder who face communicative challenges (Mäkelä et al., 2013). Beyond ATC there are further safety critical workplaces as potential application domains where observed users are generally able to communicate but must not be distracted through thinking aloud or interruptive questioning such as pilots, surgeons, or industrial operators. In mixed reality research one of the challenges is to collect data from users with the least possible interruption to their immersion or sense of presence (e.g., Frommel et al., 2015; Wienrich et al., 2018). For evaluation settings where observers' view on facial expressions is not impeded by head worn displays, Proxemo may be the method of choice. For each of the domains listed above, the expected set of emotional responses differs

and evaluators first need to define it before deploying Proxemo. The meta-evaluation presented here focused on reminiscence interventions. When using Proxemo in a new context, emotion categories need to be validated or replaced. Even for the small step from the care home to the lab we needed to leave out the category *agency* and relabelled *general alertness* as “interest” and *wistfulness* as “nostalgia” to better suit the target audience. The quality criteria may vary when Proxemo is adapted to and implemented in another context.

Regarding the third stage of the Proxemo pipeline, we used Proxemo annotations already as navigation and interpretation aid for video analysis. The efficiency effect of Proxemo as pre-filter increases with the cost of video analysis methods. For instance, Proxemo annotations could cater for a pre-selection of few critical instances in a long video recording where researchers code users’ emotions from facial expressions (Ekman & Rosenberg, 2005). Similar to debriefings with expensive users, Proxemo annotations facilitate an efficient navigation and selection by individual emotion categories deemed relevant for detailed inspection.

Beyond specific application domains, we seek to deepen the understanding of constraints in proxy evaluations as well as observers’ needs. Qualitative data from our last study (chapter 8) indicates how participants found it difficult during the retrospective review to simultaneously watch their own mimic and the happenings of the game screen capture. This poses several questions such as how do observers handle the impression overload? In the case of emotionally ambiguous situations: Is a user or their context more important for emotion classification if only one chance is given for a critical observation in real time? How do observers divide their attention between different sources of information? Therefore, would additional information support or distract human evaluators if presented simultaneously? Truly understanding an emotional event requires context knowledge which facial expression tracking does not offer (Ellis & Tucker, 2020). We expect that observers need both sources to maintain documentation quality and plan to answer those research questions in lab studies. As common video conference platforms allow streams of the webcam and screen sharing in parallel (compare also the setup of the study using multiple rooms as described in chapter 8), an implication for practitioners is that Proxemo is suitable for remote testing which has gained in great popularity during the COVID-19 pandemic.

References

- Abdollahi, H., Mollahosseini, A., Lane, J. T. & Mahoor, M. H. (2017). A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 541–546. <https://doi.org/10.1109/HUMANOIDS.2017.8246925>
- Abeele, V. V. & Zaman, B. (2009). Laddering the user experience. *User Experience Evaluation Methods in Product Development (UXEM'09)-Workshop*.
- Agarwal, A. & Meyer, A. (2009). Beyond usability: evaluating emotional response as an integral part of the user experience. *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 2919–2930. <https://doi.org/10.1145/1520340.1520420>
- Ahlstrom, U. & Arend, L. (2005). Color Usability on Air Traffic Control Displays. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(1), 93–97. <https://doi.org/10.1177/154193120504900121>
- Aizpurua, A., Harper, S. & Vigo, M. (2016). Exploring the relationship between web accessibility and user experience. *International Journal of Human-Computer Studies*, 91, 13–23. <https://doi.org/10.1016/j.ijhcs.2016.03.008>
- Alarcão, S. M. (2017). Reminiscence therapy improvement using emotional information. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 561–565. <https://doi.org/10.1109/ACII.2017.8273655>
- Alarcão, S. M. & Fonseca, M. J. (2019). Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing*, 10(3), 374–393. <https://doi.org/10.1109/TAFFC.2017.2714671>
- Algar, K., Woods, R. T. & Windle, G. (2014). Measuring the quality of life and well-being of people with dementia: A review of observational measures. *Dementia*, 15(4), 832–857. <https://doi.org/10.1177/1471301214540163>
- Allen-Walker, L. & Beaton, A. A. (2015). Empathy and perception of emotion in eyes from the FEEST/Ekman and Friesen faces. *Personality and Individual Differences*, 72, 150–154. <https://doi.org/10.1016/j.paid.2014.08.037>

- Alm, N., Dye, R., Gowans, G., Campbell, J., Astell, A. & Ellis, M. (2003). Designing an interface usable by people with dementia. *ACM SIGCAPH Computers and the Physically Handicapped*, 156–157.
- Alshammari, T., Alhadreti, O. & Mayhew, P. (2015). When to ask participants to think aloud: A comparative study of concurrent and retrospective think-aloud methods. *International Journal of Human Computer Interaction*, 6(3), 48–64.
- Analytics, S. (2021). Strategy Analytics: Global Smartwatch Shipments Leap 47 Percent to Pre-Pandemic Growth Levels in Q2 2021. Retrieved September 10, 2021, from <https://news.strategyanalytics.com/press-releases/press-release-details/2021/Strategy-Analytics-Global-Smartwatch-Shipments-Leap-47-Percent-to-Pre-Pandemic-Growth-Levels-in-Q2-2021/default.aspx>
- Anderson, S. P. (2011). *Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences, Portable Document*. Pearson Education.
- Astell, A. J., Smith, S. K., Potter, S. & Preston-Jones, E. (2018). Computer Interactive Reminiscence and Conversation Aid groups — Delivering cognitive stimulation with technology. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 4(1), 481–487. <https://doi.org/10.1016/j.trci.2018.08.003>
- Averty, P., Athenes, S., Collet, C. & Dittmar, A. (2002). Evaluating a new index of mental workload in real ATC situation using psychophysiological measures. *Proceedings. The 21st Digital Avionics Systems Conference*, 2, 7A4–7A4. <https://doi.org/10.1109/DASC.2002.1052916>
- Badea, A. C. (2021). *Gamification: Improving Supervisory Control Performance in Highly Automated Air Traffic Control* (Thesis). Technische Universiteit Delft, Netherlands. <https://repository.tudelft.nl/islandora/object/uuid:d679bb36-9d6f-42cf-b470-7588138606bc>
- Balconi, M., Cotelli, M., Brambilla, M., Manenti, R., Cosseddu, M., Premi, E., Gasparotti, R., Zanetti, O., Padovani, A. & Borroni, B. (2015). Understanding Emotions in Frontotemporal Dementia: The Explicit and Implicit Emotional Cue Mismatch. *Journal of Alzheimer's Disease*, 46, 211–225. <https://doi.org/10.3233/JAD-142826>
- Baldwin, M., Biernat, M. & Landau, M. J. (2015). Remembering the real me: Nostalgia offers a window to the intrinsic self. *Journal of Personality and Social Psychology*, 108(1), 128–147. <https://doi.org/10.1037/a0038033>
- Bandyopadhyay, A., Sarkar, S., Mukherjee, A., Bhattacharjee, S. & Basu, S. (2020). Identifying emotional Facial Expressions in Practice: A Study on Medical Students. *Indian Journal of Psychological Medicine*, 43(1), 51–57. <https://doi.org/10.1177/0253717620936783>
- Bargas-Avila, J. A. & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. <https://doi.org/10.1145/1978942.1979336>

- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251. <https://doi.org/10.1017/S0021963001006643>
- Barrett, L. F. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. *Journal of Personality and Social Psychology*, *87*(2), 266–281. <https://doi.org/10.1037/0022-3514.87.2.266>
- Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, *10*(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. <https://doi.org/10.1177/1529100619832930>
- Batty, M. & Taylor, M. J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*, *17*(3), 613–620. [https://doi.org/10.1016/S0926-6410\(03\)00174-5](https://doi.org/10.1016/S0926-6410(03)00174-5)
- Baumeister, R. F. & Leary, M. R. (1997). Writing narrative literature reviews. *Review of general psychology*, *1*(3), 311–320.
- Baumer, E. P. & Tomlinson, B. (2011). Comparing activity theory with distributed cognition for video analysis: beyond “kicking the tires”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 133–142.
- Bayne, H. B. & Hankey, M. S. (2020). Exploring Cognitive Empathy: Further Validation of the Empathic Counselor Response Scale and Application to Practice. *The Journal of Humanistic Counseling*, *59*(3), 219–239. <https://doi.org/10.1002/johc.12146>
- Bejan, A., Gündogdu, R., Butz, K., Müller, N., Kunze, C. & König, P. (2017). Using multimedia information and communication technology (ICT) to provide added value to reminiscence therapy for people with dementia. *Zeitschrift für Gerontologie und Geriatrie*, *51*(1), 9–15. <https://doi.org/10.1007/s00391-017-1347-7>
- Bejan, A., Wieland, M., Murko, P. & Kunze, C. (2018). A Virtual Environment Gesture Interaction System for People with Dementia. *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, 225–230. <https://doi.org/10.1145/3197391.3205440>
- Bentley, F. & Murray, J. (2016). Understanding Video Rewatching Experiences. *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, 69–75. <https://doi.org/10.1145/2932206.2932213>
- Beveridge, A. (2002). Time to abandon the subjective–objective divide? *Psychiatric Bulletin*, *26*(3), 101–103. <https://doi.org/10.1192/pb.26.3.101>

- Bhattacharya, S. & Gupta, M. (2019). A Survey on: Facial Emotion Recognition Invariant to Pose, Illumination and Age. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 1–6. <https://doi.org/10.1109/ICACCP.2019.8883015>
- Biopac Systems. (2015). EDA Introduction Guide. Retrieved January 1, 2022, from <https://www.biopac.com/wp-content/uploads/EDA-Guide.pdf>
- Blessing, L. T. & Chakrabarti, A. (2009). *DRM: A design reseach methodology*. Springer.
- Blom, H. (2019). Symbolic-numeric methods in reasoning about the design of future air traffic management. *Proceedings of the Fifth International Workshop on Symbolic-Numeric methods for Reasoning about CPS and IoT*, 1–2. <https://doi.org/10.1145/3313149.3313373>
- Bødker, S. (2015). Third-wave HCI, 10 years later—participation and sharing. *interactions*, 22(5), 24–31. <https://doi.org/10.1145/2804405>
- Boehner, K., DePaula, R., Dourish, P. & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, 65(4), 275–291.
- Bölte, S. (2005). *Reading mind in the eyes test - Erwachsenenversion* (Report).
- Bornoe, N. & Stage, J. (2017). Active involvement of software developers in usability engineering: two small-scale case studies. *IFIP Conference on Human-Computer Interaction*, 159–168.
- Bortz, J. & Schuster, C. (2011). *Statistik für Human-und Sozialwissenschaftler: Limitierte Sonderausgabe [Statistics for Humanities and Social Sciences: Limited Special Edition]*. Springer-Verlag.
- Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E. & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, 49(8), 1017–1034. <https://doi.org/10.1111/j.1469-8986.2012.01384.x>
- Bouvier, D. J., Hinz, J. G. & Schmidt, C. A. (2016). Pilot Study: User Acceptance of a Virtual Coach in a Mirror by Elderly Persons with Dementia. *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, Article 89. <https://doi.org/10.1145/2910674.2935843>
- Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1), 49–59.
- Breves, P. & Schramm, H. (2021). Bridging psychological distance: The impact of immersive media on distant and proximal environmental issues. *Computers in Human Behavior*, 115, 106606. <https://doi.org/10.1016/j.chb.2020.106606>
- Brill, M. & Schwab, F. (2020). T-pattern analysis and spike train dissimilarity for the analysis of structure in blinking behavior. *Physiology & Behavior*, 227, 113163. <https://doi.org/10.1016/j.physbeh.2020.113163>

- Brown, E. L., Agronin, M. E. & Stein, J. R. (2020). Interventions to enhance empathy and person-centered care for individuals with dementia: a systematic review. *Research in gerontological nursing*, 13(3), 158–168.
- Brunett, G., Eibl, M., Hamker, F., Ohler, P. & Protzel, P. (2018). *Schlussbericht StayCentered – Methodenbasis eines Assistenzsystems für Centerlotsen (MACeLot) [Final report StayCentered - Methodological bases for an assistive system for centre controllers]* (Report). Technische Universität Chemnitz.
- Bruun, A., Law, E. L.-C., Heintz, M. & Eriksen, P. S. (n.d.). Asserting Real-Time Emotions through Cued-Recall: Is it Valid? *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, Article 37. <https://doi.org/10.1145/2971485.2971516>
- Burmester, M., Mast, M., Jäger, K. & Homans, H. (2010). Valence method for formative evaluation of user experience. *Proceedings of the 8th ACM conference on Designing Interactive Systems*, 364–367.
- Buxbaum, J. (2019). Erkenntnisse aus dem Projekt "StayCentered - Methodenbasis eines Assistenzsystems für Centerlotsen" [Insights from the project "StayCentered" - Methodological bases for an assistive system for centre controllers]. *Innovation im Fokus*, 1. https://www.dfs.de/dfs_homepage/de/Flugsicherung/Forschung%20&%20Entwicklung/Servicebereich/Forschungszeitschrift%20%E2%80%9EInnovation%20im%20Fokus%E2%80%9C/Innovation%20im%20Fokus%201901%20-%20Artikel%20StayCentered.pdf
- Byers, J. C. (1989). Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? *Advances in Industrial Ergonomics and Safety I*.
- Byrt, T., Bishop, J. & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5), 423–429.
- Card, S. K., Moran, T. P. & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7), 396–410.
- Carol, A. E. N. (1997). A Note On "A Concordance Correlation Coefficient to Evaluate Reproducibility". *Biometrics*, 53(4), 1503–1507. <https://doi.org/10.2307/2533516>
- Carré, A., Stefaniak, N., D'ambrosio, F., Bensalah, L. & Besche-Richard, C. (2013). The Basic Empathy Scale in Adults (BES-A): Factor structure of a revised form. *Psychological assessment*, 25(3), 679.
- Carstensen, L. L., Pasupathi, M., Mayr, U. & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of personality and social psychology*, 79(4), 644.
- Carstensen, L. L., Turan, B., Scheibe, S., Ram, N., Ersner-Hershfield, H., Samanez-Larkin, G. R., Brooks, K. P. & Nesselroade, J. R. (2011). Emotional experience improves with age: evidence based on over 10 years of experience sampling. *Psychology and aging*, 26(1), 21–33. <https://doi.org/10.1037/a0021285>

