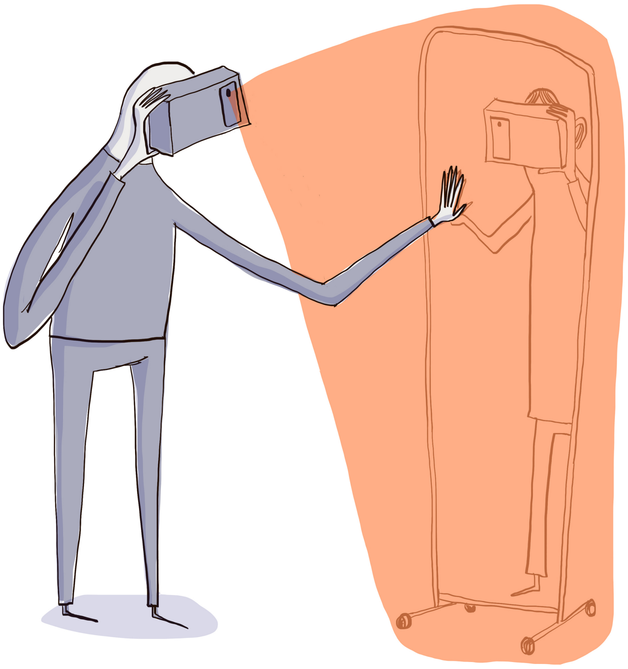
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD Thesis

Aske Mottelson

Computer-Cognition Interfaces:

Sensing and Influencing Mental Processes with Computer Interaction



Advisor: Kasper Hornbæk

Submitted: December 31st, 2018

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Colophon

Author
Aske Mottelson
Human-Centered Computing (HCC)
Department of Computer Science (DIKU)
University of Copenhagen

Title of PhD Thesis
Computer-Cognition Interfaces

Submitted / Defended
December 31st 2018 / ?

Advisor
Kasper Anders Søren Hornbæk, Professor, University of Copenhagen

Assessment Committee
Christian Holz, Assistant Professor, ETH Zürich
Tanzeem Choudhury, Associate Professor, Cornell University
Daniel Lee Ashbrook, Associate Professor, University of Copenhagen (chairman)

Frontpage Illustration
Mia Mottelson, http://mottelson.dk

Print
Typeset with Concourse and Equity, http://practicaltypography.com
Printed by Toptryk Grafisk

Abstract

The variety of information about users hidden in the details of interaction data is increasingly being utilized for recognizing complex mental processes. Digital systems can correspondingly influence mental processes of users, paving the way for new interactive systems that interface with the human mind. This thesis presents advances to such interfaces: through four papers I show how human affect and cognition can be sensed and influenced computationally.

Paper 1 presents two studies that together show that affect influences mobile interaction, which allows for binary discrimination between neutral and positive affect using sensor led machine learning classification. Paper 2 builds upon the methods presented in Paper 1 and extends the classification domain to dishonesty, also using mobile interaction data. The paper shows across three studies how dishonesty and honesty vary in interactional details, and how this difference can be utilized for estimating the veracity of user behavior based on features that are engineered by mobile interaction data.

Paper 3 presents a feasibility study of conducting virtual reality studies outside a laboratory, to increase heterogeneity and power. The paper shows through two studies how a range of VR tasks can be conducted without the use of an immediate experimenter, with participants carrying out experiments themselves. In Paper 4 I apply this methodology, and conduct a VR study with more than 200 participants to study how manipulations to avatars can influence affect responses. The paper presents evidence supporting the link between affect and avatars, and additionally discusses the interplay between positive affect and body ownership.

Dansk Resumé

Mellem detaljerne i interaktionsdata er skjult information om brugere, som i stigende grad bliver udnyttet til at genkende komplekse mentale processer. Digitale systemer kan samtidigt påvirke brugeres mentale processer, hvilket baner vejen for nye interaktive systemer med grænseflader til det menneskelige sind. Denne afhandling præsenterer fremskridt for sådanne grænseflader. Gennem fire artikler viser jeg, hvordan menneskelig affekt og kognition kan genkendes og påvirkes digitalt.

Artikel 1 præsenterer to studier, der sammen viser, at affekt påvirker mobil interaktion, hvilket muliggør binær diskrimination mellem neutral og positiv affekt ved hjælp af sensordrevet maskinlæringsklassifikation. I forlængelse af Artikel 1 udvider Artikel 2 klassifikationsdomænet til løgn, også her ved brug af mobil interaktionsdata. På tværs af tre studier viser artiklen, hvordan løgn og ærlighed varierer i interaktionsdetaljer, og hvordan denne forskel kan udnyttes til at estimere oprigtigheden af brugeradfærd baseret på variabler, som er konstrueret af mobil interaktionsdata.

Artikel 3 præsenterer en gennemførlighedsundersøgelse i at udføre brugerstudier med virtual reality uden for et laboratorium for at øge heterogenitet og styrkefunktion. Artiklen viser gennem to studier, hvordan en række VR-opgaver, under hvilke deltagerne selv eksekverer opgaverne uden en tilgængelig eksperimentator, kan udføres. I Artikel 4 anvender jeg denne metode og gennemfører et VR-studie med mere end 200 deltagere for at undersøge, hvordan manipulationer af avatarer kan påvirke deltagernes affekt. Artiklen fremlægger evidens for forbindelsen mellem affekt og avatarer, og diskuterer desuden samspillet mellem positiv affekt og virtuelt kropsejerskab.

Acknowledgements

This PhD thesis not only concludes three years of ivory tower solitude, but is the culmination of eight years having the *-student* suffix at the Department of Computer Science at University of Copenhagen (DIKU). Since my initial enrollment at DIKU in 2010 I have had the pleasure of receiving schooling and guidance from, as well as collaborating with, many incredibly devoted researchers. First a thank you to my supervisor Kasper Hornbæk; without your supervision I probably would have ended up as a miserable programmer. All members of BODY-UI also deserve recognition for sharing endless scientific (and even more non-scientific) discussions, last-minute submission panic, laughs, and in general for being thirsty researchers (literally and figuratively) and wonderful colleagues: thank you to Joanna Bergström-Lehthovirta, Jarrod Knibbe, Paul Strohmeier, Henning Pohl, Klemen Lilija, and Jess McIntosh.

I would also like to thank the great colleagues of the HCC section at DIKU, to Erik Frøkjær and Sebastian Boring who have both challenged my ideas and helped realize my potential, as well as other inspiring colleagues: Xiaoyi Wang, Carlos Tejada, Daniel Ashbrook, María Menéndez, Naja Holten Møller, and Pernille Bjørn.

In the summer of 2017 I had the pleasure of visiting the MIT Media Lab, under the guidance of Pattie Maes. I would like to thank her and the entire Fluid Interfaces group, that with incredible creativity continue to push the boundaries of computing. I am especially thankful for the productive collaboration with Misha Sra and Xuhai Xu.

In closing, I would like to extend my gratitude to a number of close friends and my family who have encouraged me throughout my doctoral research. First and foremost Gabrielle for supporting me unconditionally, while I continuously travel abroad for extended periods of work. A special thank you is also due to the Mottelson klan, and the best friends one could wish for in *Hej Drenge*; in particular for lively conversations during our Thursday dinners.

Lastly, to you reader, thank you for showing interest. I hope reading the rest of my thesis will provide just a fraction of the joy it has making it.

Aske Mottelson
Copenhagen, December 2018

The research presented in this thesis has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement 648785. I have additionally received travel grants from Augustinus Fonden, Rudolph Als Fondet, Christian og Ottilia Brorsons Rejselegat, Oticon Fonden, and The Swedish Research Institute in Istanbul.

Contents

1 Publications

Below I list the publications which I have (co-)authored during my PhD studies. The list separates papers that serve directly as the foundation of this thesis, while the rest are research digressions that have inspired, or are inspired by, the conducted research, and are as such not directly within the scope of the thesis; part of these are used for perspectives on the core content of the thesis. All the publications are peer-reviewed full papers published at ACM conferences (except [96] which is a journal paper; [175] which is in review at alt.chi; [152] which is unpublished). The list is ordered chronologically by publication date; the numbers refer to the indexes in the bibliography (which is ordered alphabetically).

Included in thesis

Published

[150] Aske Mottelson and Kasper Hornbæk. An Affect Detection Technique Using Mobile Commodity Sensors in the Wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 781–792, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971654`

[153] Aske Mottelson and Kasper Hornbæk. Virtual Reality Studies Outside the Laboratory. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, 9:1–9:10, Gothenburg, Sweden. ACM, 2017. ISBN: 978-1-4503-5548-3. DOI: `10.1145/3139131.3139141`

[154] Aske Mottelson, Jarrod Knibbe, and Kasper Hornbæk. Veritaps: Truth Estimation from Mobile Interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 561:1–561:12, Montreal QC, Canada. ACM, 2018. ISBN: 978-1-4503-5620-6. DOI: `10.1145/3173574.3174135`

Manuscript

[152] Aske Mottelson and Kasper Hornbæk. Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body. In *Manuscript*, 2018

Not included in thesis

Published

[155] Aske Mottelson, Christoffer Larsen, Mikkel Lyderik, Paul Strohmeier, and Jarrod Knibbe. Invisiboard: Maximizing Display and Input Space with a Full Screen Text Entry Method for Smartwatches. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '16, pages 53–59, Florence, Italy. ACM, 2016. ISBN: 978-1-4503-4408-1. DOI: `10.1145/2935334.2935360`

[208] Misha Sra, Aske Mottelson, and Pattie Maes. Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users. In *Proceedings of the 2018 Designing Interactive Systems Conference*,

DIS ’18, pages 85–97, Hong Kong, China. ACM, 2018. ISBN: 978-1-4503-5198-0. DOI: 10.1145/3196709.3196788

[209] Misha Sra, Xuhai Xu, Aske Mottelson, and Pattie Maes. VMotion: Designing a Seamless Walking Experience in VR. in *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS ’18, pages 59–70, Hong Kong, China. ACM, 2018. ISBN: 978-1-4503-5198-0. DOI: 10.1145/3196709.3196792

[17] Joanna Bergstrom-Lehtovirta, Aske Mottelson, Andreea-Anamaria Muresan, and Kasper Hornbæk. Tool Extension in Human–Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, Glasgow, UK. ACM, To appear in 2019

In review

[96] Kasper Hornbæk, Aske Mottelson, Jarrod Knibbe, and Dan Vogel. What Do We Mean by ‘Interaction’? An Empirical Analysis of 35 Years of CHI. *ACM Transactions on Computer-Human Interaction*. TOCHI, In review

[175] Henning Pohl and Aske Mottelson. How we Guide, Write, and Cite at CHI. in *Proceedings of the 2019 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’19, Glasgow, UK. ACM, In review

Part I

Introduction

2 Preface

The general rules and guidelines for the PhD programme at the Faculty of Science at University of Copenhagen [62] defines the objective of the PhD study as the following:

> *The PhD programme is a research programme aiming to train PhD students at an international level to independently undertake research [...]* [62].

The PhD program is equivalent to 180 ECTS credits, which corresponds to three years of full-time study. I initiated my PhD studies January 1st 2016, and submitted December 31st 2018. The intention with present thesis is to show that independent research at an international level was undertaken throughout this period.

2.1 Structure of thesis

The thesis is organized in four parts: *Introduction*, *Sensing*, *Influencing*, and *Perspectives*. As indicated in the title of the thesis, the main contribution of the presented research is in methods and applications of using computer interaction to sense and influence mental processes, namely affect and cognition. These are presented in the two middle parts; *Sensing* and *Influencing*. To prepare the reader for the content of the thesis, the *Introduction* part provides some brief context and overview. The *Perspectives* part is intended for zooming out a bit, and it considers methodological, ethical, and future directions of the work.

2.2 Selection of papers

The thesis' core content origins from four independent papers I have first-authored, of which three are published [150, 153, 154], and one is currently in manuscript [152]. These papers are quite different, both in terms of technology used (VR, mobile, web), methodology (crowdsourcing, laboratory study), and analysis (machine learning, statistical hypothesis testing). There are however also similarities that bind the individual work together. The papers all share the ideal of designing interactive systems that interface with users' thinking and feeling; the breadth of the research is a testament to the various applications of cognitive aspects of both applied and theoretical HCI. I have also had the opportunity to collaborate on several research projects that I have not spearheaded myself. Some of these works are used to provide some other perspectives to the thesis.

2.3 Abstracts of papers

Paper 1: An Affect Detection Technique Using Mobile Commodity Sensors in the Wild

Current techniques to computationally detect human affect often depend on specialized hardware, work only in laboratory settings, or require substantial individual training. We use sensors in commodity smartphones to estimate affect in the wild with no training time based on a link between affect and movement. The first experiment had 55 participants do touch interactions after exposure to positive or neutral emotion-eliciting films; negative affect resulted in faster but less precise interactions, in addition to differences in rotation and acceleration. Using off-the-shelf machine learning algorithms we report 89.1% accuracy in binary affective classification, grouping participants by their self-assessments. A follow up experiment validated findings from the first experiment; the experiment collected naturally occurring affect of 127 participants, who again did touch interactions. Results demonstrate that affect has direct behavioral effect on mobile interaction and that affect detection using common smartphone sensors is feasible.

Paper 2: Veritaps: Truth Estimation from Mobile Interaction

We introduce the concept of Veritaps: a communication layer to help users identify truths and lies in mobile input. Existing lie detection research typically uses features not suitable for the breadth of mobile interaction. We explore the feasibility of detecting lies across all mobile touch interaction using sensor data from commodity smartphones. We report on three studies in which we collect discrete, truth-labelled mobile input using swipes and taps. The studies demonstrate the potential of using mobile interaction as a truth estimator by employing features such as touch pressure and the inter-tap details of number entry, for example. In our final study, we report an F_1-$score$ of $.98$ for classifying truths and $.57$ for lies. Finally we sketch three potential future scenarios of using lie detection in mobile applications; as a security measure during online log-in, a trust layer during online sale negotiations, and a tool for exploring self-deception.

Paper 3: Virtual Reality Studies Outside the Laboratory

Many user studies are now conducted outside laboratories to increase the number and heterogeneity of participants. These studies are conducted in diverse settings, with the potential to give research greater external validity and statistical power at a lower cost. The feasibility of conducting virtual reality (VR) studies outside laboratories remains unclear because these studies often use expensive equipment, depend critically on the physical context, and sometimes study delicate phenomena concerning body awareness and immersion. To investigate, we explore pointing, 3D tracing, and body-illusions both *in-lab* and *out-of-lab*. The in-lab study was carried out as a traditional experiment with state-of-the-art VR equipment; 31 completed the study in our laboratory. The out-of-lab study was conducted by distributing commodity cardboard VR glasses to participants; 57 completed the study anywhere they saw fit. The effects found in-lab were comparable to those found out-of-lab, with much larger variations in the settings in the out-of-lab condition. A follow-up study showed that performance metrics are mostly governed by the technology used, where more complex VR phenomena depend more critically on the internal control of the study. We argue that conducting VR studies outside the laboratory is feasible, and that certain types of VR studies may advantageously be run this way. From the results, we discuss the implications and limitations of running VR studies outside the laboratory.

Paper 4: Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body

Human bodies influence the owners' affect through posture, facial expressions, and movement. It remains unclear whether similar links between virtual bodies and affect exist. Such links could present design opportunities for virtual environments and advance our understanding of fundamental concepts of embodied VR.

An initial outside-the-lab between-subjects study using commodity equipment presented 207 participants with seven avatar manipulations, related to posture, facial expression, and speed. We conducted a lab-based between-subjects study using high-end VR equipment with 41 subjects to clarify affect's impact on body ownership.

The results show that some avatar manipulations can subtly influence affect. Study I found that facial manipulations emerged as most effective in this regard, particularly for positive affect. Also, body ownership showed a moderating influence on affect: in Study I body ownership varied with valence but not with arousal, and Study II showed body ownership to vary with positive but not with negative affect.

3 Background

While the individual chapters of this thesis are meant to be understood on their own, I will in this chapter briefly present some general research paradigms that the presented research builds upon.

3.1 Mental Processes

A mental process or mental function is an umbrella term for all the things individuals are capable of doing with their minds, such as memory, perception, and emotions. The work presented in this thesis relates to interactive systems that interface with the human mental processes affect and cognition.

What is Affect?

The concepts of emotion, mood, and affect are sometimes mistakenly used interchangeably. This thesis proceeds from a view of affect inspired by Ekkekakis's efforts to untangle the conceptualization and measurement of affect [57]. He argued for distinguishing among three forms of affect: core affect, mood, and emotion. The most fundamental of these is core affect, underpinning moods and emotions. This is an evaluative feeling always available to the consciousness. Pleasure offers a clear example. Emotion, in contrast, depends on appraisal, and involves an object toward which the emotion is directed. Anger is one example. Finally, moods are long-term affective states in comparison to emotions; they are also less intense. Irritation is an apt example of a mood. In addition to presenting these basic categories, Ekkekakis argued that specific metrics employed in studies must operate from a particular view of affect (i.e., the entity composed of core affect, mood, and emotion). In particular, it must be clear whether the researcher is interested in dimensional or in categorical measures of affect.

3.2 Ubiquitous Computing and Mobile Sensing

In the 1999 article *The Computer for the 21st Century*, Weiser [237] presented his vision of a next *Ubiquitous Computing* paradigm. Weiser envisioned that computers move into the background, operating without our explicit awareness:

> *[...] specialized elements of hardware and software, connected by wires, radio waves and infrared, will be so ubiquitous that no one will notice their presence.* [237]

A prevalent research agenda within ubiquitous computing is sensing. Sensing refers to the activity of computationally inferring (often human) context from real life situations, such as assessing the amount of people in a room based on sounds [244], or predicting how tired a person is based on their phone activity [80]. Approaches vary in both sensing domains (e.g., physical activity, cognitive), sensor types (e.g., accelerometer, microphone), time frames (e.g., real-time, weeks), active or passive sensing (e.g., direct manipulation, background sensing), as well as in modeling approaches (e.g., correlation, regression, classification). Many sensing approaches require only basic sensors and computing units to function. Because of the prevalence of smart phones with a range of embedded sensors, powerful CPUs, and always-availability, mobile sensing have become the defacto standard sensing approach; notably being referred to as 'cognitive phones' [32]:

> *Because people carry their phone as they navigate through the day, phones are well situated to go beyond simple inference of classes by building up knowledge of the user's life patterns and choices. What if a phone could not only build lifelogs but also predict outcomes and assist the user? We argue the next step in the evolution of the phone is the cognitive phone.* [32]

In 2014, Liu and colleages [133] reported on the development of the Ubiquitous Computing field, by looking at author provided keywords in papers published at the UbiComp conference, which has existed since 1999. The analysis showed, among other things, that UbiComp is increasingly focusing on mobile devices [133].

To shed light on where the (increasing) UbiComp sub field of mobile sensing is heading, I have conducted a small semi-automatic paper review. The intention is to create an overview of the sensing domains (and their complexity) through all UbiComp sensing papers over time, thus showing large scale changes throughout the field's relative short history. The intention is not to make a comprehensive review (instead see for instance [176], for a review on sensing affect with mobile devices), which would require multiple paper venues, a stricter paper inclusion criteria, and a formal way of assessing the complexity of these sensing domains.

The UbiComp conference was chosen, because novel sensing techniques often are presented at this venue. Many other venues both within, and outside HCI, would be relevant for this exercise too, but would make a succinct (and somewhat automated) review difficult.

From the ACM Digital Library, I scraped the titles of the 1671 papers (full and short, not posters or demos) ever published at the UbiComp conference from 1999-2018. Note that the conference has had many names throughout its existence, such as *Handheld and Ubiquitous Computing*, *Ubiquitous Information Management and Communication*, *Ubiquitous Intelligence and Computing*, *Joint conference on Pervasive and ubiquitous computing*, etc. In recent time, the publication format has changed to a journal (*IMWUT*) which is also included in this analysis.

To find UbiComp papers about mobile sensing I searched paper titles for keywords. Any paper published at UbiComp with **two or more** of the following string patterns within the title were included for subsequent analysis: 'wearable*', '*phone*', 'mobile*', 'sens*', 'predict*', 'recogni*', 'classif*', 'detect*', 'estimati*', 'distinguish*', and 'identif*'; yielding 185 papers. Obviously this is not guaranteed to be the exhaustive list of mobile sensing papers at UbiComp, it should give a fair estimate, however. These papers were manually checked to see if they were about mobile sensing, interpreted broadly as any sort of data driven human context recognition using phones or equivalent sensors (thus excluding, e.g., capacitive sensing for interactional purposes, biometrics, device-device sensing, etc.). That resulted in 73 UbiComp papers. The domain of sensing was also noted during this part of the process.

Last, I ranked the sensing domains from 'cognitively simple' (e.g., device position, walking) to 'cognitively complex' (e.g., emotions, depression). The resulting plot is shown in Figure 3.1. The plot shows how the domains that mobile sensing papers at UbiComp have targeted are increasing in cognitive complexity over time. Where UbiComp historically has been a venue for presenting breakthroughs to sensing techniques of simpler human activity, such as device pose, walking, running, or other physical activity; the field has seen a shift towards sensing domains relating to complex human thinking. The work presented in this thesis contributes to the latter ideal.

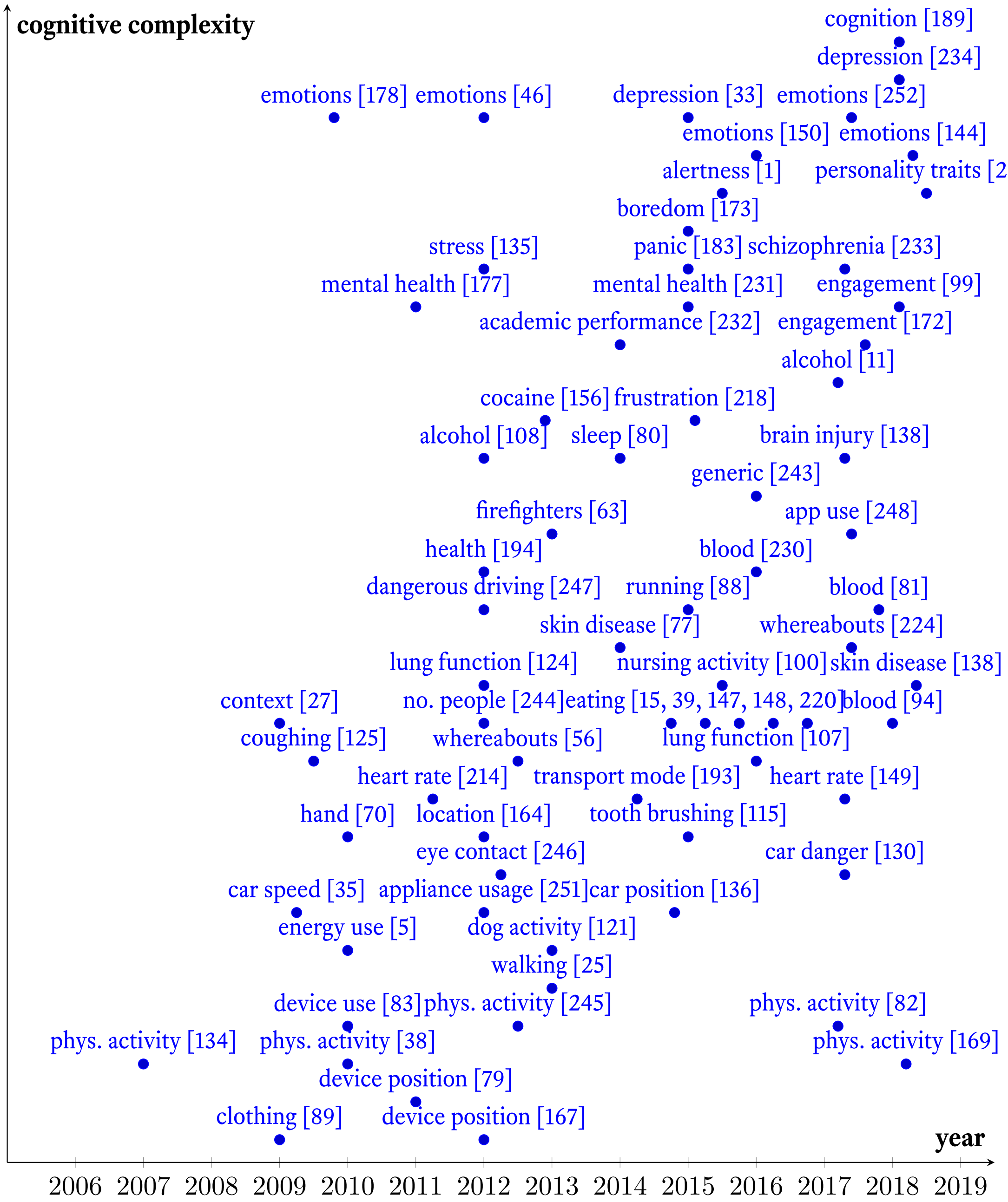


Figure 3.1: The domains of mobile sensing papers published at UbiComp 1999-2018. The cognitive complexity is a subjective assessment; overlaps have been handled by slightly moving domains with similar complexities. The analysis shows that sensing domains are increasingly cognitive complex.

3.3 Affective Computing

Affective computing is *computing that relates to, arises from, or influences emotions* [171]. It is an interdisciplinary field spanning computer science, psychology, neuroscience, engineering, linguistics and other disciplines. Even within computer science, affective computing draws attention in various fields such as human-computer interaction (HCI) and machine learning (ML). As such, it spans diverse research areas such as computerized simulation, recognition and induction of human emotions, as well as design of systems that interpret and respond to human emotion.

The first computerized emotion detection systems saw the light in the early 70's when Williams and Stevens [239] presented emotion detection using vocal features. Notably Suwa and colleagues [217] then presented an early attempt at inferring affect from facial features. Since, then numerous approaches to affect recognition have been pursued, mostly using audio and/or vision (see review by Zeng et al. [250]). While facial and vocal features provide convincing accuracies for affect detection, they pose privacy as well as availability issues.

Affect detection techniques that employ UbiComp methodologies have seen quite an increase within the last decade (e.g., [46, 71, 150, 178, 252]). Instead of relying on problematic data sources such as video or audio, there are increasingly more papers that report promising accuracies of affect detection using off-the-shelf sensors from smart phones (that may be privacy invasive, too), from passive (e.g., battery level) or active (e.g., touch force) interaction data. A recent review lists 42 of such studies [176]; across different affect conceptualizations and sensor types, different accuracies have been reported, with many above 80%.

3.4 Crowdsourcing

Crowdsourcing in research context refers to obtaining user responses from unsupervised participants. This practice is used extensively for surveys, online experiments, and data labeling. For many tasks, crowdsourcing divides the work into smaller tasks, and distributes work between participants to achieve an aggregate result. User studies are typically crowdsourced by giving participants small amounts of payment for conducting experiments on online crowdsourcing platforms such as Amazon Mechanical Turk [112]. Research shows that crowdsourcing, compared to laboratory studies, gives a higher diversity of participants [140, 166, 181] and can be done low-cost [28, 112, 140], reliably [28, 49, 182], and quickly [112].

With the rise of consumer-oriented AI systems, the need for throngs of labelled data has increased. With the increase of focus on external validity with user studies, and because of the maturity of online micro market platforms, crowdsourcing has received enormous attention within recent years. Papers at the CHI conference referred to 'crowdsourcing' first time in 2008 (through three papers/1%); at CHI' 18, 123 or 18% of papers mentioned the word 'crowdsourcing' (see Figure 3.2). This rapid increase in interest in crowdsourcing is not isolated to HCI; many other fields from social sciences to economics publish extensively with emperical data sourced from such models.

Crowdsourcing is a relatively broadly used term within HCI, that describe unsupervised study participation. Alternatives to micro market platforms include LabintheWild [179], which is a scalable way of conducting unsupervised and uncompensated experimentation. LabintheWild is an online experimental platform that provides participants with information about themselves in exchange for participation in experiments. In-the-wild mobile experiments have also been conducted. Henze et al. [91] showed how to crowdsource user studies for mobile games, by carrying out mobile gaming-based experiments. These games were distributed on the Android market store, and participation was thus unsupervised and uncompensated. Lafreniere et al. [122] also showed how fabrication can be crowdsourced, by having

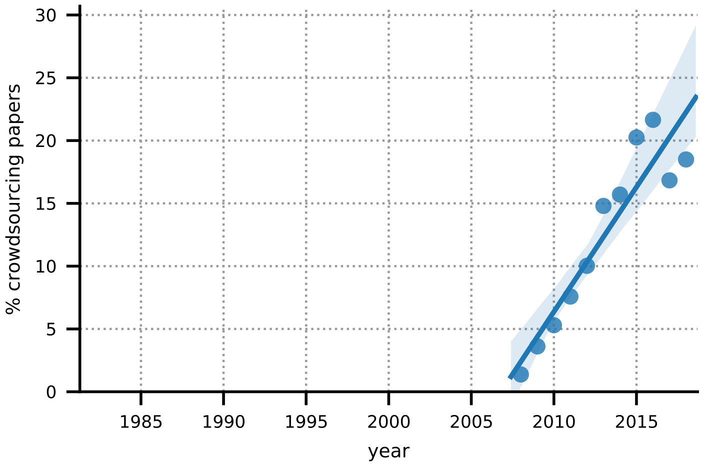


Figure 3.2: The percentage of papers at CHI including the term 'crowdsourcing'. Year 2008 was the first time at CHI where such a paper was published; in 2018 this amounted to 18.5% or 123 papers.

museum guests collaborative build structures as they visit an exhibition. Overall the use of crowdsourcing has become vast within HCI, but the term is also diversifying in its use.

Part II

Sensing

4 Affect Detection from Mobile Interaction

UBICOMP '16, SEPTEMBER 12–16, 2016, HEIDELBERG, GERMANY

An Affect Detection Technique using Mobile Commodity Sensors in the Wild

Aske Mottelson & Kasper Hornbæk
Department of Computer Science, University of Copenhagen
Njalsgade 128, DK-2300 Copenhagen, Denmark
{amot, kash}@di.ku.dk

ABSTRACT

Current techniques to computationally detect human affect often depend on specialized hardware, work only in laboratory settings, or require substantial individual training. We use sensors in commodity smartphones to estimate affect in the wild with no training time based on a link between affect and movement. The first experiment had 55 participants do touch interactions after exposure to positive or neutral emotion-eliciting films; negative affect resulted in faster but less precise interactions, in addition to differences in rotation and acceleration. Using off-the-shelf machine learning algorithms we report 89.1% accuracy in binary affective classification, grouping participants by their self-assessments. A follow up experiment validated findings from the first experiment; the experiment collected naturally occurring affect of 127 participants, who again did touch interactions. Results demonstrate that affect has direct behavioral effect on mobile interaction and that affect detection using common smartphone sensors is feasible.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g., HCI)

Author Keywords
Affective computing; affect detection; smartphone; touch; crowdsourcing

INTRODUCTION

Affect influences cognitive abilities and motor skill; it also influences human-human and human-computer interaction. As a result thereof, the research field of how computers can assess and respond to human affect has grown. Picard [29] popularized this research field of *Affective Computing*, and since then numerous systems that detect and respond to affect have been proposed (e.g., [6, 12, 23, 26, 34]).

Contemporary techniques for detecting human affect, however, have several limitations. Often these techniques are verified in laboratory experiments with few participants equipped with costly hardware. Another approach has been to study participants in office-settings over long periods of time, resulting in techniques that require long individual training to function.

This paper departs from findings in experimental psychology that provide evidence for a link between affect and movement; Coombes et al. [10] for instance found that exposure to unpleasant images caused greater error and faster performance in a subsequent square-tracing task. We use these findings to present an affect detection technique inspired by emotion psychology theory using commodity sensors, that works in the wild, without per-user training. We therefore address the following limitations in current affect detection techniques:

Specialized Hardware: Several techniques for inferring affect have been proposed, including audio/video approaches, or using specialized hardware, for instance heart rate variability or galvanic skin response (e.g., [27, 34, 36]). These techniques are often either intrusive to user privacy or less suitable for widespread adoption because of the need to acquire and wear custom sensors. We propose an affect detection technique using less invasive measurement methods, namely sensors already present in most commercial smartphones.

Controlled Laboratory Experiments: Previous studies concerning human affect and computer interaction have mostly conducted experiments using artificial tasks in controlled laboratory settings with relatively few participants (e.g., [4, 10, 18, 23]). The external validity of these studies makes it difficult to reason about how effective the proposed techniques are in more real-life settings. We present a crowdsourced method of gathering touch interactions and affective assessments, thus increasing generalizability.

Extensive Individual Training: Previous studies have used commodity hardware sensors to detect affect, such as using keystroke dynamics [14] or smartphone usage [26, 30]. However, these studies conducted extensive experiments lasting weeks to months, resulting in idiosyncratic models that require substantial per-user training to estimate affect effectively. We present an approach that requires 140 seconds of user interaction to assess affect, without any previous training with data from that user.

We report findings from two experiments, where participants recruited through crowdsourcing conducted general touch tasks on their own devices after being emotionally primed using video clips. The results show that the affective impact on touch interaction corroborates psycho-motor theory: Speed

UbiComp '16, September 12-16, 2016, Heidelberg, Germany
©2016 ACM. ISBN 978-1-4503-4461-6/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2971648.2971654

781

This chapter is almost identical to the paper shown to the left [150], presented at UbiComp '16. The paper was an immediate continuation of the work presented in my Master's thesis *Detecting and understanding the impact of affect in touch-based computer interaction* (2015).

In this chapter, I present a technique to assess smart phone users' affect using only built-in sensors. The raison d'être for the introduced approach is to rely on non-specialized hardware sensors that are available to most users, enabling affect detection as a commodity tool. The research presented here draws on the literature from activity sensing, that has seen an increase in complexity in recent years; namely applying techniques and computational approaches that have previously been applied for distinguishing between physical activities, to foster predictions about cognitive activities.

In this chapter, I describe the data acquisition, setting up experiments, and making models for inferring user affect. We did, however, also gather other labels from the participants including gender and handedness. While neither of these ideally should be considered binary, the distribution of the data does suggest that classifying participants in these categories is easier than for instance considering affect.

The biggest struggle with building the models that are presented in this chapter arose from differences across participants' devices. While two IMUs operating with completely different scales by definition are incomparable, this work also did not result in comparable representations for swipes, touches, and taps because of differences across touch screen sizes and phone form factors. Figuring out a reasonable cross-device model representation of interaction thus presents itself as an open research opportunity, specifically for normalization between screen estates.

This chapter is based on a collaborative effort as described below.

Title
An Affect Detection Technique Using Mobile Commodity Sensors in the Wild

Authors
Aske Mottelson and Kasper Hornbæk

Journal
Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing

DOI
10.1145/2971648.2971654

What was the role of the PhD student in designing the study?
The PhD student was the first author of the paper,
and responsible for the design of the described studies.

How did the PhD student participate in data collection and/or development of theory?
The PhD student was responsible for study implementation,
execution, data collection, and theory development.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed to all parts of the manuscript.

Did the PhD student read and comment on the final manuscript?
Yes.