- Caruelle, D., Gustafsson, A., Shams, P. & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research*, *104*, 146–160. <https://doi.org/10.1016/j.jbusres.2019.06.041>
- Casali, J. G. & Wierwille, W. W. (1983). A Comparison of Rating Scale, Secondary-Task, Physiological, and Primary-Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load. *Human Factors*, *25*(6), 623–641. <https://doi.org/10.1177/001872088302500602>
- Casali, J. G. & Wierwille, W. W. (1984). On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics*, *27*(10), 1033–1050. <https://doi.org/10.1080/00140138408963584>
- Casey, A.-N., Low, L.-F., Goodenough, B., Fletcher, J. & Brodaty, H. (2014). Computer-Assisted Direct Observation of Behavioral Agitation, Engagement, and Affect in Long-Term Care Residents. *Journal of the American Medical Directors Association*, *15*(7), 514–520. <https://doi.org/10.1016/j.jamda.2014.03.006>
- Cassé-Perrot, C., Fakra, E., Jouve, E. & Blin, O. (2007). Conceptualisation et validation factorielle d'un questionnaire mesurant le profil émotionnel : Emotional State Questionnaire (ESQ) [Conceptualisation and validation of the "Emotional State Questionnaire (ESQ)": evaluation of an emotional profile]. *Encephale*, *33*(2), 169–78. [https://doi.org/10.1016/s0013-7006\(07\)91547-x](https://doi.org/10.1016/s0013-7006(07)91547-x)
- Chang, W.-L., Šabanovic, S. & Huber, L. (2014). Observational study of naturalistic interactions with the socially assistive robot PARO in a nursing home. *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 294–299. <https://doi.org/10.1109/ROMAN.2014.6926268>
- Charmaz, K. & Mitchell, R. G. (2001). Grounded Theory in Ethnography. In P. Atkinson, A. Coffey, S. Delamont, J. Lofland & L. Lofland (Eds.), *Handbook of Ethnography* (pp. 160–174). SAGE Publications Ltd. <https://doi.org/10.4135/9781848608337>
- Chawana, T. & Adebesein, F. (2021). The current state of measuring return on investment in user experience design. *South African Computer Journal*, *33*(1), 22–36. <https://doi.org/10.18489/sacj.v33i1.950>
- Chen, Y.-H., Dammers, J., Boers, F., Leiberg, S., Edgar, J. C., Roberts, T. P. L. & Mathiak, K. (2009). The temporal dynamics of insula activity to disgust and happy facial expressions: A magnetoencephalography study. *NeuroImage*, *47*(4), 1921–1928. <https://doi.org/10.1016/j.neuroimage.2009.04.093>
- Chopra, S., Dixon, E., Ganesh, K., Pradhan, A., Radnofsky, M. L. & Lazar, A. (2021). Designing for and with People with Dementia using a Human Rights-Based Approach. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 44). Association for Computing Machinery. <https://doi.org/10.1145/3411763.3443434>

- Chu, M.-T., Khosla, R., Khaksar, S. M. S. & Nguyen, K. (2017). Service innovation through social robot engagement to improve dementia care quality. *Assistive Technology*, 29(1), 8–18. <https://doi.org/10.1080/10400435.2016.1171807>
- Chung, J. C. C. (2009). An intergenerational reminiscence programme for older adults with early dementia and youth volunteers: values and challenges. *Scandinavian Journal of Caring Sciences*, 23(2), 259–264. <https://doi.org/10.1111/j.1471-6712.2008.00615.x>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4), 284.
- Cockburn, A., Quinn, P. & Gutwin, C. (2015). Examining the Peak-End Effects of Subjective Experience. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 357–366. <https://doi.org/10.1145/2702123.2702139>
- Cockburn, A., Quinn, P. & Gutwin, C. (2017). The effects of interaction sequencing on user experience and preference. *International Journal of Human-Computer Studies*, 108, 89–104. <https://doi.org/10.1016/j.ijhcs.2017.07.005>
- Cohen, G. D., Firth, K. M., Biddle, S., Lloyd Lewis, M. J. & Simmens, S. (2008). The First Therapeutic Game Specifically Designed and Evaluated for Alzheimer’s Disease. *American Journal of Alzheimer’s Disease & Other Dementias*, 23(6), 540–551. <https://doi.org/10.1177/1533317508323570>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37–46.
- Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, 1(3), 98–101.
- Cohen-Mansfield, J., Dakheel-Ali, M., Jensen, B., Marx, M. S. & Thein, K. (2012). An analysis of the relationships among engagement, agitated behavior, and affect in nursing home residents with dementia. *International Psychogeriatrics*, 24(5), 742–752. <https://doi.org/10.1017/S1041610211002535>
- Colibaba, A., Colibaba, S., Gheorghiu, I., Ursa, O., Colibaba, C. & Ionel, A. (2015). The digital timelines course to maintain the quality of life and help people with memory loss: Multimedia applications for medical and healthcare education and e-learning. *2015 E-Health and Bioengineering Conference (EHB)*, 1–4. <https://doi.org/10.1109/EHB.2015.7391485>
- Connolly, H. L., Lefevre, C. E., Young, A. W. & Lewis, G. J. (2020). Emotion recognition ability: Evidence for a supramodal factor and its links to social cognition. *Cognition*, 197, 104166. <https://doi.org/https://doi.org/10.1016/j.cognition.2019.104166>
- Conte, S., Brenna, V., Ricciardelli, P. & Turati, C. (2018). The nature and emotional valence of a prime influences the processing of emotional faces in adults and children. *International Journal of Behavioral Development*, 42(6), 554–562. <https://doi.org/10.1177/0165025418761815>

- Conversy, S., Gaspard-Boulinç, H., Chatty, S., Valès, S., Dupré, C. & Ollagnon, C. (2011). Supporting air traffic control collaboration with a TableTop system. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 425–434. <https://doi.org/10.1145/1958824.1958891>
- Cook, A. (2007). *European air traffic management: principles, practice, and research*. Ashgate Publishing, Ltd.
- Cowen, A. S. & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38). <https://doi.org/10.1073/pnas.1702247114>
- Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H. & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841), 251–257. <https://doi.org/10.1038/s41586-020-3037-7>
- Critchley, M. (1964). The neurology of psychotic speech. *The British Journal of Psychiatry*, 110(466), 353–364.
- Cruz-Sandoval, D., Morales-Tellez, A., Sandoval, E. B. & Favela, J. (2020). A Social Robot as Therapy Facilitator in Interventions to Deal with Dementia-Related Behavioral Symptoms. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 161–169. <https://doi.org/10.1145/3319502.3374840>
- Cruz-Sandoval, D., Penaloza, C. I., Favela, J. & Castro-Coronel, A. P. (2018). Towards Social Robots that Support Exercise Therapies for Persons with Dementia. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 1729–1734. <https://doi.org/10.1145/3267305.3267539>
- Csikszentmihalyi, M. & Larson, R. (2014). *Flow and the foundations of positive psychology*. Springer. <https://doi.org/10.1007/978-94-017-9088-8>
- Czech, E., Shibasaki, M., Tsuchiya, K., Peiris, R. L. & Minamizawa, K. (2020). Discovering Narratives: Multi-sensory Approach Towards Designing with People with Dementia. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.
- Dang, J., King, K. M. & Inzlicht, M. (2020). Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends in cognitive sciences*, 24(4), 267–269. <https://doi.org/10.1016/j.tics.2020.01.007>
- Dan-Glauser, E. S. & Gross, J. J. (2015). The temporal dynamics of emotional acceptance: Experience, expression, and physiology. *Biological Psychology*, 108, 1–12. <https://doi.org/10.1016/j.biopsycho.2015.03.005>
- Darwin, C. (1872). *The expression of the emotions in man and animals by Charles Darwin*. John Murray.

- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1), 113.
- De Raadt, A., Warrens, M. J., Bosker, R. J. & Kiers, H. A. L. (2019). Kappa Coefficients for Missing Data. *Educational and Psychological Measurement*, 79(3), 558–576. <https://doi.org/10.1177/0013164418823249>
- Demir, E., Desmet, P. M. & Hekkert, P. (2009). Appraisal patterns of emotions in human-product interaction. *International Journal of Design*, 3(2).
- Desmet, P. & Fokkinga, S. (2020). Beyond Maslow's pyramid: introducing a typology of thirteen fundamental needs for human-centered design. *Multimodal Technologies and Interaction*, 4(3), 38.
- Desmet, P. & Hekkert, P. (2007). Framework of product experience. *International journal of design*, 1(1), 57–66.
- Desmet, P., Overbeeke, K. & Tax, S. (2001). Designing products with added emotional value: Development and application of an approach for research through design. *The design journal*, 4(1), 32–47.
- Dewey, J. (1929). *Experience and nature*. George Allan & Unwin.
- Dewey, J. (1934). *Art as experience*. Putnam.
- Dewing, J. (2007). Participatory research: A method for process consent with persons who have dementia. *Dementia*, 6(1), 11–25. <https://doi.org/10.1177/1471301207075625>
- Doble, N. A. & Hansman, R. J. (2004). Experimental evaluation of portable electronic flight progress strips. *The 23rd Digital Avionics Systems Conference (IEEE Cat. No.04CH37576)*, 1, 5C4.1–5C4.11. <https://doi.org/10.1109/DASC.2004.1391340>
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und evaluation [Research methods and evaluation]*. Springerverlag. <https://doi.org/10.1007/978-3-642-41089-5>
- Dourish, P. (2004). *Where the action is: the foundations of embodied interaction*. MIT press.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.
- Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J. W., de Gelder, B., Deonna, J., de Sousa, R., ... Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*, 5(7), 816–820. <https://doi.org/10.1038/s41562-021-01130-8>
- Dul, J. (2021). *Revisiting the future of ergonomics; Three triennials later* [Panel contribution at IEA2021: 21st Triennial Congress of the International Ergonomics Association].
- Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W. S., Wilson, J. R. & van der Doelen, B. (2012). A strategy for human factors/ergonomics: developing the discipline and profession. *Ergonomics*, 55(4), 377–395.

- Edmeads, J. & Metatla, O. (2019). Designing for Reminiscence with People with Dementia. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper LBW1721. <https://doi.org/10.1145/3290607.3313059>
- Eggemeier, F. T., Crabtree, M. S. & LaPointe, P. A. (1983). The Effect of Delayed Report on Subjective Ratings of Mental Workload. *Proceedings of the Human Factors Society Annual Meeting*, 27(2), 139–143. <https://doi.org/10.1177/154193128302700205>
- Eggemeier, F. T., Wilson, G. F., Kramer, A. F. & Damos, D. L. (1991). Workload assessment in multi-task environments. *Multiple-task performance*, 207–216.
- Eickers, G. & Prinz, J. (2020). Emotion Recognition as a Social Skill. *The Routledge Handbook of Philosophy of Skill And Expertise* (pp. 347–361). Routledge.
- Ekman, P. & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, 3(4), 364–370. <https://doi.org/10.1177/1754073911410740>
- Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 124–129.
- Ekman, P., Friesen, W. V. & Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of personality and social psychology*, 49(5), 1416. <https://doi.org/10.1037/0022-3514.49.5.1416>
- Ekman, P. & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)* (2nd ed.). Oxford University Press.
- Elfenbein, H. A. & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2), 203–235. <https://doi.org/10.1037/0033-2909.128.2.203>
- Elliott, R., Bohart, A. C., Watson, J. C. & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, 48(1), 43–49. <https://doi.org/10.1037/a0022187>
- Ellis, D. (2018). Social media, emoticons and process. *Affect and Social Media: Emotion, Meditation, Anxiety and Contagion*. Maryland: Rowman & Littlefield, 19–25.
- Ellis, D. & Tucker, I. (2020). *Emotion in the Digital Age: Technologies, Data and Psychosocial Life*. Routledge.
- Ellsworth, P. C. & Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of affective sciences*. (pp. 572–595). Oxford University Press.
- Engeström, Y. (2015). *Learning by expanding*. Cambridge University Press.
- Eugenio, B. D. & Glass, M. (2004). The Kappa Statistic: A Second Look. *Computational Linguistics*, 30(1), 95–101. <https://doi.org/10.1162/089120104773633402>
- Ewald, O. (1908). German Philosophy in 1907. *The Philosophical Review*, 17(4), 400–426. <https://doi.org/10.2307/2177913>

- Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Feng, Y., Barakova, E. I., Yu, S., Hu, J. & Rauterberg, G. W. M. (2020). Effects of the Level of Interactivity of a Social Robot and the Response of the Augmented Reality Display in Contextual Interactions of People with Dementia. *Sensors*, *20*(13), 3771. <https://www.mdpi.com/1424-8220/20/13/3771>
- Feng, Y., Yu, S., van de Mortel, D., Barakova, E., Hu, J. & Rauterberg, M. (2019). LiveNature: Ambient Display and Social Robot-Facilitated Multi-Sensory Engagement for People with Dementia. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1321–1333. <https://doi.org/10.1145/3322276.3322331>
- Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P. & Baron-Cohen, S. (2013). Test-retest reliability of the ‘Reading the Mind in the Eyes’ test: a one-year follow-up study. *Molecular autism*, *4*(1), 1–6.
- Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system* (Report). National Research Council, Division of Anthropology and Psychology, Committee on Aviation Psychology.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin*, *51*(4), 327.
- Fokkinga, S. & Desmet, P. (2012). Meaningful mix or tricky conflict? A categorisation of mixed emotions and their usefulness for design. *Out of Control: Proceedings of the 8th International Conference on Design and Emotion*. <http://resolver.tudelft.nl/uuid:dca821fedb93-4bdb-b783-588a098d55db>
- Foley, S., Pantidi, N. & McCarthy, J. (2019). Care and Design: An Ethnography of Mutual Recognition in the Context of Advanced Dementia. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 610. <https://doi.org/10.1145/3290605.3300840>
- Foley, S., Welsh, D., Pantidi, N., Morrissey, K., Nappey, T. & McCarthy, J. (2019). Printer Pals: Experience-Centered Design to Support Agency for People with Dementia. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 404. <https://doi.org/10.1145/3290605.3300634>
- Forlizzi, J. & Battarbee, K. (2004). Understanding experience in interactive systems. *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, 261–268. <https://doi.org/10.1145/1013115.1013152>
- Fowler, R. D. (1969). An Overview of Human Factors in Europe. *Human Factors*, *11*(1), 91–94. <https://doi.org/10.1177/001872086901100113>
- Franz, R. L., Neves, B. B., Epp, C. D., Baecker, R. & Wobbrock, J. O. (2019). Why and How Think-Alouds with Older Adults Fail: Recommendations from a Study and Expert Interviews. In S. Sayago (Ed.), *Perspectives on Human-Computer Interaction Research with*