4.1 Abstract

Current techniques to computationally detect human affect often depend on specialized hardware, work only in laboratory settings, or require substantial individual training. We use sensors in commodity smartphones to estimate affect in the wild with no training time based on a link between affect and movement. The first experiment had 55 participants do touch interactions after exposure to positive or neutral emotion-eliciting films; negative affect resulted in faster but less precise interactions, in addition to differences in rotation and acceleration. Using off-the-shelf machine learning algorithms we report 89.1% accuracy in binary affective classification, grouping participants by their self-assessments. A follow up experiment validated findings from the first experiment; the experiment collected naturally occurring affect of 127 participants, who again did touch interactions. Results demonstrate that affect has direct behavioral effect on mobile interaction and that affect detection using common smartphone sensors is feasible.

4.2 Introduction

Affect influences cognitive abilities and motor skill; it also influences human-human and human-computer interaction. As a result thereof, the research field of how computers can assess and respond to human affect has grown. Picard [171] popularized this research field of *Affective Computing*, and since then numerous systems that detect and respond to affect have been proposed (e.g., [36, 52, 109, 131, 190]).

Contemporary techniques for detecting human affect, however, have several limitations. Often these techniques are verified in laboratory experiments with few participants equipped with costly hardware. Another approach has been to study participants in office-settings over long periods of time, resulting in techniques that require long individual training to function.

This paper departs from findings in experimental psychology that provide evidence for a link between affect and movement; Coombes et al. [44] for instance found that exposure to unpleasant images caused greater error and faster performance in a subsequent square-tracing task. We use these findings to present an affect detection technique inspired by emotion psychology theory using commodity sensors, that works in the wild, without per-user training. We therefore address the following limitations in current affect detection techniques:

Specialized Hardware: Several techniques for inferring affect have been proposed, including audio/video approaches, or using specialized hardware, for instance heart rate variability or galvanic skin response (e.g., [135, 190, 215]). These techniques are often either intrusive to user privacy or less suitable for widespread adoption because of the need to acquire and wear custom sensors. We propose an affect detection technique using less invasive measurement methods, namely sensors already present in most commercial smartphones.

Controlled Laboratory Experiments: Previous studies concerning human affect and computer interaction have mostly conducted experiments using artificial tasks in controlled laboratory settings with relatively few participants (e.g., [30, 44, 71, 109]). The external validity of these studies makes it difficult to reason about how effective the proposed techniques are in more real-life settings. We present a crowdsourced method of gathering touch interactions and affective assessments, thus increasing generalizability.

Extensive Individual Training: Previous studies have used commodity hardware sensors to detect affect, such as using keystroke dynamics [59] or smartphone usage [131, 173]. However, these studies conducted extensive experiments lasting weeks to months, resulting in idiosyncratic models that require substantial per-user training to estimate affect effectively. We present an approach that requires 140 seconds of user interaction to assess affect, without any previous training with data from that user.

We report findings from two experiments, where participants recruited through crowdsourcing conducted general touch tasks on their own devices after being emotionally primed using video clips. The results show that the affective impact on touch interaction corroborates psycho-motor theory: Speed and precision of motor control varies with affective states. Using participants' touch data it was possible to model affect using 140 seconds of smartphone sensor data, with 89.1% accuracy for binary (high/low) self-assessed affect, and 1.33 RMSE on a 1-7 positive-negative scale. An additional study using participants' natural occurring affect showed a similar effect, although with less confidence; it was possible to detect binary affect with 69.0% accuracy (1.32 RMSE, 1-7 positive-negative), binary valence with 81.7% (1.61 RMSE, 1-9 SAM), and binary arousal with 67.5% (1.88 RMSE, 1-9 SAM).

4.3 Background and Related Work

Research on how emotions influence physical expression has been treated extensively, starting with Darwin's work in the $19th$ century [51]. Darwin proposed that emotions are products of evolution; discrete emotions trigger actions that have been favorable to survival [51, 129]. This widely supported view suggests that emotions are organized around a motivational base such that our state of mind motivates beneficial physical expression. For instance, when a negative or threatening situation occurs, a fast reaction with less emphasis on precision optimizes chances of survival.

A multitude of emotional modalities and their respective physiological responses have been studied. It has been shown that moods influence cognitive performance, general health and well-being, creativity, decision-making processes, and social relationships [31, 102]. Most commonly studied is the relation between emotions and facial expressions [31, 52], but studies have also shown that affect has a significant impact on both motor skills [21, 44] and voice intonation [36, 65], in addition to body movements and body postures [229].

Models of Affect

Popular models of emotions are Plutchik's emotion wheel [174], that offers a hybrid between emotional dimensions and discrete emotions, and Russel's circumplex model of affect [184] which describes linear combinations of two dimensions, valence and arousal, as varying degrees of stimulus (valence) and intensity (arousal). Sometimes these two dimensions are extended by a third dimension (the PAD model [143]), dominance, which describes the degree of control exerted by a stimulus.

The proposed emotional models are rather complex, and their respective self-assessment measures are therefore extensive, making them less suitable for an in-the-wild mobile experiment. Also self-assessment measures such as PANAS [236] or SAM [24] may reveal the purpose of the study to participants filling them out, distorting the elicited affect [238]. In this study we are interested in the direct physiological response to affective stimuli, and we therefore employ the term *affect*, measured on a positive-negative scale, as proposed by Isen et al. [102].

Motion and Emotion

Emotions change our physical behavior; we smile when we are happy and our bodies tremble when angry. Drawing on the Darwinian view that emotions cause biological determined reactions, Ekman [58] proposed his theory of basic emotions. From cross-cultural field studies he found six discrete emotions to cause similar physical response in facial expressions across cultures. Body postures and movements have in a similar way shown to be influenced by emotions [229], which is essential to the core theory of the emergent field of embodied cognition; that cognition as well as affective aspects go beyond the brain and manifest themselves physically in our bodies, such that emotions provide embodied information. This paper draws upon this view: If motor behavior, including physical interaction with mobile devices, encodes affective information, this should be detectable by analyzing the user behavior patterns of interactions with mobile devices.

Previous studies also examined the affective aspects of computer interaction. Cairns et al. [30] studied the influence of emotions on a simple number entry task. The preliminary study showed that participants who were in a more positive emotional state were more accurate at entering numbers on a touch-based number pad. A study investigating the impact of emotions on the performance of computerized motor tasks was carried out by Coombes et al. [44]. The authors had 40 participants perform a computerized square-tracing task after being exposed to affective imagery. The authors concluded that exposure to affective pictures has direct behavioral consequences on speed and precision of performance on motor control.

These studies together show the correlation between movements and emotions, and provide evidence for the link between computer interaction and affect.

Affect Detection

Among real-time affective predictors, audio and vision-based techniques are by far the most common and robust (see Zeng et al. [250] for a review). Emotional states can also be inferred using physiological sensors; equipment measuring for instance heart rate variability or galvanic skin response have been used to infer stress levels [135, 215] and emotional states [190].

These sensors may be disturbing to their users (such as common galvanic skin response sensors), and the equipment is generally absent from home and office settings. Therefore, several studies have employed non-specialized equipment available at most home or office settings to detect affective states; such as using keystroke dynamics [59], touch-based gameplay strokes [71], computer mouse tasks [215] or smartphone usage [131, 173].

Gao et al. [71] studied mobile touch activity as an indicator of emotional states by extracting finger-stroke features from 15 participants during a *Fruit Ninja* game. Self-assessed emotional states coupled with touch strokes led to an 89.7% accuracy in binary arousal classification, with almost similar rates for valence. The small sample size and the specific task studied limit the generalizability of the results, and thus further work is needed to shed light on the implication of affect on general touch interaction.

Another example of utilizing commodity computer equipment for affective computing was conducted by Sun et al. [215] who inferred stress measurements through common computer mouse operations. In a study with 49 participants, physiological measurements and stress self-reports were measured, and the data collected were used to train a stress detection system with a stress rate detection accuracy of 70%.

LiKamWa et al. [131] leveraged smartphone usage to estimate participants' affective states. Thirty-two participants partook in a field study, where self-assessed mood was linked to phone activity. The authors failed to create a generic robust affective model, but reported 93% accuracy in affective classification using a personalized model with two months of training data.

By analyzing the rhythms of 12 participants' typing patterns on a standard keyboard in a field study, Epp et al. [59] reported the correlation between emotions and keyboard typings. The authors reported 77-88% binary classification accuracy for 15 emotional states.

Limitations of Earlier Work & Our Approach

Although the emotional influence on human motor aspects has been studied extensively, few studies concern the influence of affect on human-computer interaction. Promising results in affective detection depend on either specialized hardware or personalized models that require prolonged per-user training to function. Also, a prevalent shortcoming of the previous work on affective modeling stems from the use of relatively few participants in controlled experimental settings.

The intention of this paper is to study the affective impact on HCI on more immediate use, in more ecologically valid settings, using common sensing hardware, with more participants than related research. To do so, we report findings from two crowdsourced user studies providing evidence for the feasibility of in-the-wild affect detection from mobile interaction, in addition to an analysis of interaction patterns and their relation to affect. The overall reasoning behind the experimental approach used is to increase the external validity, and thereby the generalizability of the findings in comparison to previous affect and HCI related papers. To do so, we designed a set of general purpose mobile touch tasks covering the bulk of mobile interaction strategies employed in most touch based graphical user interfaces. To increase quantity and representativeness of the participants, we did online recruiting; participants installed our experimental software on their own devices, and followed on-screen instructions.

4.4 Experiment I: Emotion Elicitation

Based on psycho-motor theory we envisioned that the physical properties of mobile interactions would vary with affect. To manipulate emotion as an independent variable to study the contribution to mobile interaction made by affect, we employed emotion elicitation (see Coan and Allen [41]). This way we also enforce a bigger variance in affect among participants. The purpose of the experiment was therefore to gather information about participants' mobile interactions and couple them to their elicited affective states. The collected data was then used to train a classifier.

To gather data from mobile interactions in the wild, we developed a mobile application that participants installed on their own devices. The application collected demographic information, and elicited either neutral or positive affect using video. Subsequent to elicitation, participants conducted three touch tasks. We refrained from eliciting negative affect, due to ethical concerns.

Participants

We crowdsourced 276 participants who partook in the experiment for US $1. Half of the participants were from the USA, the rest were from other native English speaking countries or Western Europe. Ages ranged from 18-70 (M = 30.5), with 54% males, and 92% right-handed.

Apparatus

The application was implemented as an Android application, targeting Android ≥ 4.0. The app sent relevant user metrics over HTTP every 30 seconds to a server application created using the Python-based web application framework `webapp2` deployed at Google App Engine. The application forced a full-screen landscape orientation.

Procedure

Participants installed our experimental application on their own smartphones, and followed the same experimental procedure (see Figure 4.1). Half of the participants were placed in the positive group, and the other half in the neutral. The order of the touch tasks was randomized.

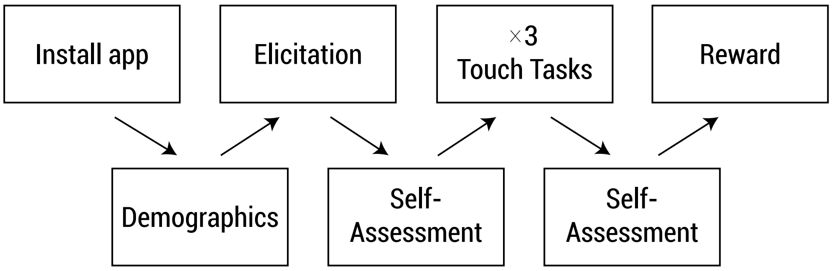


Figure 4.1: Overview of study procedure.

Design

The experiment used a between-subject design where participants, after being elicited with either positive or neutral affect, conducted three touch tasks. We collected affective self-assessments following both elicitation and touch tasks. We employed the self-assessment protocol proposed by Isen et al. [102]: five affective differentials, measured on 7-point likert scales, with four filler items (refreshed vs. tired, calm vs. anxious, alert vs. unaware, and amused vs. sober) and one deliberate item (positive vs. negative).

Emotion Elicitation

Emotion elicitation techniques or mood induction procedures (MIPs), are methods that allow for scientific investigations of emotions through experimentally controlling emotions [41]. A comparative study of MIPs by Westermann et al. [238] found that showing movie clips had a larger effect size compared to other procedures: *Film/Story + Instruction is significantly more effective than all other MIPs* [238]. It is also fairly simple to include video content into a mobile application, making movie clips a suitable MIP for this experiment.

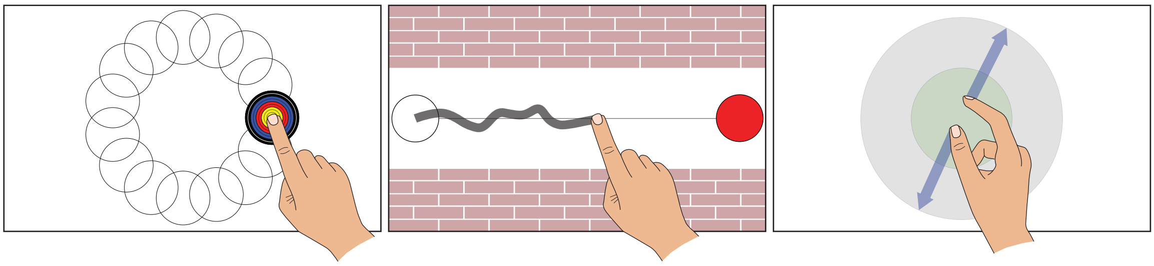

(a) (b) (c)

Figure 4.2: Three touch tasks: (a) Tapping, (b) Steering, and (c) Scaling.

Choice of Movies

A study by Schaefer et al. [188] reported mean affective assessments followed by watching a large variety of movie clips. We conducted a small-scale between-subjects movie-survey (N = 43) surveying four movie clips eliciting the highest mean positive affect from [188]. The results indicated that a scene from *There's Something About Mary* elicited the most positive affect (M = 6.13, SD = 1.13, on a 1-7 likert scale, 1=negative, 7=positive), and that the neutral movie clips (two clips from *Three Colors: Blue* and one from *The Lover*) scored significantly lower (M = 4.33, SD = 1.14), Cohen's d = 1.40; consequently, this movie configuration was chosen. The duration of the positive and neutral clips were 01:48 and 01:52, respectively. The neutral clips were shown with a 2 seconds black still in between. The order of the neutral movies was randomized.

Tasks

We wanted to design tasks that characterize the bulk of common operations on touch devices. Studies on human psycho-motor modeling using general computer mouse tasks include point-and-click, drag-and-drop, and steering through straight, narrowing, and spiral tunnels [2, 67, 206, 215]. Touch differs from mouse interaction as drag-and-drop is almost identical to steering and as touch interaction rarely requires complex steering. In addition, touch interfaces commonly employ multi-finger interactions. We therefore end up with three tasks: tapping [206], steering [2, 215], and scaling [222].

1. Tapping

This task (see Figure 4.2a) presented a series of circles one at a time, located at different locations in a circular formation. The participant had to tap the circular targets as fast and accurately as possible. The task is a common Fitts's Law exercise described by ISO 9241-400 [103], used in numerous previous studies. The specific setup, such as order of sequence and number of targets, adhered to MacKenzie [206] who reported best practices for this task. This touch interactivity corresponds to regular taps, frequent when users dial numbers using a number pad or enter text using a soft keyboard.

2. Steering

In this task participants were asked to draw a line through a straight tunnel from left to right (see Figure 4.2b), as fast and accurately as possible. The trajectory of the participant corresponds to a drag-and-drop activity, as [215]. Steering behavior is used in mobile contexts for instance when reordering home screen applications, panning in maps or scrolling web pages.

3. Scaling

In this task (see Figure 4.2c) participants were asked to scale a circle as accurately and fast as possible. Scaling was done by expanding the distance from the origin of the two fingers' initial positions, similar to [222]. Two-finger behavior is common when browsing the internet (zooming), or navigating maps (rotating and zooming).

Task Repetitions

To ensure the same task difficulty for all participants regardless of phone size, we used Fitts's Law [67] to calculate appropriate target sizes using constant ID's, hardcoded in the application. We used Fitts's Law settings as described by MacKenzie [206]. To use as much of the limited screen size as possible, and thus also ensuring largest possible target widths, we maximized the distance, D, according to the phones' screen sizes. This means that we could calculate appropriate target width sizes, by solving the Shannon formulation for W:

$$ID = log_2\left(\frac{D}{W}+1\right) \Rightarrow W = \frac{D}{2^{ID}-1}$$

To keep the experiment as short as possible, while ensuring validity we chose to present participants for the same condition 15-16 times per task, as shown in Table 4.1.

Task	*IDs*	*Targets*	*Rep./ID*	*Actions*
Tapping	8 (2 - 4.1)	15	1	120
Steering	8 (2 - 4.1)	2	8	128
Scaling	8 (2 - 3.05)	2	8	128

Table 4.1: Summary of experimental settings used to ensure the same level of difficulty across devices.

Experimental Conditions

The independent variable was binary affect (positive, neutral), ensured through emotion elicitation.

While some previous studies proposed behavioral predictors trained on any available mobile data; such as time of day, network-strength, battery-level and so on, we collected measurements related to the physical properties of mobile device interactions, devised from the theoretical link between movement and affect. We measured among other things speed and precision of participants' touch tasks, see Table 4.2 for a complete list of dependent variables.

Hypotheses

Coombes et al. [44] reported an effect of affective stimuli on both speed and precision in a square-tracing task. Their findings suggest that negative valence causes greater haste and/or reduced precision in motor tasks, which corresponds to Fitts's Law [67]; that speed and precision are inversely proportional. Taken above in to consideration in regards to the outcome of present experimentation, we hypothesize the following:

H1: Exposure to positive affective stimuli will decrease participants' speed when performing touch tasks compared to exposure to neutral affective stimuli.

Sensor	*Measurement*	*Unit*
Touchscreen	Finger position	(x, y)
Touchscreen	Touch area	mm^2
Touchscreen	Precision	$]0, 1] \in \mathbb{R}$
Pressure sensor	Pressure applied	$]0, 1] \in \mathbb{R}$
Timer	Duration of action	ms
Accelerometer	Acceleration	m/s^2
Gyroscope	Change of orientation	rad/s
Screen	Width × height	pixels
Phone	Brand, model	name
Questionnaire	Age	years
Questionnaire	Handedness	left/right
Questionnaire	Gender	male/female
Self-assessment	5 differentials	1-7 (×2)

Table 4.2: Data from mobile sensors collected during the experiment, as well as user-reported measures.

H2: Exposure to positive affective stimuli will increase participants' precision when performing touch tasks compared to exposure to neutral affective stimuli.

H3: Tasks completed immediately after exposure to affective stimuli cause bigger variance across the experimental groups, than tasks conducted later in the experiment.

Due to the lack of previous studies in this field, it is difficult to hypothesize about whether, and to what extent, mobile sensor data such as acceleration and rotation correlate with affect.

Analysis

Anomalies

We removed participants with zero variance in their self-assessments (i.e., all differentials were answered the same). We also removed participants who due to technical issues spent unreasonable long time watching the movies. After removing disqualified participants, 194 participants' data remained.

Emotion Elicitation

The effect of the elicitation was less significant compared to our movie survey, but the positive group did report higher positive-negative self-assessments (*M* = 5.10, *SD* = 1.48), than the neutral group (*M* = 4.77, *SD* = 1.46); Cohen's d = 0.22. Figure 4.3 depicts the variance in participants' self-assessments for every differential.

Participants' self-assessments showed to differ immediately after elicitation, but ended up almost identical post experiment (about 10 minutes, see Figure 4.4), showing a relatively fast decrease of the emotional effect after elicitation. The relative modest, although consistent, increase in self-assessments compared to the positive group are attributable to the absence of a negative elicited group.

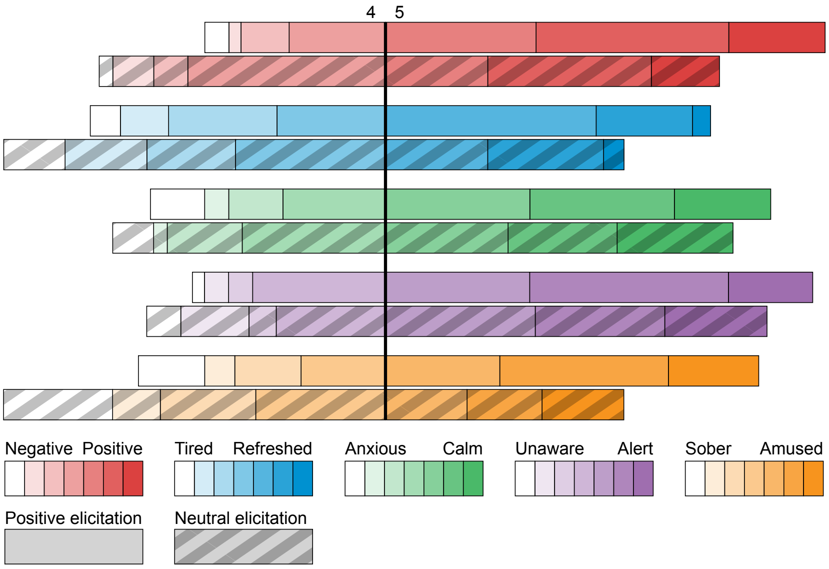


Figure 4.3: Self-assessments immediately after elicitation: The positive group reported higher assessments for all five differentials (1-7). The width of the bars represent the percentage of participants who reported the same value. The graph is fixed at the value 4 (mid of 1–7). The neutral group (striped) reported lower values for all differentials, and in general had a larger percentage of self-reports at the value 4.

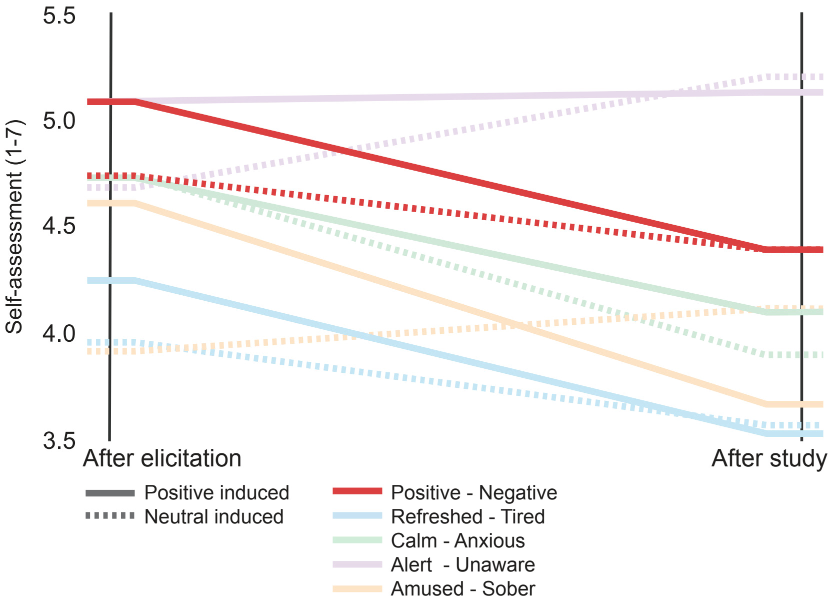


Figure 4.4: Temporal development of self-assessments: Difference between affective differentials (1=negative, 7=positive) immediately after elicitation, and after the task completion (with emphasis on the positive-negative differential). The trend is that the groups on average self-report differently immediately after elicitation, but have similar assessments towards the end of the study, showing the temporal equalization to the affective base level.

Feature	*Description*	*Interpretation*
speed	Distance traversed divided by duration, in px/ms	Higher values equal faster finger interactions
speedID	Speed divided by index of difficulty	Speed normalized by task difficulty
precision	Precision of activity: tap task uses distance to center, steering task average distance to center line, two fingers task distance to target scaling	1.0 equals perfect hit, 0.0 exactly on target edge, and $\leq$-1.0 completely off target
precisionID	Precision divided by index of difficulty	Precision normalized by task difficulty
accelerationY	Horizontal acceleration force, in m/s^2	Movements to left or right
accelerationX	Vertical acceleration force, in m/s^2	Up- and downward movements
Δacceleration	$\Delta\sqrt{x^2+y^2+z^2}$	Difference in aggregated acceleration, high values suggest constant shaking
rotationα	Rotation around z-axis, in rad/s	Rotations around axis pointing towards the participant
rotationβ	Rotation around x-axis, in rad/s	Rotations around the short edge of the phone
rotationγ	Rotation around y-axis, in rad/s	Rotations around the long edge of the phone
Δrotation	$\Delta\sqrt{\alpha^2+\beta^2+\gamma^2}$	Difference in aggregated rotation, high values suggest constant rotation
pressure	Applied finger pressure, from 0-1	High values indicate harder pressure
pressureDecline	Difference in pressure between beginning and end of interaction	Higher values indicate bigger pressure differences
devAngle	Difference in angle between fingers and centroid in beginning and end of interaction	Higher values indicate bigger differences in angles
centerAngle	Angle between horizontal line intersecting the centroid and line intersecting centroid and tap, $\Rightarrow \cos^{-1}(x/r)$	0-180 indicates activity on top half of target, and 180-360 on lower half.
approachDirection	Position of tap corrected for approach direction	Same as centerAngle but corrected for direction from last interaction
tapMovement	Movement of finger during tap	Usually very low, indicates slippage in pixels of finger during tapping
fingerDistance	Distance between two fingers	Distance between the two fingers in pixels

Table 4.3: The best features selected using recursive feature elimination. Features may be represented by their maximum, minimum, average, median and standard deviations measured throughout the experiment. The above 18 feature types represent a total of 46 features, out of the initial 352.

Durations

The neutral group on average completed all three tasks slower than the positive group. There was a close to significant difference in the task durations for the positive group (*M* = 420s, *SD* = 119s) and neutral group (*M* = 472s, *SD* = 255s); $t(190) = -1.859, p = 0.065$.

Data Reduction

The data anomaly removal resulted in 194 participants out of initial 251. On this data set we were able to achieve a binary classification accuracy of 67%. However, data analysis and visual inspection led us to believe that differences in hardware among participants distorted the analysis because of non-comparable scales and granularity of sensors. Another difficulty we encountered was normalizing touch interactions between phones with difference in pixel depth and aspect ratios. We therefore limited the data analysis to only include common phone models with similar sized screens, thus removing all tablets and any phone used by less than 3% of the participants. The resulting data contained 55 participants using seven phone manufacturers: Samsung, Asus, Google, LG, Motorola, and Sony. Whereas this dramatically reduced the sample, it allowed us to compare measurements obtained from phone sensors.

Feature Selection

Participants accounted for roughly 6MB of raw data each, primarily because of the very comprehensive capturing of motion, touch, and timing data. To facilitate the use of machine learning on this data, every participant's data needed to be represented by a number of features. The strategy was to include features found in related work, features derived from emotion theory, in addition to conceivable features computable using the gathered data.

Some features are applicable for all three tasks (such as applied finger pressure or finger size), while some only apply to a specific task (such as distance between fingers in the scaling task). Each feature is represented by several sub-features: the minimum, maximum, average, and median value, with some variation based on the applicability of the specific feature. Distances were normalized over screen sizes. We extracted a total of 352 features, predominantly computed using motion sensor measurements.

We used the checklist for optimizing variable and feature selection provided by Guyon and Elisseeff [84], which amongst other things suggests normalization (using l_2-norm), variable ranking, and outlier detection.

Feature Relevance

There are several reasons to estimate the individual relevance of the set of features: To facilitate visualization and understanding of the data, reduce storage requirements, reduce training time, and improve prediction performance [84].

We used recursive feature elimination with five-fold cross-validation on a linear SVM to do automatic tuning of the number of selected features. This yielded 46 as the optimal number of features. Table 4.3 lists the features with the highest discriminative powers.

Classifier Selection and Optimization

We tried a variety of classification methods; both linear, non-linear, and ensemble methods to classify affect. Specifically we compared k-Nearest Neighbor (k-NN), Support Vector Machines (SVM) with Radial Basis Function (RBF) and Linear kernels, Decision Tree (DT), Random Forest (RF), AdaBoost, Naive Bayes (NB), and Linear- and Quadratic Discriminant Analysis (LDA, QDA). The RBF kernel SVM showed to be the most promising predictor of the inspected algorithms. To find optimal parameter values, we used grid search on some bandwidth parameters calculated using the Jaakkola's heuristic [104]. The decision boundary created using the before-mentioned SVM can be seen in Figure 4.5.

Results

Classification

Table 4.4 shows the classification accuracies. We conducted three classifications: (1) a binary classification using the elicited affect, (2) a regression using the 7-point self-assessments protocol, and (3) a binary classification grouping participants by their self-assessments. Using the assessed affect (above/below median of assessment) results in a better accuracy than using the experimental groups. That is not very surprisingly, since participants' assessments should be closer to the actual affective

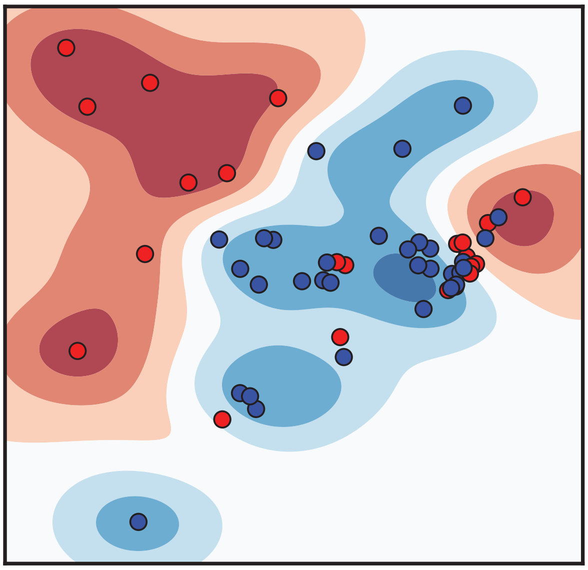


Figure 4.5: Data reduced to two dimensions using PCA with the decision boundary, created using an RBF-kernel SVM. Red and blue dots represent neutral and positive elicited participants, respectively.

state, than due to the elicitation. Overall the results are rather promising, showing that mobile device interactions are fairly efficient indicators of affect.

Variable	*Classes*	*Chance*	*Result*
Elicited affect (P/N)	2	60.0 %	87.3 %
Self-assessment (1-7)	7		1.33 RMSE
Grouped by median	2	56.4 %	89.1 %

Table 4.4: Classification accuracies for participants with similar hardware, N = 55. Results are obtained using an SVM with an rbf-kernel (regression or binary classification, measured using RMSE or Leave-one-out).

Affective Impact on Touch Performance

Analysis showed that 11 features had a significant difference among elicited groups, see Table 4.5. A t-test showed that all tasks provided measures that are influenced by affect. These results are consistent with the theoretically predicted directions of the parameters, specifically that speed decreases for positive affect, and that precision increases (see Table 4.5).

Speed

The neutral elicited participants on average had faster interactions speeds for several tasks and measures (see Table 4.5). On the contrary, the neutral group on average finished all tasks slower than the positive induced participants due to lower precision (and therefore an increased amount of repeated interactions), most significant for the scaling task.

Feature	*task*	*t*	*df*	*p*
speedID_min	3	3.556	39.748	.001
deltaAccelerationAgg_max	3	2.854	45.128	.006
rotationB_max	3	-2.669	38.076	.011
speed	3	2.618	45.192	.012
speedID_avg	3	2.518	43.919	.016
accelerationY_avg	3	2.313	34.883	.027
speedID_med	3	2.250	45.703	.029
accelerationY_med	3	2.233	39.937	.031
speedID_std_dev	1	2.080	21.431	.050
precision_avg	1	-2.077	21.497	.050
precision_min	2	-2.010	51.332	.050

Table 4.5: 2-tailed t-test results for features with significance at .05, with equal variances not assumed.

Acceleration

Neutral elicited, compared to positive elicited participants, performed tasks with higher aggregated acceleration $t(45) = 2.854, p = .006$. This was particular dominating on the y-axis. This implies that participants exposed to positive stimuli performed the tasks with smaller horizontal movements; that is positive affect led to steadier control of the device.

Rotation

Positive elicited participants account for higher max values of β-rotations (roll), $t(38) = -2.669, p = .011$. This means that the highest sudden rotational value around the x-axis for participants, were higher on average for participants exposed to positive stimuli. Rotations around the other axes followed same direction, but were not found significant.

Precision

Participants in the positive group performed tasks with higher precision than the neutral, most predominant for the tapping task, $t(21) = -2.077, p = .05$.

Two fingers

Positive elicited participants performed scaling tasks with bigger distance between their fingers. The angle between the two fingers and the center of the scaling target was also higher, although not statistically significant, $t(38) = 1.918, p = .063$.

Taps

The amount of pixels the finger drifted while performing taps showed not to have a statistical significant difference, $t(28) = 1.057, p = .3$. The position of the tap showed not to significantly differ, nor if it was normalized for approach direction.

Pressure

The data from force sensors showed no significant correlation to affect. The difference in decline of force throughout experimentation was not significant either.

The above descriptions presented several significant differences of features caused by affect, with speed, acceleration and precision as the most dominating affective indicators. The analysis showed that speed and precision of mobile interaction follow psycho-motor theory; that positive stimuli cause slower and more accurate motor behavior. Additionally, the results show that positive stimuli led to bigger movement (although less change of orientation) with the devices.

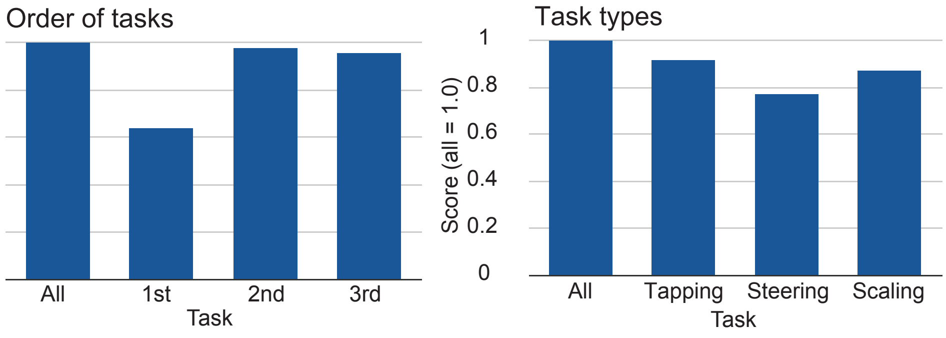
Insights

To get insights into the contribution to classification accuracy of each task type and order of tasks, we compared classification accuracies from predictors trained with different subsets of the data.

As evident in Figure 4.6a, all tasks individually provided classification accuracies over chance level, with tapping as the most robust task of estimating affect, and steering the worst. The combination of data from all tasks provided higher accuracy than any of the tasks alone.

Since the order of the tasks was randomized, each task was conducted at different times, relative to the other tasks. Analysis showed that the first task regardless of type of task, contrary to the hypothesis **H3**, scored relatively lower than the second and third (see Figure 4.6b). An explanation to this could be that the first task imposed a bigger challenge since participants were not completely aware of the application's mechanics, and therefore caused noisier data.

The accumulated classification accuracy did not increase when increasing the number of tasks from two to three, suggesting limiting the number of tasks for pragmatic affect detection.



(a) (b)

Figure 4.6: The relative contribution to classification accuracy of (a) task order, and (b) type of task (1.0=standardized classification accuracy of a predictor trained with data from all tasks combined).

A reduction in affect over time was evident, probably both due to the natural temporal reduction of the elicited affect and because the tasks themselves influenced participants' affective states (towards neutral). It is therefore natural to question which time window of data provides the optimal affective model, at least in terms of classification accuracy. Comparing classification accuracies of 80 different amounts of aggregated participant data, showed the highest classification accuracy at 89.1% for 140 seconds of data (see Figure 4.7).