- Older People* (pp. 217–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-06076-3_14
- Fredrickson, B. L. & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55. <https://doi.org/10.1037/0022-3514.65.1.45>
- Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1).
- Frommel, J., Rogers, K., Brich, J., Besserer, D., Bradatsch, L., Ortinau, I., Schabenberger, R., Riemer, V., Schrader, C. & Weber, M. (2015). Integrated Questionnaires: Maintaining Presence in Game Environments for Self-Reported Data Acquisition. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 359–368. <https://doi.org/10.1145/2793107.2793130>
- Gall, D., Preßler, J., Hurtienne, J. & Latoschik, M. E. (2020). Self-organizing knowledge management might improve the quality of person-centered dementia care: A qualitative study. *International Journal of Medical Informatics*, *139*, 104132. <https://doi.org/10.1016/j.ijmedinf.2020.104132>
- Gentsch, K., Grandjean, D. & Scherer, K. R. (2014). Coherence explored between emotion components: Evidence from event-related potentials and facial electromyography. *Biological Psychology*, *98*, 70–81. <https://doi.org/10.1016/j.biopsycho.2013.11.007>
- Gertler, J., Novotny, S., Poppe, A., Chung, Y. S., Gross, J. J., Pearson, G. & Stevens, M. C. (2020). Neural correlates of non-specific skin conductance responses during resting state fMRI. *NeuroImage*, *214*, 116721. <https://doi.org/10.1016/j.neuroimage.2020.116721>
- Gibson, A., McCauley, C., Mulvenna, M., Ryan, A., Laird, L., Curran, K., Bunting, B., Ferry, F. & Bond, R. (2016). Assessing Usability Testing for People Living with Dementia. *Proceedings of the 4th Workshop on ICTs for Improving Patients Rehabilitation Research Techniques*, 25–31. <https://doi.org/10.1145/3051488.3051492>
- Gilfoyle, M., Krul, J. & Oremus, M. (2021). Developing practice standards for engaging people living with dementia in product design, testing, and commercialization - a case study. *Assist Technol*, 1–9. <https://doi.org/10.1080/10400435.2021.1968069>
- Gill, T. M. & Feinstein, A. R. (1994). A critical appraisal of the quality of quality-of-life measurements. *Jama*, *272*(8), 619–626.
- Goeleven, E., De Raedt, R., Leyman, L. & Verschuere, B. (2008). The Karolinska directed emotional faces: a validation study. *Cognition and emotion*, *22*(6), 1094–1118. <https://doi.org/10.1080/02699930701626582>
- Golland, Y., Hakim, A., Aloni, T., Schaefer, S. & Levit-Binnun, N. (2018). Affect dynamics of facial EMG during continuous emotional experiences. *Biological Psychology*, *139*, 47–58. <https://doi.org/10.1016/j.biopsycho.2018.10.003>

- Gooch, D., Mehta, V., Price, B., McCormick, C., Bandara, A., Bennaceur, A., Bannasar, M., Stuart, A., Clare, L. & Levine, M. (2020). How are you feeling? Using Tangibles to Log the Emotions of Older Adults. *Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI'20)*, (In Press).
- Goodman, E., Kuniavsky, M. & Moed, A. (2012). *Observing the user experience: A practitioner's guide to user research*. Elsevier.
- Google, S. (2021). Top publications: Human Computer Interaction. Retrieved November 16, 2021, from https://scholar.google.com/citations?view_op=top_venues&hl=en&venue=0QObN5JHY_MJ.2021&vq=eng_humancomputerinteraction
- Gowans, G., Campbell, J., Alm, N., Dye, R., Astell, A. & Ellis, M. (2004). Designing a multimedia conversation aid for reminiscence therapy in dementia care environments. *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, 825–836.
- Gramlich, J., Pauli, S., Huber, S., Baur, C. & Hurtienne, J. (2022). Fin, Whale, Coin and Flatterer: Exploring Tangibles for Air Traffic Control. *TEI 2022*, (In Press). <https://doi.org/10.1145/3490149.3502260>
- Grandjean, D. & Scherer, K. R. (2008). Unpacking the cognitive architecture of emotion processes. *Emotion*, 8(3), 341–351. <https://doi.org/10.1037/1528-3542.8.3.341>
- Group, I. D. (2021). The Global Smartphone Market Grew 13.2% in the Second Quarter Despite Supply Concerns and Vendor Shakeups, According to IDC. Retrieved September 10, 2021, from <https://www.idc.com/getdoc.jsp?containerId=prUS48120021>
- Grundgeiger, T., Hurtienne, J. & Happel, O. (2020). Why and How to Approach User Experience in Safety-Critical Domains: The example of healthcare. *Human Factors*. <https://doi.org/10.1177/0018720819887575>.
- Grundgeiger, T., Sanderson, P. M. & Dismukes, R. K. (2014). Prospective memory in complex sociotechnical systems. *Zeitschrift für Psychologie*, 222(2), 100–109.
- Gündogdu, R., Bejan, A., Kunze, C. & Wölfel, M. (2017). Activating people with dementia using natural user interface interaction on a surface computer. *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 386–394.
- Gunes, H. & Hung, H. (2016). Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block. *Image and Vision Computing*, 55, 6–8. <https://doi.org/10.1016/j.imavis.2016.03.013>
- Guntuku, S. C., Li, M., Tay, L. & Ungar, L. H. (2019). Studying cultural differences in emoji usage across the east and the west. *Proceedings of the International AAAI Conference on Web and Social Media*, 13, 226–235.
- Hagemann, K., Slotty, M. & Albrecht, T. (2020). Speech Recognition for ATCO assistance. *Innovation im Fokus*, 1. https://www.dfs.de/dfs_homepage/de/Flugsicherung/Forschung%20&%20Entwicklung/Servicebereich/Forschungszeitschrift%20%E2%80

- 9EInnovation%20im%20Fokus%E2%80%9C/Innovation%20im%20Fokus%201901%20-%20Artikel%20StayCentered.pdf
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23.
- Hamada, T., Naganuma, M., Kagawa, Y., Hashimoto, T., Onari, H. & Yoneoka, T. (2016). Study on transition of elderly people's reactions in robot therapy. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 431–432. <https://doi.org/10.1109/HRI.2016.7451791>
- Hammar, L. M., Emami, A., Götell, E. & Engström, G. (2011). The impact of caregivers' singing on expressions of emotion and resistance during morning care situations in persons with dementia: an intervention in dementia care. *Journal of Clinical Nursing*, 20(7-8), 969–978. <https://doi.org/10.1111/j.1365-2702.2010.03386.x>
- Harper, D. (2002). Online Etymology Dictionary: empathy (n.) Retrieved February 15, 2021, from <https://www.etymonline.com/search?q=empathy>
- Harrison, S., Tatar, D. & Sengers, P. (2007). The three paradigms of HCI. *Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA*, 1–18.
- Hartson, R., Andre, T. S. & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International journal of human-computer interaction*, 13(4), 373–410.
- Hartson, R. & Pyla, P. (2018). *The UX book: Agile UX design for a quality user experience*. Morgan Kaufmann.
- Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, 13(4), 481–499.
- Hassenzahl, M. (2010). Experience design: Technology for all the right reasons. *Synthesis lectures on human-centered informatics*, 3(1), 1–95.
- Hassenzahl, M., Burmester, M. & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttrakDiff: A questionnaire for measuring perceived hedonic and pragmatic quality]. *Mensch & Computer 2003* (pp. 187–196). Springer.
- Hassenzahl, M., Diefenbach, S. & Göritz, A. (2010). Needs, affect, and interactive products—Facets of user experience. *Interacting with computers*, 22(5), 353–362. <https://doi.org/10.1016/j.intcom.2010.04.002>
- Hassenzahl, M. & Tractinsky, N. (2006). User experience – a research agenda. *Behaviour and Information Technology*, 25(2), 91–97.
- Hassenzahl, M., Platz, A., Burmester, M. & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 201–208.

- Hattink, B., Droes, R.-M., Sikkes, S., Oostra, E. & Lemstra, A. W. (2016). Evaluation of the Digital Alzheimer Center: Testing Usability and Usefulness of an Online Portal for Patients with Dementia and Their Carers. *JMIR Res Protoc*, *5*(3), e144. <https://doi.org/10.2196/resprot.5040>
- Haugg, E. & Konopka, J. (2022). Spacing-Assistant for Leipzig and Munich Approach. *Transportation Research Procedia*, *66*, 292–303. <https://doi.org/https://doi.org/10.1016/j.trpro.2022.12.029>
- Hayes, A. F. & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Hayes, G. S., McLennan, S. N., Henry, J. D., Phillips, L. H., Terrett, G., Rendell, P. G., Pelly, R. M. & Labuschagne, I. (2020). Task characteristics influence facial emotion recognition age-effects: A meta-analytic review. *Psychology and Aging*, *35*(2), 295–315. <https://doi.org/10.1037/pag0000441>
- Hemetsberger, P. (2021). dict.cc Deutsch-Englisch Wörterbuch [German-English dictionary]: lustig. Retrieved June 10, 2021, from <https://www.dict.cc/?s=lustig>
- Hendriks, N., Huybrechts, L., Wilkinson, A. & Slegers, K. (2014). Challenges in doing participatory design with people with dementia. *Proceedings of the 13th Participatory Design Conference: Short Papers, Industry Cases, Workshop Descriptions, Doctoral Consortium papers, and Keynote abstracts*, *2*, 33–36. <https://doi.org/10.1145/2662155.2662196>
- Hernandez-Cruz, N., Garcia-Constantino, M., Beltran-Marquez, J., Cruz-Sandoval, D., Lopez-Nava, I. H., Cleland, I., Favela, J., Nugent, C., Ennis, A., Rafferty, J. & Synnott, J. (2019). Study Design of an Environmental Smart Microphone System to Detect Anxiety in Patients with Dementia. *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 383–388. <https://doi.org/10.1145/3329189.3329234>
- Hertzum, M., Hansen, K. D. & Andersen, H. H. K. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165–181. <https://doi.org/10.1080/01449290701773842>
- Hertzum, M., Molich, R. & Jacobsen, N. E. (2014). What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33*(2), 144–162. <https://doi.org/10.1080/0144929X.2013.783114>
- Hildebrandt, M. K., Jauk, E., Lehmann, K., Maliske, L. & Kanske, P. (2021). Brain activation during social cognition predicts everyday perspective-taking: A combined fMRI and ecological momentary assessment study of the social brain. *NeuroImage*, *227*, 117624. <https://doi.org/10.1016/j.neuroimage.2020.117624>
- Hill, A. P. & Bohil, C. J. (2016). Applications of Optical Neuroimaging in Usability Research. *Ergonomics in Design*, *24*(2), 4–9. <https://doi.org/10.1177/1064804616629309>

- Hinderks, A., Schrepp, M. & Thomaschewski, J. (2018). A Benchmark for the Short Version of the User Experience Questionnaire. *WEBIST*, 373–377.
- Hodge, J., Montague, K., Hastings, S. & Morrissey, K. (2019). Exploring Media Capture of Meaningful Experiences to Support Families Living with Dementia. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3290605.3300653>
- Hogan, R. (1969). Development of an empathy scale. *Journal of consulting and clinical psychology*, 33(3), 307.
- Holleis, P., Otto, F., Hussmann, H. & Schmidt, A. (2007). Keystroke-level model for advanced mobile phone interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1505–1514. <https://doi.org/10.1145/1240624.1240851>
- Hollnagel, E. (1999). Keep cool: the value of affective computer interfaces in a rational world. *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I-Volume I*, 676–680.
- Holthe, T., Halvorsrud, L., Karterud, D., Hoel, K.-A. & Lund, A. (2018). Usability and acceptability of technology for community-dwelling older adults with mild cognitive impairment and dementia: a systematic literature review. *Clinical Interventions in Aging*, 13, 863–886. <https://doi.org/10.2147/CIA.S154717>
- Holtzblatt, K. & Beyer, H. (2016). *Contextual Design: Design for Life*. Morgan Kaufmann.
- Holtzblatt, K., Wendell, J. B. & Wood, S. (2004). *Rapid contextual design: a how-to guide to key techniques for user-centered design*. Elsevier.
- Houben, M., Brankaert, R., Bakker, S., Kenning, G., Bongers, I. & Eggen, B. (2019). Foregrounding Everyday Sounds in Dementia. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 71–83. <https://doi.org/10.1145/3322276.3322287>
- Houben, M., Brankaert, R., Bakker, S., Kenning, G., Bongers, I. & Eggen, B. (2020). The role of everyday sounds in advanced dementia care. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Houben, M., Lehn, B., van den Brink, N., Diks, S., Verhoef, J. & Brankaert, R. (2020). Turn-around: exploring care relations in dementia through design. *2020 CHI Conference on Human Factors in Computing Systems*, LBW351.
- Hsieh, H.-F. & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9), 1277–1288.
- Hsu, W.-Y., Hsieh, L.-L., Su, Y.-H., Su, M.-J., Su, L., Chen, M.-C. & Chan, H.-T. (2019). Establishment of a Music Care System for the Elderly in a Long-term Care Facility. *2019 E-Health and Bioengineering Conference (EHB)*, 1–4. <https://doi.org/10.1109/EHB47216.2019.8970095>