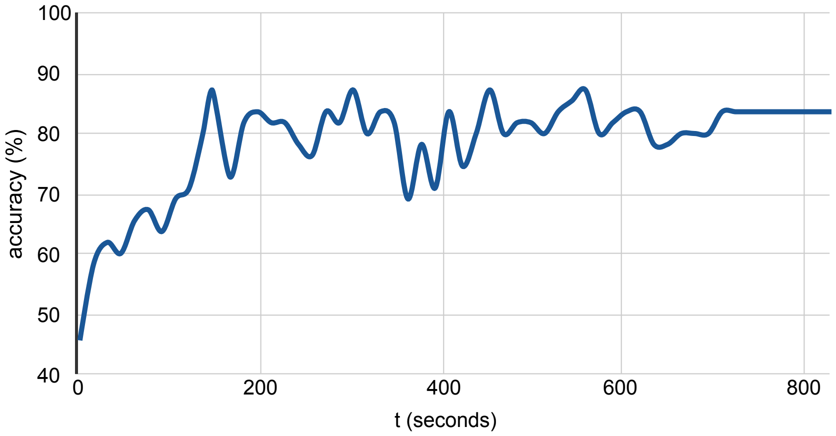


Figure 4.7: Classification accuracy as a function of the temporal amount of data the predictor is trained with. Classification accuracy peaks at 140 seconds with 89.1%.

4.5 Experiment II: Naturally Occurring Affect

The purpose of the second experiment was to understand the influence of naturally occurring affect on touch interaction in the light of the insights from the first experiment, with more participants, in order to validate the features used.

Design

The first experiment showed that an optimal classification accuracy was found using 140 seconds of participant data, and that the tapping task accounted for the highest discriminative power among the tasks; consequently this experiment employed only the Fitts's Law tapping task. To study the detection of participants' natural occurring affective states, we ran this experiment without emotion elicitation.

We extended the affective assessments to also include valence and arousal assessment using the pictorial 9-point Self-Assessment Manikin (SAM) [24] to allow higher dimensional affective assessments. Also, this assessment protocol would not sensitize participants to the study purpose because of the absence of elicitation in this study design. Experiment II allowed only a specific list of comparable phone models, ensured through strict settings at the Android app market. In summary Experiment II constituted the following:

- Within subjects design
- 1 task: Fitts's Law tapping task
- 140 seconds intended duration
- No induction; naturally occurring affect
- Removed demographics questions
- Five affective differentials and additionally SAM (9-point)

- Only comparable devices

Participants

127 participants participated for US $0.5, of which 29 also participated in the first experiment 75 days before. Seven were discarded because of zero variance in their self-assessments (5.5%), leaving N = 120.

Procedure

Because this experiment did not involve emotion elicitation and only contained one touch task, the procedure was simpler than the first experiment (see Figure 4.8).

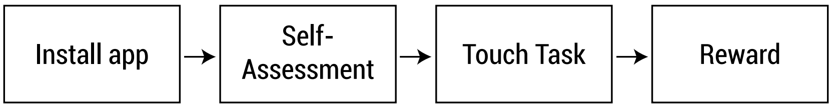


Figure 4.8: Overview of study procedure of Experiment II.

Results

Again an RBF-kernel SVM showed the most promising accuracies, with results over chance level for all protocols. Table 4.6 shows the accuracies of classifying binary affect ($\geq$ median) for the three protocols. This way, both groups are of approximately equal size in all classifications. Additionally results from predicting affects on the full likert scales, using a RBF-kernel Support Vector Regression (SVR) is shown. The best accuracy was found for valence detection, with 81.7%, or 1.61 RMSE for regression (valence of 1-9).

Variable	*Classes*	*Chance*	*Result*
Positive-negative (1-7)	7		1.32 RMSE
Valence (1-9)	9		1.61 RMSE
Arousal (1-9)	9		1.88 RMSE
Binary affect	2	54.2 %	69.0 %
Binary valence	2	51.7 %	81.7 %
Binary arousal	2	50.8 %	67.5 %

Table 4.6: Classification accuracies from Experiment II.

Kory and D'Mello [31] noted that affect detection using natural affect usually results in less accurate models than those constructed from elicited affect; results from this study reflect this finding. It is encouraging that it was possible to reach classification accuracies above chance level for all protocols, when training the model on data only from self-assessments of non-elicited participants.

This follow-up study showed that a rather short generic mobile touch task can estimate human affect with above chance-level accuracy, although less robustly than having a set of general tasks. It also showed that natural occurring affect is detectable using measurements from mobile interaction, although with less accuracy than elicited. We conclude that it is possible to design a generic affect detection task that detects affect above chance level, and that a standard Fitts's law task with data collection using mobile sensors is likely a good candidate for such a task.

4.6 Discussion

Experiment I showed that the way humans interact with mobile devices does encode affective information. The affect is detectable by statistically analyzing features representing physical behavior, inspired by the literature on psycho-motor theory [21, 44]; speed, acceleration, and precision of touch operations showed to be indicators of affective states. Using this information we developed an affect detection technique that showed an 89.1% accuracy in binary classification of affect. Experiment II showed that the insights from Experiment I could be used to develop a more generic affect detection mechanism with only one touch task. Without emotional elicitation we were able to link touch activity to participants' self-assessed affective states with accuracies well above chance level for all assessed protocols.

Together the findings show that movement during mobile interaction corroborates psycho-motor theory, and that in-the-wild and affordable affect detection can be implemented using already common mobile sensors. The technique described in this paper differs from related research by training a model on data gathered in the wild using commodity hardware, which in turn provides a more generic affect detection protocol that does not require per-user training.

Hypotheses

Based on findings in experimental psychology (e.g., [21, 44]) we hypothesized that (1) positive affect would decrease speed of touch tasks, and (2) positive affect would increase precision of touch tasks, and (3) tasks completed immediately after elicitation would better indicate affect than tasks completed later.

H1: Positive Affect Decreases Speed

Results corroborated this hypothesis: The speed of finger movement was significantly slower for participants who were exposed to positive eliciting stimuli. Normalizing the speeds with the index of difficulty of the tasks resulted in better indicators of affect. We also found that the overall task completion times for the positive elicited group were shorter than the neutral group; due to a higher tendency among positive elicited participants in completing tasks in the first attempt.

H2: Positive Affect Increases Precision

Results corroborated this hypothesis: Precision of tasks (distance to ideal touch activities) was found significantly higher for participants who were exposed to positive eliciting stimuli. This follows Darwinian emotion theory: That accuracy of movement decreases with more negative affect to rapidly prepare the organism to respond to threatening behavior [129].

H3: Encoded Affect is Stronger Immediately after Stimuli

This hypothesis was not confirmed: Features computed from the first task conducted immediately after exposure to emotion eliciting stimuli provided worse affective detection accuracies than data from the two subsequent tasks. We believe this is likely caused by the effect of learning and the difficulty in conducting tasks that are non-familiar.

Towards a Generic and Unobtrusive Technique

For generic affect detection we pursued a technique that ideally works independent of people, situations and devices. The participants we recruited showed a non-significant difference to the population of the world: comparing age, gender, and handedness with *The World Factbook* [40], we find $\chi^2 = 1.36$, $p < .51$, in line with crowdsourcing literature [66]. The proposed technique was verified using data gathered from uncontrolled settings, such that the whereabouts of participants were unknown, suggesting an at least wider situational applicability than the laboratory. By normalizing interactions for screen size, and index of difficulty we envisioned that cross-device comparison of interactions would be possible. Nevertheless we struggled to model affect robustly device-independently. Between devices comparability of mobile interactions is highly dependent on the mobile devices' physical properties and form factors. And since different sensing hardware offer different frequency and granularity of measurements, a device-independent affect model based on touch interactions is cumbersome to achieve. The results from the studies presented in this paper suggest that obtaining high accuracies in affect detection from mobile interaction requires predictors trained individually by phone model.

We employed established HCI practices [2, 67, 93, 206, 222] in designing the mobile tasks that worked as affective indicators. We envision that the the proposed affect detection technique could be implemented in virtually any mobile interface that offers different IDs: The validity of Fitts's Law has been verified at numerous occasions and interfaces both in experimentally controlled settings and in the wild [37]. As the proposed technique uses interactions originating from artificial touch tasks, it is uncertain to what extent the experimental tasks conducted by the participants in our study represent the actual bulk of touch interactions performed on touch devices. Therefore, it would be interesting to implement the affect detection technique proposed in this paper in the background of common applications such as text messaging or other popular applications such as social media or news applications, making the detection completely subtle. This way learning and boredom of conducting the tasks would not influence results either, since the affect detection would run on top of existing interfaces users are already engaging with. Work is still needed on the proposed affect prediction technique in order to seamlessly and robustly provide applications with information about the user's current affective state, needed by most real-life applicative scenarios. Our technique showed to peak in accuracy at 140 seconds, although providing above chance-level accuracies from 15 seconds of touch interaction and onwards (60% accuracy at 15 seconds, see Figure 4.7). While faster than existing models that employ related subtle affective predictors (e.g., [59, 131]), the duration of the calibration is relatively long for many real life scenarios. Ideally the technique would deliver instant affective predictions – further work is needed to achieve this accurately.

Future Work

Experiment I showed that accelerations on the y-axis (horizontal in landscape) and rotations around the x-axis (vertical in landscape) were significantly different across emotion elicitation. The reason why activity at these axes stood out, and conversely – why the others did not – is still a question. There is much room to deeper understand the mechanics of moving/shaking/rotating the devices and its attribute to affect. Because the affective computing literature previously has not concerned movements in these dimensions, the relation remains uncertain.

Results from Experiment II showed that valence classification resulted in a higher accuracy compared to arousal classification (81.7% vs 67.5%), contrary to related work (e.g., [71, 114]), probably due to the previous studies' domains being clearer indicators of arousal: Gao et al. [71] inferred arousal and valence from touch strokes in a gaming context, and Kleinsmith et al. [114] recognized affect from body movements and postures. Both studies reported classification accuracies contrary to the results presented in this paper – arousal classifications rates were consistently higher than valence. Intuitively one would also think that the level of arousal (intensity of emotion) would be stronger encoded in human movement or device interaction than the level of valence (pleasure of emotion), and thus easier detectable. Since previous work, contrary to this, confirmed this intuition, future investigations in the detectability of the emotional dimensions are needed: Is this difference due to the boredom of the tasks studied or the approach to the detection technique?

Out of an ethical concern we decided not to elicit negative affect in the wild, and Experiment I therefore grouped participants by either positive or neutral elicitation. We believe that having the full spectrum of affective elicitation would enforce a bigger discrepancy between mobile interaction behaviors, thus likely providing better classification results.

4.7 Conclusion

This paper presented an affect detection technique departing from psycho-motor theory, that equip smartphones with the ability to assess human affect, thereby allowing the devices to employ more human-human like interaction styles. The presented technique addresses limitations in contemporary affect detection techniques: (1) by using only commodity mobile sensors, the proposed technique avoids requiring specialized hardware; (2) by doing experimentation in the wild instead of in a laboratory with more participants, the external validity increases; (3) by employing a more generic affect detection technique, we achieve a model that does not require per-user training. In this paper we reported findings from two crowdsourced experiments that studied the implications of human affect on general purpose touch-based mobile interaction. The results of the two empirical studies presented in this paper, reflect findings in experimental psychology, and together confirm that affect has direct behavioral consequences for interactions with mobile devices. We show that it is possible to detect mobile users' affective states using off-the-shelf machine learning techniques. Results show encouraging affect detection accuracies, revealing at most 89.1% accuracy for binary affect classification.

5 Truth Estimation from Mobile Interaction

CHI 2018 Paper CHI 2018, April 21–26, 2018, Montréal, QC, Canada

Veritaps: Truth Estimation from Mobile Interaction

Aske Mottelson, Jarrod Knibbe, Kasper Hornbæk
Department of Computer Science
University of Copenhagen
DK-2300 Copenhagen, Denmark
{amot, jarrod, kash}@di.ku.dk

ABSTRACT

We introduce the concept of Veritaps: a communication layer to help users identify truths and lies in mobile input. Existing lie detection research typically uses features not suitable for the breadth of mobile interaction. We explore the feasibility of detecting lies across all mobile touch interaction using sensor data from commodity smartphones. We report on three studies in which we collect discrete, truth-labelled mobile input using swipes and taps. The studies demonstrate the potential of using mobile interaction as a truth estimator by employing features such as touch pressure and the inter-tap details of number entry, for example. In our final study, we report an F_1*-score* of .98 for classifying truths and .57 for lies. Finally we sketch three potential future scenarios of using lie detection in mobile applications; as a security measure during online log-in, a trust layer during online sale negotiations, and a tool for exploring self-deception.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation: (e.g., HCI)

Author Keywords

Lie detection; Polygraph; Dishonesty; Deception; Mobile Input; Smartphones

INTRODUCTION

We frequently lie, whether to advance our own aims or to protect others [13]. Consequently, we are also subject to many lies. Though this provides ample opportunity for practice, humans are only slightly better than chance at detecting lies and exhibit a positive bias in assessing the truth [3]. This deficiency has led to a century-long interest in lie detection. Visual, vocal, and physiological features of communication have all been explored [25], but, to date, natural language processing leads the way in identifying lies in digital communication. Through linguistic, psychological, and personal features, research has demonstrated success in classifying dishonest prose such as spam and deceptive reviews [18].

This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada
Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5620-6/18/04. . . $15.00
DOI: https://doi.org/10.1145/3173574.3174135

However, writing prose covers only a small part of our digital input. As we increasingly use our mobile devices for digital communication, our input also comes to include individual taps and swipes, such as button clicks, checkbox selection, and number entry. This leaves much digital activity open for deceptive behaviour with our approximately chance-level truth assessments. To this end, we explore a content-agnostic approach to mobile lie detection, ignoring the content of the input (i.e., input text), and enabling lie detection across a much wider spectrum of input.

To enable content agnostic lie detection, we draw on research demonstrating that a variety of information is hidden in the details of mobile input, such as stress [14], boredom [19], and affective states [16]. We explore whether dishonesty and deception are similarly hidden. Research suggests the presence of physiological responses to lying, such as increased hand/finger activity [25]. We hypothesize that these responses, although subtle, can be identified through smartphone sensors.

We test this across three crowdsourced smartphone studies. In Study I, we verify that lying on a smartphone exhibits similar behavioural cues to lying in conversation, and that this can support the separation of honest and dishonest responses. In this study, following the paradigm of Williams et al. [26], participants are instructed to tell the truth or lie, and the cues are identified through response time. Visible trends in other sensor data, such as input speed, motivate a second study using a more natural, spontaneous lying paradigm. The results from Study II show that acceleration, rotation, and inter-key-press duration can drive lie classification with an F_1-score of .77. Finally, we validate this result with an additional study, where we explore our identified features from Study II with a dice paradigm. In Study III we show 98% precision, and 97% recall for truths (F_1 = .98), and 65% precision and 59% recall for lies (F_1 = .57).

Following the studies, we sketch the concept of Veritaps, an additional layer of communication to assist mobile device users in their own lie detection accuracy. Veritaps enables users to automatically share a belief state indicator alongside their input. With high accuracy, Veritaps can label truthful input ✓. We can also label inconclusive taps and swipes ?, informing the user that they should use caution or seek further information in assessing this input. We illustrate the opportunities of Veritaps across a range of example scenarios, including (i) automated lie analysis when completing online

Paper 561 Page 1

Upon publishing the paper on commodity sensor-led affect prediction [150], I became interested in whether our methodology could be applied to other cognitive domains, as is an ongoing trend in UbiComp (e.g., towards boredom [173], stress [135], moods [131]). It became clear from analyzing the data from the affect study, that the specific cognitive construct (e.g., affect dimensions; conceptualizations of affect) resulted in quite diverse dependent outcomes (disregarding the task context, or subject variability). Absolute metrics used in the model that employed a two dimensional affect conceptualization (e.g., valence and arousal), does not necessarily generalize across a broader 'How do you feel' from positive–negative conceptualization (e.g., [102]).

The intention with this paper was thus not to deploy a variation of the same model to predict other cognitive aspects; rather we were interested if the content-agnostic approach (deploying an app that generalizes touch interaction; crowdsourcing user studies; feature engineering based on touch data) was reliant across multiple cognitive aspects. This would not only allow us to claim the methodology as more robust (i.e., that it can be applied to multiple domains), but would also increase its application.

This chapter is identical to the paper shown to the left [154], presented at CHI' 18. It presents the extension of the previous chapter's methods to prediction of lying, rather than affect. While lying is obviously a different domain than affect, both are mental processes closely related to cognition; lying has, for instance, consistently shown to require increased cognitive load, compared to telling the truth. This stems from the cognitive pressure of creating a real-world consistent story in parallel to expressing it.

Humans have a bias in assessing the veracity of an utterance, showing a sub-par 54% accuracy in classifying truthful statements, and only 61% accuracy at classifying lies [23]. As this chapter shows across three crowdsourced smartphone-based studies, lies are also harder than truths to accurately predict computationally. To emphasize that the approach works significantly better for classifying truthful behavior, the paper employs the term truth estimation, rather than lie detection.

This chapter is based on a collaborative effort as described below.

Title
Veritaps: Truth Estimation from Mobile Interaction

Authors
Aske Mottelson, Jarrod Knibbe, and Kasper Hornbæk

Journal
Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems

DOI
10.1145/3173574.3174135

What was the role of the PhD student in designing the study?
The PhD student was the first author of the paper,
and responsible for the design of the described studies.

How did the PhD student participate in data collection and/or development of theory?
The PhD student was responsible for study implementation,
execution, data collection, and theory development.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed to all parts of the manuscript.

Did the PhD student read and comment on the final manuscript?
Yes.

5.1 Abstract

We introduce the concept of Veritaps: a communication layer to help users identify truths and lies in mobile input. Existing lie detection research typically uses features not suitable for the breadth of mobile interaction. We explore the feasibility of detecting lies across all mobile touch interaction using sensor data from commodity smartphones. We report on three studies in which we collect discrete, truth-labelled mobile input using swipes and taps. The studies demonstrate the potential of using mobile interaction as a truth estimator by employing features such as touch pressure and the inter-tap details of number entry, for example. In our final study, we report an F_1-$score$ of $.98$ for classifying truths and $.57$ for lies. Finally we sketch three potential future scenarios of using lie detection in mobile applications; as a security measure during online log-in, a trust layer during online sale negotiations, and a tool for exploring self-deception.

5.2 Introduction

We frequently lie, whether to advance our own aims or to protect others [128]. Consequently, we are also subject to many lies. Though this provides ample opportunity for practice, humans are only slightly better than chance at detecting lies and exhibit a positive bias in assessing the truth [23]. This deficiency has led to a century-long interest in lie detection. Visual, vocal, and physiological features of communication have all been explored [228], but, to date, natural language processing leads the way in identifying lies in digital communication. Through linguistic, psychological, and personal features, research has demonstrated success in classifying dishonest prose such as spam and deceptive reviews [163].

However, writing prose covers only a small part of our digital input. As we increasingly use our mobile devices for digital communication, our input also comes to include individual taps and swipes, such as button clicks, checkbox selection, and number entry. This leaves much digital activity open for deceptive behaviour with our approximately chance-level truth assessments. To this end, we explore a content-agnostic approach to mobile lie detection, ignoring the content of the input (i.e., input text), and enabling lie detection across a much wider spectrum of input.

To enable content agnostic lie detection, we draw on research demonstrating that a variety of information is hidden in the details of mobile input, such as stress [135], boredom [173], and affective states [151]. We explore whether dishonesty and deception are similarly hidden. Research suggests the presence of physiological responses to lying, such as increased hand/finger activity [228]. We hypothesize that these responses, although subtle, can be identified through smartphone sensors.

We test this across three crowdsourced smartphone studies. In Study I, we verify that lying on a smartphone exhibits similar behavioural cues to lying in conversation, and that this can support the separation of honest and dishonest responses. In this study, following the paradigm of Williams et al. [240], participants are instructed to tell the truth or lie, and the cues are identified through response time. Visible trends in other sensor data, such as input speed, motivate a second study using a more natural, spontaneous lying paradigm. The results from Study II show that acceleration, rotation, and inter-key-press duration can drive lie classification with an F_1-score of .77. Finally, we validate this result with an additional study, where we explore our identified features from Study II with a dice paradigm. In Study III we show 98% precision, and 97% recall for truths (F_1 = .98), and 65% precision and 59% recall for lies

(F_1 = .57).

Following the studies, we sketch the concept of Veritaps, an additional layer of communication to assist mobile device users in their own lie detection accuracy. Veritaps enables users to automatically share a belief state indicator alongside their input. With high accuracy, Veritaps can label truthful input ✓. We can also label inconclusive taps and swipes ?, informing the user that they should use caution or seek further information in assessing this input. We illustrate the opportunities of Veritaps across a range of example scenarios, including (i) automated lie analysis when completing online forms, (ii) increased richness of trust in mobile messaging, and (iii) as a prompt to prevent self-deception.

We present the following contributions:

1. An exploration of lie detection across mobile devices, regardless of the input content.
2. Results from three studies, showing dishonesty affects user interaction with mobile devices.
3. Convincing classification rates of lies in mobile entry, potentially improving a user's ability to judge the veracity of others' mobile input.
4. Veritaps: a concept that allows users to share their belief states with other users and applications.

5.3 Related Work

Our work explores lie detection in mobile input. Specifically, we are interested in classifying lying through sensor data, rather than actual user input, in order to make lie detection available for a broader range of mobile input types.

Classifying Behaviors from Mobile Sensors

Research shows that complex cognitive and affective phenomena can be inferred using commodity sensors. The linearity of swiping, for example, correlates with emotions during game-play [71]. Similarly, speed, acceleration, and precision in touch input are indicative of affective states [151]. Mobile activity can also provide insight into a user's thinking, where app activity, battery level, and time of day are strong correlates of boredom [173].

Based on the idea of using mobile sensor data to support real-time inferences about human cognition, we explore indicators of lying in mobile sensor data.

Lie Detection

Deceptive behavior carries a range of verbal and nonverbal cues, and research has explored various strategies for using such cues to uncover deception. Among the most famous of these strategies is the polygraph. Polygraphs examine the subject's heart rate, galvanic skin response, respiration, and blood pressure as physiological markers of deception. It is widely accepted, however, that the interpretation of physiological responses and, thus, polygraph results, is 'a complex clinical task' [186]. The debate continues regarding the accuracy and applicability of polygraph testing. For example, a large body of research assessing the validity of polygraph techniques uses 'mock crime' scenarios, which inherently lack the consequences of real crime scenarios, and thus call into question the validity of their results [45].

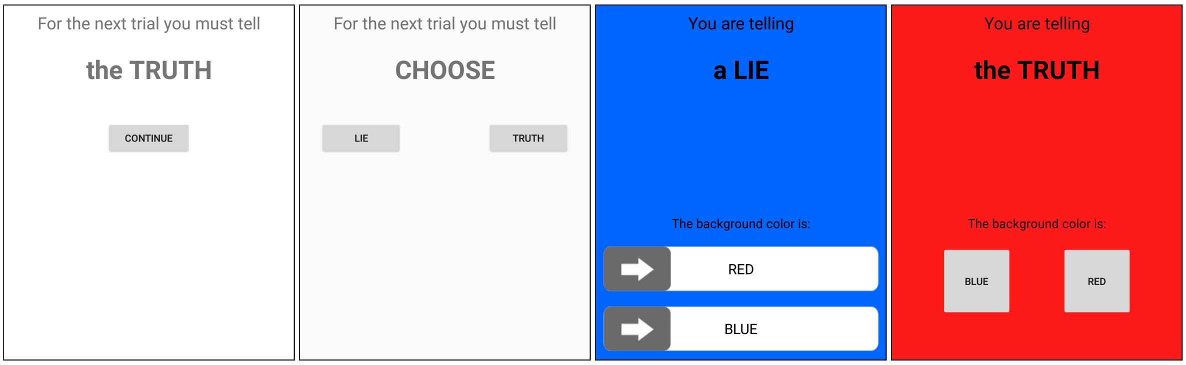
Other work has provided evidence on verbal, visual, and vocal cues to deception (e.g., [228]). Zuckerman et al. [253], for example, suggested that lying is a more cognitively complex task than telling the truth, requiring liars to formulate internally and externally consistent events. These greater cognitive challenges result in greater response latency, more hesitations, increased pupil dilations, and fewer heartbeats.

More recently, research has shown that lies include more complex imagery, longer words, and a greater number of pauses than truths [7, 87, 228]. This has led to automatic lie detection in text. Mihalcea et al. [146], for example, reported 71% accuracy in lie detection across three text corpuses. Ott et al. [163] used linguistic features (such as average word length or misspelling rate), psychological features (such as social or emotional clues), and personal features (such as references to money or religion), to classify spam and deceptive reviews.

Lying has also become a subject of exploration in crowdsourcing studies. Gino et al. [74] asked participants to report the outcome of random events (such as dice rolling or coin tossing). They identify lying across all of the input based on the deviation from the expected mean, offering an insight into lying across an entire study.

Opportunities for Lie Detection in Smartphones

While current research points towards physiological- and content-based lie detection, a common and robust strategy to lie detection has yet to be derived. We look for a commodity, content-agnostic approach to lie detection, that can be used to identify deception in basic mobile input; taps and swipes. We hypothesize that the bodily influences of deception can be measured using sensors available in consumer smartphones, making commodity lie detecting feasible.


For the next trial you must tell

the TRUTH

For the next trial you must tell

CHOOSE

You are telling

a LIE

The background color is:

You are telling

the TRUTH

The background color is:

(a) (b) (c) (d)

Figure 5.1: Example screens from the experimental application used for Study I. The figures show (a) a directed trial pre-screen, (b) a choice trial pre-screen, (c) a trial with sliders - where the participant should slide the ‘RED’ slider, in this case, and (d) a trial with buttons - where the participant should select the ‘RED’ button, in this case.

5.4 Study I: Simple Lies

Research shows that lying takes longer than telling the truth [227]. A common explanation is that the construction of a lie forces additional cognitive load compared to telling the truth, and thus causes longer response times.

Study I had two goals: (i) to establish whether lying through touch interaction on mobile devices produces results that are consistent with verbal responses in a laboratory, and (ii) to demonstrate the feasibility of separating honest and dishonest activity using mobile interaction data. We ran a mobile crowdsourced study of an experimental paradigm originally developed Williams et al. [240]. The participants were asked to either lie or tell the truth about the color of the screen, using common mobile UI elements. This paradigm offers an experimental procedure for studying both instructed and voluntary lies, while maintaining an even distribution of lies and truths. This provides a simple method for initially investigating differences in interaction patterns between telling lies and truths using mobile devices.

Task

The experiment progressed as a series of random-ordered trials, each beginning with an objective: TRUTH, LIE, or CHOICE. Directed trials (where participants were told to LIE or tell the TRUTH) presented a continue button, and the CHOICE trials had two buttons prompting the user to choose between lying or telling the truth (see Figures 5.1a and 5.1b).

Upon establishing the objective, participants were presented with a screen with a red or blue background. The participant's objective was written as a visual reminder at the top of the screen. The UI controls (button or slider) appeared at the bottom of the screen (see Figure 5.1d). The order of UI controls was randomised. Participants then had to activate the correct UI control according to (a) the color of the background, (b) the text on the UI control, and (c) the trial's objective (see Figure 5.1c). Trials were separated with a white screen for 1s. Participants were instructed to respond as quickly and accurately as possible. Participants were asked to lie and tell the truth half of the time each in the choice condition.

The task was similar to the original study [240], with the exceptions that: (i) instead of a lab-based study, participants were recruited online and completed the experiment on their own phones, (ii) vocal responses were replaced with selections using buttons or sliders, and (iii) the colors were changed to be visible for color blind (red and blue, instead of red and green).

Design

The study used a $2 \times 2 \times 2$ within-subjects design. The independent variables were honesty of response (lie vs. truth), type of instruction (directed vs. choice), and UI (button vs. slider). The dependent variable was response time. Each participant did a total of 192 trials, with 64 from the directed to lie condition, 64 from the directed to tell the truth condition, and 64 from the choice condition. In half of the trials participants responded by tapping a button, and the other half by dragging a slider. The order of trials was randomized. The study took 15 minutes on average.

Participants

We recruited 100 participants from Mechanical Turk, aged 19-59 (M = 31), 33 females. Participants installed our experimental application on their own Android smart phones (Android version $\geq$ 6.0), and followed onscreen instructions. Participants were reimbursed with $2.00 USD.

To ensure only qualified participation, we (i) required 90% HIT approval, (ii) had participants pass a qualification test about the task before starting the HIT, (iii) stored a unique device ID to avoid multiple participations, and (iv) ensured that app and MTurk HIT participation count matched.

Data

Of the 100 participants, ten never lied and one never told the truth in the choice trials, and were therefore removed. The remaining 89 participants performed a total of 16,671 trials. We removed (i) the first 10 trials per participant as warm-up rounds, (ii) 460 trials (2.9%) that lasted more than 4 seconds, and (iii) 623 incorrectly answered trials (3.9%). The analysis is made on a resulting data set comprising 14,788 trials.

Results

Lying took longer than telling the truth, both when answering with a button and a slider, and when being told whether to lie or when given the option to choose (see Figure 5.2).

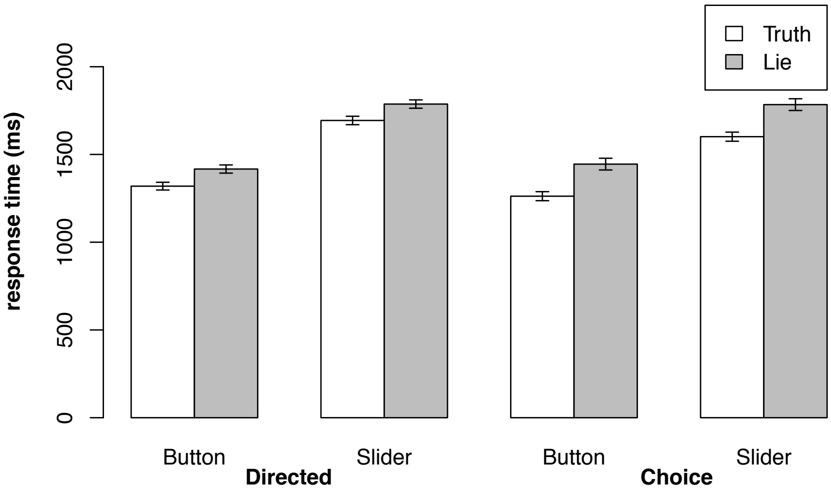


Figure 5.2: Response times for telling truths and lies using two different UIs both for the directed and choice conditions. Error bars show 95% confidence intervals. It took on average longer to tell a lie for both UIs.

Except for directed trials with the slider, participants took significantly longer when lying (on .5, see Table 5.1). The effect was larger when participants chose whether to lie or not.

Condition	*UI*	*F*	*df*	*p*		*Cohen's d*
Directed	Button	5.05	176	.026	*	0.34
Directed	Slider	3.03	176	.084		0.26
Choice	Button	8.94	176	.003	**	0.45
Choice	Slider	7.00	176	.009	**	0.40

Table 5.1: Results from an ANOVA comparing truths and lies. Lying caused significantly longer responses for close to all conditions, with the largest effect for choice trials.

Summary

The results show that constructing a lie is a cognitively harder task than simply telling the truth, reflected by the increased response time when participants were asked to lie about the background color of a mobile UI. This corroborates the findings of Williams et al. [240]. While the data sourced do not allow for an effective binary discrimination of truth and lies per entry, the results imply the feasibility of separating honest and dishonest activity using mobile interaction data, specifically timing in this case. We further investigate if this difference can be observed for other parameters in Study II-III.

Although not statistically significant, we observed that slider interactions were performed faster (by 4.4%) when telling the truth; $F(1, 175) = 2.16, p = 0.14$. Although mean response times pertaining to honest and dishonest behaviour were distinguishable in this study, we were keen to explore whether additional features become more prominent with (a) spontaneous lying, and (b) a more natural distribution of truths and lies (i.e., [225]).

5.5 Study II: Ultimatum Game

We ran a second study to analyze natural deceptive behavior. We employed a mobile version of the Ultimatum game, a commonly studied task in behavioral economics. In this variant the participants are offered an incentive to lie.

In the Ultimatum Game, the first participant (the proposer) receives a sum of money and proposes a division of the money between themselves and the second participant (the responder). The responder then either accepts the division, giving both participants the proposed funds, or rejects it altogether resulting in no payout for any of the participants. In the variant developed by Besancenot et al. [18], which we use, the proposer is given the opportunity to lie about the amount of allocated funds. Therefore, for each trial, the proposer to the responder (i) declares the amount that was allocated and (ii) proposes a division. This provides a monetary incentive for the participant to understate the provided funds, enabling the study of naturally occurring dishonest behavior.

Participants

We recruited 41 participants from the USA from Mechanical Turk, aged 22–63 (M = 33); 18 females, 36 right-handed. Participants were told that they were taking part in an economics experiment. Participants installed our experimental application on their own Android smart phones (Android version $\geq$ 6.0), and followed onscreen instructions. Participants were reimbursed \$1.00 USD, in addition to the money collected throughout the experiment, which ranged from \$1.74-\$4.52 (M = \$3.35). The experiment took at most 10 minutes. We employed the same qualification standards as for Study I.

Design

Each participant did 10 trials of proposals, excluding a warm-up round. The independent variable was funds allocated (25-99¢). The dependent variables were declared allocated funds and the proposed division of money. Additionally, throughout the trials, the mobile application collected data related to interaction with the UI using touch, pressure, accelerometer, and gyro sensors.

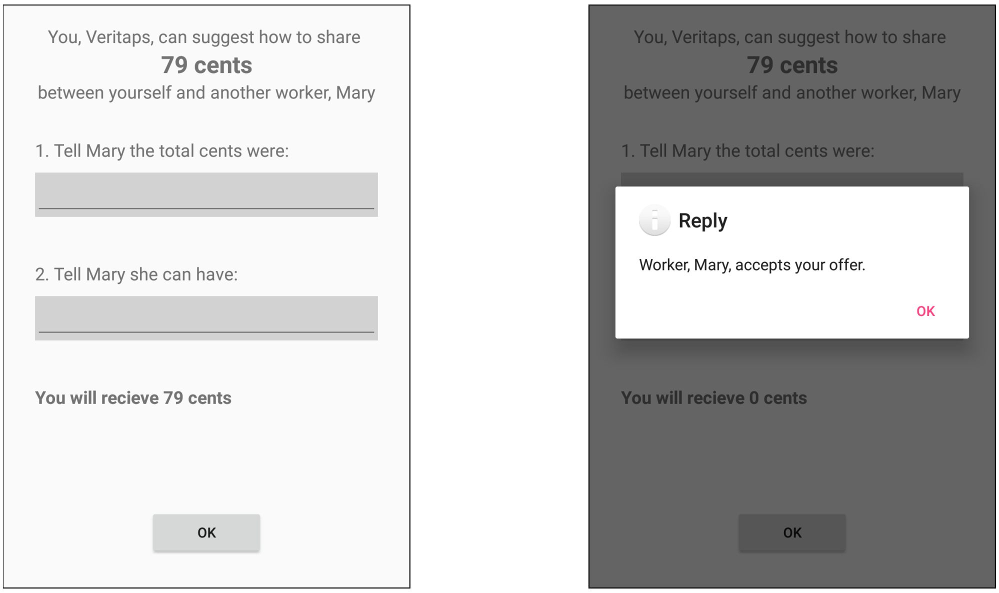
All participants had the role of the proposer. Participants were paired with an AI in the responder role, presented as the human worker *Mary* with a fictional worker ID. Mary would simulate human latency when responding to proposals, and would accept or reject proposals based on the available heuristics and basic economic and moral behavior: greed was punished while fair divisions were rewarded.