- Huber, S., Bejan, A., Radzey, B., Berner, R., Murko, P. & Hurtienne, J. (2018). UX-Evaluationen in der Erinnerungspflege bei Demenz [UX evaluations in reminiscence sessions for people with dementia]. *Mensch und Computer 2018-Workshopband*.
- Huber, S., Bejan, A., Radzey, B. & Hurtienne, J. (2019). Proxemo or How to Evaluate User Experience for People with Dementia. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. <https://doi.org/10.1145/3290607.3313018>
- Huber, S., Berner, R., Ly-Tung, N., Preßler, J. & Hurtienne, J. (2017). Evaluation eines Public Displays für Menschen mit Demenz [Evaluation of a public display for people with dementia]. *Mensch und Computer 2017-Workshopband*.
- Huber, S., Berner, R., Uhlig, M., Klein, P. & Hurtienne, J. (2019). Tangible Objects for Reminiscing in Dementia Care. *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 15–24. <https://doi.org/10.1145/3294109.3295632>
- Huber, S., Gramlich, J. & Grundgeiger, T. (2020). From Paper Flight Strips to Digital Strip Systems: Changes and similarities in air traffic control work practices. *Proceedings of ACM Human-Computer Interaction*, 4(CSCW1), Article 028. <https://doi.org/10.1145/3392833>
- Huber, S., Gramlich, J., Pauli, S., Mundschenk, S., Haugg, E. & Grundgeiger, T. (2022). Toward User Experience in ATC: Exploring Novel Interface Concepts for Air Traffic Control. *Interacting with Computers*, iwac032. <https://doi.org/10.1093/iwc/iwac032>
- Huber, S., Preßler, J. & Hurtienne, J. (2016). Anpassung von Contextual Design für den Kontext Demenz [Adaption of Contextual Design for the Dementia Context]. *Mensch und Computer 2016-Tagungsband*.
- Huber, S., Preßler, J. & Hurtienne, J. (2017). Proxemo: A Demo-Presentation. *DementiaLab*, 2, 161–164.
- Huber, S., Preßler, J., Tung, N. L. & Hurtienne, J. (2017). Evaluating Interaction-Triggered Emotions in People with Dementia. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2659–2667.
- Huber, S. & Rathß, N. (in press). Empathic Accuracy and Mental Effort During Remote Assessments of Emotions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23-28, 2023*, 1–6. <https://doi.org/10.1145/3544548.3580824>
- Huisman, G., van Hout, M., van Dijk, E., van der Geest, T. & Heylen, D. (2013). LEMtool: measuring emotions in visual interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 351–360.
- Hussain, J., Khan, W. A., Hur, T., Bilal, H. S. M., Bang, J., Hassan, A. U., Afzal, M. & Lee, S. (2018). A Multimodal Deep Log-Based User Experience (UX) Platform for UX Evaluation. *Sensors*, 18(5), 1622. <https://www.mdpi.com/1424-8220/18/5/1622>

- Hyers, L. L. (2018). *Diary methods*. Oxford University Press.
- Ijsselsteijn, W., De Kort, Y., Poels, K., Jurgelionis, A. & Bellotti, F. (2007). Characterising and measuring user experiences in digital games. *International conference on advances in computer entertainment technology*, 2, 27.
- Ijsselsteijn, W., de Kort, Y. A. W. & Poels, K. (2013). *The Game Experience Questionnaire*. Technische Universiteit Eindhoven.
- Ijsselsteijn, W., Tummers-Heemels, A. & Brankaert, R. (2020). Warm Technology: A Novel Perspective on Design for and with People Living with Dementia. In R. Brankaert & G. Kenning (Eds.), *HCI and Design in the Context of Dementia* (pp. 33–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-32835-1_3
- Imants, P. & Greef, T. d. (2011). Using eye tracker data in air traffic control. *Proceedings of the 29th Annual European Conference on Cognitive Ergonomics*, 259–260. <https://doi.org/10.1145/2074712.2074769>
- Innes, A. & Surr, C. (2001). Measuring the well-being of people with dementia living in formal care settings: The use of Dementia Care Mapping. *Aging & mental health*, 5(3), 258–268.
- International Civil Aviation Organization. (2020). *Aircraft Accident and Incident Investigation: Annex 13 to the Convention on International Civil Aviation*. ICAO. https://www.emsa.europa.eu/retro/Docs/marine_casualties/annex_13.pdf
- ISO. (2018). ISO 9241-11:2018. Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and concepts. <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en>
- ISO. (2019). 9241-210:2019. Ergonomics of Human System Interaction-Part 210: Human-centred design for interactive systems (formerly known as 13407). <https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en>
- Iwamoto, M., Kuwahara, N. & Morimoto, K. (2015). Evaluation of the Impact on the Emotion of Young People Listening Attentively in at the Time of Using a Photograph of the Memory of the Elderly. *2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence*, 195–200. <https://doi.org/10.1109/ACIT-CSI.2015.43>
- Jainendra, S., Barreda-Angeles, M., Oliver, J., Nandi, G. C. & Puig, D. (2019). Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Transactions on Affective Computing*, 1–1. <https://doi.org/10.1109/TAFFC.2019.2901673>
- Jansen, A. M., Giebels, E., van Rompay, T. J. L. & Junger, M. (2018). The Influence of the Presentation of Camera Surveillance on Cheating and Pro-Social Behavior. *Frontiers in Psychology*, 9, 1937. <https://www.frontiersin.org/article/10.3389/fpsyg.2018.01937>
- JASP Team. (2020). JASP (Version 0.14.1)[Computer software]. <https://jasp-stats.org/>
- Jeannot, E., Kelly, C. & Thompson, D. (2003). *The development of situation awareness measures in ATM systems* (Report). Eurocontrol.

- Jessen, S. & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage*, *58*(2), 665–674. <https://doi.org/10.1016/j.neuroimage.2011.06.035>
- John, B. E. & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, *16*(4-5), 188–202.
- Jönsson, K.-E., Ornstein, K., Christensen, J. & Eriksson, J. (2019). A reminder system for independence in dementia care: a case study in an assisted living facility. *Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments*, 176–185. <https://doi.org/10.1145/3316782.3321530>
- Joyekurun, R. (2007). Weather hazards in ATM: designing for resilient operations. *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, 285–288. <https://doi.org/10.1145/1362550.1362610>
- Jung, T., Kaß, C., Schramm, T. & Zapf, D. (2017). So what really is user experience? An experimental study of user needs and emotional responses as underlying constructs. *Ergonomics*, *60*(12), 1601–1620. <https://doi.org/10.1080/00140139.2017.1341555>
- Jütten, L. H., Mark, R. E. & Sitskoorn, M. M. (2019). Empathy in informal dementia caregivers and its relationship with depression, anxiety, and burden. *International Journal of Clinical and Health Psychology*, *19*(1), 12–21. <https://doi.org/10.1016/j.ijchp.2018.07.004>
- Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J. P. P., Karvonen, H., Keskinen, T., Koskinen, H., Lu, Y., Saariluoma, P., Tokkonen, H. & Turunen, M. (2015). Defining user experience goals to guide the design of industrial systems. *Behaviour & Information Technology*, *34*(10), 976–991. <https://doi.org/10.1080/0144929X.2015.1035335>
- Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Prentice-Hall Inc.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A. & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological science*, *4*(6), 401–405.
- Kamp, I. & Desmet, P. M. (2014). Measuring product happiness. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 2509–2514. <https://doi.org/10.1145/2559206.2581274>
- Kashimoto, Y., Firouziyan, A., Asghar, Z., Yamamoto, G. & Pulli, P. (2016). Twinkle megane: Near-eye LED indicators on glasses in tele-guidance for elderly. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 1–6. <https://doi.org/10.1109/PERCOMW.2016.7457134>
- Kawakita, J. (1991). The original KJ method. *Tokyo: Kawakita Research Institute*, 5.
- Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of personality and social psychology*, *68*(3), 441.
- Khosla, R., Nguyen, K. & Chu, M.-T. (2014). Assistive robot enabled service architecture to support home-based dementia care. *2014 IEEE 7th International Conference on Service-Oriented Computing and Applications*, 73–80. <https://doi.org/10.1109/SOCA.2014.53>

- Kirouac, G. & Doré, F. Y. (1983). Accuracy and Latency of Judgment of Facial Expressions of Emotions. *Perceptual and Motor Skills*, 57(3), 683–686. <https://doi.org/10.2466/pms.1983.57.3.683>
- Kitwood, T. & Bredin, K. (1992). Towards a theory of dementia care: personhood and well-being. *Ageing and society*, 12(03), 269–287.
- Kleinsmith, A. & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*, 4(1), 15–33. <https://doi.org/10.1109/T-AFFC.2012.16>
- Klüber, S., Maas, F., Schraudt, D., Hermann, G., Happel, O. & Grundgeiger, T. (2020). Experience Matters: Design and evaluation of an anesthesia support tool guided by user experience theory. *2020 ACM Designing Interactive Systems Conference*, 1523–1535. <https://doi.org/10.1145/3357236.3395552>
- Kok, R. D., Rothweiler, J., Scholten, L., Zoest, M. v., Boumans, R. & Neerincx, M. (2018). Combining Social Robotics and Music as a Non-Medical Treatment for People with Dementia. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 465–467. <https://doi.org/10.1109/ROMAN.2018.8525813>
- Konstan, D. (2015). Affect and emotion in Greek literature. In G. Williams (Ed.), *Oxford Handbooks Online: Classical Studies*. <https://doi.org/10.1093/oxfordhb/9780199935390.013.41>
- Kosiński, J., Szklanny, K., Wieczorkowska, A. & Wichrowski, M. (2018). An Analysis of Game-Related Emotions Using EMOTIV EPOC. *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 913–917.
- Koutsabasis, P., Spyrou, T. & Darzentas, J. (2007). Evaluating usability evaluation methods: criteria, method and a case study. *International Conference on Human-Computer Interaction*, 569–578.
- Kox, M., van Eijk, L. T., Zwaag, J., van den Wildenberg, J., Sweep, F. C. G. J., van der Hoeven, J. G. & Pickkers, P. (2014). Voluntary activation of the sympathetic nervous system and attenuation of the innate immune response in humans. *Proceedings of the National Academy of Sciences*, 111(20), 7379–7384. <https://doi.org/10.1073/pnas.1322174111>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage publications.
- Krüger, F., Heine, C., Bader, S., Hein, A., Teipel, S. & Kirste, T. (2017). On the applicability of clinical observation tools for human activity annotation. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 129–134. <https://doi.org/10.1109/PERCOMW.2017.7917545>
- Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E. & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. *Interacting with Computers*, 23(5), 473–483. <https://doi.org/10.1016/j.intcom.2011.06.005>

- Kumfor, F., Hazelton, J. L., Rushby, J. A., Hodges, J. R. & Piguet, O. (2019). Facial expressiveness and physiological arousal in frontotemporal dementia: Phenotypic clinical profiles and neural correlates. *Cognitive, Affective, & Behavioral Neuroscience*, 19(1), 197–210. <https://doi.org/10.3758/s13415-018-00658-z>
- Kynast, J., Polyakova, M., Quinque, E. M., Hinz, A., Villringer, A. & Schroeter, M. L. (2021). Age- and Sex-Specific Standard Scores for the Reading the Mind in the Eyes Test. *Frontiers in Aging Neuroscience*, 12. <https://doi.org/10.3389/fnagi.2020.607107>
- Lakens, D., Scheel, A. M. & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.
- Lanius, C., Weber, R. & Robinson, J. (2021). User Experience Methods in Research and Practice. *Journal of Technical Writing and Communication*, 00472816211044499. <https://doi.org/10.1177/00472816211044499>
- Laugwitz, B., Held, T. & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. *Symposium of the Austrian HCI and usability engineering group*, 63–76.
- Laurans, G. & Desmet, P. (2012). Introducing PREMO2: New directions for the non-verbal measurement of emotion in design. *Out of Control: Proceedings of the 8th International Conference on Design and Emotion, London, UK, 11-14 September 2012*.
- Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Sage publications.
- Law, E. L.-C. (2006). Evaluating the downstream utility of user tests and examining the developer effect: A case study. *International Journal of Human-Computer Interaction*, 21(2), 147–172.
- Law, E. L.-C., Vermeeren, A. P. O. S., Hassenzahl, M. & Blythe, M. (2007). Towards a UX manifesto. *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21, 1–2*.
- Law, E. L.-C., Brühlmann, F. & Mekler, E. D. (2018). Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 257–270. <https://doi.org/10.1145/3242671.3242683>
- Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S. & David, A. S. (2004). Measuring empathy: reliability and validity of the Empathy Quotient. *Psychological medicine*, 34(5), 911.
- Lawton, M. P., Van Haitsma, K. & Klapper, J. (1996). Observed affect in nursing home residents with Alzheimer's disease. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 51(1), P3–P14.

- Lawton, M. P., Van Haitsma, K. & Klapper, J. (1999a). Observed affect and quality of life in dementia: Further affirmations and problems. *Journal of mental Health and Aging*, 5(1), 69–82.
- Lawton, M. P., Van Haitsma, K. & Klapper, J. (1999b). Observed Emotion Rating Scale. Retrieved June 10, 2021, from <https://abramsonseniorcare.org/media/1199/observed-emotion-rating-scale.pdf>
- Lazar, A., Thompson, H. & Demiris, G. (2014). A systematic review of the use of technology for reminiscence therapy. *Health Educ Behav*, 41(1 Suppl), 51S–61S. <https://doi.org/10.1177/1090198114537067>
- Lazar, A., Edasis, C. & Piper, A. M. (2017a). A Critical Lens on Dementia and Design in HCI. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2175–2188. <https://doi.org/10.1145/3025453.3025522>
- Lazar, A., Edasis, C. & Piper, A. M. (2017b). Supporting People with Dementia in Digital Social Sharing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2149–2162. <https://doi.org/10.1145/3025453.3025586>
- Leibetseder, M., Laireiter, A.-R. & Köller, T. (2007). Structural analysis of the E-scale. *Personality and Individual Differences*, 42(3), 547–561.
- Lewis, G. J., Lefevre, C. E. & Young, A. W. (2016). Functional architecture of visual emotion recognition ability: A latent variable approach. *Journal of Experimental Psychology: General*, 145(5), 589–602. <https://doi.org/10.1037/xge0000160>
- Li, D., Wang, Z., Wang, C., Liu, S., Chi, W., Dong, E., Song, X., Gao, Q. & Song, Y. (2019). The Fusion of Electroencephalography and Facial Expression for Continuous Emotion Recognition. *IEEE Access*, 7, 155724–155736. <https://doi.org/10.1109/ACCESS.2019.2949707>
- Li, S. & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 1–20. <https://doi.org/10.1109/TAFFC.2020.2981446>
- Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International Journal of Industrial Ergonomics*, 36(12), 1069–1074. <https://doi.org/10.1016/j.ergon.2006.09.007>
- Lipps, T. (1903). Einleitendes zur Frage der Einfühlung [Introduction on the issue of Empathy]. In T. Lipps (Ed.), *Psychologie des Schönen und der Kunst* (1st ed., pp. 96–223). Verlag von Leopold Voss.
- Liu, W., Batchelor, M. & Williams, K. (2020). Ease of use, feasibility and inter-rater reliability of the refined Cue Utilization and Engagement in Dementia (CUED) mealtime video-coding scheme. *Journal of Advanced Nursing*, 76(12), 3609–3622.
- Liu, Y., Lan, Z., Traspilawati, F., Sourina, O., Chen, C.-H. & Müller-Wittig, W. (2019). EEG-Based Human Factors Evaluation of Air Traffic Control Operators (ATCOs) for Optimal

- Training. *2019 International Conference on Cyberworlds (CW)*, 253–260. <https://doi.org/10.1109/CW.2019.00049>
- Loeffler, D., Hess, A., Maier, A., Hurtienne, J. & Schmitt, H. (2013). Developing intuitive user interfaces by integrating users' mental models into requirements engineering. *27th International BCS Human Computer Interaction Conference (HCI 2013)* 27, 1–10.
- Loft, S. (2014). Applying Psychological Science to Examine Prospective Memory in Simulated Air Traffic Control. *Current Directions in Psychological Science*, 23(5), 326–331. <https://doi.org/10.1177/0963721414545214>
- Lohse, T. & Qari, S. (2018). Video recordings in experiments – Are there effects on self-selection or the outcome of the experiment? *Economics Bulletin*, 38(3), 1381–1394.
- Losonczy, M. E. & Brandt, L. J. (2003). Latency and intensity of discrete emotions: are discrete emotions differentiated by latency and/or intensity of expression? *Annals of the New York Academy of Sciences*, 1000(1), 193–196. <https://doi.org/10.1196/annals.1280.023>
- Lotze, R. H. (1858). *Mikrokosmos*. <https://books.google.de/books?id=mb0IAAAAQAAJ>
- Lourties, S., Léger, P.-M., Sénécal, S., Fredette, M. & Chen, S. L. (2018). Testing the Convergent Validity of Continuous Self-Perceived Measurement Systems: An Exploratory Study. In F. F.-H. Nah & B. S. Xiao (Eds.), *HCI in Business, Government, and Organizations* (pp. 132–144). Springer International Publishing.
- Luckmann, T. (2012). Some remarks on scores in multimodal sequential analysis. In H. Knoblauch, B. Schnettler, J. Raab & H.-G. Soeffner (Eds.), *Video Analysis. Methodology and Methods. Qualitative Audiovisual Data Analysis in Sociology* (pp. 29–34). Peter Lang.
- Lufthansa Services. (2020). *Aviation – You Up There, We Down Here: Flight Communication – Lufthansa*. Youtube. Retrieved December 12, 2021, from <https://www.youtube.com/watch?v=2G4EFZGFVwc>
- Ma, K., Wang, X., Yang, X., Zhang, M., Girard, J. M. & Morency, L.-P. (2019). ElderReact: A Multimodal Dataset for Recognizing Emotional Response in Aging Adults. *2019 International Conference on Multimodal Interaction*, 349–357. <https://doi.org/10.1145/3340555.3353747>
- Mackay, W. E. (1999). Is Paper Safer? The role of paper flight strips in air traffic control. *ACM Transactions of Computer-Human Interaction*, 6(4), 311–340. <https://doi.org/10.1145/331490.331491>
- Magai, C., Cohen, C., Gomberg, D., Malatesta, C. & Culver, C. (1996). Emotional Expression During Mid- to Late-Stage Dementia. *International Psychogeriatrics*, 8(3), 383–395. <https://doi.org/10.1017/S104161029600275X>
- Mäkelä, S., Bednarik, R. & Tukiainen, M. (2013). Evaluating User Experience of Autistic Children through Video Observation. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 463–468. <https://doi.org/10.1145/2468356.2468438>

- Malakis, S., Kontogiannis, T. & Psaros, P. (2014). Monitoring and evaluating failure-sensitive strategies in air traffic control simulator training. *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments*, Article 43. <https://doi.org/10.1145/2674396.2674462>
- Malone, T. W. (1982). Heuristics for Designing Enjoyable User Interfaces: Lessons from Computer Games. *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, 63–68. <https://doi.org/10.1145/800049.801756>
- Manning, C. D., Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Martin, A. C., Qi, W., Kazunori, H. & Shamsul, H. (2005). A Cross-Cultural Investigation of Autobiographical Memory: On the Universality and Cultural Variation of the Reminiscence Bump. *Journal of Cross-Cultural Psychology*, 36(6), 739–749. <https://doi.org/10.1177/0022022105280512>
- Martinez, M., Multani, N., Anor, C. J., Misquitta, K., Tang-Wai, D. F., Keren, R., Fox, S., Lang, A. E., Marras, C. & Tartaglia, M. C. (2018). Emotion Detection Deficits and Decreased Empathy in Patients with Alzheimer’s Disease and Parkinson’s Disease Affect Caregiver Mood and Burden. *Frontiers in Aging Neuroscience*, 10, 120. <https://doi.org/10.3389/fnagi.2018.00120>
- Maslow, A. H. (1971). *The farther reaches of human nature*. Penguin Books Ltd.
- Matthews, G., Reinerman-Jones, L. E., Barber, D. J. & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, 57(1), 125–143.
- Maybury, M. T. (2012). Usable Advanced Visual Interfaces in Aviation. *International Working Conference on Advanced Visual Interfaces*, 2–3. <https://doi.org/10.1145/2254556.2254558>
- Mayring, P. & Fenzl, T. (2019). Qualitative Inhaltsanalyse [Qualitative Content Analysis]. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 633–648). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_42
- McCarthy, J. & Wright, P. (2003). The Enchantments of Technology. In J. Karat & J. Vanderdonck (Eds.), *Funology* (pp. 81–90). Springer.
- McCarthy, J. & Wright, P. (2004). Technology as experience. *Interactions*, 11(5), 42–43.
- Meiland, F., Innes, A., Mountain, G., Robinson, L., van der Roest, H., García-Casal, J. A., Gove, D., Thyrian, J. R., Evans, S., Dröes, R.-M., Kelly, F., Kurz, A., Casey, D., Szcześniak, D., Dening, T., Craven, M. P., Span, M., Felzmann, H., Tsolaki, M. & Franco-Martin, M. (2017). Technologies to Support Community-Dwelling Persons With Dementia: A Position Paper on Issues Regarding Development, Usability, Effectiveness and Cost-Effectiveness, Deployment, and Ethics. *JMIR Rehabil Assist Technol*, 4(1), e1. <https://doi.org/10.2196/rehab.6376>

- Meiland, F. J. M., Bouman, A. I. E., Sävenstedt, S., Bentvelzen, S., Davies, R. J., Mulvenna, M. D., Nugent, C. D., Moelaert, F., Hettinga, M. E., Bengtsson, J. E. & Dröes, R.-M. (2012). Usability of a new electronic assistive device for community-dwelling persons with mild dementia. *Aging & Mental Health*, 16(5), 584–591. <https://doi.org/10.1080/13607863.2011.651433>
- Mekler, E. D. & Hornbæk, K. (2016). Momentary Pleasure or Lasting Meaning? Distinguishing eudaimonic and hedonic user experiences. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4509–4520. <https://doi.org/10.1145/2858036.2858225>
- Mentler, T. & Herczeg, M. (2016). On the Role of User Experience in Mission-or Safety-Critical Systems. *Mensch und Computer 2016–Workshopband*. <https://doi.org/10.18420/muc2016-ws01-0001>
- Micallef, L. & Rodgers, P. (2014). euler APE: Drawing area-proportional 3-Venn diagrams using ellipses. *PloS one*, 9(7), e101717.
- Mill, A., Allik, J., Realo, A. & Valk, R. (2009). Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion*, 9(5), 619.
- Minge, M. & Thüning, M. (2009). *Dynamics of User Experience. Judgments of Attractiveness, Usability, and Emotions Over Time* (Report). Erhältlich als Technical Report 10-2009, Berlin: TU Berlin.
- Minge, M., Thüning, M., Wagner, I. & Kuhr, C. V. (2017). The meCUE questionnaire: a modular tool for measuring user experience. *Advances in Ergonomics Modeling, Usability & Special Populations* (pp. 115–128). Springer.
- Molich, R. & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, 27(3), 263–281. <https://doi.org/10.1080/01449290600959062>
- Moors, A., Ellsworth, P. C., Scherer, K. R. & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, 5(2), 119–124. <https://doi.org/10.1177/1754073912468165>
- Morrissey, K. & McCarthy, J. (2015). Creative and Opportunistic Use of Everyday Music Technologies in a Dementia Care Unit. *2015 ACM SIGCHI Conference on Creativity and Cognition*, 295–298. <https://doi.org/10.1145/2757226.2757228>
- Morrissey, K., McCarthy, J. & Pantidi, N. (2017). The Value of Experience-Centred Design Approaches in Dementia Research Contexts. *2017 CHI Conference on Human Factors in Computing Systems*, 1326–1338. <https://doi.org/10.1145/3025453.3025527>
- Morrissey, K., Wood, G., Green, D., Pantidi, N. & McCarthy, J. (2016). 'I'm a Rambler, I'm a Gambler, I'm a Long Way from Home': The Place of Props, Music, and Design in Dementia Care. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 1008–1020. <https://doi.org/10.1145/2901790.2901798>

- Morville, P. (2005). Experience design unplugged. *ACM SIGGRAPH 2005 Web program*, 10–es. <https://doi.org/10.1145/1187335.1187347>
- Moyle, W., Jones, C., Dwan, T. & Petrovich, T. (2018). Effectiveness of a Virtual Reality Forest on People With Dementia: A Mixed Methods Pilot Study. *The Gerontologist*, *58*(3), 478–487. <https://doi.org/10.1093/geront/gnw270>
- Multani, N., Galantucci, S., Wilson, S. M., Shany-Ur, T., Poorzand, P., Growdon, M. E., Jang, J. Y., Kramer, J. H., Miller, B. L., Rankin, K. P., Gorno-Tempini, M. L. & Tartaglia, M. C. (2017). Emotion detection deficits and changes in personality traits linked to loss of white matter integrity in primary progressive aphasia. *Neuroimage Clin*, *16*, 447–454. <https://doi.org/10.1016/j.nicl.2017.08.020>
- Muñoz, D., Favilla, S., Pedell, S., Murphy, A., Beh, J. & Petrovich, T. (2021). Evaluating an App to Promote a Better Visit Through Shared Activities for People Living with Dementia and their Families. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3411764.3445764>
- Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482.
- Murphy, B. A. & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment*, *31*(8), 1062–1072. <https://doi.org/10.1037/pas0000732>
- Nacke, L. (2009). *Affective ludology: Scientific measurement of user experience in interactive entertainment* (Thesis).
- Naumann, A. & Hurtienne, J. (2010). Benchmarks for intuitive interaction with mobile devices. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, 401–402.
- Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 152–158.
- Nielsen, J. (1994b). *Usability engineering*. Morgan Kaufmann.
- Nielsen, L. & Kaszniak, A. W. (2007). Conceptual, theoretical, and methodological issues in inferring subjective emotion experience. *Handbook of emotion elicitation and assessment*, 361–375.
- Niewiadomski, R. & Sciutti, A. (2021). Multimodal Emotion Recognition of Hand-Object Interaction. *26th International Conference on Intelligent User Interfaces*, 351–355. <https://doi.org/10.1145/3397481.3450636>
- Nilsen, E. L. (1996). *Perceptual-motor control in human-computer interaction* (Report). University of Michigan, Division of Research and Development, Ann Arbor, MI.
- Noldus. (2021). Tools for the Facial action coding system (FACS). Retrieved December 9, 2021, from <https://www.noldus.com/applications/facial-action-coding-system>

- Norman, D., Miller, J. & Henderson, A. (1995). What you see, some of what's in the future, and how we go about doing it: HI at Apple Computer. *Conference Companion on Human Factors in Computing Systems*, 155. <https://doi.org/10.1145/223355.223477>
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology*, 3(1), 1–18.
- Nur, A. I., Santoso, H. B. & Putra, P. O. H. (2021). The Method and Metric of User Experience Evaluation: A Systematic Literature Review. *2021 10th International Conference on Software and Computer Applications*, 307–317. <https://doi.org/10.1145/3457784.3457832>
- Obrist, M., Law, E., Väänänen-Vainio-Mattila, K., Roto, V., Vermeeren, A. & Kuutti, K. (2011). UX Research: What Theoretical Roots Do We Build on – If Any? *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 165–168. <https://doi.org/10.1145/1979742.1979526>
- Obrist, M., Roto, V., Vermeeren, A., Väänänen-Vainio-Mattila, K., Law, E. L.-C. & Kuutti, K. (2012). In search of theoretical foundations for UX research and practice. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, 1979–1984. <https://doi.org/10.1145/2212776.2223739>
- Obrist, M., Wright, P. C., Kuutti, K., Rogers, Y., Höök, K., Pyla, P. S. & Frechin, J.-L. (2013). Theory and practice in ux research: uneasy bedfellows? *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2433–2438. <https://doi.org/10.1145/2468356.2468795>
- O'Donnell, R. & Eggemeier, T. (1986). Workload assessment methodology. In T. J. Boff KR Kaufman L (Ed.), *Handbook of perception and human performance*. Wiley-Interscience.
- Osgood, C. E., May, W. H. & Miron, M. S. (1975). *Cross-cultural universals of affective meaning* (Vol. 1). University of Illinois Press.
- Osterwalder, A., Pigneur, Y., Bernarda, G. & Smith, A. (2014). *Value proposition design: How to create products and services customers want* (Vol. 2). John Wiley & Sons.
- Otto, J. H., Döring-Seipel, E., Grebe, M. & Lantermann, E.-D. (2001). Entwicklung eines Fragebogens zur Erfassung der wahrgenommenen emotionalen Intelligenz. Aufmerksamkeit auf, Klarheit und Beeinflussbarkeit von Emotionen. [Development of a questionnaire for measuring perceived emotional intelligence: Attention to, Clarity, and Repair of emotions.] *Diagnostica*, 47(4), 178–187. <https://doi.org/10.1026/0012-1924.47.4.178>
- Parekh, V., Foong, P. S., Zhao, S. & Subramanian, R. (2018). AVEID: Automatic Video System for Measuring Engagement In Dementia. *23rd International Conference on Intelligent User Interfaces*, 409–413. <https://doi.org/10.1145/3172944.3173010>
- Pérez-Sáez, E., Pérez-Redondo, E. & González-Ingelmo, E. (2020). Effects of dog-assisted therapy on social behaviors and emotional expressions: a single-case experimental design in 3 people with dementia. *Journal of geriatric psychiatry and neurology*, 33(2), 109–119.