The AI was implemented as nine simple steps that would accept offers deemed favorable, or refuse offers that were either directly too low (< 25¢), or too unfair ($3P < F$), where P is the proposal, and F the declared funds. The AI would also reject offers when they repeatedly showed lower declared funds than expected from a random sample. If all steps passed, a 75% chance of acceptance was returned, to introduce some degree of unpredictable behavior.

Mary did not know whether the participant was in fact honest or dishonest, but instead reasoned based on the distribution of declared allocations from all trials. Mary accepted 86% of all proposals made (very similar to human behaviour observed in other of the Ultimatum game studies [160]).

Procedure

Upon installing and opening the experimental application, participants were informed that they were playing the proposer and were paired with our AI (under the guise of another crowdworker). For each of the 11 rounds, an amount of US cents between 25 and 99 were allocated to the participant. The participant would then, using num-pads, first state the amount of allocated funds (about which they could lie), and then propose a division (see Figure 5.3a).



You, Veritaps, can suggest how to share
79 cents
between yourself and another worker, Mary

1. Tell Mary the total cents were:

2. Tell Mary she can have:

You will recieve 79 cents

OK

(a)

You, Veritaps, can suggest how to share
79 cents
between yourself and another worker, Mary

1. Tell Mary the total cents were:

Reply

Worker, Mary, accepts your offer.

OK

You will recieve 0 cents

OK

(b)

Figure 5.3: The experimental application used in Study II. Figures show (a) the screen where participants declare the allocated funds and propose a division, and (b) a positive response from the AI, Mary, acting as another human worker.

Shortly hereafter, the participant would receive a notification of whether the responder (Mary) had accepted the division (see Figure 5.3b). Participants collected money throughout the trials, and were paid according to their final score to create a monetary incentive to lie.

Data

An entry was defined as the window of time between when the proposal screen would appear (see Figure 5.3a), and until the participant hit *OK*. The resulting data set comprised 41 participants and 410 entries.

Participants lied about the available funds on average 35.1% of time; this was most prominent when the allocated funds were high. Seventeen participants never understated the available funds (59% lied at least once). Three participants understated at every entry. Participants discounted the actual endowment by 17.4% on average. The crowdsourced participants appear more loyal than laboratory participants (Besancenot et al. found that on average 88.5% of the proposers discount the actual endowment by 20.5% [18]); in this study we observe that 41% of the participants never lied at all, consistent with some feedback we received, such as:

> *It seemed fair to me to split the money evenly. I don't believe in dishonesty so I did not want to lie*
>
> – Crowdworker

Classification

We built a binary truth/lie classifier based on the data obtained. We defined a lie as an entry where the declared funds were lower than the allocated.

Choice of Classifier

We tried a range of classification algorithms, including ensemble methods. An SVM with a radial basis function kernel provided the most promising classification accuracy. Hyper parameters were selected using grid search. The classifier was developed in Python using the ML library Scikit-learn.

Feature Generation

Features were chosen based on previous work in classification of human factors using mobile devices (e.g., [151, 173]), such as speed, precision, rotation, and acceleration (sampled at 50 Hz). We also included features from empirical observations of deception (e.g., [228]), such as immediacy and response length.

Feature Selection

We clustered our features in related groups (see Table 5.2), and handpicked the effective predictors for truth classification. The feature groups *acceleration* and *num-pad* presented the most viable features for classifying truths and lies, and were thus shown in our final classifier (i.e., manual feature selection).

Performance

We measure how well our predictor works, by reporting the average binary F_1-score obtained over a randomized 5-fold cross validation. The F_1-score can be interpreted as a weighted average of the precision and recall, where an F_1-score reaches its best value at 1 and worst score at 0, and is defined as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Feature Group	*Features*	*Description*
Timing	immediacy response	t before first event entry duration
Finger size	touch area	finger contact size
Num-pad	key dynamics hold-time tap precision	see [59] button hold-down time distance to target center
Button clicks	hold-time click area backspaces	button hold-down time quadrant activated number of deletions
Done-button	taps precision hold-time pressure click area	number of times distance to target center button hold-down time screen pressure quadrant activated
Acceleration	x-, y-, and z	a for all axes
Rotation	α-, β-, and γ	ω around all axes
Signal Magnitude	$\sqrt{x^2 + y^2 + z^2}$	for both a and ω

Table 5.2: Feature groups and specific features for each group.

where precision and recall relate to true positives (TP), false positives (FP), and false negatives (FN) as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad , \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Using a randomized 5-fold cross validation we obtain precisions of 81% and 66% for truths and lies respectively. The rates for recall are 88% and 52% for truths and lies respectively. This yields an average F_1-score of .77; .81 for truths and .66 for lies. These performances are well over both chance level (.50), the baseline (.65), and human performance [23].

How Lies and Truths Differ

Next we report on how interaction with the mobile UI differed between honest and dishonest entries, in particular features that varied with the honesty of the interaction. We inspect the distribution of features using density plots: blue areas represent the prevalence of honest entries; red areas represent dishonest entries (those with deflated declared funds). Note that a single feature seldom alone is enough to support classification. Instead combinations of features make up the decision, which is not clear from a single feature's distribution.

Acceleration

We observed that a low mean acceleration was most frequent among honest entries. This suggests that honest entries resulted in less hand movement by the device-holding hand (their non-dominant hand). This was true both on the x-axis, and the z-axis (see Figure 5.4). This follows findings from an existing study of non-phone deceit [228], which showed that dishonesty causes increased hand/finger activity.

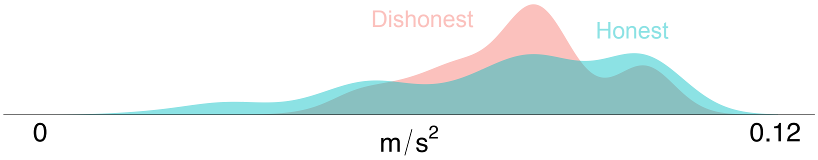


Figure 5.4: Mean acceleration on the z-axis during an entry. Entries towards the low spectrum are predominately honest.

Num-pad

For each entry an amount of cents between 25 and 99 was allocated, requiring participants to input a two-digit number in the declared input field using a num-pad. We observe that the duration between the first key event and the second key event is higher for dishonest entries (see Figure 5.5). This suggests that participants decide whether to lie, and by how much, per individual digit, rather than per input.

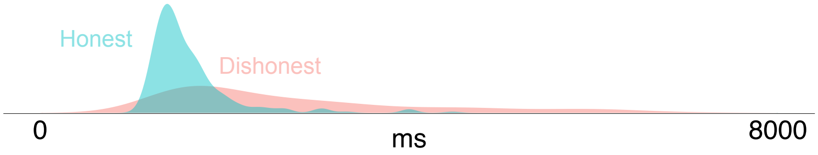


Figure 5.5: Duration between first and second num-pad key event. Truthful entries show shorter durations between the first two num-key presses.

Our num-pad dialog implementation could be dismissed by tapping outside of the num-pad area (instead of clicking 'OK'). Additionally, if, after having entered a number, the participants decided to correct their entry, additional 'OK' taps could be performed. The more taps on the 'OK' button in the num-pad, the more likely an entry was to be honest (see Figure 5.6); we almost exclusively observe dialog dismissal amongst dishonest entries, and we almost only find honest entries for high number of taps on 'OK'[1].

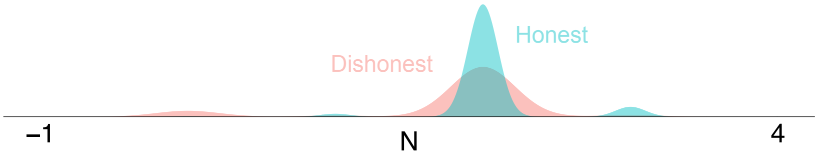


Figure 5.6: Total number of taps on the 'OK' button in the num-pad. Honest entries tend to contain more taps on 'OK'; almost only dishonest entries closed the dialog without confirming 'OK'; almost only honest entries reopened the dialogue and pressed 'OK' again.

[1] This may also suggest that honest users lied initially, before correcting their input to the truth. Dishonest users may show reluctance to 'confirm' their lie, and thus avoid pressing 'OK'. Further research is needed to verify this behaviour.

Summary

Study II shows that the way people interact with their mobile UI can change with the level of honesty of the action. Specifically, movement of the phone (acceleration) and num-pad interactions varied. This increases our confidence in the feasibility of using sensor data to estimate the veracity of input. We built a classifier based on smartphone sensor data and achieved an average F_1-score of .77. This classification accuracy shows that mobile sensor data can be a promising path towards lie detection. To validate these results, and to assess whether the results generalize to other settings, we ran a third study.

5.6 Study III: Yatzy Game

Both Study I and Study II showed that we can observe differences in interaction data between lies and truths using mobile UIs. In Study I, participants were instructed to lie and response time was the only distinguishing feature. In Study II, participants were made aware that they could lie without punishment, resulting in a higher proportion of lies than expected in everyday interaction [225]. From this study, a wider spectrum of mobile input became valuable features for classification.

In order to validate the classification results from study II, we ran a third study. This study still facilitated spontaneous lying, but made no reference to dishonesty in its description. The study required participants to play a dice-based game on a mobile device, inspired by a widely used experimental task in dishonesty research. The task supported spontaneous lying, and allowed for automatic labeling of discrete trials as either honest or dishonest. The participants were rewarded based on their reported score, thereby making lying profitable. We did not encourage participants to lie, and given that all participants passed an initial qualification test about the rules, we can assume that participants were aware of their wrongdoings. Overstating scores could provoke both moral dissonance and fear of not having the crowdwork approved (and thus not getting paid); we hypothesize that this manifests itself in the participant's mobile interaction.

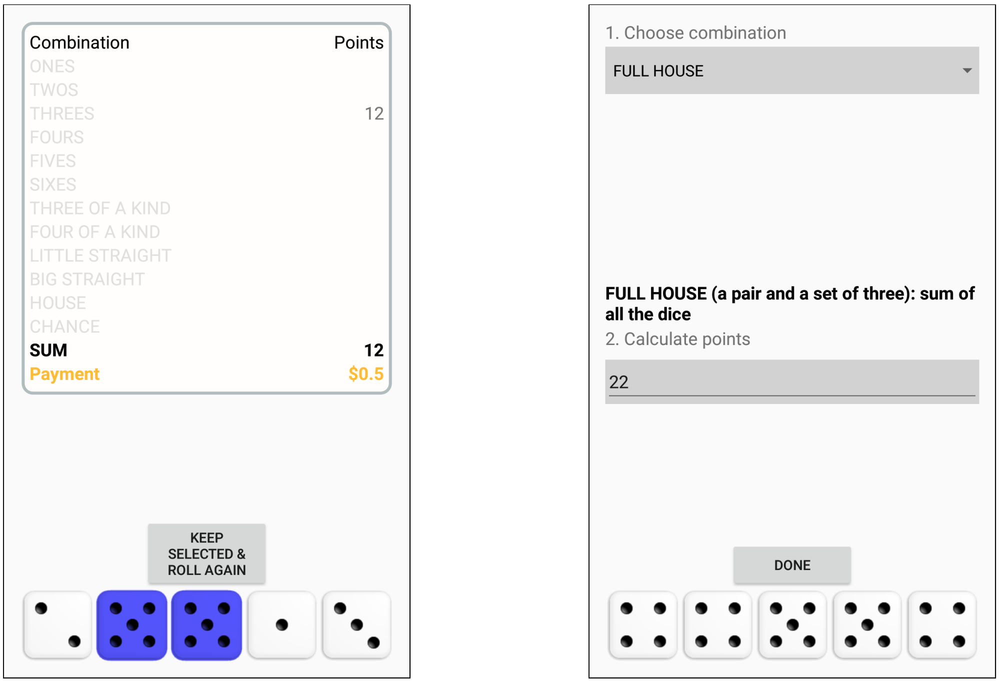
Task

A commonly used task in studying deceit and dishonest behavior requires participants to report on the outcome of randomized events such as rolling a die, or tossing a coin (see [92] for an overview). To encourage lying, participants are rewarded relative to the reported outcomes. The actual outcomes of the events are only known to the participants. This paradigm supports inferences about deceit across all reports (based on deviation from the expected mean) but the individual reports cannot be labeled as honest or dishonest. To support the training of a classifier, we used a dice rolling paradigm, but made changes to allow for labeling of discrete events. Additionally, we wished to collect data across a range of taps and swipes, so as to cover a wider spectrum of typical mobile input. The application required participants to *swipe* through lists, *tap* desired selections, and *tap* numbers on a num-pad.

We developed a mobile dice game, similar to the popular game Yatzy. The game consisted of 12 rounds of rolls with five dice. Each round required an initial roll, and two potential re-rolls of selected dice (see Figure 5.7a). Participants then chose from a list of possible combinations (such as sixes, or three-of-a-kind) and entered the score that a certain combination would yield (see Figure 5.7b). This was typically the sum of the dice. There was a total of 12 combinations; one for each round. Each combination could only be selected once. If the final dice of a round did not equate to a combination, then any combination could be selected and a score of 0 should be entered. The game recorded both the participants' actual score and their reported score. The participants were rewarded based on the sum of their reported scores, providing an incentive to lie:

$0.50 : below 150 points
$1.00 : between 150 and 200 points
$2.00 : more than 200 points

Participants were briefed about the rules and scoring system of the game. Prior to taking part, participants did a qualification test, to ensure that they understood the rules. A help text was available throughout the game for assistance. After completing the experiment, a debriefing screen explained the actual research agenda.



Combination	Points
ONES	
TWOS	
THREES	12
FOURS	
FIVES	
SIXES	
THREE OF A KIND	
FOUR OF A KIND	
LITTLE STRAIGHT	
BIG STRAIGHT	
HOUSE	
CHANCE	
SUM	**12**
Payment	$0.5

1. Choose combination

FULL HOUSE

FULL HOUSE (a pair and a set of three): sum of all the dice

2. Calculate points

22

(a)

(b)

Figure 5.7: The experimental application used in Study III. Figures show (a) the home screen where participants roll and select dice (selected dice are blue), and see the score board, and (b) the entry screen where combinations and amount of points are entered. The entry screen appears after finishing three rolls and pressing 'select combination'.

What Constitutes an Entry

In order to train our classifier, we labeled each entry as either a lie or a truth. When beginning a round, the participants were presented with the home screen (see Figure 5.7a). After rolling the dice the third time, and pressing *Select Combination*, they were presented with the entry screen (see Figure 5.7b). We define an entry as the time frame from when participants were presented with the entry screen, until and including they hit *Done*. During an entry, the user had to pick a dice combination from a list, and enter the amount of points that the combination and the dice roll amounted to. Swipes were recorded when scrolling the list of combinations; taps were recorded when entering the amount of points on a num-pad. IMU sensors recorded motion data throughout the entry.

We expected three possible outcomes of an entry in the game:

1. The participant reports their score accurately (Truth).
2. The participant purposefully inflates their score (Lie)
3. The participant unintentionally inflates the score (Truth - the participant does not intend to deceive)

In an attempt to differentiate (2) and (3), lies were defined as $score_{enter} - score_{real} > 4$. This was informed by the mean negative deviation from the real score (i.e., when participants under-reported their score, M = -3.4).

Participants

We recruited 51 participants from Mechanical Turk, aged 22-57 ($M = 31.5$), 20 females. Participants were told that they were reviewing a mobile game before its launch. Participants were paid according to their score, ranging from $0.50 US to $2.00 US, to incentivize lying. We employed the same qualification standards as for Study I.

Apparatus

We developed the application for Android version 6.0 and higher. To obtain comparable data between participants, we excluded tablets and other large-screen devices. A pilot study identified touch pressure level as a good predictor of truthful input, so for the final study we invited only participants who had phones with pressure sensors. This limited the phones to specific models from Google, LG, Motorola, HTC, and OnePlus. We also excluded mobile devices that could not report rotation or acceleration data.

Results

Fifty-one participants took part in the study, completing 561 unique entries, with 44 labeled as lies (8%); 31% of the participants lied at least once. The average lie provided the participant with 15.6 surplus points. Conversely, nine entries reported scores below the actual score, with an average shortfall of 3.4 points.

Our classification results show 98% precision, and 97% recall for truths (F_1 = .98), and 65% precision and 59% recall for lies (F_1 = .57).

Classification

The classifier was built using the same approach as Study II.

Data Cleaning

We removed participants whose entries indicated that they did not understand the rules, or deliberately rushed the game to optimize payment (amounting to four participants). No participant lied on every single entry.

We removed entries with entered points lower than the actual score (amounted to nine entries, mean shortfall -3.4 points). While they lack an intention to deceive, they could represent either miscalculations or lack of attention with the task. We remove them because correct classification is impossible. We also removed all first entries to account for participants learning to use the interface.

The final data set comprised 51 participants, and 561 entries. The lies covered 44 entries, amounting to 8%. We normalized features per participant (using L_2), and standardized the data set along all axes.

Feature Generation

We generated the same features as in Study II (see Table 5.2). Additionally, we computed features originating from interactions with the list of dice combinations as well as pressure data (see Table 5.3).

Feature Group	*Features*	*Description*
List	clicks on list	n clicks
Swipe	distance	$d(p_0, p_n)$
	duration	$t_n - t_0$
	length	$\sum d(p_i, p_{i+1})$
	linearity	r^2, linear regression
	slope	linear regression
	speed	length / duration
	number	n swipes
Pressure	swipe pressure	screen pressure
	button pressure	screen pressure

Table 5.3: Additional features used for the classifier in Study III.

Feature Selection

Again we handpicked feature groups; *acceleration*, *pressure*, and *num-pad* presented the most viable feature groups for the classification task.

We used recursive feature reduction to eliminate specific bland features within each feature group. From the initial set of features, 11 remained:

- *Num-pad*: button precision (mean, min)
- *Pressure*: button pressure (mean, max, *SD*, pressure)
- *Pressure*: swipe pressure (mean)
- *Acceleration*: x-acceleration (mean, max, *SD*)
- *Acceleration*: z-acceleration (*SD*)

Both Study I and previous work explicitly consider timing as a key predictor of lying [207]. While promising in an administered setting, timing is not robust to the practicalities of day-to-day mobile device usage, where distractions can easily occur mid-input. For this reason, we did not use timing, or response length, as features in our classifier. Additionally, our focus for this study is on lie classification through physiological factors present in sensor data.

There is, however, a temporal dimension within the accelerometer data. Lies took, on average, longer to enter than truths, resulting in more accumulated acceleration data for dishonest input. The acceleration statistics that we computed go some way towards normalizing the effect of this increase in data. To reduce the effect further, we checked for entries longer than three standard deviations of the mean (there were none).

Performance

As Table 5.4 shows, we achieve high performance in classifying truths ($F_1 = .98$) and above-chance accuracy for lies ($F_1 = .57$). To clarify our results, we provide classification metrics for two other "classifiers". *Coin-toss* demonstrates classification at random (i.e., tossing a coin), and *Naïve* reports truth for every input (i.e., the most common observation; ZeroR). We observe that our classifier performs well above the random and the naïve approach.

Classifier		*Precision*	*Recall*	*F_1-score*
Veritaps	Truth	.98	.97	.98
	Lie	.65	.59	.57
	Avg	.96	.95	.95
Coin-toss	Truth	.92	.50	.65
	Lie	.08	.50	.14
	Avg	.85	.50	.61
Naïve	Truth	1.0	.50	.67
	Lie	0.0	0.0	0.0
	Avg	.92	.46	.62

Table 5.4: Performance metrics of classification results. To compare, we report the theoretical scores from a randomized/Coin-toss and a naïve/ZeroR classifier. The scores show the mean score from a 5-fold cross validation. Average is computed with respect to the skewed distribution of truths and lies.

How Lies and Truths Differ

A truth took on average $13.0s$ (*SD* = 12.0) to complete. A lie took on average $20.8s$ (*SD* = 30.6) to complete. Lies were most prominent in the beginning of the experiment; two thirds of all lies were made in the first half of the experiment.

To understand the fundamental differences between an average lie and an average truth, we pick a representative entry from both groups. The entries chosen are the two observations closest to the centroids of two k-means clusters. Here, we explain how the most influential features varied.

Num-pad

We analyzed a range of num-pad entry features, including key dynamics, precision, and hold-down time. Only precision proved to be an effective predictor - the truthful entry records taps with closer proximity to the button's center.

Pressure

Most entries comprise two num-pad taps, excluding an additional tap on the done button. The truthful entry showed a higher average pressure, and also an increase in pressure between the taps. The lie showed a lower average pressure, and a decrease in pressure between taps.

Acceleration

Acceleration varies between lies and truthful entries, mainly on the x-axis. Specifically, the mean, max, and *SD* of x-axis acceleration contribute effective indication of truth in input. For these examples of entries, the mean x-acceleration is higher for the lie, which hints that the honest entry enforced a more steady hand during interaction, as in Study II.

Summary

Our results demonstrate an F_1-score of .98 in classifying truths. We also achieve an F_1-score of .57 in identifying lies. While promising, the recall rate of lies (59%) renders the technique impractical for binary lie-detection. This is in-line with other so-called lie detectors, such as the polygraph, that report indicators associated with lying for interpretation by a practitioner, rather than a binary classification [186].

Research has shown that we are only slightly better than chance at identifying lies when judging statements with an evenly distributed truth-value [23, 228]. Instead of acting as a standalone lie detector then, our results can provide cues to assist us in improving our lie detection accuracy (as in polygraph tests). Whereas, typically, we would need to assess the truth of any input information, the sensor data associated with mobile input allows us to identify only the subset of information that needs further consideration. We can pre-identify a large part of input as true. This can reduce the space of statements that require approximately chance-level lie analysis and make us significantly more accurate at lie-detection overall.

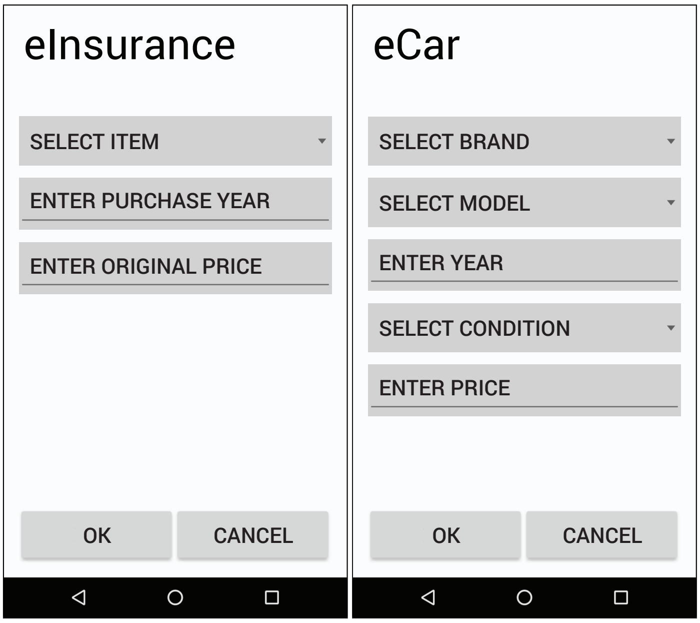
5.7 Veritaps for Mobile Input

We playfully propose to use the results as an additional layer of communication, *Veritaps*, helping a recipient determine the veracity of information from a sender. Veritaps marks both truthful input ✓, and questionable input ?. If the input is questionable, then the recipient can choose whether to request additional information from the sender. In this way, Veritaps can limit the space of interaction that requires further consideration and reduce veracity uncertainty in communication.

We sketch the use of Veritaps across three different styles of mobile interaction: data entry, interpersonal interaction, and personal reflection. Across these three domains, we provide concept use cases, based on the styles of interaction we explored in our studies, and highlight the potential benefits of increased accuracy in veracity judgment.

Mobile Data Entry

We perform many tasks on our mobile devices for which security is paramount, such as online banking. These tasks involve interactions that do not rely on prose, but instead focus on taps and swipes for navigating menu items and entering codes. This renders natural language processing techniques for lie detection inapplicable. Using Veritaps, however, can provide additional layers of security.



(a) (b)

Figure 5.8: Example Veritaps applications: (a) Veritaps can be used to verify the veracity of an insurance claim, (b) Veritaps may verify the declared condition of a vehicle upon creating an online listing.

Veritaps could provide services with an additional layer of scrutiny for online forms. For example, by adding a mobile entry step to the submission process insurance companies could use Veritaps to flag submissions that need further attention or further supporting documents (see Figure 5.8a). Online marketplaces could in a similar way use Veritaps to flag suspicious classified advertisements (see Figure 5.8b).

Interpersonal Communication

We envision that Veritaps could also afford a layer of inter-personal communication, such as increasing confidence in conversations with strangers, or as a playful dimension between friends. For example, after engaging the seller of a car, Veritaps could assess the veracity of the chat messages, ensuring that information presented privately beyond the initial listing is also verified.

Personal Reflection

Self-deception is common and natural, and believed to relate closely to *ethical fading*; the decisions we make, justified by self-deception, that are ethically questionable [219]. Veritaps can provide prompts against self-deception. For example, you could install a Veritaps browser plugin that prompts you every time you exhibit deceptive behavior. The plugin could help make you aware of how often you are finding excuses for canceling on your trainer, or neglect your diet.

5.8 Discussion

The results of the empirical studies, as well as the Veritaps concept, raise several discussion points. They concern the studies, the concept, and the ethical concerns of lie detection.

Our classification accuracies across a broader range of mobile input are relevant only for spontaneous lying. In our directed lying study (Study I), only response time provided a distinguishable feature. For this reason, we cannot speculate about Veritaps' accuracy for habituated lying. Classification accuracy would likely be low here, however, as we believe the spontaneity and guilt of lying creates the physiological features that empower our classification. We also assume that participants made their decision to lie during the entry step of the task and that we, therefore, capture this moment. Currently, we cannot separate the effects of this decision moment from the entry itself, and thus cannot be certain of the efficacy of one without the other. This needs further exploration.

Because pressure data contributed an important feature in our classification, we required participants to have phones with pressure sensors. This limited the applicable participants, as only recent Android phones have pressure sensors. As a result, we found little variation in the phones used in the study. This assisted our classification accuracy, as it reduced requirements for preprocessing of data. As smartphone chipsets are not standardized, a production setup of our proposed technique would optimally require a per-phone-model training process.

The lying we are able to classify has a number of characteristics that limit the generalizability of the findings. Three types of lying may be differentiated [228]: outright lies, exaggeration, and subtle lies. The experimental task in Study I and II examined outright lies, while Study III considered exaggeration. We do not know how our findings generalize to subtle lies. Also, interpersonal lies as present in Study II, might have caused participants to react quite differently compared to interaction without another human being. This could be an explanation for the difference in classification accuracies for Study II-III. Verifying this remains an avenue for future work.

As for the *Veritaps concept*, we have sketched simple example mock-up scenarios. There are practical challenges that arise with implementing Veritaps on smartphones, however.

Ideally, the smartphone would receive a steady stream of labelled ground truth input, to help train the classifier. In practice, however, this would lead to repeated training interruptions on the device and likely prevent adoption.The alternative would be for the phone to come pre-packaged with a trained model, however, without per-user training this could under-perform.

Machine learning based estimators require considerable labelled data to perform well. We evaluated to what extent our proposed method would work without per-user calibration. To do this, we ran Leave-One-Subject-Out (LOSO) cross validation [85] using the classifiers from Study II-III. This caused an increase in performance for Study II, but a decrease for Study III. This shows us the more open-ended nature of the task in Study III works poor without per-user training, while other tasks are more suitable to use with pre-trained models.

Within our presented Veritaps scenarios, there remains an opportunity to learn the features that suggest truthful input and, therefore, trick the system. To reduce this risk, we propose that the user should not be shown their own Veritaps assessments, rather they should only be made available to the receiver of the information. In this way, it is important that both parties consent to engaging in a Truth-Verified interaction. Future work should attempt to implement these to study the users' reactions and adaptations based on feedback from our algorithm.

Further, the polygraph has been banned from use in courts in most justice systems because of negligible reliability. In the same way, we do not propose the use of Veritaps as any means of assessment of objective truth. The predictive insights provided by Veritaps should be used with caution, and we cannot recommend critical reliance on Veritaps in any system.

The Veritaps concept raises a number of *ethical questions*. First, lying is an important social lubricant. For instance, small lies play an important role in computer-mediated conversations (e.g., [86]). Also, many lies are simply ignored (the so-called ostrich effect [228]). Therefore, making lies explicit, as in some of the design concepts we discussed, threaten to undermine those functions and introduce mistrust into computer-mediated communication. We call for empirical studies of the Veritaps concept to understand how the availability of truth verification might impact the experience and outcomes of digital communication. Second, our algorithm, and even improved algorithms, are likely to misclassify. This might challenge the basis of human conversation [78]; that the information we communicate is accurate and truthful. One reaction to this is to have both parties opt-in to having that basis challenged; this would work for several of the concepts we discussed.

Based on the feature analyses, we believe that user interfaces made up of standard UI elements as input fields and buttons are likely to perform best. Across Study I-III we found the details of simple user actions such as taps to carry more reliable information of the veracity of an action, than for instance sliding and scrolling gestures. Additionally, if the methods described in this paper were combined with content-based features, it could likely outperform the performances presented.

Our data do not suggest that smaller lies are harder to detect that bigger lies. For Study II, a binary distinction compared to a scale of deflated scores yielded the clearest division. In other words; an entry with a deflation of one cent, on average held more similar interactional properties with dishonest entries than honest entries.

5.9 Conclusion

We are frequently subject to lying, and to date lack means of classifying lies on mobile devices beyond written text and speech. This leaves a large space of interaction open to deception. We explore the feasibility of a content-agnostic, sensor-led approach to lie detection on smartphones that considers only taps and swipes. Through three studies we presented empirical evidence for the feasibility of commodity lie detection using mobile interaction.

First, we found significant differences in response times between lies and truths for simple mobile interactions.

Next, we reported on the individual interaction differences observed between lying and truth telling in a mobile version of the Ultimatum game that encouraged lying. The study showed that some features of mobile interaction varies with the honesty of an action. Specifically, properties of number entry were good indicators of deceit.

Last, we reported on a study where participants took part in a mobile dice game that incentivized lying. We trained a classifier on mobile sensor data that ignores the input data itself. We achieved 96% precision and 95% recall in truth detection, and 65% precision and 59% recall for lie detection. While promising, these results do not support reliable binary lie classification. Instead, we suggest their use a means of improving peoples' own near-chance level lie classification.

Based on the findings, we introduced Veritaps: an optional layer in mobile interaction, allowing users to share truth assessments of their input. We presented three potential use cases of Veritaps, across online form-filling, inter-personal communication, and personal reflection.

6 Crowdsourcing Virtual Reality Studies

Virtual Reality Studies Outside the Laboratory

Aske Mottelson
Department of Computer Science
University of Copenhagen
amot@di.ku.dk

Kasper Hornbæk
Department of Computer Science
University of Copenhagen
kash@di.ku.dk

ABSTRACT

Many user studies are now conducted outside laboratories to increase the number and heterogeneity of participants. These studies are conducted in diverse settings, with the potential to give research greater external validity and statistical power at a lower cost. The feasibility of conducting virtual reality (VR) studies outside laboratories remains unclear because these studies often use expensive equipment, depend critically on the physical context, and sometimes study delicate phenomena concerning body awareness and immersion. To investigate, we explore pointing, 3D tracing, and body-illusions both *in-lab* and *out-of-lab*. The in-lab study was carried out as a traditional experiment with state-of-the-art VR equipment; 31 completed the study in our laboratory. The out-of-lab study was conducted by distributing commodity cardboard VR glasses to participants; 57 completed the study anywhere they saw fit. The effects found in-lab were comparable to those found out-of-lab, with much larger variations in the settings in the out-of-lab condition. A follow-up study showed that performance metrics are mostly governed by the technology used, where more complex VR phenomena depend more critically on the internal control of the study. We argue that conducting VR studies outside the laboratory is feasible, and that certain types of VR studies may advantageously be run this way. From the results, we discuss the implications and limitations of running VR studies outside the laboratory.

CCS CONCEPTS

•**Human-centered computing →Virtual reality;** *User studies;* Empirical studies in HCI;

KEYWORDS

Consumer VR; Google Cardboard; User Studies; Crowdsourcing

ACM Reference format:
Aske Mottelson and Kasper Hornbæk. 2017. Virtual Reality Studies Outside the Laboratory. In *Proceedings of VRST '17, Gothenburg, Sweden, November 8–10, 2017,* 10 pages.
DOI: 10.1145/3139131.3139141

1 INTRODUCTION

The recent advance in consumer technology has accelerated research in virtual reality (VR). In particular, a host of VR user studies are being conducted. They include both evaluations of the usability and user experience of particular VR applications, as well as behavioral research using VR. The former includes evaluating games and educational applications (e.g., [Bolton et al. 2014; von Zadow et al. 2013]). The latter includes simulating environments to conduct experiments that would otherwise be difficult (e.g., [Pan et al. 2016; Slater et al. 2013]), impossible (e.g., [Banakou et al. 2013; Kilteni et al. 2012; Slater et al. 2010]) or even unethical (e.g., [Slater et al. 2006]) to carry out using classical experimental paradigms.

Conducting VR studies, however, faces similar decisions about practical matters and research validity as running studies using non-VR technology; those decisions and their associated trade-offs are well described (e.g., [Hornbæk 2013; McGrath 1995]). For instance, much planning goes into recruiting people, managing schedules, selecting environments in which to conduct studies, and running the actual studies. Nevertheless, VR studies are almost exclusively done in laboratories using specialized equipment (e.g., for tracking) and few and homogeneous participants (e.g., typically fewer than 25 participants recruited through university mailing lists). In that respect, VR studies are similar to studies from other parts of HCI [Caine 2016; Hornbæk et al. 2014].

For non-VR technologies, many of these studies are now done outside the laboratory, for instance with crowdsourcing or as in-the-wild studies. In crowdsourcing, user studies are conducted as micro tasks giving small amounts of payment on crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower [Kittur et al. 2008]. Research shows that crowdsourcing often give a higher diversity of participants [Mason and Suri 2012; Paolacci and Chandler 2014; Ross et al. 2010] and that they can be done at a low cost [Buhrmester et al. 2011; Kittur et al. 2008; Mason and Suri 2012], reliably [Buhrmester et al. 2011; Crump et al. 2013; Rouse 2015], and quickly [Kittur et al. 2008].

Although out-of-lab experimentation has been applied in many computing areas (e.g., [Carter et al. 2007; Heer and Bostock 2010; Kittur et al. 2008; Mottelson and Hornbæk 2016; Reinecke and Gajos 2015]), it is not clear if that is feasible or valid for VR. Earlier work has suggested that this type of experimental practice is ill suited for tasks that depend on the physical environment [Heer and Bostock 2010]. Also, many VR studies depend on headsets still not in common use and, sometimes, other equipment that is not widely available (e.g., for tracking or physical stimulation). Finally, participants in unsupervised experiments are not always paying attention, switch tasks frequently [Gould et al. 2016], and may decide to pause an experiment; all of these behaviors could interfere with goals of VR studies, such as generating perception of presence. Goodman et al. [Goodman et al. 2013] stressed that *"MTurk participants are less likely to pay attention to experimental materials"*, which could reduce the effects of experimental manipulations.