- Perrault, S. T., Lecolinet, E., Eagan, J. & Guiard, Y. (2013). Watchit: simple gestures and eyes-free interaction for wristwatches and bracelets. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1451–1460). Association for Computing Machinery. <https://doi.org/10.1145/2470654.2466192>
- Petrie, H. & Precious, J. (2010). Measuring user experience of websites: think aloud protocols and an emotion word prompt list. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 3673–3678. <https://doi.org/10.1145/1753846.1754037>
- Petta, P., Pelachaud, C. & Cowie, R. (2011). *Emotion-oriented systems: the HUMAINE handbook*. Springer.
- Peute, L. W. P., de Keizer, N. F. & Jaspers, M. W. M. (2015). The value of Retrospective and Concurrent Think Aloud in formative usability testing of a physician data query tool. *Journal of Biomedical Informatics*, 55, 1–10. <https://doi.org/10.1016/j.jbi.2015.02.006>
- Pfaltz, M. C., McAleese, S., Saladin, A., Meyer, A. H., Stoecklin, M., Opwis, K. & Martin-Soelch, C. (2013). The Reading the Mind in the Eyes Test: Test-retest reliability and preliminary psychometric properties of the German version. *International Journal of Advances in Psychology*, 2(1), 1–9.
- Plattner, H., Meinel, C. & Weinberg, U. (2009). *Design-thinking*. Springer.
- Plutchik, R. (1991). *The emotions*. University Press of America.
- Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4), 344–350. <http://www.jstor.org/stable/27857503>
- Pöllänen, S. H. & Hirsimäki, R. M. (2014). Crafts as Memory Triggers in Reminiscence: A Case Study of Older Women with Dementia. *Occupational Therapy In Health Care*, 28(4), 410–430. <https://doi.org/10.3109/07380577.2014.941052>
- Proske, M. (2021). Was wir von Menschen mit Demenz lernen können [What we can learn from people with dementia]. Retrieved September 13, 2021, from <https://www.pflege-durch-angehoerige.de/von-demenzkranken-lernen/>
- Prpa, M., Fdili-Alaoui, S., Schiphorst, T. & Pasquier, P. (2020). Articulating Experience: Reflections from Experts Applying Micro-Phenomenology to Design Research in HCI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376664>
- Qiao, H., Li, Y., Zhang, J., Xiaotian, E., Zou, X., Jiang, Y., Xiong, L. & Sun, X. (2018). The peak-end effects in controllers' mental workload evaluation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 87–91. <https://doi.org/10.1177/1541931218621020>
- Quesque, F. & Rossetti, Y. (2020). What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science*, 15(2), 384–396. <https://doi.org/10.1177/1745691619896607>

- Quince, T. A., Parker, R. A., Wood, D. F. & Benson, J. A. (2011). Stability of empathy among undergraduate medical students: a longitudinal study at one UK medical school. *BMC medical education*, *11*(1), 1–9.
- Rasquin, S. M., Willems, C., De Vlieger, S., Geers, R. & Soede, M. (2007). The use of technical devices to support outdoor mobility of dementia patients. *Technology and disability*, *19*(2, 3), 113–120.
- Recio, G., Schacht, A. & Sommer, W. (2014). Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials. *Biological Psychology*, *96*, 111–125. <https://doi.org/10.1016/j.biopsycho.2013.12.003>
- Recio, G., Wilhelm, O., Sommer, W. & Hildebrandt, A. (2017). Are event-related potentials to dynamic facial expressions of emotion related to individual differences in the accuracy of processing facial expressions and identity? *Cognitive, Affective, & Behavioral Neuroscience*, *17*(2), 364–380. <https://doi.org/10.3758/s13415-016-0484-6>
- Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A. & Kupfer, D. J. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses [PMID: 23111466]. *American Journal of Psychiatry*, *170*(1), 59–70. <https://doi.org/10.1176/appi.ajp.2012.12070999>
- Reinhardt, D. (2020). *IntuiBeat: Formative und summative Evaluation intuitiver Benutzung [IntuiBeat: Formative and Summative Evaluation of Intuitive Use]* (Thesis). <https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/21759%20https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-217599>
- Reinhardt, D., Haesler, S., Hurtienne, J. & Wienrich, C. (2019). Entropy of Controller Movements Reflects Mental Workload in Virtual Reality. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 802–808. <https://doi.org/10.1109/VR.2019.8797977>
- Reinhardt, D., Hurtienne, J. & Wienrich, C. (2020). Measuring Mental Effort via Entropy in VR. *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 43–44. <https://doi.org/10.1145/3379336.3381493>
- Reisman, R. & Kaliouby, R. e. (2007). *Facial expression affective state recognition for air traffic control automation concept exploration*. Association for Computing Machinery. <https://doi.org/10.1145/1280720.1280901>
- Remington, N. A., Fabrigar, L. R. & Visser, P. S. (2000). Reexamining the circumplex model of affect. *Journal of personality and social psychology*, *79*(2), 286. <https://doi.org/10.1037//0022-3514.79.2.28>
- Rezaei, S., Moturu, A., Zhao, S., Prkachin, K. M., Hadjistavropo, T. & Taati, B. (2020). Ambient Pain Monitoring in Older Adults with Dementia to Improve Pain Management in Long-Term Care Facilities. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 352–353. <https://doi.org/10.1145/3395035.3425353>

- Rice, A. D. & Lartigue, J. W. (2014). Touch-level model (TLM): evolving KLM-GOMS for touch-screen and mobile devices. *Proceedings of the 2014 ACM Southeast Regional Conference*, Article 53. <https://doi.org/10.1145/2638404.2638532>
- Richter, D. & Kunzmann, U. (2011). Age differences in three facets of empathy: Performance-based evidence. *Psychology and aging*, *26*(1), 60.
- Riediger, M., Voelkle, M. C., Ebner, N. C. & Lindenberger, U. (2011). Beyond “happy, angry, or sad?”: Age-of-poser and age-of-rater effects on multi-dimensional emotion perception. *Cognition and Emotion*, *25*(6), 968–982. <https://doi.org/10.1080/02699931.2010.540812>
- Riehle, M., Kempkensteffen, J. & Lincoln, T. (2017). Quantifying Facial Expression Synchrony in Face-To-Face Dyadic Interactions: Temporal Dynamics of Simultaneously Recorded Facial EMG Signals. *Journal of Nonverbal Behavior*, *41*(2), 85–102. <https://doi.org/10.1007/s10919-016-0246-8>
- Rinn, W. E. (1984). The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, *95*(1), 52.
- Rokicki, S. M. (1987). Heart Rate Averages as Workload/Fatigue Indicators during OT&E. *Proceedings of the Human Factors Society Annual Meeting*, *31*(7), 784–785. <https://doi.org/10.1177/154193128703100721>
- Rose Addis, D. & Tippett, L. (2004). Memory of myself: Autobiographical memory and identity in Alzheimer’s disease. *Memory*, *12*(1), 56–74.
- Roto, V. (2007). User experience from product creation perspective. *Towards a UX Manifesto*, 31.
- Rozzi, S., Amaldi, P., Wong, W. & Field, B. (2007). Operational potential for 3D displays in air traffic control. *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, 179–183. <https://doi.org/10.1145/1362550.1362586>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145. <https://doi.org/10.1037/0033-295X.110.1.145>
- Rutten, E., van den Bogaert, L. & Geerts, D. (2021). From Initial Encounter with Mid-Air Haptic Feedback to Repeated Use: the Role of the Novelty Effect in User Experience. *IEEE Transactions on Haptics*, *14*(3), 591–602. <https://doi.org/10.1109/TOH.2020.3043658>
- Salekin, A., Wang, H. & Stankovic, J. (2020). Demo: KinVocal: Detecting Agitated Vocal Events. *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 459–460. <https://doi.org/10.1145/2809695.2817853>
- Salmon, P. M., Read, G. J., Walker, G. H., Stevens, N. J., Hulme, A., McLean, S. & Stanton, N. A. (2020). Methodological issues in systems Human Factors and Ergonomics: Perspectives

- on the research–practice gap, reliability and validity, and prediction. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 32(1), 6–19.
- Salovey, P., Mayer, J. D., Goldman, S. L., Turvey, C. & Palfai, T. P. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the Trait Meta-Mood Scale. *Emotion, disclosure, & health*. (pp. 125–154). American Psychological Association. <https://doi.org/10.1037/10182-006>
- Sanderson, P. M., Mooij, M. & Neal, A. (2007). Investigating sources of mental workload using a high-fidelity ATC simulator. *International Symposium on Aviation Psychology*, 618–623.
- Santos, C. P., Gaans, N. C. M. F. v., Khan, V.-J. & Markopoulos, P. (2019). Effects of advertisements and questionnaire interruptions on the player experience. *2019 IEEE Conference on Games (CoG)*, 1–8. <https://doi.org/10.1109/CIG.2019.8848023>
- Sauer, J., Sonderegger, A. & Schmutz, S. (2020). Usability, user experience and accessibility: towards an integrative model. *Ergonomics*, 1–23. <https://doi.org/10.1080/00140139.2020.1774080>
- Savage-Knepshield, P., Hullinger, D., Lund, R., Manning, C., Pierce, L., Seely, O. & Thomas, J. (2016). The Challenges of Measuring Human Performance in Complex Operational Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1), 2053–2057. <https://doi.org/10.1177/1541931213601466>
- Schaaff, K. & Adam, M. T. P. (2013). Measuring Emotional Arousal for Online Applications: Evaluation of Ultra-short Term Heart Rate Variability Measures. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 362–368. <https://doi.org/10.1109/ACII.2013.66>
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. <https://doi.org/10.1177/0539018405058216>
- Scherer, K. R., Dieckmann, A., Unfried, M., Ellgring, H. & Mortillaro, M. (2021). Investigating appraisal-driven facial expression and inference in emotion communication. *Emotion*, 21(1), 73. <https://doi.org/10.1037/emo0000693>
- Schirmer, A. (2014). *Emotion*. Sage Publications.
- Schirmer, A. (2017). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing. *Social Cognitive and Affective Neuroscience*, 13(1), 1–13. <https://doi.org/10.1093/scan/nsx142>
- Schlör, D., Zehe, A., Kobs, K., Veseli, B., Westermeier, F., Brübach, L., Roth, D., Latoschik, M. E. & Hotho, A. (2020). Improving Sentiment Analysis with Biofeedback Data. *Proceedings of LREC2020 Workshop “People in language, vision and the mind”(ONION2020)*, 28–33.
- Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2), 81.
- Schrepp, M., Hinderks, A. & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *IJIMAI*, 4(6), 103–108.

- Schuler, H. (2002). Emotionale Intelligenz—ein irreführender und unnötiger Begriff [emotional intelligence - a misleading and unnecessary term]. *Zeitschrift für Personalpsychologie*, *1*(3), 138–140.
- Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, *14*(4), 433–440. <https://doi.org/10.1080/026999300402745>
- Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. *The science of real-time data capture: Self-reports in health research*, *11*, 26.
- Scriven, M. (1972). Die Methodologie der Evaluation [The methodology of Evaluation]. In C. Wulf (Ed.), *Evaluation: Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen* (pp. 60–91). R. Piper & Co. Verlag.
- Sears, A. (1997). Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, *9*(3), 213–234.
- Sheppard, B., Sarrazin, H., Kouyoumjian, G. & Dore, F. (2018). *The business value of design* (Report). Retrieved December 7, 2021, from <https://www.mckinsey.com/business-functions/mckinsey-design/our-insights/the-business-value-of-design>
- Siegel, S. & Castellan, N. J. (1981). *Nonparametric Statistics for the Behavioral Sciences* (2, Ed.).
- Siriaraya, P. & Ang, C. S. (2014). Recreating living experiences from past memories through virtual worlds for people with dementia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3977–3986.
- Slater, M. & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603–616.
- Sloane, P. D., Brooker, D., Cohen, L., Douglass, C., Edelman, P., Fulton, B. R., Jarrott, S., Kasayka, R., Kuhn, D., Preisser, J. S., Williams, C. S. & Zimmerman, S. (2007). Dementia care mapping as a research tool. *International Journal of Geriatric Psychiatry*, *22*(6), 580–589. <https://doi.org/10.1002/gps.1721>
- Smeenk, W., Sturm, J. & Eggen, B. (2018). Empathic handover: how would you feel? Handing over dementia experiences and feelings in empathic co-design. *CoDesign*, *14*(4), 259–274. <https://doi.org/10.1080/15710882.2017.1301960>
- Solé, C., Mercadal-Brotons, M., Galati, A. & De Castro, M. (2014). Effects of Group Music Therapy on Quality of Life, Affect, and Participation in People with Varying Levels of Dementia. *Journal of Music Therapy*, *51*(1), 103–125. <https://doi.org/10.1093/jmt/thu003>
- Span, M., Hettinga, M., Groen-van de Ven, L., Jukema, J., Janssen, R., Vernooij-Dassen, M., Eefsting, J. & Smits, C. (2018). Involving people with dementia in developing an interactive web tool for shared decision-making: experiences with a participatory design approach. *Disabil Rehabil*, *40*(12), 1410–1420. <https://doi.org/10.1080/09638288.2017.1298162>

- Stamen. (2020). Atlas of Emotions. Retrieved November 5, 2021, from <https://stamen.com/work/atlas-of-emotions/>
- Stanton, N. A. (2016). On the reliability and validity of, and training in, ergonomics methods: a challenge revisited. *Theoretical Issues in Ergonomics Science*, 17(4), 345–353.
- Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C. & Jenkins, D. P. (2017). *Human factors methods: a practical guide for engineering and design*. CRC Press.
- Steinert, L., Putze, F., Küster, D. & Schultz, T. (2020). Towards Engagement Recognition of People with Dementia in Care Settings. *Proceedings of the 2020 International Conference on Multimodal Interaction*, 558–565. <https://doi.org/10.1145/3382507.3418856>
- Stemmler, G. (2004). Physiological processes during emotion. *The regulation of emotion* (pp. 48–85). Psychology Press.
- Stoeckle, M. & Freund, L. (2016). A proof of concept personalized music player for persons with alzheimer’s disease. *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, 1–4.
- Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.
- Taati, B., Snoek, J. & Mihailidis, A. (2011). Towards Aging-in-Place: Automatic Assessment of Product Usability for Older Adults with Dementia. *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, 205–212. <https://doi.org/10.1109/HISB.2011.43>
- Taati, B., Zhao, S., Ashraf, A. B., Asgarian, A., Browne, M. E., Prkachin, K. M., Mihailidis, A. & Hadjistavropoulos, T. (2019). Algorithmic Bias in Clinical Populations—Evaluating and Improving Facial Analysis Technology in Older Adults With Dementia. *IEEE Access*, 7, 25527–25534. <https://doi.org/10.1109/ACCESS.2019.2900022>
- Tabbaa, L., Ang, C. S., Rose, V., Siriaraya, P., Stewart, I., Jenkins, K. G. & Matsangidou, M. (2019). Bring the Outside In: Providing Accessible Experiences Through VR for People with Dementia in Locked Psychiatric Hospitals. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 236. <https://doi.org/10.1145/3290605.3300466>
- Testkuratorium. (2018). Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen, rev. Fassung vom 03. Januar 2018 [Test assessment system of the Diagnostic and Test Board of German Psychological Associations, revised version of January 03, 2018]. *Psychologische Rundschau*, 69(2), 109–148.
- Thackray, R. I. (1980). *Boredom and monotony as a consequence of automation: a consideration of the evidence relating boredom and monotony to stress* (Report). Civil Aerospace Medical Institute.
- The Unicode Consortium. (2019). Emoji Frequency. Retrieved October 30, 2021, from <https://home.unicode.org/emoji/emoji-frequency/>