VRST '17, Gothenburg, Sweden
© 2017 ACM. 978-1-4503-5548-3/17/11…$15.00
DOI: 10.1145/3139131.3139141

Virtual reality (VR) allows researchers to investigate subjects in life threatening situations, very specific or uncommon events, or with modifications to worlds or bodies without obeying the laws of physics or human morphology. This makes VR, as a technology, an especially interesting platform for HCI research, and for investigating psychological phenomena. Researchers have, in particular within recent years, adopted this technology to investigate aspects of how our bodies (or other bodies) influence psychological constructs, such as racism, attraction, empathy, etc.

In the two previous chapters I have shown how to overcome the issue of ensuring reliable cognitive labelled data from many participants by crowdsourcing user studies: I have recruited participants online who then install experimental research applications onto their own devices, that then report activity data back alongside user or condition based labels. In this chapter I show how to apply the methodology for VR studies.

This chapter contains the contents of the paper shown to the left [153], published at VRST '17. I apply the same methodological approach of conducting unsupervised experiments online to virtual reality user studies: we handed out some 100 Google cardboard commodity VR glasses together with the link to a mobile-based VR application. This application featured three seminal VR environments, each depicting three levels of VR complexity (Fitts's Law, 3D tracing, and a body ownership illusion). Additionally we conducted a laboratory-based VR user study, with the same task, to identify which VR paradigms can be conducted without supervision.

While unsupervised VR experiments are still cumbersome to conduct due to the modest adoption of VR enabled devices, the paper shows that it can be done, and that reliable quantitative metrics can be acquired. This is especially true for the more performance related data, rather than the experiential.

This chapter is based on a collaborative effort as described below.

Title
Virtual Reality Studies Outside the Laboratory

Authors
Aske Mottelson and Kasper Hornbæk

Journal
Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology

DOI
10.1145/3139131.3139141

What was the role of the PhD student in designing the study?
The PhD student was the first author of the paper,
and responsible for the design of the described studies.

How did the PhD student participate in data collection and/or development of theory?
The PhD student was responsible for study implementation,
execution, data collection, and theory development.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed to all parts of the manuscript.

Did the PhD student read and comment on the final manuscript?
Yes.

6.1 Abstract

Many user studies are now conducted outside laboratories to increase the number and heterogeneity of participants. These studies are conducted in diverse settings, with the potential to give research greater external validity and statistical power at a lower cost. The feasibility of conducting virtual reality (VR) studies outside laboratories remains unclear because these studies often use expensive equipment, depend critically on the physical context, and sometimes study delicate phenomena concerning body awareness and immersion. To investigate, we explore pointing, 3D tracing, and body-illusions both *in-lab* and *out-of-lab*. The in-lab study was carried out as a traditional experiment with state-of-the-art VR equipment; 31 completed the study in our laboratory. The out-of-lab study was conducted by distributing commodity cardboard VR glasses to participants; 57 completed the study anywhere they saw fit. The effects found in-lab were comparable to those found out-of-lab, with much larger variations in the settings in the out-of-lab condition. A follow-up study showed that performance metrics are mostly governed by the technology used, where more complex VR phenomena depend more critically on the internal control of the study. We argue that conducting VR studies outside the laboratory is feasible, and that certain types of VR studies may advantageously be run this way. From the results, we discuss the implications and limitations of running VR studies outside the laboratory.

6.2 Introduction

The recent advance in consumer technology has accelerated research in virtual reality (VR). In particular, a host of VR user studies are being conducted. They include both evaluations of the usability and user experience of particular VR applications, as well as behavioral research using VR. The former includes evaluating games and educational applications (e.g., [22, 249]). The latter includes simulating environments to conduct experiments that would otherwise be difficult (e.g., [165, 198]), impossible (e.g., [12, 111, 199]) or even unethical (e.g., [195]) to carry out using classical experimental paradigms.

Conducting VR studies, however, faces similar decisions about practical matters and research validity as running studies using non-VR technology; those decisions and their associated trade-offs are well described (e.g., [95, 142]). For instance, much planning goes into recruiting people, managing schedules, selecting environments in which to conduct studies, and running the actual studies. Nevertheless, VR studies are almost exclusively done in laboratories using specialized equipment (e.g., for tracking) and few and homogeneous participants (e.g., typically fewer than 25 participants recruited through university mailing lists). In that respect, VR studies are similar to studies from other parts of HCI [29, 97].

For non-VR technologies, many of these studies are now done outside the laboratory, for instance with crowdsourcing or as in-the-wild studies. In crowdsourcing, user studies are conducted as micro tasks giving small amounts of payment on crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower [112]. Research shows that crowdsourcing often give a higher diversity of participants [140, 166, 181] and that they can be done at a low cost [28, 112, 140], reliably [28, 50, 182], and quickly [112].

Although out-of-lab experimentation has been applied in many computing areas (e.g., [34, 90, 112, 151, 179]), it is not clear if that is feasible or valid for VR. Earlier work has suggested that this type of experimental practice is ill suited for tasks that depend on the physical environment [90]. Also, many VR studies depend on headsets still not in common use and, sometimes, other equipment that is not widely available (e.g., for tracking or physical stimulation). Finally, participants in unsupervised experiments are not always paying attention, switch tasks frequently [76], and may decide to pause an experiment; all of these behaviors could interfere with goals of VR studies, such as generating perception of presence. Goodman et al. [75] stressed that *"MTurk participants are less likely to pay attention to experimental materials"*, which could reduce the effects of experimental manipulations.

We explore the possibility of conducting out-of-lab VR studies, and compare experiments in uncontrolled settings using commodity VR technology to doing them in the laboratory. We distributed VR cardboard glasses to 57 participants for use with participants' own smartphones in exchange for their participation in the study involving three canonical experimental VR tasks. The results show that it is a feasible way to conduct affordable, ecologically valid, and large-scale VR studies outside the laboratory. Additionally we discuss potential directions for out-of-lab VR experimentation, and how crowdsourcing is an interesting platform for future VR studies.

6.3 Related Work

VR studies have been organized in a variety of ways, including analytic evaluation techniques (e.g., [10, 216]) as well as empirical ones (e.g., [12, 111, 199]). The literature also contains evaluations of usability issues and user experience with VR head sets [141], and Marsh [139] discussed some issues in evaluating the usability of VR. Here we focus on empirical user studies using VR and first discuss those briefly. Then we review types of user studies conducted outside of laboratories, and outline the potential of that methodology for VR.

VR Studies

Virtual reality research has for long been engaged in user studies, and with the availability of consumer oriented VR technology a host of VR studies are being conducted. Those studies include evaluations of the usability and user experience of particular VR applications, such as games (e.g., [22]) and educational applications (e.g., [249]). Another line of behavioral research uses VR to study phenomena relating to body perception and body schema. These often employ *body ownership illusions*, which are studies where participants perceive non-bodily objects, or alterations of their own body to be parts of their own body [110]. These illusions are usually made feasible by means of synchronous stimulation of the virtual and physical body [196, 197]. With this, perceptions of objects sizes have been shown to be influenced by hand size alterations [132], and racial attitudes are found to be influenced by ownership of an other skin-toned body [137]. Also, ownership of a child body has been shown to cause faster identification of child-like attributes [12].

Researchers in virtual reality studies face many of the same concerns that go into doing any user study, such as recruiting people, managing schedules, selecting environments, and running experiments. Other concerns relate to validity and research methodology; similar to concerns about user studies in other domains (e.g., [95, 142, 192]). While locomotion has historically been a critical topic within VR research [201, 223], many VR studies are conducted stationary (e.g., sitting, standing, lying). Most consumer oriented VR applications are also used stationary: the most popular consumer VR system that supports walking for locomotion (HTC Vive), reports standing as the most common configuration amongst its users [210]. This could indicate that many VR user studies could be straightforward for participants to conduct without the guidance of a human evaluator. If the user studies conducted in VR research bear at least some similarities to user studies conducted in other parts of HCI, might a shift in experimental practice from laboratory to out of laboratory (which is widely successful in other parts of HCI) be beneficial for VR studies?

Out-of-lab Studies in HCI

External validity concerns whether a causal relationship holds over persons, settings, treatments, and outcomes [192]; that is, to which extent findings are generalizable to a broader domain. In an attempt to increase the external validity of a study's results, researchers may conduct their research outside of a laboratory. Unlike observational research such as field studies, some out-of-lab research practices allow researchers to control experimental conditions and manipulate independent variables. These unsupervised experimental practices, such as crowdsourcing and in-the-wild experiments, have been an ongoing endeavor within HCI for a while [26, 34, 113].

Crowdsourcing. In crowdsourcing, user studies are conducted as micro-tasks giving small amounts of payment on crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower [112]. Crowdsourcing has shown to be a valuable experimental practice that allows for fast and low-cost experimentation, with high diversity of participants [28, 50, 112, 140, 166, 181, 182].

In-the-wild. Conducting in-the-wild experiments has long been a research agenda within the ubiquitous computing community, and as such a variety of protocols for conducting unsupervised experimental research has evolved. An alternative to the popular micro-task platforms includes LabintheWild [179], a highly scalable way of conducting studies with widespread, uncompensated, and unsupervised participation. The authors created an online experimental platform that provides participants with information about themselves in exchange for their participation in studies. In-the-wild mobile experiments have also been conducted, Henze et al. [91] for instance did in-the-wild mobile experiments, with mobile app store distributed gaming-based user studies.

Potential of out-of-lab studies for VR

When is high external validity key to VR research? For some VR studies, high external validity is of less concern than others. This is the case when VR is being used to mitigate arachnophobia [72], estimate general practitioners' susceptibility to prescribing antibiotics [165] - and in general studies with homogeneous participants and few experimental settings, especially for within-subjects designs. For studies concerning a heterogeneous population, using subtle differences between conditions, with more experimental settings, the external validity is of much higher concern for the integrity of the research. This is often true when employing a between-subjects design. Could out-of-lab experimentation be a worthwhile methodology to consider for such studies? Unfortunately, it is not clear if it is feasible to use widespread out-of-lab experimental practices to conduct VR studies, or whether these approaches are valid. While crowdsourcing could give larger samples of more varied participants, the widespread adoption of VR consumer devices has yet to happen which makes if difficult to recruit participants for a crowdsourced VR study.

While the potential of conducting VR studies out-of-lab is promising, the associated questions are severe. Although increasing the generalizability of VR research has been an ongoing agenda (e.g., [116]), we are only aware of one study that conducted out-of-lab VR experimentation; Steed et al. [211] ran a mobile app-based experiment, gathering user stories from owners of household VR devices such as Google Cardboard and Samsung Gear to study presence and embodiment in a bar with a singer. The authors did a between-subjects study with eight conditions, exploring among other things the prevalence of a self-avatar on presence, hand-tapping, and eye contact with a virtual person. While this study is a valuable example of using consumer VR technology to conduct studies using mobile app stores, it does not conclude on the validity or feasibility of using that approach. Also, the authors employed an untried procedure, which makes it difficult to separate effects from the experimental procedure and the methodology.

How to study VR outside the laboratory

We surveyed crowdworkers to understand the types of VR equipment at their disposal. We asked 250 people (for 0.05$ USD pay, 92% validated using a verifiable control question) about their ownership of computer equipment using a randomized ordered checklist of household computer technology. We found that at the time of writing, 3% of crowdworkers own one or more devices capable of VR. In particular participants reported ownership of the following: Google Cardboard (2.2%), Samsung Gear (1.3%), HTC Vive (0.9%), and Oculus Rift (0.4%). In comparison, 83.4% of the respondents reported ownership of an Android smart phone. The effective size of the active MTurk population has been estimated to be about 7300 workers [212]. If our sample is representative of the MTurk population, we should expect at most 226 crowdworkers to own a VR device. Thus, the modest share of crowdworkers who own VR equipment at the time of writing make it unrealistic to use popular micro-task markets to crowdsource VR experimentation. Because of the widespread adoption of consumer smart phones combined with Google Cardboard as a cheap alternative to other VR technology, we see a contemporary opportunity for inexpensive large-scale out-of-lab VR experimentation. To provide insights about the feasibility and validity of conducting out-of-lab VR studies, we propose a study protocol where participants are recruited online, pre-screened prior to participation, and provided with commodity cardboard VR glasses to participate in the study.

Crowdsourcing is arguably not the correct term for the approach tried in this work, because we pre-screen participants and require them to visit our premises. Accordingly, the term *in-the-wild* seems inaccurate, since the experimental nature of our setup will enforce an artificial controlled setting, not expected to occur completely in-the-wild (we did not expect any voluntary participation from regular app store users). When HMDs become more prevalent, it will be possible to completely crowdsource VR studies without requiring participants to visit the premises. In the remainder of this work, we employ the term *out-of-lab* to describe the method of equipping pre-screened participants with cardboard VR glasses, and have them conduct experiments in non-controlled settings. This is opposed to when we speak about *in-lab* studies, which covers experimentation in controlled settings, at our research facilities. We conducted experiments using canonical VR paradigms, all previously verified in laboratories. We study the implications of conducting out-of-lab VR experiments by directly comparing the participants' differences across settings and technology.

6.4 Experiment

The purpose of the experiment was to validate the potential of conducting VR studies outside the laboratory, and to do so we compare in-lab to out-of-lab VR experiments over a range of VR phenomena. The studies were set up to be representative of how VR studies are usually done (i.e., laboratory VR studies usually do not use commodity technology).

Participants

We posted the invitation to participate in our experiment on large group for locals on Facebook, in addition to sending invitations using our internal e-mail list. Participants signed up online for either the in-lab or out-of-lab study. In both cases, participants came to our premises; either to pick up a set of cardboard glasses, or to participate in our lab-study.

In-lab. Thirty-one people, aged 20-50 (*SD* = 10.3 years) participated in the laboratory study and were reimbursed with a gift worth the equivalent of 15$ US, our regular minimum rate for lab-study participation. Of these participants were 12 male.

Out-of-lab. 100 participants were given a Google cardboard for participating in the study. Fifty-seven participants, aged 20-40 (*SD* = 7.0 years) completed the study within 20 days; of these were 35 males, with 34 using iOS and 23 using Android.

Apparatus

We developed the VR tasks using Unity 5.3. The VR application was identical for in-lab and out-of-lab, except that the VR equipment held by the avatar was substituted to match the visuals of the actual VR equipment. The applications would send relevant user metrics to a server application written in Python. The application contained on-screen instructions that blended in with the VR environments.

In-lab. The VR studies in-lab used an HTC Vive.

Out-of-lab. We deployed the VR application with the Google Cardboard SDK, distributed at relevant application stores for both Android (version $\geq$5.0) and iOS (version $\geq$7.0).

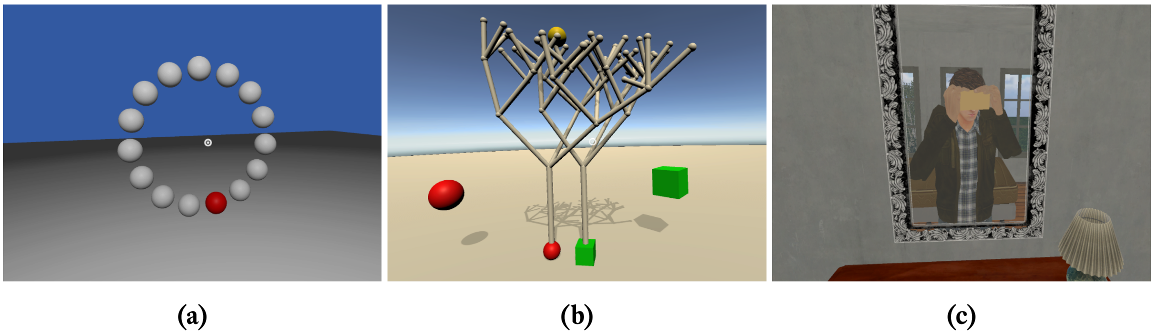


Figure 6.1: The tasks. (a) Pointing: participants targeted the red spheres as fast and accurate as possible by moving their head. (b) 3D tracing: participants selected which tree a yellow leaf belonged to. Participants could either inspect the trees dynamically by moving their head around, or could only see the trees from one angle. (c) Body Ownership Illusion: participants were immersed in a virtual bedroom, with their bodies substituted with sex-matched avatars. In half of the cases avatars mapped the participants' movements real-time. A mirror that reflected the virtual body was present in the bedroom as shown.

Design

Participants conducted three independent tasks; one without any experimental variation, and two with a between-subject two-condition design (see Table 6.1). Participants were randomly assigned to the experimental conditions on a per-task basis. The three tasks were administered in a randomized order. The study procedure was carried out both in-lab and out-of-lab with different participants.

		Task	*Conditions*	*Dependent variables*
In-lab	*Out-of-lab*	Pointing		Movement time, accuracy, throughput
		3D Tracing	Dynamic	Duration, estimations
			Static	
		Body Ownership Illusion	Consistent	Body ownership, presence
			Inconsistent	

Table 6.1: The three tasks employed with their corresponding experimental conditions and dependent variables.

The intention of this design was to combine tasks where absolute performance values could be compared (pointing task) and tasks with expected experimental effects that differ between multiple conditions (3D tracing, body ownership illusion). For the latter we are interested in comparing the outcomes of experimental conditions for in-lab and out-of-lab. To validate the feasibility of conducting VR experimentation outside the laboratory, we hypothesize that out-of-lab studies yield similar effects and effect sizes to those conducted in-lab.

Tasks

The participants were presented with three VR tasks as below. The intention was to get insights of VR experimentation in-lab and out-of-lab in a broad range of tasks; we therefore employed three different complexities of VR experimentation: a pointing task, a 3D tracing task, and a body illusion task. The tasks are based on established experimental protocols about performance aspects of HCI and perception in VR. The tasks do not require the participant to travel across physical space, making them suitable to conduct virtually anywhere and without the need of a human evaluator.

Pointing Task. The goal of this task was to show the feasibility of collecting performance metrics from elementary VR navigation. We studied participants' performance with 2D navigation within a VR environment, using a common Fitts's Law task (see Figure 6.1a). Most aspects of the task adhered to Soukoreff and MacKenzie [206], but to alleviate fatigue we used 15 targets with four IDs (range 2-4), and two repetitions per ID. Thus every participant pointed at 120 targets, excluding a warm-up round. Translational movements were ignored for this task.

3D Tracing Task. This task measured participants' performance in judging depth and navigating in VR. The task by Arthur et al. [9] compares users' performance on distinguishing 3D objects in different viewing conditions. Two trees composed of straight lines were placed next to each other (see Figure 6.1b). Each tree consisted of three levels of branches, resulting in 27 branches for each tree excluding the root. For each trial, one of the trees would contain a yellow leaf (the leaf was placed on the branch with the x coordinate nearest to the center), and the participant then distinguished which of the trees the leaf belonged to. Participants were randomly assigned to either of two conditions: (1) participants could inspect the 3D spatial properties of the trees dynamically by moving their head around, or (2) participants were presented with a static view, requiring them to determine the origin of the leaves having seen the trees from one angle only. Textual feedback (correct/incorrect) was provided after each selection for one second. This task was intended to test if out-of-lab participants could use the 3D spatial capabilities of VR to increase depth judgment accuracy. For each trial of the 40 trials, a leaf was randomly placed on a branch belonging to either the left or right tree.

Q#	*Question*	*Scale*	*Purpose*	*Ref.*
Q1	How much did you feel that the virtual body you saw when you looked down at yourself was your own body?	Not at all .. Very much	Body Ownership	[12]
Q2	How much did you feel that the virtual body you saw when you looked at yourself in the mirror was your own body?	Not at all .. Very much	Body Ownership	[12]
Q3	How much did you feel that your virtual body resembled your own (real) body in terms of shape, skin tone or other visual features?	Not at all .. Very much	Body Ownership	[12]
Q4	How much did you feel as if you had two bodies?	Not at all .. Very much	Body Ownership	[12]
Q5	Did the room that you saw seem bigger, smaller or about the same as what you would expect from your everyday experience?	Smaller .. Larger	Body Ownership	[12]
Q6	Did the virtual body you owned seem bigger, smaller or about the same as what you would expect from your everyday experience?	Smaller .. Larger	Body Ownership	[12]
Q7	While being in the virtual room, did you feel your unseen real body being:	Smaller .. Larger	Body Ownership	[12]
Q8	Did you feel that the virtual {sex} you saw compared with your virtual body was:	Smaller .. Larger	Body Ownership	[12]
Q9	Did you feel that the virtual {sex} you saw compared with your real felt body was:	Smaller .. Larger	Body Ownership	[12]
Q10	How much did you feel that the {sex} you saw was aware of your presence?	Not at all .. Very much	Body Ownership	[12]
Q11	During the experience did it feel as if you moved across the bed room?	Not at all .. Very much	Control	[200]
Q12	Please rate your sense of being in the bed room, where 3 represents the normal experience of being in a place.	Not normal .. Normal	Presence	[200]
Q13	To what extent were there times during the experience when the virtual reality became the ’reality’ for you, and you almost forgot about the ‘real world’ in which the whole experience was really taking place?	Not at all .. All the time	Presence	[200]
Q14	During the time of the experience, which was strongest on the whole, your sense of being in the virtual room, or of being in the real world	Real world .. Virtual room	Presence	[200]

Table 6.2: Body ownership illusion task post-questionnaire measuring among other things participants' body ownership and presence (from [12, 200]).

Body Ownership Illusion Task. Body ownership illusions refer to the class of illusions where participants perceive virtual bodies to be their own [110]. The illusion of ownership of virtual bodies has been shown feasible by means of consistent stimulation of the virtual and physical body (e.g., using a rod) [196, 197]. To study the feasibility of conducting out-of-lab VR tasks involving more complex VR phenomena, we designed a body ownership illusion task inspired by Banakou et al. [12]. The intention was to induce the illusion of body ownership, and the feeling of being present in a virtual room. The participants were asked to looked around and take notice of the room for two minutes. The participants could look down and see a sex-matched virtual body. Additionally, a mirror was present where the participant's avatar could be seen (see Figure 6.1c). The task employed two conditions: consistent and inconsistent visuomotor stimuli. In the consistent condition, participants' movements were mapped real-time to the avatar, both when looking down at the virtual body and when looking in the mirror. With participants hands fixed in a binocular pose, we only mapped the upper torso using a simple inverse kinematics system. In the inconsistent condition, the avatar's body did not reflect participants' movements. We expected the consistent condition to result in higher degrees of body ownership and presence, as previous studies (e.g., [12, 185]). We employed a questionnaire that quantified body ownership and presence on a $[-3, 3]$ Likert scale. We employed a mix of two questionnaire protocols, the first with questions about body ownership by Banakou et al. [12]; in addition to the Slater-Usoh-Steed (SUS) questionnaire [200] about presence (see questionnaire at Table 6.2).

Procedure

All tasks started with a textual description of the task. Participants would have to trigger a begin button to initiate a task. All button selections were done by dwelling two seconds at a target using a cross hair that triggered visual feedback.

In-lab. Participants first signed an informed consent form, and were then placed standing in the middle of a 4×4m room. To minimize effect of disturbing noise from our laboratory, a noise cancellation headset was put on, and together with the HMD placed on the participants' heads by the evaluator. To minimize effects of body posture compared to the out-of-lab study, we asked the participants to keep a posture similar to that when using a Google cardboard (arms and hands in binocular pose) during the entire study. In the same fashion, although VR applications for the HTC Vive are usually controlled using hand-carried controllers, we employed a dwell-based head controlling system to gather comparable data to the out-of-lab cardboard study. An evaluator stayed in the room with the participant during the study and made sure directions were adhered to.

Out-of-lab. Participants were invited to come by our premises to pick up their Google cardboard (if their phone was capable of Android version $\geq$5.0 or iOS version $\geq$7.0). Together with the commodity VR equipment, participants were provided with instructions on how to acquire the experimental application and carry out the experiment, in addition to descriptions of the extent of the data collection and associated privacy concerns.

Ethical Concerns

The immersiveness of VR combined with participants conducting the study protocol on their own give rise to a number of concerns, and out of ethical concerns, some VR experiments should be avoided as out-of-lab studies. Additionally collecting data for publicly open studies requires some consideration. First, opposite to a laboratory-based experiment, an evaluator is not present to help subjects, for instance in case of motion sickness or falling over objects. Second, our experimental application also took two photos we used to analyze participants' surroundings, and to confirm that the phone was in fact correctly placed in the cardboard VR glasses. How to ethically log user data from mobile experiments depends on several factors, and a simple solution to this question does not seem to exist. We however followed directions proposed by Henze et al. [91] and informed users prior to participation about the logging and thus implicitly get consent from the user by continuing use of the application.

Data Validation

As in other studies, attention and compliant participation is key. This is especially difficult to verify in out-of-lab studies because of the absence of a human evaluator during the study. In addition to exclusion criteria based on earlier work [112, 211], we used the front-facing camera of participants' phones to take a photo that was later used to determine whether the phone was accurately placed inside the cardboard equipment (this was mentioned in the experiment invitation). In summary, we used the following criteria to disqualify participants:

- Erroneous placement of phone in VR equipment (front camera)
- Zero variance in questionnaire responses
- High response to control question ($Q5$) (> 2)
- Too slow completion time ($> M + 3\,SD$)
- Too fast response to questionnaire ($< M - 3\,SD$)

Four participants were discarded from the in-lab study, one because of zero variance in the questionnaire answers, three because of the control question ($Q5$); thus 27 participants remained. For the out-of-lab study, we also discarded four participants, one for not placing the phone in the VR glasses, one from taking too long, and two because of the control question; 53 participants remained.

6.5 Results

The overall purpose of the result section is to present differences in dependent variables due to experimental conditions, and to do comparison between the in-lab study and the out-of-lab study. We report results from each of the three tasks, divided by the type of analysis. Where nothing else is indicated, statistical tests were done with a one-way ANOVA.

Pointing Task

Movement time. While both in-lab and out-of-lab participants finished the pointing task without issues, in-lab participants finished significantly faster: in-lab: M = 98s (*SD* = 14s), out-of-lab: M = 149s (*SD* = 62s), $F(1, 78) = 17.5, p < .001$. We plotted the linear fits of participants' movement times as a function of ID (see Figure 6.2), which shows that Fitts's Law (Shannon formulation) is indeed a very accurate model for pointing in VR (both in controlled and uncontrolled settings). An ANCOVA shows a significant effect of ID and place on movement time, but no significant interaction $F(1, 316) = 0.72, ns$, hence the slopes, or the influence on movement time by ID is comparable for in-lab and out-of-lab VR pointing. The distribution of movement times between two targets were similar for in-lab and out-of-lab, both ranging between roughly 500 and 2000 milliseconds (see Figure 6.2). The distribution hints that participants in both studies did not encounter any difficulties nor took any noticeable breaks in the middle of trials.

Accuracy. The coordinates of each trial's movements between to targets were fitted to a linear regression. An r^2-value of 1.0 would thus show a perfect linear movement, and 0.0 a non-linear movement

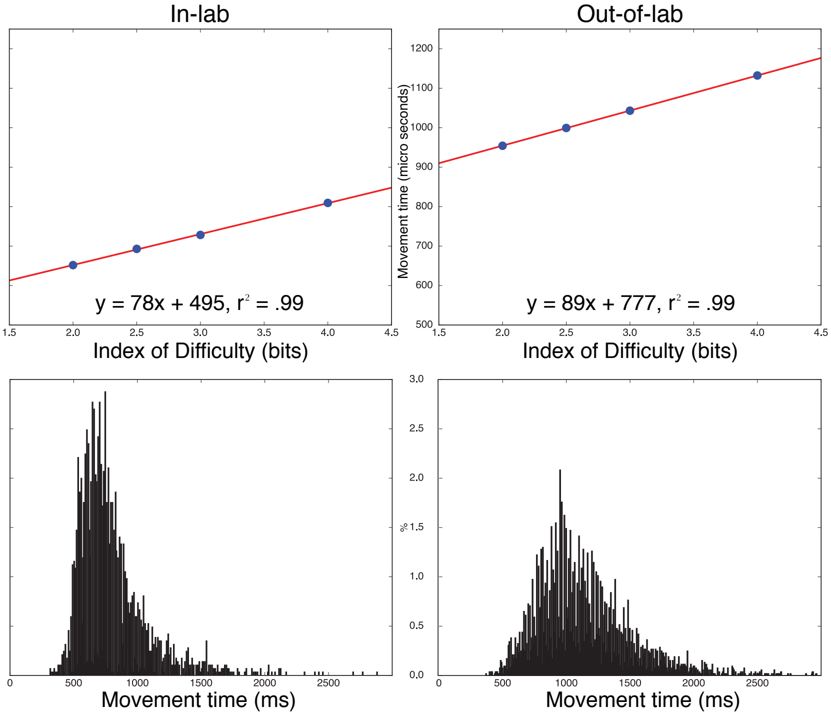


Figure 6.2: Top: MT as a function of ID. Bottom: the distribution of movement times. Left: In-lab; Right: Out-of-lab.

between targets. This metric thus represents how optimal the movement was. The in-lab study showed an average linearity of .75 (*SD* = .07), and the out-of-lab study .63 (*SD* = .04). The accuracy was significantly higher for the in-lab study: $F(1, 78) = 99.4, p < .001$.

Throughput. We find a higher throughput (computed using the formula from [206]), TP, for the in-lab study: $TP = 3.96$ (*SD* $= .61$), compared to out-of-lab: $TP = 2.85$ (*SD* $= .57$). In-lab participants had significantly higher TP: $F(1, 78) = 63.4,\ p < .001$.

Summary. The results show that participants had no difficulties conducting VR pointing without an evaluator present, and that VR pointing tasks can be conducted virtually anywhere. However, task completion times, throughput, and accuracy were significantly better when the experiment was conducted in-lab.

3D Tracing Task

Completion Time. The manipulation did not have a significant effect on task completion times (see Table 6.3). There was not a significant difference on completion time between in-lab and out-of-lab either.

	Dynamic	*Static*	*p*
In-lab	240.9s ± 62.4	247.8s ± 79.0	
Out-of-lab	292.4s ± 92.1	269.8s ± 91.2	

Table 6.3: Completion times for the 3D Tracing task, $\pm SD$.

Correct Estimations. Both the in-lab and out-of-lab study showed that participants in the dynamic condition estimated the origin of the leaves significantly better (see Table 6.4). In-lab participants increased estimations by 10.2%; Cohen's $d = .35$; out-of-lab participants increased estimations by 9.1%; Cohen's $d = .39$. The effect of the experimental condition was significant for both studies; for in-lab this effect was $F(1, 25) = 13.1,\ p < .01$, for out-of-lab it was $F(1, 51) = 16.0,\ p < .001$.

	Dynamic	*Static*	*p*
In-lab	86.8% ± 8.4	76.6% ± 4.6	**
Out-of-lab	83.1% ± 10.6	74.0% ± 9.5	***

Table 6.4: Correct estimations for the 3D Tracing task, $\pm SD$.

In-lab vs. Out-of-lab. The rates of correct answers to which tree the leaves originated were negligible between in-lab and out-of-lab. In-lab participants on average estimated 81.7% correct; out-of-lab 78.6%. This difference was not significant. Confidence intervals for the effect sizes for this task show that the effects due to experimental conditions are within the same range: out-of-lab, $d = 1.11$, 95% CIs $[.45, 1.80]$; in-lab, $d = 1.40$, 95% CIs $[.40, 2.77]$.

Summary. Results from both the in-lab and out-of-lab study showed that participants increase their rate of correctly estimating the origin of a leaf in VR, given the ability to inspect trees spatially. The data show that the effect of the experimental condition was the same between in-lab and out-of-lab, and with comparable effect sizes. Also, in-lab and out-of-lab participants did not perform significantly different in terms of speed. The data therefore provide evidence for the validity of conducting VR studies that entail 3D navigation outside the laboratory.

Body Ownership Illusion Task

Body Ownership. Body ownership is *different sensory cues unified into the perception of my body* [110]. We asked four questions from [12] to quantify body ownership (see Table 6.2): VRBody ($Q1$), Mirror ($Q2$), Features ($Q3$), and TwoBodies ($Q4$). Figure 6.3 gives an overview of the responses. Based on previous work (e.g., [12]), we expected the visuomotor consistency to increase the degree of body ownership.

For the in-lab study, all means but TwoBody were higher in the consistent condition; for the out-of-lab study all means but VRBody were higher in the consistent condition. Using a Wilcoxon rank-sum test on these questions across the consistent and inconsistent conditions, we found that only Mirror showed a significant difference, and only for the in-lab study, $Z = -1.92, p = .05$.

Presence. Presence is the sense of *being there*, distinguished from *immersion*, as being the participants's response to the environment, and not relating the fidelity of technology used [202]. We asked three questions originating from [211], relating to presence (see Table 6.2): BeingThere ($Q6$), VRExp ($Q7$), and VRWorld ($Q8$). Figure 6.4 gives an overview of the responses. As evident from Figure 6.4, the differences between conditions for both studies are negligible; the medians are similar across both conditions both for the in-lab and out-of-lab study. No significant differences for any of the presence questions were found for the two conditions using a Wilcoxon rank-sum test. Steed et al. [211] did also not find any difference in degrees of presence across conditions, except for synchronous tapping which had the opposite effect to the expected (presence decrease).

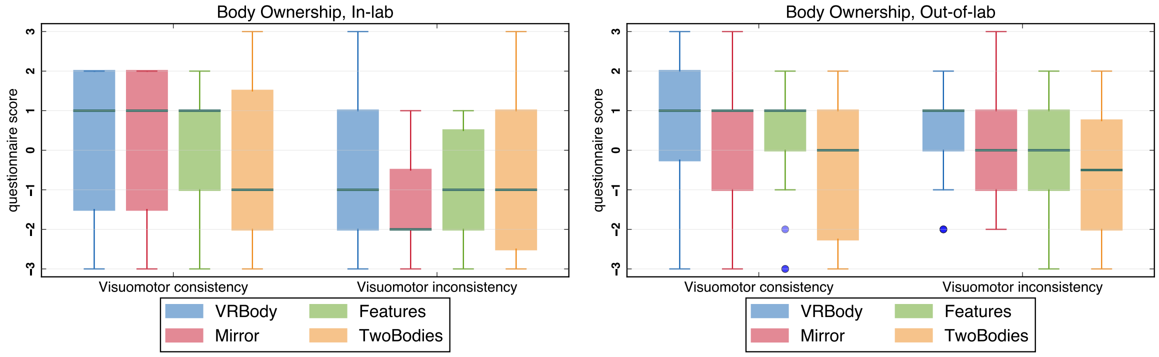


Figure 6.3: Box plots of body ownership data ($Q1 - Q4$) from in-lab (left) and out-of-lab (right). Solid lines show medians, boxes show interquartile ranges, circles show outliers.

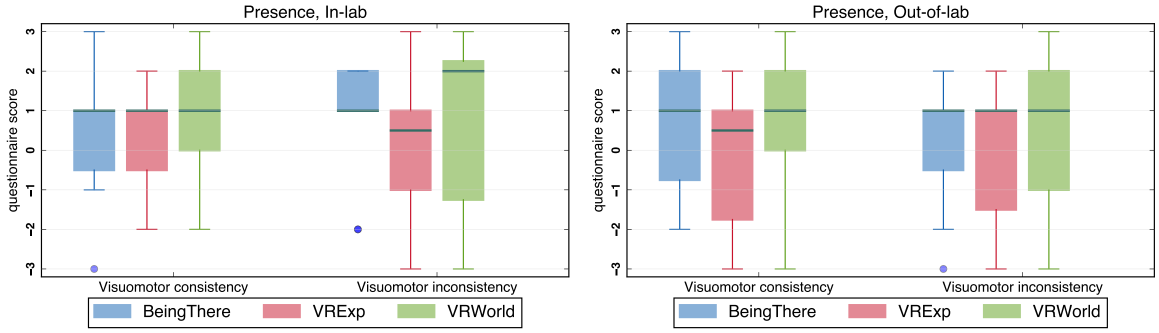


Figure 6.4: Box plots of presence data ($Q6 - Q8$).

Probing Individual Differences. Steed et al. [211] wrote that *"An obvious route for in the wild studies would be to probe individual differences in presence response"*. We did so and found no differences attributable to study place or VR technology: a Wilcoxon rank-sum test comparing participants' responses to the questionnaire showed no significant differences for any questions across in-lab/out-of-lab; Z-values ranging from $[-1.81, -.16]$ and p-values from $[.07, .92]$. While this does not directly verify the feasibility of studying complex VR phenomena out-of-lab, it shows that for the employed experiments more advanced VR technology combined with higher experimental rigor, did not cause significant changes to responses to body ownership and presence.

Summary. Similar to [12], the body ownership means in the consistent conditions were higher for all questions, but one, in both the in-lab and out-of-lab study (one significant). There were no differences attributable to study place, with similar responses in-lab and out-of-lab. This shows that it is feasible to obtain comparable data to laboratory experiments, even for more complex VR phenomena when conducting them out-of-lab.

6.6 Follow-up Study: In-lab/Low-tech

While the in-lab and out-of-lab studies by and large yielded comparable results, it is hard to attribute the observed differences between the two studies to condition (lab, out-of-lab) or the hardware (low-tech, high-tech). We find this confound natural because in-lab studies would typically be conducted with high-tech and out-of-lab studies are currently only feasible with low-tech VR. However, it leaves us unable to discuss the relative influence of condition and hardware. We therefore conducted a follow-up study. We speculate that some of the differences observed in our user studies could be explained by the setting, where other differences could be explained by the employed hardware.

To explore whether the differences observed were due to experimental setting (lab vs. non-lab), or fidelity of VR apparatus (HTC Vive vs. Google cardboard), we ran a follow-up study. We conducted an in-lab study, using the out-of-lab technology from the first study. We thus did a new study with an additional condition: in-lab/low-tech. The follow-up study showed that differences in absolute performances are likely related to employed technology, where more complex VR phenomena such as immersion scores showed to differentiate with experimental control.

Participants

Twenty-two people, aged 20-40 (*SD* = 5.8 years) participated in this laboratory study, and were reimbursed with a gift equivalent of 15$ US. None of the participants had previously participated in any of our studies. Seven of the participants were male. One participant was discarded due to the control question. The experimental design and apparatus followed the out-of-lab condition of the first study.

Pointing Task

While the first study showed that a simple pointing task can easily be completed by participants both in-lab and out-of-lab, significant differences in performances were reported. The results from the second study provides evidence, that this performance gap is most likely due to fidelity of the employed VR technology. Figure 6.5 shows comparable performances between the out-of-lab and in-lab/low-tech conditions, with significant differences to the in-lab performances. Neither movement time, accuracy, or throughput varied significantly between the out-of-lab and in-lab/low-tech conditions. This shows that the differences observed in the first study for this task, are most likely due to the employed apparatus.

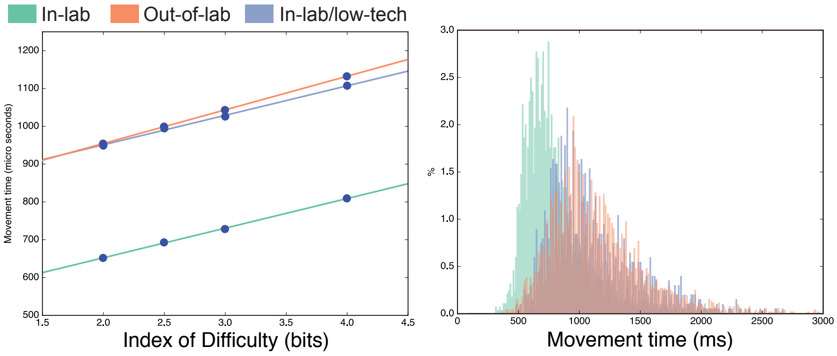


Figure 6.5: Results from the pointing task in-lab, out-of-lab, and in-lab/low-tech: (left) MT as a function of ID, and (right) histograms of movement times. Differences in pointing performances are likely due to the technology.

3D Tracing Task

In the first study we observed that the experimental condition had the same effect across study places; both participants in-lab and out-of-lab estimated origins of leaves better using a dynamic view. We did not observe significant differences in completion times.

	Dynamic	*Static*	*p*
In-lab	240.9s ± 62.4	247.8s ± 79.0	
Out-of-lab	292.4s ± 92.1	269.8s ± 91.2	
In-lab/low-tech	277.0s ± 155s	276.0 ± 72.6	

Table 6.5: Completion times for the 3D Tracing task, $\pm SD$.

As Table 6.5 shows, the experimental manipulation did not cause changes to completion time in the in-lab/low-tech study, as with the two previous conditions. Through-out all three conditions, completion time for the 3D tracing task did not vary significantly between experimental conditions.

The experimental manipulation caused better estimations for both in-lab and out-of-lab. As evident from Table 6.6 this pattern was also true for the follow-up study; the dynamic condition caused significantly better estimations of origins, $F(1, 20) = 14.4, p < .001$.

	Dynamic	*Static*	*p*
In-lab	86.8% ± 8.4	76.6% ± 4.6	**
Out-of-lab	83.1% ± 10.6	74.0% ± 9.5	***
In-lab/low-tech	87.0% ± 7.5	75.0% ± 7.0	***

Table 6.6: Correct estimations for the 3D Tracing task, $\pm SD$.

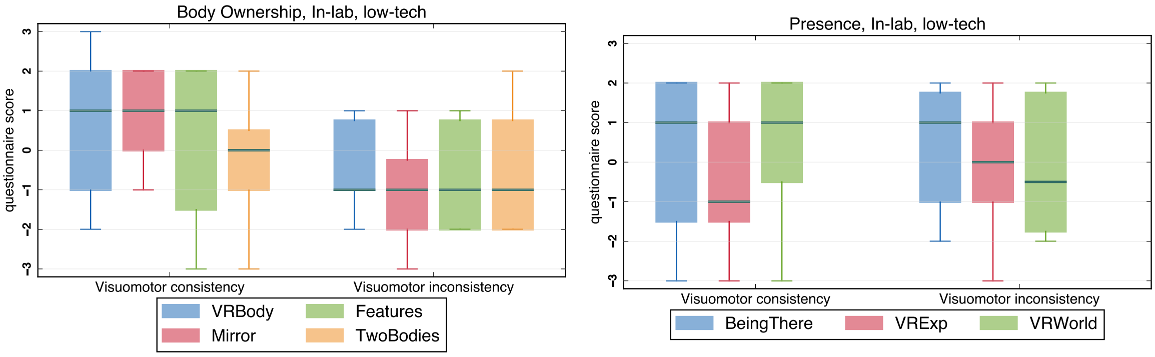
The in-lab/low-tech condition did not significantly differ from the the other conditions. Confidence intervals for the effect sizes for the in-lab/low-tech condition showed comparable to the first study: $d = 1.66$, 95% CIs $[.61, 2.56]$, showing that the intervals for out-of-lab and in-lab/low-tech are contained in the in-lab interval.

Body Ownership Illusion Task

The third task had the most complex VR phenomena of the three. It studied if an avatar's motor consistency with the participant varies the participant's sense of body ownership and presence.

Body Ownership. The mean of all body ownership scores were higher in the condition with visuomotor consistency (see Figure 6.6a), consistent with the first study. Exactly as with the in-lab condition for the first study, only *Mirror (Q2)* showed to significantly differ between the conditions, $Z = -2.45, p = .01$. This tells us, that even though we observe the same overall trends in body ownership throughout the studies, the two laboratory studies did cause more similar ownership responses.

Presence. As with the first study the observed differences in presence due to experimental manipulation are very negligible, making it difficult to make any inferences of the experimental variation on presence scores.



(a)

(b)

Figure 6.6: Box plots of questionnaire results on (a) body ownership ($Q1 - Q4$), and (b) on presence ($Q6 - Q8$). The results fit the laboratory condition from the first study.

Summary of Follow-up Study. The follow-up study showed that the differences in simple performance metrics (e.g., accuracy, speed of pointing), between and the in-lab and out-of-lab conditions of the VR study were likely due to the hardware used. High-end VR equipment often deployed for in-lab studies caused faster interactions, with higher accuracy, compared to commodity VR technology.

Conversely, in the more delicate spectrum of dependent variables for the VR studies; the place of study seems to be of more concern than the fidelity of technology. This is indicated by our results, in the body ownership illusion task; results from the follow-up study, all-though by and large follow the same trend as both previous condition, match the in-lab better. That is, that the mean body ownership showed higher for the condition with visuomotor consistency for most questions, but the same question ($Q2$) showed statistically different by experimental manipulation in the in-lab and in-lab/low-tech conditions.

6.7 Other data

In addition to the tasks' dependent variables we logged other measurements to get insights in the uncertain factors of conducting studies without a human evaluator. We here look at the physical surroundings and the differences in technology.

Setting

During the out-of-lab study, participants' phones stored a photo using the back-facing camera, to provide insights into to the contexts in which participants carried out the study. We later printed all photos (see example photos in Figure 6.7), and categorized them. The categorization of the photos resulted in five non-exclusive groups:

- **Place**: where participant was during the study
- **Locale**: type of place (home, public or office)
- **Barriers**: near surroundings contained physical obstacles
- **Activity**: surroundings showed signs of co-occurring activity
- **Social**: other humans were present

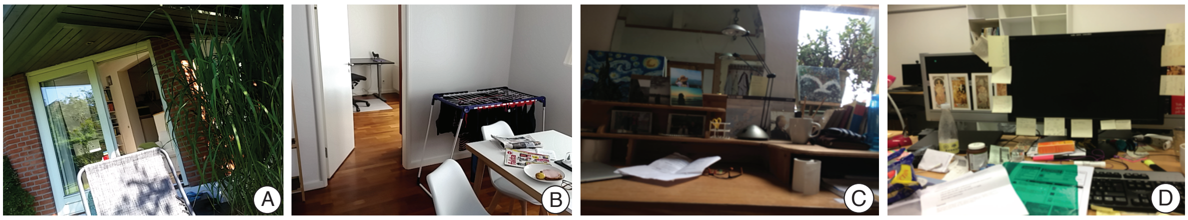


Figure 6.7: Examples of out-of-lab study settings: outside (a), inside (b,c,d), home (a,b,c), office (d), standing (a,b), sitting (c,d). Photos printed with permission from the participants.

Note that thirteen of the photos did not contain much information, for instance when showing a close-up wall, and hence did not provide further insights other than it was taken inside. Additional, nine photos were indecipherable, for instance very blurry or dark.

Out-of-lab Technology

We analyzed the effect of two parameters relating to participants' equipment: phone brand and screen size, to see if the technology used had an impact on the performance of the participants. We looked at all the dependent variables from each task (see Figure 6.1). Of the 53 participants, 31 participated with an iOS device and 22 with an Android device. We found no significant difference on any of the dependent variables: $F(1, 51) = [.01, 2.7]$, $p = [.11, .92]$. We compared area of screen to the same parameters, and also here found no significant effects attributable to screen size: $F(5, 47) = [.42, 1.7]$, $p = [.15, .84]$.

Participation in Out-of-lab Study

We distributed 100 cardboard VR glasses over the course of 20 days. 80 participants installed our experimental application, and 57 completed the study. The data show that throngs of participants do not come for free, but that it is possible to recruit subjects with the modest reimbursement of a pair of cardboard glasses. This presumable only works for first time VR users; as VR equipment becomes more commonplace, other reimbursements should be offered.

6.8 Discussion

We have explored if it is feasible to conduct virtual reality (VR) user studies outside the laboratory. Potentially, this would give access to more varied physical and social settings, and higher participation, which in turn could give VR studies higher external validity at a lower cost. In particular, across three tasks we investigated if the performance parameters obtained (e.g., task completion times) compare to a laboratory condition and if the findings of experimental comparisons compared across in-lab and out-of-lab. For the two tasks containing experimental conditions, we did find significant effects (i.e., not null results). We found that similar effect sizes can be found using an easier and cheaper study method.

We believe that this gives a good first indication of how VR could be crowdsourced. Because VR equipment is currently not widely available we ran the out-of-lab experiments by giving cardboards to participants. We decided against mailing out cardboards because popular crowdsourcing platforms currently do not have the population required for an out-of-lab VR study, and we are not in a country with a significant crowdworker population. Nonetheless, the results show that valid data can be acquired from VR studies without supervision, across a range of VR phenomena and complexity.

The literature contains numerous comparisons of in-lab and out-of-lab studies (e.g., [28, 50, 73, 140, 166, 181, 182]); to our knowledge, this paper is the first to do such comparison for VR studies. Steed et al. [211] provided a first exploration of whether VR studies could be conducted in-the-wild, in this paper we have explored whether the results of in-lab and out-of-lab studies are comparable and indeed whether out-of-lab is a valid methodology for VR.

Our results showed that the absolute differences in performance between the in-lab and out-of-lab study were substantial; participants in the laboratory study performed better in most of the absolute performance metrics (throughput, accuracy, completion time, and depth estimation). A follow-up study however revealed that this difference is likely due to technology used, and to a lesser extent due to the experimental design.

Data from all tasks confirms the feasibility of out-of-lab VR: there were no significant differences between effects of experimental conditions for tasks when comparing the in-lab and out-of-lab studies. We show that even complex VR phenomena entailing body ownership are possible to conduct out-of-lab with comparable results to in-lab studies, although the effects indicate that levels of body ownership were likely higher for the laboratory-based studies.

Recommendations

Although our setup is limited in a number of ways to be discussed, we can still provide a first set of recommendations on VR studies outside the laboratory.

- Pre-screen participants for the technology accessible to them to avoid recruiting unqualified people
- Expect roughly half of the participants to complete the study
- 15 minutes seems like the maximum tolerable duration for keeping the pose required to use the VR cardboard system
- Validate the integrity of participants, for instance using verifiable control questions, context photos, or user performance.
- Design experiments well-suited for both standing and sitting
- Expect simpler dependent variables (speed, accuracy, throughput) to vary with technology, but complex phenomena (body ownership, presence) to depend more on internal control

Open Questions

The results in this paper were obtained with Google cardboards. First of all, it raises the question of how to achieve the large-scale participation typically seen in out-of-lab research. We believe it is possible to mail cardboards directly to participants who have signed up online or possibly have them buy them and be reimbursed. The current rate of cardboard adoption (about 2%-3% of the crowdworkers surveyed) makes recruiting on crowdsourcing platforms infeasible for anything but small studies. Second, of course it raises the question what happens when more crowdworkers have high-end equipment (e.g., Oculus, HTC Vive). We do not see that as imminent but the differences in settings, movement of participants, and absolute performance values would be interesting to observe.

The current approach, mainly due to how cardboard VR glasses work sets several limitations on the task design. People must hold the same posture (binocular pose) during the entire study, and it is therefore infeasible to actively use the hands for anything, as you normally would in more advanced VR immersions. Additionally, the use of vibrotactile feedback (such as synchronous stimulation with a rod on virtual and physical body), as seen in many body ownership illusion studies (e.g., [195, 197, 199]), is infeasible. We foresee that these limitations could be resolved in the future, due to advances in commodity VR and wearable technology. This will also open up to longer studies, where fatigue will not be a factor in task design.

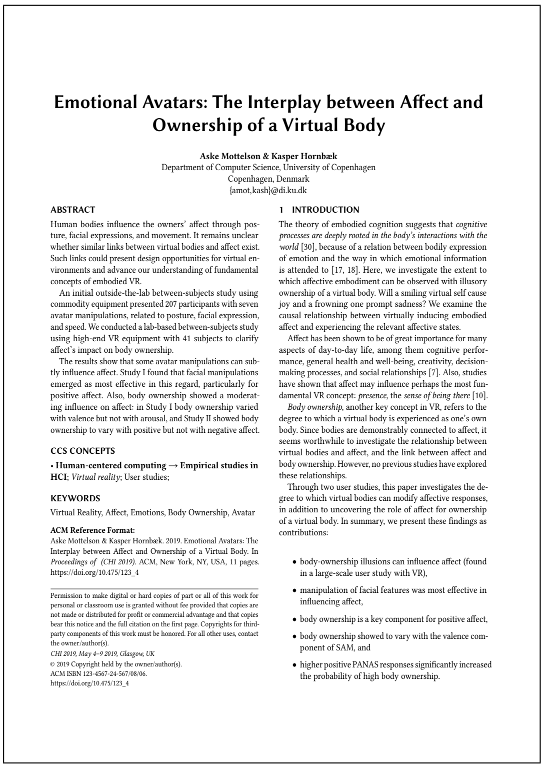
Conclusion

This paper shows that for VR studies concerning a heterogeneous population, out-of-lab experimentation is a worthwhile and valid methodology to consider. We compared VR tasks concerning pointing, 3D tracing and body ownership illusions, both as in-lab and out-of-lab studies. We showed that it is feasible to get reliable data by conducting VR user studies outside the laboratory, across a range of tasks and VR phenomena. This study is the first to validate VR experimentation outside the laboratory, and provides a first set of suggestions on how to crowdsource VR user studies.

Part III

Influencing

7 Affective Avatars



Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body

Aske Mottelson & Kasper Hornbæk
Department of Computer Science, University of Copenhagen
Copenhagen, Denmark
{amot,kash}@di.ku.dk

ABSTRACT

Human bodies influence the owners' affect through posture, facial expressions, and movement. It remains unclear whether similar links between virtual bodies and affect exist. Such links could present design opportunities for virtual environments and advance our understanding of fundamental concepts of embodied VR.

An initial outside-the-lab between-subjects study using commodity equipment presented 207 participants with seven avatar manipulations, related to posture, facial expression, and speed. We conducted a lab-based between-subjects study using high-end VR equipment with 41 subjects to clarify affect's impact on body ownership.

The results show that some avatar manipulations can subtly influence affect. Study I found that facial manipulations emerged as most effective in this regard, particularly for positive affect. Also, body ownership showed a moderating influence on affect: in Study I body ownership varied with valence but not with arousal, and Study II showed body ownership to vary with positive but not with negative affect.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *Virtual reality*; User studies;

KEYWORDS

Virtual Reality, Affect, Emotions, Body Ownership, Avatar

ACM Reference Format:
Aske Mottelson & Kasper Hornbæk. 2019. Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body. In *Proceedings of (CHI 2019)*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.475/123_4

CHI 2019, May 4–9 2019, Glasgow, UK
© 2019 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

The theory of embodied cognition suggests that *cognitive processes are deeply rooted in the body's interactions with the world* [30], because of a relation between bodily expression of emotion and the way in which emotional information is attended to [17, 18]. Here, we investigate the extent to which affective embodiment can be observed with illusory ownership of a virtual body. Will a smiling virtual self cause joy and a frowning one prompt sadness? We examine the causal relationship between virtually inducing embodied affect and experiencing the relevant affective states.

Affect has been shown to be of great importance for many aspects of day-to-day life, among them cognitive performance, general health and well-being, creativity, decision-making processes, and social relationships [7]. Also, studies have shown that affect may influence perhaps the most fundamental VR concept: *presence*, the *sense of being there* [10].

Body ownership, another key concept in VR, refers to the degree to which a virtual body is experienced as one's own body. Since bodies are demonstrably connected to affect, it seems worthwhile to investigate the relationship between virtual bodies and affect, and the link between affect and body ownership. However, no previous studies have explored these relationships.

Through two user studies, this paper investigates the degree to which virtual bodies can modify affective responses, in addition to uncovering the role of affect for ownership of a virtual body. In summary, we present these findings as contributions:

- body-ownership illusions can influence affect (found in a large-scale user study with VR),
- manipulation of facial features was most effective in influencing affect,
- body ownership is a key component for positive affect,
- body ownership showed to vary with the valence component of SAM, and
- higher positive PANAS responses significantly increased the probability of high body ownership.

Affective computing, as defined by Picard in 1997 is *computing that relates to, arises from, or influences emotions* [171]. The perhaps most commonly researched topics within Affective Computing relates to making computers understand human affect; affect recognition. Many different methods have been suggested in the research field's roughly twenty years existence. The most robust approaches for this use physiological sensors (e.g, pulse, hear rate, skin response) and video.

As presented in the previous part of this thesis, some research is focusing cognitive sensing around interaction data alone, showing promising classification accuracies for boredom, affect, and truth estimation, among others. Little research has, however, been conducted in *influencing* affect using human-computer interaction. Certainly there has been extensive research in emotion elicitation; manipulating affect for instance using film, stories, or imagery. These methods are, however, quite generic and usually make it obvious for the subject what is going on. Also, they have seldomly been implemented using interactive technology.

In this chapter I report how to use VR avatars to influence participants' affect; specifically we hypothesized that body-affect links, might have the same effect for virtual bodies and affect. The chapter is identical to [152], as shown to the left, which is unpublished. In addition to experimenting with VR bodies and their influence on affect, the paper reports on how affect and body ownership relates.

This chapter is based on a collaborative effort as described below.

Title
Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body

Authors
Aske Mottelson and Kasper Hornbæk

Journal
-

DOI
-

What was the role of the PhD student in designing the study?
The PhD student was the first author of the paper,
and responsible for the design of the described studies.

How did the PhD student participate in data collection and/or development of theory?
The PhD student was responsible for study implementation,
execution, data collection, and theory development.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed to all parts of the manuscript.

Did the PhD student read and comment on the final manuscript?
Yes.

7.1 Abstract

Human bodies influence the owners' affect through posture, facial expressions, and movement. It remains unclear whether similar links between virtual bodies and affect exist. Such links could present design opportunities for virtual environments and advance our understanding of fundamental concepts of embodied VR.

An initial outside-the-lab between-subjects study using commodity equipment presented 207 participants with seven avatar manipulations, related to posture, facial expression, and speed. We conducted a lab-based between-subjects study using high-end VR equipment with 41 subjects to clarify affect's impact on body ownership.

The results show that some avatar manipulations can subtly influence affect. Study I found that facial manipulations emerged as most effective in this regard, particularly for positive affect. Also, body ownership showed a moderating influence on affect: in Study I body ownership varied with valence but not with arousal, and Study II showed body ownership to vary with positive but not with negative affect.

7.2 Introduction

The theory of embodied cognition suggests that *cognitive processes are deeply rooted in the body's interactions with the world* [241], because of a relation between bodily expression of emotion and the way in which emotional information is attended to [157, 158]. Here, we investigate the extent to which affective embodiment can be observed with illusory ownership of a virtual body. Will a smiling virtual self cause joy and a frowning one prompt sadness? We examine the causal relationship between virtually inducing embodied affect and experiencing the relevant affective states.

Affect has been shown to be of great importance for many aspects of day-to-day life, among them cognitive performance, general health and well-being, creativity, decision-making processes, and social relationships [31]. Also, studies have shown that affect may influence perhaps the most fundamental VR concept: *presence*, the *sense of being there* [54].

Body ownership, another key concept in VR, refers to the degree to which a virtual body is experienced as one's own body. Since bodies are demonstrably connected to affect, it seems worthwhile to investigate the relationship between virtual bodies and affect, and the link between affect and body ownership. However, no previous studies have explored these relationships.

Through two user studies, this paper investigates the degree to which virtual bodies can modify affective responses, in addition to uncovering the role of affect for ownership of a virtual body. In summary, we present these findings as contributions:

- body-ownership illusions can influence affect (found in a large-scale user study with VR),
- manipulation of facial features was most effective in influencing affect,
- body ownership is a key component for positive affect,
- body ownership showed to vary with the valence component of SAM, and
- higher positive PANAS responses significantly increased the probability of high body ownership.

7.3 Overview

Measuring Affect

Both physiological and subjective measurements are commonly employed for estimation of affective responses. Especially popular among the latter are the Positive and Negative Affect Schedule (PANAS) [236] and the Self-Assessment Manikin (SAM) [24]. The former considers positive and negative affect on independent scales, producing two summary scores from participants' rating of their emotional fit on a five-point scale with 10 positive and 10 negative words, while the SAM typically presents two scales, valence (pleasantness) and arousal (activation), using a pictorial manikin for representing the scores. There are versions with a five-, seven-, and nine-point scale (Figure 7.1 shows a nine-point SAM).

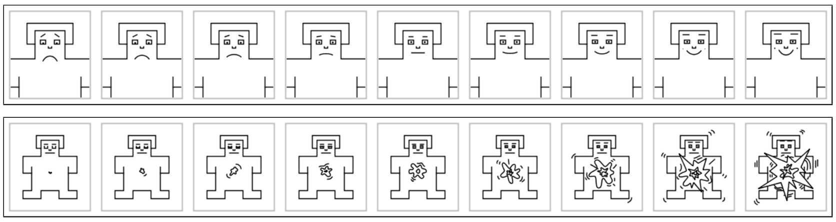


Figure 7.1: The nine-point Self-Assessment Manikin, measuring valence (unpleasant–pleasant, at the top) and arousal (deactivation–activation, at bottom).

Manipulating Affect

The process of altering affect experimentally is known as affect manipulation. It is sometimes colloquially referred to as mood induction or emotion elicitation (although this informal use is undesirable). Affect manipulation allows researchers to enforce specific affective responses from a sample than would have arisen if subjects had merely reported on their current affective state. Hence, this manipulation is an important psychological tool for understanding how mood, emotion, and core affect modify human cognition and behavior within the constraints of experimental scrutiny.

There are several ways of manipulating affect (Westermann et al. give an overview [238]). Among the most commonly employed are film, IAPS, and Velten. These respectively involve showing affective movie clips, imagery, or asking participants to immerse themselves in an emotional story.

Virtual Reality As an Affective Medium

Research covering virtual reality and affect is surprisingly scarce, notwithstanding the general consensus on affect's highly important influence on cognition and behavior, and the recent proliferation of behavioral research using virtual reality. That said, some scholars have attempted to manipulate affect via virtual environments [14, 64, 105, 180, 221]. We review that work below.

Baños et al. [14] created a virtual environment capable of eliciting four discrete emotions. The environments depicted alterations of a city park that incorporated a multitude of classical emotion-elicitation procedures, such as Velten, IAPS, and movie clips. The full procedure, lasting 30 minutes, showed effective for discrete emotion elicitation.

Another emotion-eliciting virtual park was created by Felnhofer et al. [64]. Five park scenarios were designed, each eliciting a specific emotion: joy, anger, boredom, anxiety, or sadness. The scenarios differed in lighting, coloring, and sound, along with their plant and animal types. The authors concluded that virtual environments can be used to induce emotional states.

Riva et al. [180] constructed their own park scenarios to elicit emotions in VR. Their three parks shared the same structure but differed in their aural and visual experience. These researchers too confirmed the efficacy of VR as an affective medium.

To our knowledge, only Jun et al. [105] have attempted to alter affect via avatar manipulations. They found that facial expressions of a virtual avatar can modulate emotions and that greater presence is associated with higher valence.

Body-Ownership Illusions

Body-ownership illusions are a class of experiment in which the participants are led to believe that they are owners of another body, or part thereof [170]. These illusions are persuasive when the replacement bodies are virtual avatars made possible by VR [110]. An optimal body-ownership illusion induces a high level of ownership of the virtual body, usually by means of visuo-motor synchrony between the physical and virtual body, sometimes alongside tactile feedback [159].

While there is ongoing scholarly debate on how far avatar manipulations can be extended without distortions to body ownership, evidence thus far suggests that ownership can be maintained even with substantial alterations to the avatar, such as scaling the arms to three times the normal length [111], rotating the body 15 degrees [20], or using a child's body for the avatar [12].

A large body of research in this area is focused on manipulating avatars to study whether humans' perception of the world can be influenced through body alteration/replacement. For instance, owning a child avatar influences estimates of objects' size and expedites association of the self with childlike attributes [12]; similarly, owning a different-skin-toned avatar affects implicit racial bias [168].

The aforementioned body of work confirms the feasibility of changing body-oriented perceptions when participants are manipulated to believe the virtual body they see in VR is, in fact, their real body.

Limitations of Previous Work

While previous studies attest that VR scenarios can alter affect, this has been confined largely to park scenarios. The conditions involved are time-consuming, explicit in their purpose, and geared for a specific narrative. Most importantly, they also neglect embodiment as a part of the immersion. This renders it hard to evaluate the processes leading to affect manipulation in embodied VR precisely. While some evidence suggests that one can influence affect and presence by manipulating the avatar's facial features [105], the connection between affect and the illusion of owning a virtual body remains unclear, as does whether other avatar manipulations may inform affect. Some authors have shed light on affect's role for presence [13]: the literature suggests a positive correlation between presence and affect.

We provide evidence that affect can be influenced by only bodies truly thought to be ours. Additionally, we present a first study showing interplay between affect and body ownership. Our findings confirm the existence of a link between (positive) affect and virtual bodies.

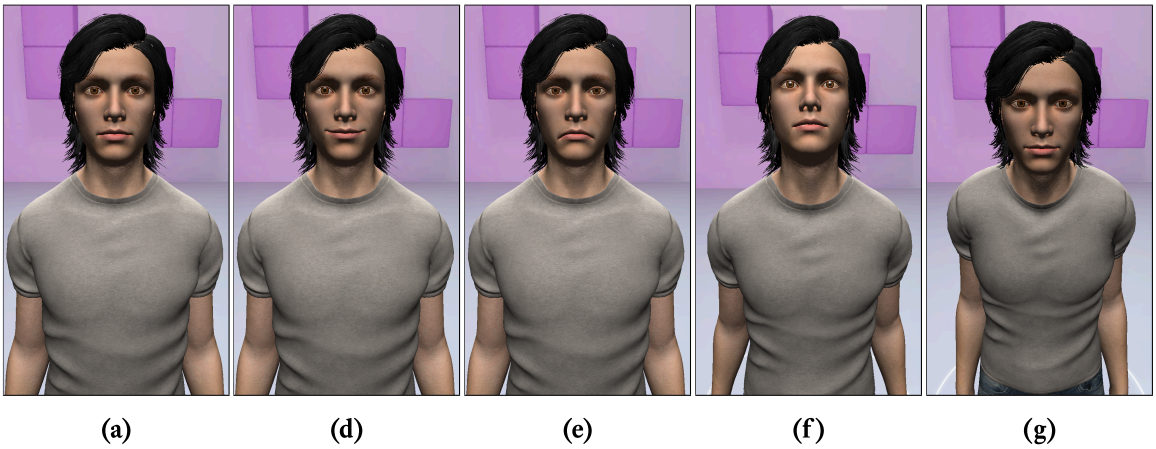


Figure 7.2: Experimental manipulations for Study I, as described in Table 7.1. Manipulations (b) and (c) used the same body as the control condition (a), with changes to speed of navigation.

7.4 Study I

We conducted an outside-the-lab VR study, as research has proposed this to be a valid method for low-cost VR user studies with high power [153, 211]. Each participant was given a commodity cardboard VR headset to use in combination with his or her smartphone. This was deemed a cost-effective approach enabling a large-scale user study that would allow us to try many experimental conditions and, thereby, many avatar manipulations. To our knowledge, no other VR user study has been done with more than 200 participants.

The overall purpose of this study was to i) validate the feasibility of influencing affect with avatar manipulations in VR, ii) experiment with several avatar manipulations with high power, and iii) identify specific questions pertaining to the body-ownership–affect link for further investigation via a laboratory study focusing on high internal validity.

With Study I, we were interested in core affect [57]. Since this is the broadest class of non-reflective affective feelings available to the consciousness, core affect is simpler to address than the underlying emotions and moods. Because it was not clear in advance how the manipulations would influence participants, we decided on a dimension-based conceptualization of core affect, involving valence and arousal (similar to Russell's circumplex model of affect [57]).

Participants

In all, 207 undergraduate computer science students took part in the study. Most were male (51 females participated), and the age range was 19–50 (M = 23, SD = 3.8). Smartphones with iOS were slightly more common than Android ones among the participants, with 110 using iOS in the study. Participation was counted toward the students' credit for a compulsory introductory HCI course.

Apparatus

The participants used their smartphones within a Google cardboard VR headset that had a head strap. A few students did not have a compatible device, so we lent them one. The application was developed with the Google VR SDK for Unity 2017, deployed for Android and iOS and distributed via the relevant app stores. We employed an inverse kinematics (IK) system for the humanoid avatar, using the head as the only tracked point, with three degrees of freedom (DoF).

Design

We used three tasks in the design of our between-subjects experiment. The independent variable was body manipulation, with seven conditions (see Table 7.1 and Figure 7.2): control, speed (*slow*/*fast*), face (*smile*/*frown*), and posture (*upright*/*hunched*)

These specific avatar manipulations were chosen based on a review of body–affect relations. Coombes et al. [44], for instance, show that speed of a task varies as a function of affect, and Wallbott [229] presents an overview of how postures are related to emotion. Finally, research has showed how facial expressions can modulate affect [105, 157, 213].

The tasks were adapted from pen-and-paper ones in Strack et al.'s work [213] to suit an embodied VR experience. In all three tasks, the subject selected spheres in turn, on the basis of the number or letter printed on them. Each sphere was initially white, turned red when the subject looked at it, and then turned blue once the subject had been dwelling on it for two seconds. While participants were performing the tasks, a large mirror was in front of them, showing them their avatar (see Figure 7.3). Between tasks, the mirror served as a screen displaying buttons with which participants answered questions, again via two-second dwell times. The specifics of the tasks are described below.

Task I

Task I was designed to give familiarity with selecting targets by dwelling on them via head position and to introduce Likert-scale use in VR. Participants selected two spheres, numbered 1 and 2, by dwelling on each in turn.

Task II

In the second task too, the participants selected numbered spheres, this time from 1 to 9.

Task III

For the final task, they were asked to select all spheres with vowels ("Y" was optional) on them, in any order. A sphere was present for each letter of the alphabet.

	Variable	Manipulation	Hypothesis
(a)	Control	None	Baseline
(b)	Rotation speed	Slow (-40%)	High affect
(c)	Rotation speed	Fast (+40%)	Low affect
(d)	Face	Smile	High affect
(e)	Face	Frown	Low affect
(f)	Posture	Upright	High affect
(g)	Posture	Hunched	Low affect

Table 7.1: The experimental manipulations in Study I (the hypotheses correspond to SAM responses).

In the control condition (a), there were no manipulations. For the speed conditions, we increased/reduced 40% rotational gain around the y-axis, such that a 100-degree movement would result in rotation of 140 degrees for (b) and 60 degrees for (c). The face conditions manipulated facial features (using built-in morphing), to produce smiling (d) and frowning (e). Postural manipulations were implemented by adding artificial head and chest targets for the IK model, resulting in either upright (f) or hunched (g) posture.