- Thoolen, M., Brankaert, R. & Lu, Y. (2020). AmbientEcho: exploring interactive media experiences in the context of residential dementia care. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1495–1508. <https://doi.org/10.1145/3357236.3395432>
- Thorgrimsen, L., Selwood, A., Spector, A., Royan, L., de Madariaga Lopez, M., Woods, R. T. & Orrell, M. (2003). Whose Quality of Life Is It Anyway?: The Validity and Reliability of the Quality of Life-Alzheimer's Disease (QoL-AD) Scale. *Alzheimer Disease & Associated Disorders*, 17(4). https://journals.lww.com/alzheimerjournal/Fulltext/2003/10000/Whose_Quality_of_Life_Is_It_Anyway_---The_Validity.2.aspx
- Thüring, M. & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International journal of psychology*, 42(4), 253–264.
- Titchener, E. B. (1909). *Lectures on the Experimental Psychology of the Thought-Processes*. Macmillan.
- Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.
- Tomkins, S. (2008). *Affect imagery consciousness: The complete edition: Two Volumes*. Springer publishing company.
- Tracy, J. L. & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion*, 8(1), 81. <https://doi.org/10.1037/1528-3542.8.1.81>
- Traoré, M. & Hurter, C. (2016). Exploratory study with eye tracking devices to build interactive systems for air traffic controllers. *International Conference on Human-Computer Interaction in Aerospace*, 1–9. <https://doi.org/10.1145/2950112.2964584>
- Troper, D. (2018). Android Wear, it's time for a new name. Retrieved September 10, 2021, from <https://www.blog.google/products/wear-os/android-wear-its-time-new-name/>
- Truschzinski, M. (2017). Modellierung und Vorhersage von mentaler Arbeitsbeanspruchung in einem Fluglotsenaufgabenexperiment [Modeling and predicting mental workload in an air traffic control task experiment]. In M. Eibl & M. Gaedke (Eds.), *INFORMATIK 2017* (pp. 2295–2300). Gesellschaft für Informatik.
- Tscharn, R. (2019). *Innovative And Age-Inclusive Interaction Design with Image-Schematic Metaphors* (Thesis). Julius-Maximilians-Universität Würzburg, Germany. <https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/17576%20https://nbn-resolving.org/urn:nbn:de:bvb:20-opus-175762>
- Tseng, K. C., Lin, B.-S., Han, C.-M. & Wang, P.-S. (2013). Emotion recognition of EEG underlying favourite music by support vector machine. *2013 1st International Conference on Orange Technologies (ICOT)*, 155–158. <https://doi.org/10.1109/ICOT.2013.6521181>
- Tuncer, S. (2016). The Effects of Video Recording on Office Workers' Conduct, and the Validity of Video Data for the Study of Naturally-Occurring Interactions. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 17(3). <https://doi.org/10.17169/fqs-17.3.2604>

- Twidale, M., Randall, D. & Bentley, R. (1994). Situated evaluation for cooperative systems. *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 441–452. <https://doi.org/10.1145/192844.193066>
- Unbehau, D., Vaziri, D. D., Aal, K., Wieching, R., Tolmie, P. & Wulf, V. (2018). Exploring the Potential of Exergames to affect the Social and Daily Life of People with Dementia and their Caregivers. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3173574.3173636>
- Vaccaro, C. & Duca, G. (2011). From identification of requirements to the operational validation of an integrated solution: Approach and issues to design an effective human machine interface for air traffic controller working position in SESAR. *2011 Tyrrhenian International Workshop on Digital Communications - Enhanced Surveillance of Aircraft and Vehicles*, 3–8.
- van den Haak, M., De Jong, M. & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, 22(5), 339–351. <https://doi.org/10.1080/0044929031000>
- van Gennip, D., van den Hoven, E. & Markopoulos, P. (2015). Things That Make Us Reminisce, 3443–3452. <https://doi.org/10.1145/2702123.2702460>
- van Rijen, K., Cobbenhagen, T., Janssen, R., Olsen, M., Brankaert, R., Houben, M. & Lu, Y. (2020). RelivRing: Reliving Social Activities for People with Dementia. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. <https://doi.org/10.1145/3334480.3383033>
- van Someren, M. W., Barnard, Y. F. & Sandberg, J. A. (1994). *The think aloud method: a practical approach to modelling cognitive*. AcademicPress.
- Vischer, R. (1873). Über das optische Formgefühl: Ein Beitrag zur Aesthetik [On the Optical Sense of Form: A Contribution to Aesthetics]. In A. v. Hildebrand (Ed.), *Das Problem der Form in der bildenden Kunst*. Hermann Credner. <https://books.google.de/books?id=mb0IAAAAQAAJ>
- Wallace, J., Thieme, A., Wood, G., Schofield, G. & Olivier, P. (2012). Enabling self, intimacy and a sense of home in dementia: an enquiry into design in a hospital setting. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2629–2638.
- Wallace, J., Wright, P. C., McCarthy, J., Green, D. P., Thomas, J. & Olivier, P. (2013). A design-led inquiry into personhood in dementia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2617–2626.
- Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C. & Munoz, D. P. (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. *Frontiers in Neurology*, 9(1029). <https://doi.org/10.3389/fneur.2018.01029>

- Wang, P., Zhao, Q., Zhou, Y., Hong, Z., Guo, Q. & Liu, J. (2021). Emotional Comparison Between Semantic Dementia and Alzheimer's Disease. *Frontiers in Psychiatry, 12*(698). <https://doi.org/10.3389/fpsy.2021.680332>
- Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063.
- Watson, D. & Stanton, K. (2017). Emotion Blends and Mixed Emotions in the Hierarchical Structure of Affect. *Emotion Review, 9*(2), 99–104. <https://doi.org/10.1177/1754073916639659>
- Watt-Smith, T. (2015). *The book of human emotions: From Ambiguphobia to umpty - 154 Words from Aroud the World for how we feel*. Profile books.
- Westgate, E. C. & Steidle, B. (2020). Lost by definition: Why boredom matters for psychology and society. *Social and Personality Psychology Compass, 14*(11), e12562. <https://doi.org/10.1111/spc3.12562>
- Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors, 50*(3), 449–455.
- Wienrich, C., Döllinger, N., Kock, S., Schindler, K. & Traupe, O. (2018). Assessing User Experience in Virtual Reality – A Comparison of Different Measurements. In A. Marcus & W. Wang (Eds.), *Design, User Experience, and Usability: Theory and Practice* (pp. 573–589). Springer International Publishing.
- Wierwille, W. W. & Connor, S. A. (1983). Evaluation of 20 Workload Measures Using a Psychomotor Task in a Moving-Base Aircraft Simulator. *Human Factors, 25*(1), 1–16. <https://doi.org/10.1177/001872088302500101>
- Wierwille, W. W., Rahimi, M. & Casali, J. G. (1985). Evaluation of 16 Measures of Mental Workload using a Simulated Flight Task Emphasizing Mediatlional Activity. *Human Factors, 27*(5), 489–502. <https://doi.org/10.1177/001872088502700501>
- Wingbermuehle, C., Bryer, D., Berg-Weger, M., Tumosa, N., McGillick, J., Rodriguez, C., Gill, D., Wilson, N., Leonard, K. & Tolson, D. (2014). Baseball Reminiscence League: A Model for Supporting Persons With Dementia. *Journal of the American Medical Directors Association, 15*(2), 85–89. <https://doi.org/10.1016/j.jamda.2013.11.006>
- Wiratanaya, A., Lyons, M. J., Butko, N. J. & Abe, S. (2007). iMime: an interactive character animation system for use in dementia care. *Proceedings of the 12th international conference on Intelligent user interfaces, 262–265*. <https://doi.org/10.1145/1216295.1216342>
- Wood, W., Womack, J. & Hooper, B. (2009). Dying of boredom: An exploratory case study of time use, apparent affect, and routine activity situations on two Alzheimer's special care units. *American Journal of Occupational Therapy, 63*(3), 337–350.
- Woods, B., O'Philbin, L., Farrell, E. M., Spector, A. E. & Orrell, M. (2018). Reminiscence therapy for dementia. *Cochrane database of systematic reviews, (3)*. <https://doi.org/10.1002/14651858.CD001120.pub3>

- World Health Organization. (2018). *International classification of diseases for mortality and morbidity statistics (11th Revision)*. Retrieved April 3, 2020, from <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/546689346>
- Wright, P. & McCarthy, J. (2010). *Experience-centered design: designers, users, and communities in dialogue* (Vol. 3). Morgan & Claypool.
- Wright, P., McCarthy, J. & Meekison, L. (2003). Making sense of experience. In J. Karat & J. Vanderdonckt (Eds.), *Funology* (pp. 43–53). Springer.
- Wurhofer, D. (2018). *Characterizing Experiential Changes: Temporal Transitions of User Experience* (Thesis).
- Yamamoto, H., Yokokohji, Y. & Ishihara, T. (2020). Practicality Assessment of the Improved ICT-Based Dementia Care Mapping Support System. *American Journal of Alzheimer's Disease & Other Dementias*, *35*, 1–14. <https://doi.org/10.1177/1533317520935716>
- Yik, M., Russell, J. A. & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705. <https://doi.org/10.1037/a0023980>
- Yokoi, T. & Okamura, H. (2013). Why do dementia patients become unable to lead a daily life with decreasing cognitive function? *Dementia (London)*, *12*(5), 551–68. <https://doi.org/10.1177/1471301211435193>
- Zeiner, K. M., Burmester, M., Haasler, K., Henschel, J., Laib, M. & Schippert, K. (2018). Designing for Positive User Experience in Work Contexts: Experience categories and their applications. *Human Technology*, *14*(2), 140–175.
- Zhang, J., Hedden, T. & Chia, A. (2012). Perspective-Taking and Depth of Theory-of-Mind Reasoning in Sequential-Move Games. *Cognitive Science*, *36*(3), 560–573. <https://doi.org/10.1111/j.1551-6709.2012.01238.x>
- Zimmer, H. (2000). Frequenz und mittlere Amplitude spontaner elektrodermaler Fluktuationen sind keine austauschbaren Indikatoren psychischer Prozesse. [Frequency and mean amplitude of spontaneous electrodermal fluctuations do not convey identical information about mental processes.] *Zeitschrift für Experimentelle Psychologie*, *47*(2), 129–143. <https://doi.org/10.1026/0949-3964.47.2.129>
- Zwaag, J., ter Horst, R., Blaženović, I., Stoessel, D., Ratter, J., Worseck, J. M., Schauer, N., Stienstra, R., Netea, M. G., Jahn, D., Pickkers, P. & Kox, M. (2020). Involvement of Lactate and Pyruvate in the Anti-Inflammatory Effects Exerted by Voluntary Activation of the Sympathetic Nervous System. *Metabolites*, *10*(4), 148. <https://doi.org/10.3390/metabo10040148>

Appendices

A.1 Chapter 4: Interview Guide for Case Study 1

- Hast du schon mal eine Session auf Papier dokumentiert? Inwiefern war die Dokumentation mit der Uhr anders als auf Papier?
- Hättest du an der Position, wo du saßt auch auf Papier dokumentiert?
- (Bei mehreren beobachteten Nutzern:) Hast du für alle oder einzelne Nutzer Emotionen geloggt? Wie bist du vorgegangen?
- Gibt es Kritik und Verbesserungsideen? [kamen aber meist von alleine]

A.2 Chapter 5: Interview Guide for Case Study

Wir fragten Fluglotsen nach jedem Debriefing:

- Stört die Videoaufnahme während des Lotsens?
- Helfen die Zeitstempel und das Video dabei, das Simulations-Erlebnis zu reflektieren?
- Welche wichtige Komponenten zur des Simulations-Erlebnisses fehlen? (korrekte Emotionen, fehlende Emotionen)

Wir fragten Supervisor nach jedem Debriefing:

- Lenkt die App von der Beobachtung ab? Ist der Formfaktor okay und erlaubt die App eine effiziente Dokumentation der Emotionen?
- Fehlen für die Evaluation wichtige Emotionskategorien?

Wir fragten Forscher und Entwickler nach der Abschlussrunde:

- Sind die gesammelten Erkenntnisse zum Nutzererleben hilfreich für Konzeption und Entwicklung des Systems?

A.4 Chapter 8: Game Screenshots

A.4.1 Flight Control HD



Figure 2: In Flight Control HD (Electronic Arts, available on https://store.steampowered.com/app/62000/Flight_Control_HD/) players use their mouse to control aircraft. Aircraft can be directed to land at appropriate runways or heliports through flight paths established via drag and drop.

A.4.2 SuperTuxKart



Figure 3: In SuperTuxKart (SuperTuxKart Team, available on <https://supertuxkart.net/Download>, chose an avatar from the Open Source universe and steer their cart using the keyboard. They can gather items and fire them at opponents which were non-player characters on easy mode in our study. On the top left is the current ranking visualised with avatars, on the bottom left is the position of all players on the map, on the top right is a lap counter along with the elapsed time and in the right bottom corner players can read their current speed and available “extra nitro” from a gauge.