Measurements

The primary dependent variables were valence and arousal, measured with a nine-point SAM (see Figure 7.1) [24] administered with the textual prompt "How do you feel?" (Table 7.2 provides the full list of dependent variables). We chose the SAM to measure the primary dependent variables because it requires few of the cumbersome in-VR button selections, which would be prevalent using common affect questionnaires (e.g., PANAS [236]). The dominance scale sometimes used as a third dimension for the SAM instrument is more emotion- than core-affect-oriented [57] so was not used in our study.

Measurement	Instrument	Category
Valence	SAM	Affect
Arousal	SAM	Affect
Difficulty	Likert scale	Post-task metric
Completion time	Clock	Performance
Error rate	Head orientation	Performance
Mirror time	Head orientation	Activity
Body ownership	Two questions	Questionnaire
VR experience	Four options	Questionnaire

Table 7.2: The dependent variables for Study I.

Also, we measured task difficulty (subjective difficulty), speed (task duration), and accuracy (error rate) for each task. In addition, we recorded the time spent looking in the virtual mirror, as the number of frames for which the participant's point of view collided with the mirror.

Procedure

Participants entered a virtual room (see Figure 7.3) after installing and opening the mobile application and placing their phone in the cardboard VR headset. In this room, a standing sex-matched avatar could be seen in the mirror. Participants' coarse body movements were mapped in real time, with head rotation used alongside IK to create a sense of body ownership over the virtual character. Each participant was randomly assigned to one of the seven experimental conditions. Each subject performed, in all, three tasks, in a predefined order. In each task, the participants looked for floating spheres with predefined locations. To create the illusion of body ownership, the mirror showed the avatar moving in visuo-motor synchrony with the participant.

The participants rated the difficulty of each task upon its completion. After the third task, they filled out a SAM form that used a nine-point Likert scale created for VR.

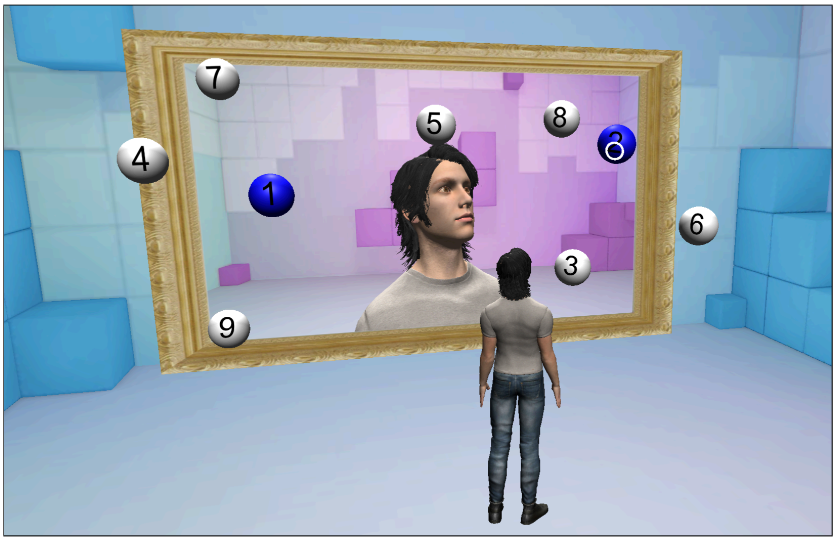


Figure 7.3: Task II shown from a third-person perspective – the participant is placed in front of a mirror in a virtual room, and numbered spheres are selected in turn via dwelling.

Debriefing

After completing the VR part of the study, participants completed a brief Web-based questionnaire, providing data on body ownership and prior VR experience. We defined their body ownership as the maximum value reported for these two questions on body ownership used by Bakakou and colleagues [12]):

- How much did you feel that the virtual body you saw when you looked down at yourself was your own body?

- How much did you feel that the virtual body you saw when you looked at yourself in the mirror was your own body?

Additionally, we asked the participants to guess the purpose of the study, for exclusion of anyone who suspected its true purpose (no one did). Each participant included the unique ID generated by the app for linking the app's study log with the post-experiment questionnaire. A week later, we revealed the purpose of the study to the student participants.

Results

We tested the responses for normality, and a Shapiro–Wilk normality test showed that body-ownership, valence, and arousal responses did not follow a normal distribution. Therefore, our reporting of results in the following section refers to non-parametric statistics. Kruskal–Wallis tests were used, with χ^2 test results reported in this section. For normally distributed data, F-scores from ANOVA tests are given.

Data

Most participants had only a little or some prior experience with VR; 21% had never tried it before, 55% had a few times, and 24% had considerable VR experience. We excluded two of the 207 participants for taking too long (>30 minutes), and 37 were removed from the sample for not experiencing body ownership (BO) with the virtual avatar (BO ≤ 2). Data from 168 participants remained.

The average time for completing the study was 297 s (SD = 197). No participants were excluded for guessing the purpose of the study; the vast majority responded with variations of "I don't know," with some speculating that the aim was to understand usability aspects of VR navigation. A little less than half of the time, 119 sec. (*SD* = 37), was spent looking in the mirror. The mean difficulty (from 1 to 9) for each respective task was 1.8, 2.9, and 4.4, showing participants' ease with completing the tasks.

Affect

The control group (without avatar manipulations) reported, on average, a score of 5.3 for valence and 3.3 for arousal. Figure 7.4 shows this group's differences in means for both valence and arousal; it is evident that the differences in valence between conditions are negligible (with means between 5.0 for frowning and 5.8 for smiling), while differences in arousal are more pronounced (means range from 3.0 for hunched posture to 4.8 for smiling).

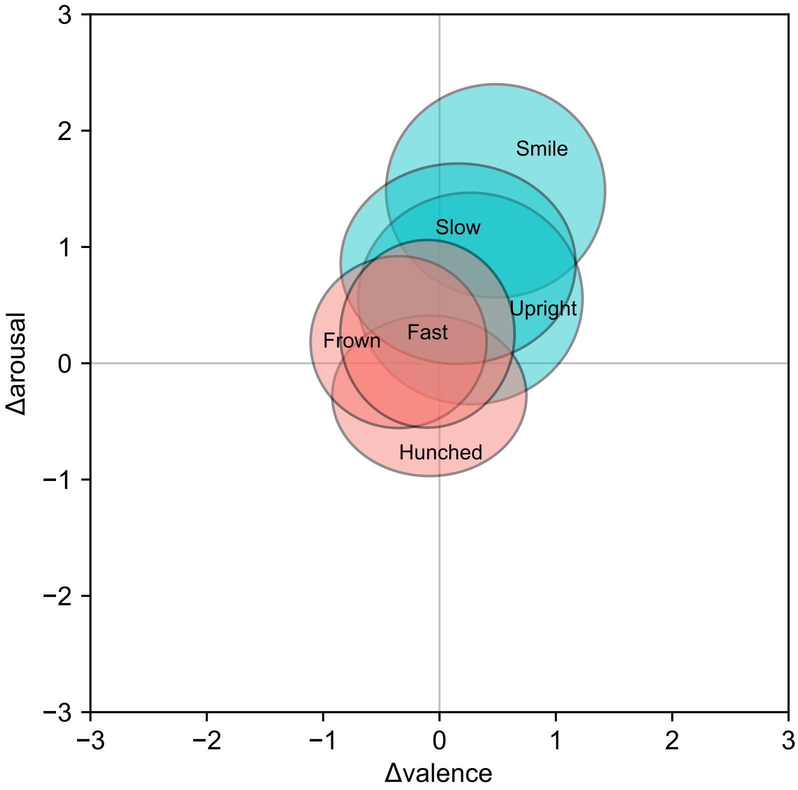


Figure 7.4: Difference in means from the six conditions, normalized in terms of the control condition. Blue ellipses represent conditions with hypotheses of high affect, and red ones represent conditions with hypotheses of low affect. The ellipses are fitted to represent 95% CIs.

We did find a significant effect of experimental condition on arousal: $\chi^2_6 = 14.61$, $p = .02$. A Bonferroni-adjusted *post hoc* Dunn's test showed that *smile* and *hunched* differed significantly, with $z = -3.32$, $p < 0.01$. We found no significant effect of experimental condition on valence: $\chi^2_6 = 2.75$, $p = .84$.

The condition *smile* showed itself to be the most effective manipulation for causing positive affect. Nonintuitively, its counterpart, *frown*, did not emerge as the most effective contributor to negative affect; rather, it seems that this and other manipulations hypothesized to induce negative affect had little to no effect.

We noticed that positive avatar manipulations (*smile*, *slow*, and *upright*) led to increased arousal relative to the control (see Figure 7.4). Comparing positive and negative manipulation reveals significance for arousal: $\chi^2_1 = 7.51,\ p < .01$.

Conversely, we observed negative manipulations (*frown*, *fast*, and *hunched*) to have limited effect, with *hunched* being the most effective. We noted that other researchers too have found positive affect easier to induce than negative affect; e.g., Schaefer et al. [188] collected affect measurements for 64 movie clips and consistently found stronger induction of positive affect than creation of negative affect with these stimuli.

Body ownership

With regard to body ownership – the degree to which sensory cues coalesce in the perception that a virtual body is "my body" [110] – we expected to find consistent levels across all the manipulations. Therefore, we checked for an effect of experimental condition on level of body ownership. Indeed, we found no such significance: $\chi^2_6 = 6.61,\ p = .36$.

The results from Study I showed that body ownership varied significantly with the time spent looking in the mirror: $\chi^2_8 = 20.8,\ p = .008$. That is, the level of reported belief in the avatar being the participants' own body rose with the amount of time looking at the avatar in the mirror in front of them. This shows the importance of mirrors in body-ownership illusions, and it confirms that cardboard VR systems, modest fidelity notwithstanding, can induce those illusions for most participants (as suggested in earlier work [153, 211]).

A Kruskal–Wallis test showed that valence varied significantly with body ownership: $\chi^2_7 = 21.3,\ p < .01$. A Spearman's $\rho = .29$ showed that body ownership correlates somewhat with valence; the same was not true for body ownership and arousal, $\rho = .08$ (see Figure 7.5).

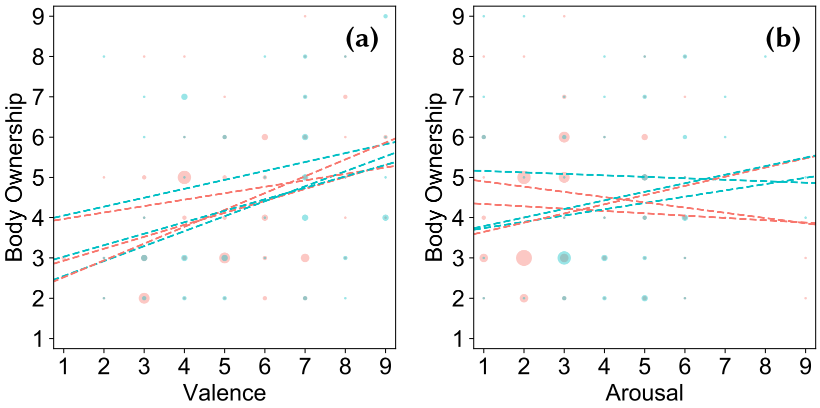


Figure 7.5: Plots showing the relationship of body ownership with (a) valence and (b) arousal, where the size of the dots indicate frequency and the dashed lines represent the trend lines for positive (blue) and negative (red) manipulations.

Performance

Coombes et al. [44] described a link between task performance (speed/precision) and affect; specifically, they reported that lower affect leads to higher speed and error rate in a motor task. Therefore, we were interested in whether task performance would vary as a function of affect in this study.

Participants completed the study most quickly in the *frown* condition ($M = 250$ sec.) and most slowly with *smile* (*M* = 353 sec.); however, the speed difference attributable to avatar manipulation was not significant: $F(6, 161) = .61,\ p = .72$.

For error rate, we used the mean error rate over the three tasks, calculated thus: for each task, we computed the Levenshtein distance, $lev(\alpha, \beta)$, between the optimal sequence of actions, α (e.g., Task II's $\alpha =$"123456789"), and the participant's sequence of actions, β. Error rate was, similarly to speed, not found to vary significantly with avatar manipulation: $F(6, 161) = 1.22,\ p = .30$.

Summary

The three avatar manipulations hypothesized to induce negative affect (*fast*, *frown*, and *hunched*) did produce a lower average affect score than each of the manipulations hypothesized to induce positive affect (*slow*, *smile*, and *upright*). This effect was statistically significant for arousal when the positive- and negative-condition groups were compared. Also, an omnibus Kruskal–Wallis test showed significance for arousal, and a Bonferroni-corrected *post hoc* test found the distributions to differ significantly for the pairing *hunched* and *smile*.

Body ownership significantly varied with valence; this was not the case for arousal. We did not find evidence suggesting that body ownership is influenced by avatar manipulation.

Neither did we find evidence for speed or error rate being influenced by the avatars' manipulation.

The results of Study I suggest that facial manipulations to avatars do alter affective responses, with posture manipulations having a similar but less pronounced effect. Manipulating speed does not seem to alter affective responses. Finally, our findings point to manipulations for positive affect as more efficient than those intended to induce negative affect.

Discussion

The mean score from the body-ownership reports (scale: 1–9) was 4.36 (*SD* = 1.91). While this is not unusually low, it does suggest that something in the circumstances of Study I limited the body-ownership scores. We believe the factors might include i) the commodity VR equipment with only 3 DoF; ii) the short study duration (five minutes, with two minutes of mirror time); and iii) the lack of internal control in outside-the-lab experimentation.

While outside-the-lab experimentation allows for rapid implementation of large-scale user studies at low cost, it imposes design constraints and creates practical limitations to the experiment. In particular, lengthy studies work poorly outside a laboratory setting, and adherence to the protocol (e.g., standing up and using a head strap) is hard to confirm. Additionally, it is rendered difficult to obtain reliable measurements of bodily aspects, such as posture and locomotion. We addressed these concerns by conducting a lab-based study, with higher internal validity and technical fidelity.

7.5 Study II

The purpose of Study II was to address unresolved questions from Study I about i) the interplay between body ownership and affect and ii) the effect of posture changes on affect.

Since valence varied with body ownership in Study I, we wanted Study II to cast more light on the connection between body ownership and affect, through measurements with high internal validity. While Study I showed only a non-significant difference between the *hunched* and *upright* conditions in terms of affect, that study was limited in a number of respects. With Study II, we hoped to ascertain whether a more elaborate setup and rigorous study protocol would reveal differences in affective responses between posture changes to the avatar.

Participants

We had 42 participants in this study, with an age range of 21–34 (*M* = 26.3, *SD* = 3.4). We recruited people to take part via an internal mailing list. The first participant was excluded on account of technical errors, so we report on analysis performed for 41 people (22 of whom were female). All participants were given a gift worth the equivalent of $20 USD for their time. Participants signed a consent form before the experiment commenced. No one who took part in Study I took part in Study II.

Apparatus

For this study, we used an HTC VIVE system (6 DoF) in combination with an OptiTrack motion-capture system for state-of-the art body-tracking. We employed a Unity scene similar to that developed for Study I, using a desktop PC (2.8 GHz Intel i7, 12 GB RAM, NVIDIA GTX 980), running Windows 10 Pro. Tracking was performed with Motive, using eight Prime 13 cameras at 120 Hz (the same frame rate used for HTC VIVE lighthouses), positioned in a semicircle (see Figure 7.6). We tracked the hands, elbows, feet, chest, and shoulders. Retroreflective markers were attached to the head-mounted display for SteamVR–Motive alignment. The chest and shoulders were not attached to the IK system but were tracked for later analysis.

Design

Study II employed a between-subjects design with avatar manipulation as the independent variable (there were three groups, with 14 people in each). Because the manipulations of avatar posture in Study

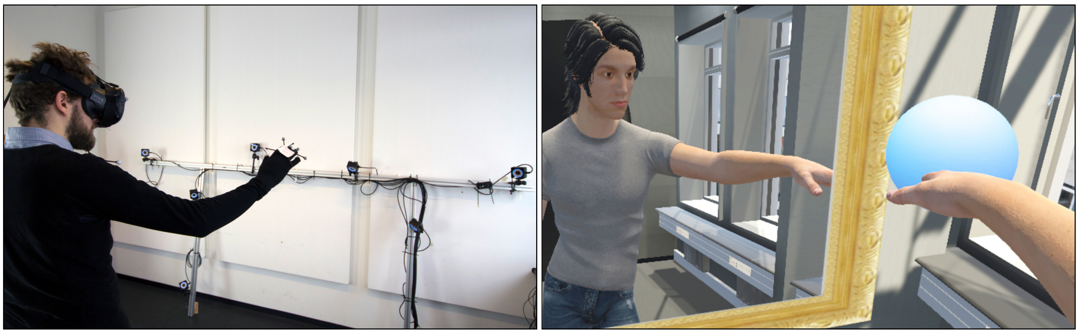


Figure 7.6: The lab study setup (top), and the view of the participant immersed in the VR world (bottom).

I exhibited no clear effect, we were interested in seeing whether this was due to the lack of technical fidelity (e.g., DoF, latency, and screen-resolution issues) and experimental control (e.g., issues with supervision, whether subject were standing up, adherence to procedure, and interruptions). On this basis, these three conditions were chosen for the independent variable: *hunched*, a control, and *upright*. The last was added halfway through the study without the experimenter's knowledge; this was done since the analysis in Study I showed little difference between the control and the *hunched*-condition group.

We chose PANAS results for the dependent variable of affect – this measure has consistently been shown to have high validity [48]. The instrument offers a broader conceptualization of affect than in Study I, in that it features items related to emotion and mood, not only core affect [57]. Thereby we hoped to gain a more nuanced picture of the influence of our manipulations while retaining a dimensional view of affect. The PANAS instrument was administered on a computer alongside the textual prompt "Please indicate to what extent you feel this way right now."

We chose a task similar to those in Study I, although this one was longer and required full-body movements instead of only head orientation and dwelling. Again, a mirror was present, in which the visuomotor synchronous, sex-matched avatar was visible. The virtual room where participants were immersed was a replica of the physical room in which the experiment took place (see Figure 7.6). The experiment was conducted by someone aware of neither the study's purpose nor of Study I.

In summary, in comparison to Study I, Study II had

- higher VR fidelity (HTC VIVE instead of cardboard),
- full-body tracking (8 × OptiTrack Prime 13),
- longer duration (17 min., as opposed to 5 min.),
- an extensive construct for affect (PANAS, not SAM),
- fewer conditions (three instead of seven), and
- fewer participants (42 instead of 207).

Procedure

After calibration in which participants' bodies were aligned with their virtual avatar, the study proper began. The study progressed with a series of floating 3D objects (spheres, cubes, and icosahedra) that disappeared when the participant tapped them with either hand. The experiment ended once the participant had tapped 200 objects. All objects were spawned in random locations between the participant and the mirror, such that the participant faced in the same direction with respect to the mirror throughout the study (see Figure 7.6). Thus, each subject was required to glance around, move about, and tap objects close to both the floor and the ceiling. This movement was reflected in the mirror placed in front of the participant. After finishing the VR task, participants filled out a computer-administered post-experiment questionnaire in which i) PANAS results, ii) body-ownership data, and iii) demographic details were collected.

Results

Data

Participants used, on average, 17.6 minutes (*SD* = 2.4) for the study, with 14 minutes spent looking in the mirror (*SD* = 1.7). The median score from the body-ownership reports (scale: 1–9) was 6 ($IQR = 4, 7$). Most subjects had little prior experience with VR: 38% had never tried it before, 38% had tried it a few times, and 24% had considerable VR experience.

Affect

The PANAS instrument covers two components, positive and negative affect, generating a score between 10 and 50 for each. These are considered two independent measures of affect. Figure 7.7 shows the negligible difference in affective responses between conditions. We were unable to find any significant effects of avatar manipulations on either PANAS component: for the positive one, $\chi^2_2 = .25,\ p = .88$; for the negative one, $\chi^2_2 = .09,\ p = .95$.

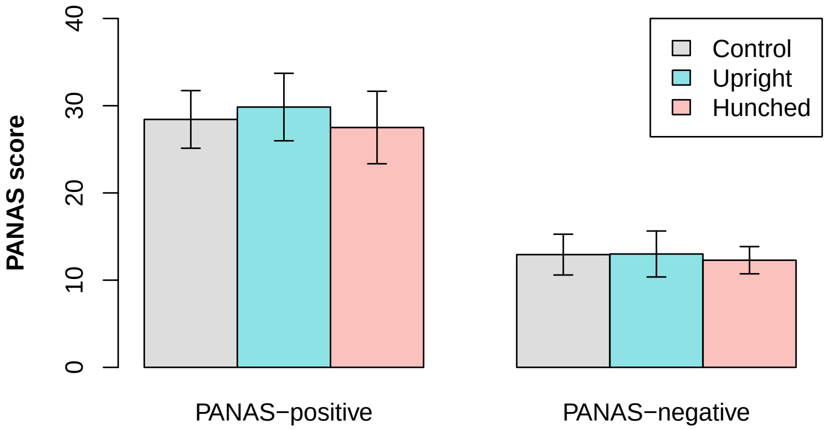


Figure 7.7: Mean PANAS scores. Error bars show 95% CIs.

Body ownership

We did not find an effect of avatar manipulation on body ownership: $\chi^2_2 = 2.93,\ p = .23$.

Since valence was found to vary with body ownership in Study I, we hoped to gain a better sense of the body-ownership–affect relationship with this study. We tested whether the PANAS components varied with body ownership and found a significant effect of body ownership on the positive one: $\chi^2_7 = 14.92,\ p = .04$. Significance was not found for the negative component: $\chi^2_7 = 8.33,\ p = .3$.

Inspired by Kilteni et al. [110], we treated body ownership as an ordered categorical value and in a combination with the numerical PANAS positive component we performed an ordinal logistic regression of body ownership. This yielded a fit value with a positive coefficient (the higher the positive-component response, the greater the likelihood of a high level of body ownership at the time). Figure 7.8 shows the estimated probabilities from the logistic fit $P($*Body ownership*$|$*Pos. PANAS*$)$. The probability of high ownership increases as PANAS values rise. For instance, the estimated probability of the ownership score being $\geq\ 7$ is $.76$ for a PANAS value of 42 but only $.05$ for a PANAS figure of 16. Ownership scores below 3 are most likely to be seen with PANAS scores under 16, with an estimated probability of .62.

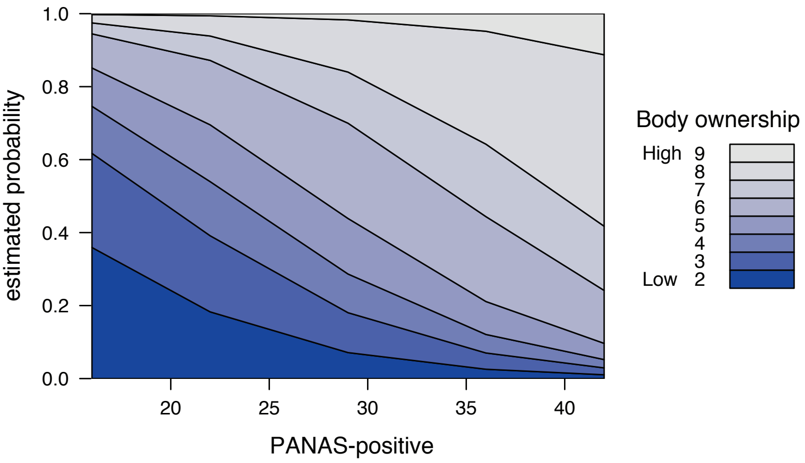


Figure 7.8: The probabilities estimated from the fitted values of an ordinal logistic regression of body ownership on the positive PANAS-response component.

Study I found valence to vary significantly with body ownership, while the foregoing analysis shows that the probability of high ownership increases with higher positive-affect scores. Together, these findings provide evidence that body ownership is an important factor in controlling the valence of affect in embodied VR experiences, particularly for inducing positive affect.

We did not find a link between negative affect and body ownership, and we speculate that the low variance between subjects for the negative PANAS component reflects this: negative-component scores varied within the range 10–26 (*SD* = 3.7), while positive-component ones ranged from 16 to 42 (*SD* = 6.4).

Summary

Study II did not reveal differences in affective responses between avatar conditions (control, *upright*, and *hunched*), just as Study I revealed no differences in affect attributable to full-body avatar manipulations. Hence, it is likely that posture changes do not have any influence on affect.

We found that body ownership varies significantly with positive affect. With higher affect scores, there was greater likelihood of high ownership values being reported: a positive-component PANAS score of 42 (out of 50) yields an estimated .76 probability of an ownership score of at least 7 (out of 9); a positive-component score of 16 yields an estimated .05 probability of an ownership score of 7 or above.

7.6 Discussion

Baños et al. [13] found that affective content has an impact on presence in virtual environments: they reported that in non-affective environments presence depended mainly on immersion, while the relationship proved more complex for environments with affective content. Jun et al. [105] too reported on presence and affect: with a body-ownership illusion they found presence to be positively correlated with valence. Specifically, the authors showed that owning an avatar with a happy face leads to higher presence estimates.

The study reported upon here expands our understanding of this interplay between virtual selves and perception. Firstly, we found that facial features are, in fact, efficient at influencing affect, while a weaker effect of this sort was found for upper-torso manipulations and movement speed. Secondly, we showed that positive affect is a good predictor of the likelihood of high body ownership. Our results suggest that valence is positively correlated with body ownership, while arousal is not. The results reveal, in addition, that positive affect is an important factor in body ownership. These findings are important for research considering affect and VR. For instance, mood-induction procedures in VR are likely to interfere with body ownership; study designs should be sensitive to this issue.

Our examination of the link between virtual bodies and users' affect was driven by a desire to inform design. We believe that the studies speak to this goal well – they may fruitfully inform the design of avatars and aid in creating a more solid foundation for subtle mood-induction techniques for virtual reality. For the domain of avatar design, our work suggests that some manipulations may influence affect, though not necessarily in great magnitude. As hinted at in prior work on VR [105], the most promising route for influencing affect seems to lie in manipulating facial expressions, especially smiling. It is far from clear that such manipulations work well for negative affect, however. Furthermore, indirect manipulations of affect seem more difficult to achieve; movement speed and posture produce unclear results. Further utility of the findings can be found in mood induction, as noted above. Existing mood-induction procedures for VR change the entire environment [14, 64, 180]. Our results show that less blatant manipulations might be feasible, although the interplay with body ownership suggests that their design might be difficult.

Study II found only the positive PANAS component, not the negative one, to vary with body ownership. While this may seem odd or counter-intuitive, the two components should be considered wholly independent. In a similar vein, Study I's manipulations intended to induce positive affect (*smile*, *slow*, and *upright*) were more effective than their negative counterparts (*frown*, *fast*, and *hunched*). Also, in Study I only valence (not arousal) was found to respond to the condition. Both studies suggest that avatar manipulations are effective primarily for influencing positive affect.

7.7 Conclusion

We examined whether it is possible to influence affective responses by altering avatars in virtual reality. Results from an outside-the-lab study with 207 participants showed that this indeed is feasible. Manipulations to the avatars' facial features proved effective in modulating valence responses.

Moreover, we tackled the seemingly harder question of how affect interacts with the illusion of owning a virtual body. Our results show that positive affect is of great importance for body ownership: in Study I, valence varied with body ownership, with positive affect being found to follow body ownership. Our analyses show that high positive-affect responses increase the probability of high body-ownership responses. Together, these findings contribute substantially to our understanding of how emotion information influences fundamental VR constructs.

8 Disfluent HCI

For reasons that will later become clear, this part of the thesis was never submitted for publication anywhere. I think, however, the undeniably clear results are rather interesting, and hence deserve to be printed.

As I was reading up on ways to manipulate cognitive load I came across the disfluency literature. Disfluency studies manipulate the ease with how humans obtain information; specifically, the hypothesis is that the harder it is to immediately comprehend a task, the higher is the probability of you doing well on that task. As this sounds counter intuitive, this experimental phenomena received quite some attention when series of studies showed how a math task with hard-to-read text yielded better scores on average, or that intentionally bad printing quality could increase memorability of words.

Many of the fantastic findings have later been found not to replicate, and studies with literally thousands of subjects have shown null-effects of disfluency on task performance. Two meta reviews [120, 145] summarize recent attempts at replication and conclude that the effects of disfluency on cognition are limited to non-existing.

Any rational empiricist would accept the critique of the studies conducted in a former troublesome era in experimental psychology. I thought, however, that there would be raison d'être for experimenting with disfluency for HCI. First of all, some evidence, even from recent studies suggest that disfluency might have an impact on memory. A HCI study also showed learning effects from increasing required effort [42]. Second, both the original disfluency study design and recent replication attempts only manipulated font readability in a binary fashion (i.e., normal vs: italicized, black vs. gray, or using a easy- or hard-to-read font), thereby leaving many ways and amounts one could manipulate information processing open, perhaps showing non-linear effects. Last, these studies mostly used tasks that are artificial (e.g., the Cognitive Reflection Test [69]).

With the work presented in this chapter I wanted to kill (at least) two birds with one stone: (i) improve the validity and integrity of the disfluency theory by experimenting with disfluency across actual tasks, modalities, with 'enough' subjects, and (ii) introduce disfluency for HCI, potentially disrupting the common belief of 'the more usability the better'.

The results, however, tell another story: disfluency systematically diminish performance, both found in decreasing quality of work, and increased completion times. I report on this phenomena in this chapter through three crowdsourced studies of applying disfluency to HCI tasks.

This chapter is based on a collaborative effort as described below.

Title
Disfluency in Human-Computer Interaction

Authors
Aske Mottelson and Kasper Hornbæk

Journal
-

DOI
-

What was the role of the PhD student in designing the study?
The PhD student was the first author of the paper,
and responsible for the design of the described studies.

How did the PhD student participate in data collection and/or development of theory?
The PhD student was responsible for study implementation,
execution, data collection, and theory development.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed to all parts of the manuscript.

Did the PhD student read and comment on the final manuscript?
Yes.

8.1 Introduction

Findings in cognitive psychology indicate that human reasoning involves two processing systems [60, 61, 106, 203]: System 1; a quick, intuitive, and effortless system, and System 2; a slow, analytical, and deliberate system. Cognitive strain can activate System 2 [6]; the slower, more deliberative, and more logical mode of thought, which allows for more elaborate logical problem solving.

Fluency is the ease with which people process information. Lack of fluency, or disfluency, introduced cognitive strain and has peculiar effects on cognition. Research shows that disfluency has effect on intelligence estimation [161], difficulty estimation [205], memory performance [53], and transcription [204], among others.

A common way to induce disfluency of prose is to present the text in a font that is difficult to read (e.g., [4, 6, 161, 162, 205]). A multitude of strategies for font manipulation have been tried; for instance italicizing [6, 161], random transformations [4], changing font-family [161, 205], changing color [6, 204], and reducing font-size [6, 162].

Thus, prior disfluency research has treated disfluency as binary (no disfluency/disfluency). This is problematic for several reasons: this assumes a linear relation between the disfluency and the potential benefits on cognition; it could very well be that the trade-off between disfluency and cognitive benefits would pose itself as non-linear. In addition, the binary treatment of disfluency makes comparisons across studies hard, for instance, if one study changed the color of a font, and another italicized. Lastly, past literature does not quantify how much previous methods make text harder to read.

In this research we treat disfluency as an interval: interpreting disfluency of 0.0 as readable, and 1.0 as unreadable (see Figure 8.1). To reach a satisfying control of disfluency as a continuous variable we apply varying amounts of blur, rotation, and white noise.

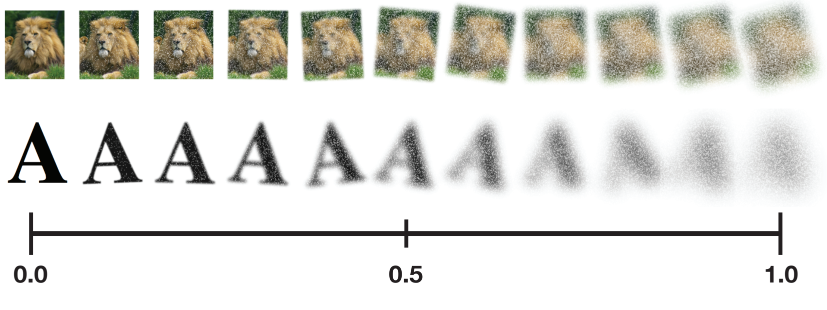
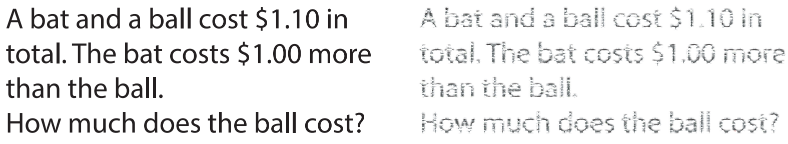


Figure 8.1: Scale of disfluency using images or text, going from fluent at 0, and disfluent at 1.

The study protocols, exclusion criteria, and the analysis methods were pre-registered before conducting any of the experiments: https://osf.io/9vksr/. That repository also contains materials for replication (dataset, code, etc.).

Related Work and Controversy of Replication Failures

In a seminal paper Alter and colleagues [6] showed how a disfluent font lead to significantly more accurate responses to questions from the Cognitive reflection test [69]. These questions are special in the sense, that they pose an immediate, but wrong, intuitive answer. To answer these correctly, one needs to overcome intuition and think a bit. One of the questions from this inventory is shown in Figure 8.2:



A bat and a ball cost $1.10 in
total. The bat costs $1.00 more
than the ball.
How much does the ball cost?

A bat and a ball cost $1.10 in
total. The bat costs $1.00 more
than the ball.
How much does the ball cost?

Figure 8.2: A question from the Cogitive reflection test: fluent version left, disfluent right.

The intuitive answer to the question is 10¢, even though the correct answer is in fact 5¢. The original paper showed that students who were exposed to the disfluent version of the question (see Figure 8.2, right) had a much higher correct response rate: "90% of participants in the fluent condition answered at least one question incorrectly, only 35% did so in the disfluent condition [...] $p < .001$".

Following this work, researchers have conducted experiments with many tasks with the potential of higher cognitive output following a disfluency condition. It has been shown that disfluency decreases intelligence estimation of authors [161], minimizes engagement, and increases difficulty estimation of tasks [205]. Disfluency can also lead to deeper processing, such as an increase in memory performance [53]. Similarly, Soboczenski et al. [204] showed how transcription errors can be minimized by reducing the presentation quality of the task.

Recently many of these studies have shown not to replicate ([120, 145]). Notably, the original study protocol using the Cognitive reflection test showed null effects for 16 independent studies in a replication attempt [145]. One of these studies crowdsourced more than 5,000 online participants (the original study had 40 participants).

Limitations of Previous Research and the Potential of Disfluency for HCI

While these replication attempts are overwhelmingly decisive, some strands of disfluency research still suggest some effects. Kuhl et al. [120] summarized recent empirical findings in disfluency research, and concluded that albeit most seminal studies fail to replicate, some studies still show significant effects of disfluency on cognition. Specifically recall, comprehension, and transfer performance seems to have some effect, under certain conditions [120]; for instance as noted for a specific group of participants, "[for] students with higher working memory capacity, a disfluent font was advantageous" [120].

So far in the disfluency research the tasks employed have either been completely artificial (e.g., [6]), or explored psychological constructs (e.g., memory, math comprehension, cognitive load), rather than exploring scenarios where this could be applied.