A.5 Chapter 8: Additional Detailed Data

A.5.1 Descriptive Data on Effectiveness

Table 1: Detailed descriptive data for thoroughness, validity and effectiveness (H1.1-H1.3), all $N = 35$

Quality criterion	Condition	Mean	SD	SE
Thoroughness flight control game	observed	.183	.148	.025
	self-documented	.086	.096	.016
Validity flight control game	observed	.177	.151	.025
	self-documented	.128	.137	.023
Effectiveness flight control game	observed	$Mdn = .014$	$IQR = .006 - .088$	
	self-documented	$Mdn = .008$	$IQR = 0 - .023$	
Thoroughness racing game	observed	.276	.237	.040
	self-documented	.086	.092	.016
Validity racing game	observed	.259	.208	.035
	self-documented	.156	.162	.027
Effectiveness racing game	observed	$Mdn = .05$	$IQR = .008 - .169$	
	self-documented	$Mdn = .011$	$IQR = 0 - .053$	

A.5.2 Emotion Frequency by Condition

Table 2: Frequency of emotional instances retrospectively reported by participants. Data is cumulated over all participants for both games and emotion categories are ordered descending by frequency.

Emotion	Condition	
	Gaming only	Self-documentation
Stress	354	293
Anger	306	305
Pride	258	220
Surprise	161	162
Boredom	72	70

A.5.3 EDA Effectiveness

Table 3: Descriptive and inferential statistical results for effectiveness of physiological and observational measures for the 35 participants, whose EDA data was usable. Tests were calculated on EDA peaks matched with all emotions. For comparison, we report additionally (1) the EDA matches without boredom and (2) the timely matches of the self-documentation condition. Note that the latter only include $n=15$ participants who had usable EDA data *and* were in the self-documentation condition, which disallows direct inferential statistics.

	Data source	Mean (SD)	Statistic
Effectiveness	Proxemo	.17 (.11)	$t(34) = -2.02,$ $p = .874, d = -.34$
	EDA	.21 (.11)	
	EDA_no boredom	.18 (.01)	
	self-documented	.09 (.08)	
Thoroughness	Proxemo	.37 (.14)	$t(34) = -5.71,$ $p > .999, d = -.97$
	EDA	.63 (.21)	
	EDA_no boredom	.57 (.19)	
	self-documented	.21 (.16)	
Validity	Proxemo	.44 (.19)	$t(34) = 4.85,$ $p > .999, d = .82$
	EDA	.36 (.18)	
	EDA_no boredom	.33 (.16)	
	self-documented	.36 (.20)	

A.5.4 GEQ Items and Factor Consistency

Several versions of the GEQ are in circulation with varying psychometric properties (Law et al., 2018). We used the longer version of the GEQ published in Nacke (2009) because it already comes with a translation to German, see table 5. However, we found similar issues with the factors as Nacke (2009), see table 4.

Table 4: As measures of reliability we calculated Guttman's $\lambda - 2$ and Cronbach's α (with 95% confidence intervals) for components of the GEQ. Resulting scores in our study compare to the values reported in Nacke (2009). Please note that we had removed item 35 from the subscale *negative affect* as its content did not apply to our games.

Factor	$\lambda - 2$	α	95% CI	α in Nacke (2009)	Items
Flow	.83	.82	[.77, .86]	.83	5, 15, 28, 31, 34
Competence	.90	.89	[.86, .92]	.90	2, 12, 17, 19, 23
Tension	.85	.84	[.80, .88]	.78	7, 9, 24, 27, 32
Challenge	.69	.66	[.56, .74]	.53	8, 13, 26, 29, 36
Positive affect	.83	.83	[.78, .87]	.89	1, 4, 6, 16, 22
Negative affect	.51	.47	[.32, .60]	.55	10, 11, 18, 25

Table 5: List of GEQ items we used with German translations taken from Nacke (2009).

Component	Item	German	English
Flow	5	I felt completely absorbed	Ich war völlig gefesselt
	15	I forgot everything around me	Ich habe alles um mich herum vergessen
	28	I lost track of time	Ich habe mein Zeitgefühl verloren
	31	I was deeply concentrated in the game	Ich habe mich sehr auf das Spiel konzentriert
	34	I lost connection with the outside world	Ich habe die Verbindung zur Außenwelt verloren
Competence	2	I felt skilful	Ich habe mich geschickt gefühlt
	12	I felt strong	Ich fühlte mich sicher
	17	I was good at it	Ich war gut
	19	I felt successful	Ich habe mich erfolgreich gefühlt
	23	I was fast at reaching the game's targets	Ich habe die Spielziele schnell erreicht
Tension	7	I felt tense	Ich war angespannt
	9	I felt restless	Ich fühlte mich ruhelos
	24	I felt annoyed	Ich habe mich verärgert gefühlt
	27	I felt irritable	Ich war reizbar
	32	I felt frustrated	Ich fühlte mich frustriert
Challenge	8	I felt that I was learning	Ich hatte das Gefühl, etwas zu lernen
	13	I thought it was hard	Ich fand es schwierig
	26	I felt stimulated	Ich fühlte mich stimuliert
	29	I felt challenged	Ich fühlte mich herausgefordert
	36	I had to put a lot of effort into it	Ich musste mich beim Spielen sehr anstrengen
Positive Affect	1	I felt content	Ich fühlte mich zufrieden
	4	I could laugh about it	Ich konnte über Sachen im Spiel lachen
	6	I felt happy	Ich habe mich glücklich gefühlt
	16	I felt good	Ich habe mich gut gefühlt
	22	I enjoyed it	Ich hatte Spaß
Negative Affect	10	I thought about other things	Ich habe an andere Dinge gedacht
	11	I found it tiresome	Ich fand es ermüdend
	18	I felt bored	Ich habe mich gelangweilt
	25	I was distracted	Ich war abgelenkt

A.6 Chapter 9: Perspectives for the Proxemo App

Our results showed that Proxemo is a thorough and effective method appropriate for collecting observational data in group sessions and is ready to be adapted to further domains.

For evaluations in the dementia context we conceptualised the Proxemo App for smartwatch usage and developed it to run on Tizen™(The Linux Foundation). Five years ago, when we started designing the app, the smartwatch market was just emerging and multiple hardware companies offered their own proprietary operating system with Tizen and Wear OS by Google⁵ (Google LLC, Mountain View, CA) as most promising competitors. Back then we chose Tizen over Android Wear, as it supported a richer variety of interactions with hardware components — such as the bezel or a crown — and hence appeared to be more future-proof regarding additional features. An early implementation of the Proxemo App on Wear OS (project available on <https://github.com/bja-engineering/EmoMem>) derived from our proposed interaction concept due to a lacking bezel and relied on multiple taps on the centre button for user switching. Lately, both worlds have been united. Beginning with the *Galaxy Watch4* released in September 2021, Samsung switched from Tizen towards an adapted version of Wear OS, to offer users access to a larger application market⁶. The *Classic* version of the Galaxy Watch4 (Samsung Electronics, Seoul, South Korea) inherits the rotatable bezel from the preceding smartwatch models Gear S2 and S3. So by porting the code from Tizen or adapting the Android implementation for Android, the initial interaction concept of the Proxemo app optimised for the dementia context can be used on the latest hardware.

As described in Chapter 5, the Proxemo app for the ATC context has been implemented for Android phones so far. The code is publicly available on GitHub (<https://github.com/bja-engineering/ProxemoTab>) and can be adapted to other domains or optimised for other screen sizes. Regarding the development of market shares in particular, the smartwatch sector (Analytics, 2021) but also phones (Group, 2021), an implementation for watchOS and iOS (Apple, Cupertino, CA) may grant broader access of the Proxemo App to UX evaluators.

Independent of the platform, a feature that allows the quick exchange of single emojis or predefined category-sets without accessing the source code might be beneficial for extreme users who regularly switch between evaluation contexts. Here, the difference in contexts does not even have to be as extreme as dementia care homes and air traffic control, serving as examples in this work. Within the context of dementia care, expected emotions may vary between evaluations of assistive technology with a more pragmatic focus and technology supporting reminiscence sessions to trigger more meaningful memories and support identity. The same is imaginable for the application domain of air traffic control: priorities, tasks, and the resulting set of likely

⁵Wear OS was formerly called *Android Wear* until September 2018 (Troper, 2018)

⁶<https://www.samsung.com/de/watches/galaxy-watch/galaxy-watch4-classic-black-bluetooth-sm-r880nzkadbt/>

experiences depend on several factors, including the controllers' working position, traffic complexity during the shift, and weather conditions. The more evaluators know about the context of the formative evaluation in advance, the better they can prepare their documentation strategy upfront and the more time they can spend on actually observing the users.

Alternatively to our arrangement of emotions based on ergonomic decisions (i.e., reachable for thumb interaction), emotion categories can also be organised based on theoretical considerations in a two-dimensional pattern (e.g., Plutchik, 2001; Scherer, 2005; Yik et al., 2011).

A.7 Description of method review criteria as proposed by Stanton (2017)

The document attached on the following two pages summarises basic information on Proxemo as a method as well as the Proxemo App. Operative details on, for instance, how to exchange the emoji files in the Tizen app are part of the published manual that can be found in the online repository.

CRITERIA	DESCRIPTION OF CRITERIA
NAME, ABBREVIATION	Proxy Documentation of Emotions, PROXEMO
AUTHORS	For questions on Proxemo contact stephan.huber@uni-wuerzburg.de
BACKGROUND & APPLICATION	<p>Available methods to measure User Experience (UX) require communicating one's current emotional state and thus the ability or resource to assess/reflect feelings and additionally to communicate them. Since cognitive decline or stressful work environments restrict those abilities, existing UX methods are not applicable to the domain of dementia.</p> <p>So far, Proxemo was used to evaluate reminiscing technology in dementia care and radar prototypes for air traffic controllers.</p>
DOMAIN OF APPLICATION	Proxemo is designed to document observed emotions in domains, where users do not have cognitive abilities or resources to self-reflect next to their primary task.
PROCEDURE & RECOMMENDATIONS	<p>Proxemo is method that requires a software for emotion documentation. The Proxemo App has been implemented for smartwatch or Android phone interfaces. By tapping on one of the pre-defined emojis, a timestamp with the associated emotion is set. The logged emotion data can be exported as CSV file and used a) directly for quantitative analysis and, more importantly b) synchronised with video data of the interaction where it serves as first annotation layer for qualitative analysis of the interaction. We recommend using Proxemo in addition to video data for formative evaluations, since it allows mapping observed emotions directly to the trigger event. For beginners and in the context of dementia we recommend to observe no more than three users at a time. Video snippets of situations annotated with Proxemo serve as basis for a thorough video analysis, or a retrospective analysis grounded on a joint video review – if the users' resources allow for it.</p>
FLOWCHART	<div data-bbox="494 1265 1396 1467" data-label="Diagram"> </div> <p>a) observe users' emotions during the interaction and capture the interaction on video b) document emotions on the Proxemo App c) synchronise emotion annotations with the video for an analysis of selected situations with the user or in the design team</p>
ADVANTAGES	<p>In preparation</p> <ul style="list-style-type: none"> • Proxemo is easy to use and learn – also by less technology affine evaluators. • Emojis can be exchanged according to the context or expected set of emotions. <p>During the observation of interactions, Proxemo allows to</p> <ul style="list-style-type: none"> • be used for evaluations of any kind of soft- or hardware, as long as the observer can view users' emotions. • log emotions quickly and discretely. • classify emotions in predefined categories. • log data for multiple users in group interactions.

	<ul style="list-style-type: none"> • Take notes in parallel or after the session. <p>During data analysis, timestamped emotions</p> <ul style="list-style-type: none"> • convey emotions to researchers/designers who were not in-situ. • facilitate the interpretation of situations where audio is corrupt, the user looks away from the camera or is hidden. • facilitate the navigation in video recordings. • provide a fast overview when analysed descriptively. • allow for comprehensive and efficient retrospective discussions with users, developers and researchers.
DISADVANTAGES	<p>Only predefined emotion categories can be used during an observation. Due to limited space we do not recommend to squeeze more than the 5+1 emojis on a smartwatch display. Yet we did not require more categories in contexts where we had sufficient space.</p> <p>Depending on the implemented features, data extraction and synchronisation with video may require manual effort and some training.</p>
RELATED METHODS	<p>The set of emotions in the dementia context was originally inspired by the <i>Observed Emotion Rating Scale</i> [2] but has been adapted to fit for interactions with technology and air traffic control. Mapping pictorials directly to their triggers has been used in the <i>LEM-tool</i> [1] already, which is restricted to web interfaces.</p>
APPROXIMATE TIME FOR TRAINING AND PROCEDURE	<p>Understanding the apps features is trivial (< 5min). Documenting an emotion takes about 1s. Retrospective analysis takes ~30% of the interaction time (compared to at least 100% in full retrospectives without annotations [3]).</p>
QUALITY CRITERIA	<p>Interrater Reliability: $\kappa > .7$ Validity: .18-.46 in lab studies with users unfamiliar to the observers; in the field likely more, if observers are familiar with the users and their context; Thoroughness: .18 - .28 in lab studies with users unfamiliar to the observers Effectiveness: .05-.12, which is higher than concurrent self-report (.01) and handwritten notes (.01), all measured in lab studies with users unfamiliar to the observers; Efficiency: subjectively lower workload (Raw TLX = 41.73) than handwritten notes (59.69) Downstream utility: <to be measured> Cost Effectiveness: <to be measured></p>
REQUIRED TOOLS	<p>Implementation projects for the Proxemo App are available for the following platforms: Tizen: https://github.com/bja-engineering/Proxemo Wear OS: https://github.com/bja-engineering/EmoMem Android: https://github.com/bja-engineering/ProxemoTab For synchronising the Proxemo watch with video data we used ELAN: https://tla.mpi.nl/tools/tla-tools/elan/download/</p>
LITERATURE	<p>[1] Gijs Huisman, Marco Van Hout, Elisabeth Van Dijk, Thea Van Der Geest, and Dirk Heylen. 2013. LEMtool: measuring emotions in visual interfaces. In <i>Proceedings of the SIGCHI Conference on Human Factors in Computing Systems</i> ACM, 351-360.</p> <p>[2] M. Powell Lawton, Kimberly Van Haitsma, and Jennifer Klapper. 1999. Observed affect and quality of life in dementia: Further affirmations and problems. <i>Journal of Mental Health and Aging</i> 5, 1, 69-82.</p> <p>[3] Jakob Nielsen. 1994. <i>Usability engineering</i>. Morgan Kaufmann.</p>