HCI research has a long tradition for blending psychological constructs with practical application design to improve user interfaces and human-computer interaction with these. We draw on research in this area, and apply methodologies for evaluating seminal computerized tasks using various amounts of disfluency. In addition to experimenting with disfluency in common UI contexts, a major limitation of previous research in this domain is the treatment of disfluency as a binary variable: we hypothesize that treating disfluency as a continuum will reveal relevant relations between the potential cognitive advantages and the relative amount of disfluency applied. Last, some previous work employ few homogeneous subjects, which we mitigate by recruiting participants online using crowdsourcing markets.

8.2 Experiment I: Image Labeling

A commonly crowdsourced task concerns image labeling for training machine learning systems to recognize future unlabelled images. Regardless of the data domain (traffic signs, types of whales, etc.), labelling tasks are quite similar: the worker has to explain what the contents of an image is. For this experiment we were interested to see how disfluency would influence label quality and performance. We hypothesized that while additional disfluency, and thus cognitive load, would increase labeling speed, it would also improve labeling accuracy.

Participants

For this experiment we invited participants from the US, with a HIT (Human Intelligence Task) acceptance rate of more than 90%. We had 121 participants from Amazon Mechanical Turk complete the task, who were reimbursed $2.50 USD. We removed 31 participants for either taking too long, being too fast, or having too many errors identified using the median absolute deviation (median $\pm 3 \times$ MAD). Of the resulting 90 participants, who were aged 20–69 ($M = 31$), 31 were female.

Apparatus

Participants completed the labeling task on a desktop computer using the Google Chrome browser, with a resolution of at least 900x700 px. We developed the web application in plain JS/HTML/CSS. Disfluency was dynamically administered with a JavaScript library developed for the purpose.

Design

We employed a within-subjects design, with disfluency (10 levels; 0.0–0.9) as the only independent variable. For the labelling task we looked for a data set containing single labelled images. A known "gold standard" would allow us to measure participants' performance under experimental manipulation. The CIFAR-10 data set [119] contains many images in 10 classes. Unfortunately, the images are only available in 32×32 px, making it unusable for this experiment. Inspired by CIFAR-10 we created our own data set with 250 images in 10 classes: bird, car, cat, deer, dog, frog, horse, monkey, ship, truck. We found the images on image-net.org using the categories provided. Images were then cropped and scaled to 300x300 pixels.

Procedure

Each participant labelled all 250 images (25 from each class), in random order using the interface shown in Figure 8.3. The order of the buttons was also randomized. The first 5 images were discarded as warm up rounds. For every trial we distorted the image with a disfluency level between 0.0 and 0.9. The disfluency consisted of blurring, rotating, and white noise. A 5 second break separated the trials.

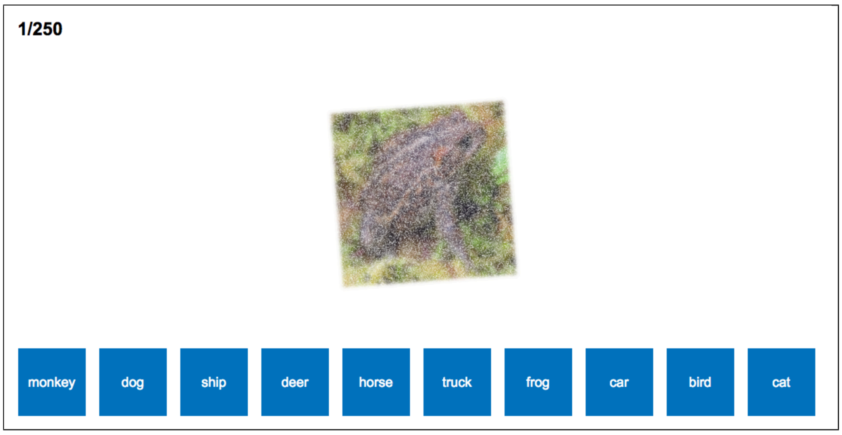
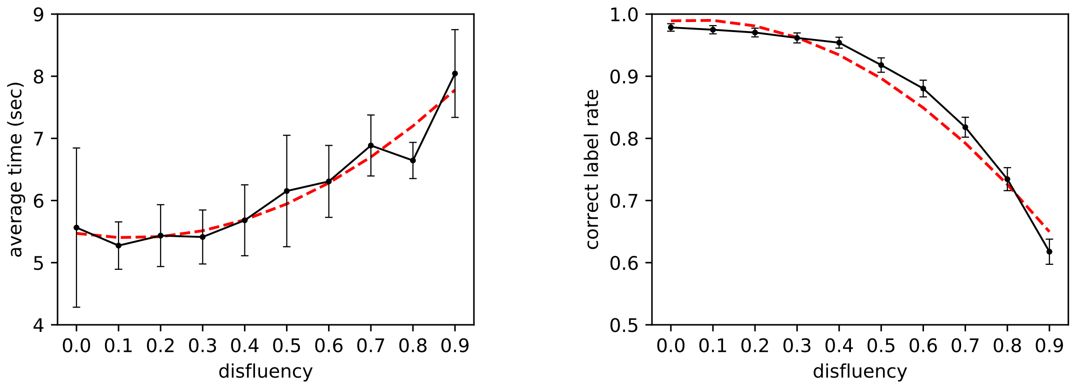


Figure 8.3: The participant labels an image, by clicking the correct button. Here is shown an image of a frog with disfluency level of .5.

Results

The results show that disfluency not only reduced speed of image classification, but also significantly reduced labeling quality (see Figure 8.4). The results from 22,050 image labels from participants, show that with increased disfluency comes higher response times, and reduced label accuracy.



(a)

(b)

Figure 8.4: The performance results for Experiment I: additional disfluency increases response time and decreases accuracy for image labeling. The red lines show 2nd degree polynomial fits. Error bars show .95 confidence intervals.

The data for both time and correct label rate suggest a polynomial relation to disfluency; we plotted the polynomial fits to Figure 8.4. For completion time this showed a good fit: $F(2, 937) = 22.5$, $p < 10^{-10}$. For correct label rate, the polynomial function was also a good fit: $F(2, 937) = 995.5$, $p < 10^{-16}$. Together this shows that the effect of disfluency on performance is close to quadratic; increasing completion time and decreasing accuracy.

8.3 Experiment II: Menu Navigation

Menu navigation is a seminal HCI task, both studied numerous times experimentally, and fundamental for deployed applications. As menu navigation requires thinking (e.g., should the user look for ‘United

Kingdom' in 'Geography' or 'European History'). We were therefore interested to see if we could enforce deeper thinking in menu navigation by applying disfluency, and thus minimize the error rate. In this experiment we employed a menu navigation task by Larson and Czerwinski [126], where participants find topics within hierarchical menus with different levels of disfluency.

Participants

We had 160 participants from Amazon Mechanical Turk partake in this experiment. We removed five for too high/low time, number of clicks, or lostness (using MAD). The remaining 155 participants (53 females) were aged 20–69 ($M = 31$). All participants came from the US, and had a HIT acceptance rate of more than 90%. We reimbursed participants $2 USD.

Apparatus

Participants completed the labeling task on a desktop computer using the Google Chrome browser, with a resolution of at least 900×700 px. We developed the web application in plain JS/HTML/CSS. Disfluency was dynamically administered with our JavaScript library.

Design

This experiment employed a within-subjects design, with disfluency as the only independent variable (10 levels; 0.0–0.9). Dependent variables were time, clicks, and lostness. Clicks refer to how many menu clicks were needed to find an item (minimum three required); lostness is a metric measuring how participants are "going around in circles" [126].

The menu used a 8×8×8 hierarchy. The menus and items were the same used in the original study; topics originated from Encarta, the discontinued digital encyclopedia. We applied a random level of disfluency to each trial. Figure 8.5 shows the interface used for this study.

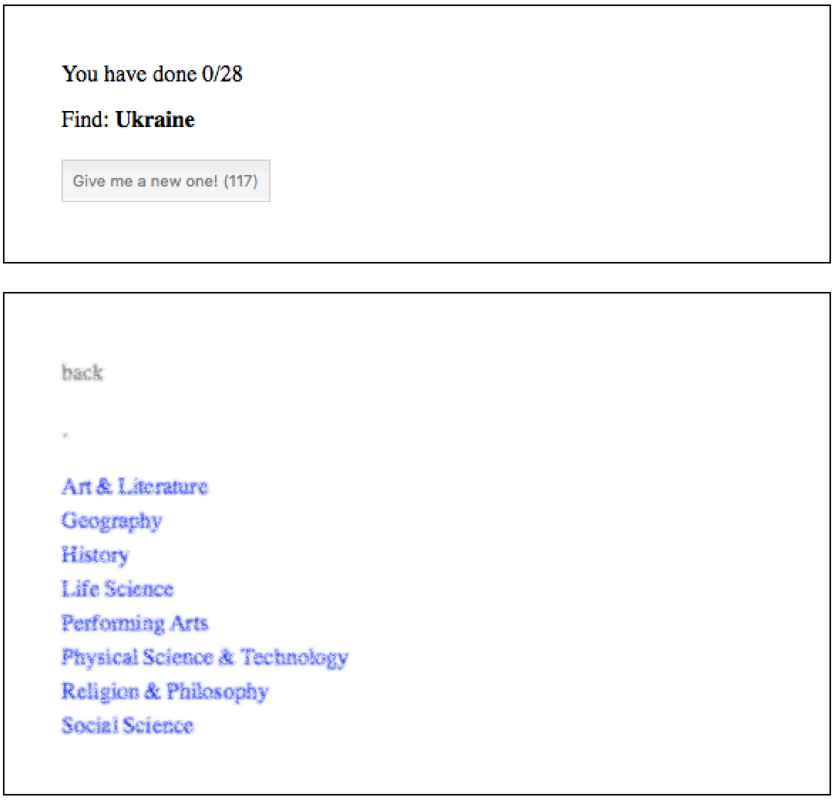
Procedure

Each participant conducted 28 trials of finding items in the menu (the first four of which were discarded as warm up rounds). Participants could request a new item to look for, if two minutes had passed without finding the item. A five second break separated the trials. The items for the user were selected randomly (from a pool of 512 items). The menus were presented alphabetically.

Results

Disfluency is thought to activate deeper reasoning and minimize intuitive reasoning. In this manner we hypothesized that some amount of disfluency could improve performance for participants finding items organized in a deep hierarchical menu. The analysis is done on a total of 3720 trials. Figure 8.6 shows the performance metrics for Experiment II.

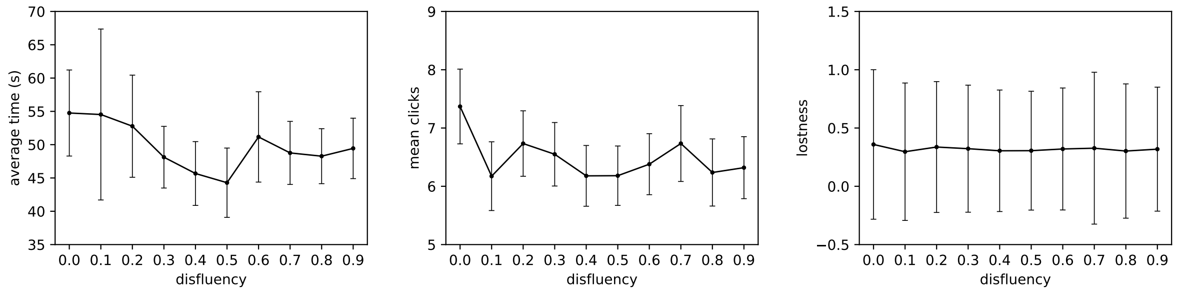
The relative high variance in the data (e.g., see the confidence intervals for lostness, Figure 8.6c), are due to the relative difference in the trials; finding 'Tango' in 'Performing Arts' → 'Dance' could be much easier than, for instance, finding 'Aesop' in 'Art & Literature' → 'Writers & Poets'.



You have done 0/28

Find: **Ukraine**

Figure 8.5: Menu navigation task used for Experiment II. In this trial the user has to find 'Ukraine' by going through Geography → Countries → Ukraine. A disfluency level of .5 is applied to the menu.



(a) **(b)** **(c)**

Figure 8.6: The performance results for Experiment II: additional disfluency does not seem to have an effect on menu navigation performance. Error bars show .95 confidence intervals.

Time

The time required to find the item was between 44–55 seconds (*SD* = 3.3). The overall trend does not show a significant decrease of reaction time on disfluency. If we, however, only look at the means for the disfluency levels of 0.0–0.5, we see a good linear fit: slope = -23.9, $r^2 = 0.94$, $p = 0.001$. This could indicate that disfluency does reduce reaction time in menu navigation, although not beyond 50% disfluency. This is in contrast to the hypothesis of slower but more deliberate activity as disfluency increases.

Clicks

The results show that the number of clicks required to find items in the $8 \times 8 \times 8$ menu range between 6.9–8.5 on average (with the perfect navigation requiring 3 clicks). The trend shows a slightly decreasing number of clicks as disfluency increases, slope $= -0.58$, however not significant: $r^2 = 0.21,\ p = .18$.

Lostness

Lostness refers to how much a participant "is going in circles" whilst navigating the structure. A lostness score higher than $.5$ is usually understood as "lost", while any score lower than $.4$ should be interpreted as "not lost". We observe that, regardless of disfluency level, participants were on average mostly "not lost", with average lostness scores between .3–.31 (*SD* = .02). The lostness is only slightly decreasing as disfluency increases, slope = $-.02$, but not significantly: $r^2 = 0.14,\ p = .29$.

Summary

This experiment showed that disfluency has little to no effect on participants' performance in a menu navigation task. We hypothesized that disfluency would increase reaction time, and reduce clicks and lostness. We found no effect on lostness and number of clicks; we did however see a reduction in reaction time, although not beyond 50% disfluency. Together, the results show that disfluency has limited effect on menu navigation, if any.

8.4 Experiment III: Spatial Recall

One of the areas where the effect of disfluency is continuously being discussed is around memory [120]: Lehmann et. al. [127] found that a disfluency condition had a significant influence on working memory capacity.

Participants

We had 311 participants from Amazon Mechanical Turk partake in this experiment. We removed 35 for taking too long/being too slow (using MAD). The remaining 276 participants (131 females) were aged 20–64 (*M* = 34). All participants came from the US, and had a HIT acceptance rate of more than 90%. We reimbursed the participants $1 USD.

Apparatus

Participants completed the spatial recall task on a desktop computer using the Google Chrome browser, with a resolution of at least 900×700 px. We developed the web application in plain JS/HTML/CSS. Disfluency was dynamically administered with our JavaScript library.

Design

This experiment employed a between-subjects design with disfluency as the only independent variable (10 levels, 0.0–0.9). The task is taken from Scarr et al. [187]: for each trial the participants find a target icon in a control panel like setup. Participants searched for the same six icons in a total of nine rounds. In each round the same six icons (for all participants) would appear as targets in a random order. We chose icons such that there would not be any overlap in columns or rows. As the original study, the first six rounds were deemed as a training phase, the last three as recall phase. We expected that the recall phase would be significantly faster than the training, because of a learning effect. Additionally we hypothesized that additional disfluency would make recall even faster, because of better memory of the locations of the icons. We used the Windows 7 Icons (instead of the Windows XP icons in the original study).

Procedure

After participants accepted the HIT on Mechanical Turk, they were redirected to our web app where the study would begin after answering a few demographics questions. The study progressed as a series of trials where a target icon was shown in a box on the right (see Figure 8.7). The user should then find and click the same icon among the 52 control panel icons, arranged in a $7 \times 7 + 3$ grid. Between every trial a large box appeared with the text "Click to begin next trial". Trials also continued if the user picked the wrong icon. These account for few of the trials as shown in Figure 8.8b.

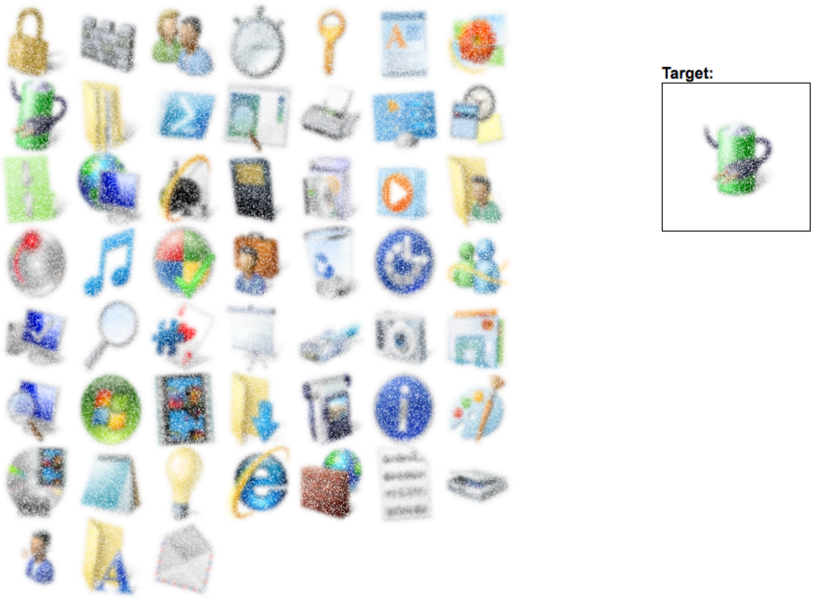


Figure 8.7: Control panel memory task: participants learn the location of six control panel items by recalling them six times. Three additional rounds of recalling the same six icons show how the memory of their locations improved recall time. Here the interface is shown with .5 disfluency, hypothesized to aid spatial memory.

Results

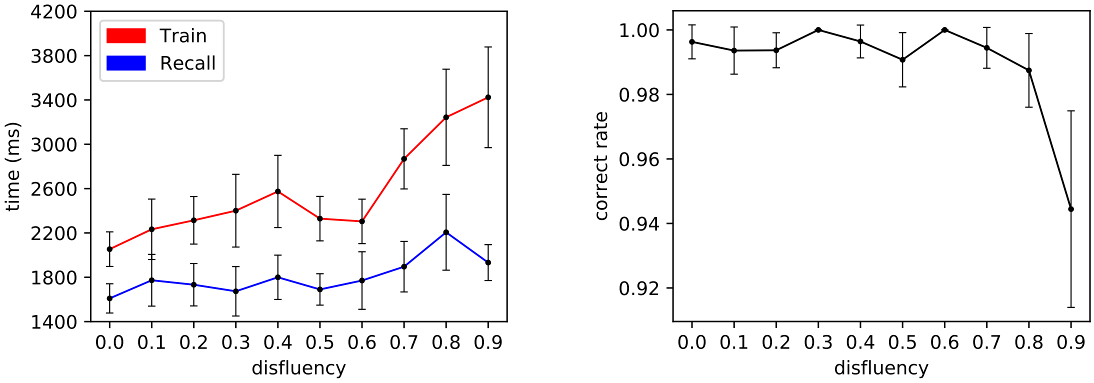
First, the data shows that accuracy only drops when applying more than .7 disfluency (see Figure 8.8a); given the simplicity of the task the vast majority found the correct icon.

It is evident that recall is significantly faster than training, as one would expect (see Figure 8.8b). This holds for any disfluency level; so regardless of additional cognitive load during training and recalling, it is much faster to recall than to learn the position of an icon in a grid.

It is also clear that more disfluency leads to both slower learning rates and recall rates. A linear fit shows a significant positive slope for training, slope $= 1311$, $r^2 = .76$, $p = .001$. This follows intuition: during learning, additional cognitive load causes prolonged learning times.

For recall we also observe an increase in response times as more disfluency is added, however not with as high a slope as for training. For recall we found a significant increase as disfluency increases: slope $= 420$, $r^2 = .56$, $p = .01$.

For spatial recalling, disfluency thus both decreases the training and recalling time; even when disfluency leads to increased training time, it does not lead to improved recall rates. For extreme disfluency levels it also reduces accuracy.



(a)

(b)

Figure 8.8: The performance results for Experiment III: (a) shows that disfluency causes additional search time both when learning and recalling icon locations; (b) shows how high levels of disfluency decreases the rate of finding the correct icon.

8.5 Conclusion

The motivation of doing this research was due to an observed deficiency in the disfluency literature. Prior to this study, disfluency had been treated binary, either present or not present. If the optimal disfluency would, for instance, reside around 60% disfluency, past literature would not have been able to detect this. Additionally, prior methods employed for making information disfluent (italicizing, coloring, or changing font family) are hard to verify that they really are harder to comprehend, which can be observed here for extreme values of disfluency in Experiment III. Also, we wanted to experiment with disfluency on other media than text (e.g., images, interfaces, interaction styles). Lastly, many tasks used in disfluency research are quite artificial and even if disfluency lead to improved comprehension, the employed tasks do not outline a path of how that can be utilized in reality.

While replication studies of disfluency studies to a large extent have shown null-results [145], this study is much clearer: two of three experiments using seminal HCI tasks showed significant *decrease* in performances (completion time and accuracy), while one study showed no difference. Even when disfluency leads to longer comprehension times, it still decreases output in terms of quality of labels, or diminished spatial memory.

We hypothesized that disfluency could be a valid UI ‘trick’ that could guide, for instance, crowdwork design or mission critical UI design, to force users to use their deliberate logical reasoning skills. However, the empirical data suggests that the application of disfluency as a design methodology is not relevant for HCI.

Part IV

Perspectives

9 What is Cognitive and Affective Interaction?

?

What Do We Mean by ‘Interaction’? An Analysis of 35 Years of CHI

KASPER HORNBÆK, Department of Computer Science, University of Copenhagen, Denmark
ASKE MOTTELSON, Department of Computer Science, University of Copenhagen, Denmark
JARROD KNIBBE, Department of Computer Science, University of Copenhagen, Denmark
DANIEL VOGEL, School of Computer Science, University of Waterloo, Canada

The notion of interaction is essential to human-computer interaction, yet rarely studied. We use quantitative and qualitative methods to investigate how this notion has been used across 35 years of proceedings from the ACM Conference on Human Factors in Computing (CHI). Using natural language processing, we extract 53,568 occurrences of the word ‘interaction’ across 4,604 papers. In these occurrences, we categorize 2,669 unique words that modify how ‘interaction’ is used in a sentence. We show that the use of ‘interaction’ is both increasing and diversifying, suggesting the importance of the notion, but also the difficulty in developing theory about interaction. Our findings show that styles of interaction are closely associated with changes in technology, and that modalities and characteristics of interaction are becoming more of a topic than specific devices or widgets. Interaction qualities, relating to structure, feel, effectiveness, and efficiency, are consistently prominent, and the quality of novelty is increasingly frequent. From this analysis, we identify open questions about interaction, including how to build knowledge across changing technologies, how to work towards a model of quality for interaction, and what the core of a science of interaction could be.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**;

Additional Key Words and Phrases: Interaction; History; Style; Quality

ACM Reference Format:
Kasper Hornbæk, Aske Mottelson, Jarrod Knibbe, and Daniel Vogel. 2018. What Do We Mean by ‘Interaction’? An Analysis of 35 Years of CHI. *ACM Trans. Comput.-Hum. Interact.* ?, ?, Article ? (June 2018), 28 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

This paper developed from a seemingly minor discussion about the history, progress, and current issues of the field of human-computer interaction (HCI). Specifically, we wondered when was interaction first labeled with ‘expressive’? Which interaction styles would be characterized this way? Are there other qualities related to this type of interaction? And is writing about expressive interaction still prominent, or is it deprecated?

These specific questions can be expanded to be about the general notion of ‘interaction’. Interaction is, of course, essential to the field of HCI, and it emphasizes how the interplay between

Authors’ addresses: Kasper Hornbæk, Department of Computer Science, University of Copenhagen, Njalsgade 128, Building 24, 5th floor, Copenhagen, DK-2300, Denmark, kash@di.ku.dk; Aske Mottelson, Department of Computer Science, University of Copenhagen, Njalsgade 128, Building 24, 5th floor, Copenhagen, DK-2300, Denmark, amot@di.ku.dk; Jarrod Knibbe, Department of Computer Science, University of Copenhagen, Njalsgade 128, Building 24, 5th floor, Copenhagen, DK-2300, Denmark, jarrod@di.ku.dk; Daniel Vogel, School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada, dvogel@uwaterloo.ca.

1073-0516/2018/6-ART? $15.00
https://doi.org/0000001.0000001

ACM Transactions on Computer-Human Interaction, Vol. ?, No. ?, Article ?. Publication date: June 2018.

A colleague of mine was struggling with writing a review paper about a specific subfield within HCI; the paper required reading many other papers, finding similarities, comparing methodologies, and summarizing the current state-of-the-art. Even just finding the set of papers that defined a subfield’s current progress seemed to cause some challenge.

Another colleague and I discussed how difficult it would be to automate this entire pipeline, or at least just parts of it. This would entail automatically collecting all HCI papers and defining each paper’s core themes, such that finding the set of papers within a theme would be straightforward. Ideally, given an abstract, the system would return with a list of DOIs of relevant papers, and a brief summary of these. We further envisioned a human-in-the-loop crowdsourced step, that would help summarize findings in a table across papers.

Admittedly this review maker never materialized. We did collect all CHI papers, tried various data science methods to summarize the field, and experimented with finding relevant papers given a topic, without convincing results. Instead, this data set has been used to write two papers that use the full CHI corpus to summarize parts of the field’s history, both of which are currently under review (about interaction at TOCHI [96]; shown left, and about readability at alt.chi [175]). The idea of this chapter is to use some of the findings presented in these two papers to zoom out a bit of this thesis’ presented work, and try to answer what constitutes interaction to be ‘cognitive’ or ‘affective’.

Throughout this thesis I have reported on empirical findings of how interaction with computers can both convey a user’s cognitive state, and also manipulate the same. I have called this ‘computer-cognition interfaces’ to imply how the interaction between man and machine can subtly work as a entrance to human mental processes, such as affect and cognition. In conclusion of this thesis I will try to distill a more formal definition of what such interaction is, based on analyses of the full text CHI corpus. This can both help understanding this HCI subfield, but may also further advance work in this area by pointing towards future strands of cognitive and affective interaction.

This chapter is based on a collaborative effort as described below.

Title
What Do We Mean by 'Interaction'? An Analysis of 35 Years of CHI

Authors
Kasper Hornbæk, Aske Mottelson, Jarrod Knibbe, and Daniel Vogel

Journal
ACM Transactions on Computer-Human Interaction

DOI
In review

What was the role of the PhD student in designing the study?
The PhD student was in charge of data acquisition and processing, and helped design the analysis process.

How did the PhD student participate in data collection and/or development of theory?
The PhD student developed the software, collected and labelled data, did data analyses, and generated figures.

Which part of the manuscript did the PhD student write or contribute to?
The PhD student contributed mainly to method and quantitative data analyses.

Did the PhD student read and comment on the final manuscript?
Yes.

9.1 Abstract

In the paper "What Do We Mean by 'Interaction'?" [96] we used quantitative and qualitative methods to investigate how the word 'interaction' has been used across 35 years of proceedings from the ACM Conference on Human Factors in Computing (CHI). We extracted 53,568 occurrences of the word 'interaction' across 4,604 papers. In these occurrences, we categorized 2,669 unique words that modify how 'interaction' is used in a sentence.

The original work led to us to divide interaction in six types: style, quality, concept, social, statistical, and other. This chapter summarizes specific findings around cognitive and affective interaction from this analysis. Both 'affective interaction' (mentioned across 284 papers, 369 sentences) and 'cognitive interaction' (mentioned across 243 papers, 368 sentences) are qualities of interaction. From the original study we expand the work with these two themes of interaction, and use the original data to distill a notion of interaction that is cognitive or affective.

9.2 Method

We are interested in understanding how the notion of cognitive and affective interaction has been used in HCI. We obtain data on this by taking the CHI proceedings as a representative, high-quality sample of work in HCI conducted over the past three-and-a-half decades. In those proceedings, we focus on the word interaction (as well as 'interactions'). Upon extracting all noun phrases including 'interaction', and manual coding these, we present findings related to those coded as cognitive or affective; find the full analysis in aforementioned paper [96].

Figure 9.1 summarizes our analysis approach, which combines natural language processing (NLP) techniques with manual classification. Next we describe each step of the process shown in Figure 9.1.

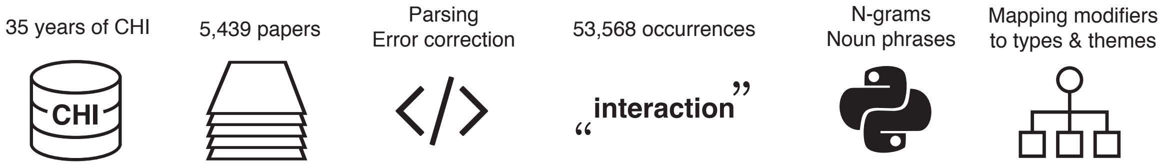


Figure 9.1: Overview of the analysis process. We analyze 35 years of CHI from 1981 to 2016, spanning 5,439 papers (4,604 of which mention 'interaction'). We then parse the text of all papers and automatically correct for OCR and other errors. In these papers we identify occurrences of 'interaction' and extract n-grams and noun phrases that contain the word interaction. This allows us to find key modifiers of interaction that we manually map to types and themes of the use of the word interaction. The analysis provided here relate only to 'cognitive' or 'affective' interaction.

Harvesting and Parsing Papers

We gathered the PDFs from all 5,349 papers or notes published at CHI 1981 through CHI 2016; note there was no CHI in 1984. We extracted the text from all PDFs containing readily available digital text using the Python PDF parser library, pdfminer[1]. Some papers had no available digital text (e.g., scanned

[1] `https://github.com/euske/pdfminer`

PDFs), poorly extracted text quality (e.g., because of partial OCR), or non-extractable text (e.g., corrupt documents). On those 1,230 papers we ran optical character recognition (OCR) using Adobe Acrobat Pro. For some older papers where OCR worked suboptimally, we manually transcribed 27 papers.

Using a mixture of string manipulations and regular expressions, we then removed paper meta data (i.e., title, author names and affiliations, conference name, conference theme, conference date and location, session title, keywords, and references) and other parts of the paper that do not contain topical content (e.g., acknowledgements). Additionally we stitched hyphenated word-parts caused by line breaks. The cleaned text representing each paper thus comprised the abstract, figure captions, tables, and the body.

Identifing Modifiers using N-grams and Noun Phrases

The complexity and number of extracted sentences make it infeasible to classify them manually. We therefore use automatic ways of identifying the uses of the word interaction. The aim of the analysis is to identify modifiers of 'interaction' that convey something about how that word is used.

To find modifiers, we build on work in automatic construction of index terms and phrases in text (e.g., [118]). Some work suggests that noun-phrases and n-grams might usefully be combined for this task [98], therefore we use both.

Noun-phrase extraction

The meaning of 'interaction' as it appears in a sentence is related to the noun phrases it occurs in. A noun phrase has a noun as its head, in our case 'interaction'. The head may then be modified in different ways, resulting in different uses of interaction. Relevant words may precede interaction, for instance, 'awkward social interaction' is both about social interaction and about that interaction being awkward. Interaction may also be followed by relevant modifiers, as in 'interaction is fluid'. To identify all of these, we automatically identify noun phrases from the sentences containing interaction. We construct grammar rules that segment the adjectives and nouns surrounding interaction, and ignore articles ('the' interaction), possessives ('their' interaction), and prepositions (interaction 'on').

We use the natural language toolkit for Python (NLTK[2]). First, a part-of-speech tagger labels every word with its grammatical function (e.g., verb, noun, preposition). While the tagger is not completely accurate, causing us to miss important modifiers or include undesired modifiers (e.g., occasional prepositions), we deal with this limitation in later steps. Having tagged all words in our corpus with their part of speech, we define a context-free-grammar rule to capture noun phrases. We assign all instances of 'interaction' to their own grammar rule (`INT`). This enables us to specify grammar rules that ensure that each noun phrase has only 'interaction' as its head. Our grammars look for arrangements of preceding and succeeding adjectives (`JJ*`), verbs (`VB*`), nouns (`NN*`), and conjunctions (`CC`, `TO`)[3]. The grammar rules are:

```
INT:        {interaction|interactions|interactive|interactional}

CONJ:       {<TO|CC>*}
```

[2] http://www.nltk.org/

[3] Grammar key: | = or, ? = optional, *= 0 or more, + = 1 or more, . = accept singular or plural forms.

```
PRINCIPLE:  {<CONJ>?<JJ.*>+}
            {<CONJ>?<NN.*>+}

INTNP:      {<PRINCIPLE>*<INT>+<VB.*>*<PRINCIPLE>*}
```

This finds 53,081 interaction noun phrases, we use every word of those phrases as a modifier (besides, of course 'interaction'), excluding common stop words and conjunctions. Note that each noun phrase can contain zero or more individual modifiers.

N-gram extraction

To mitigate against missed modifiers from the grammar rule, we compare the modifiers to non-stopwords found in both bigrams and trigrams containing interaction (i.e., those beginning and ending with interaction).

The idea here is to find modifiers of interaction using word-level n-grams that include interaction as the first or last word. We use 2- and 3-grams, expecting they would capture complex compound phrases involving 'interaction'. After stemming words and ignoring stopwords, the n-grams contributed an additional 157 modifiers missed by the noun-phrases, such as instances of 'external', 'implicit', and 'particular'.

Final set of modifiers

The combination of n-grams and noun-phrase extraction returns 6,449 modifiers. Of these modifiers, 3,780 appear in only one paper (e.g., 'prescription', 'aggressively-downloading', 'voodoo') and are removed from our keyword corpus. This leaves 2,669 modifiers. Each sentence containing the word interaction has 1.19 modifiers on average.

Mapping Modifiers to Uses of Interaction

We used a formal coding process to analyze these types of modifiers surrounding occurrences of 'interaction'. We did a series of four manual groupings of modifiers (each comprising 100 to 200 modifiers, and the sentences matched by each modifier), discussed the resulting groups, and iteratively refined a coding manual to define types of modifiers. Table 9.1 shows the resulting types.

Type	Definition	Example modifiers	Papers
Style	The form of input/output, the technology used in interaction, or the medium of interaction. Typically about the means by which interaction happens.	mobile interaction, bimanual interaction, cross-device interaction, 3D interaction, rhythmic interaction, pan interaction, zoom interaction, toy interaction	515
Quality	The valence of the interaction, capturing something that would be good or desirable in interaction (or, alternatively, bad). Typically about the experience of the person doing the interaction.	rich interaction, uncomfortable interaction, fluid interaction, playful interaction, natural interaction, meaningful interaction	371
Concept	Technical concepts in human-computer interaction that are not about style or quality; might occur in a textbook index. Typically about a thing or phenomenon and not something abstract or general	human-computer interaction, interaction designer, interaction modality, interaction technique	247
Social	Social interaction among people, or people and virtual characters and human-like agents. The interaction may be mediated by computers or not.	social interaction, face-2-face interaction, physician-patient interaction, cooperative interaction	137
Statistical	The use of interaction in a statistical sense.	three-way interaction, significant interaction, interaction effect	55
Other	Cases where the modifier only serves a grammatical function and does not add anything to the commonsense meaning of interaction.	different interaction(s), using interaction, possible interaction, many interaction(s), several interaction(s), specific interaction	1,343

Table 9.1: Six types of interaction modifiers. The *Papers* column gives the number of modifiers within each type.

To ensure reliability of the classification, the four authors of the journal paper classified all 2,669 modifiers. Each author coded half of the data set (1,338 modifiers), administered such that authors overlapped with 669 modifiers each.

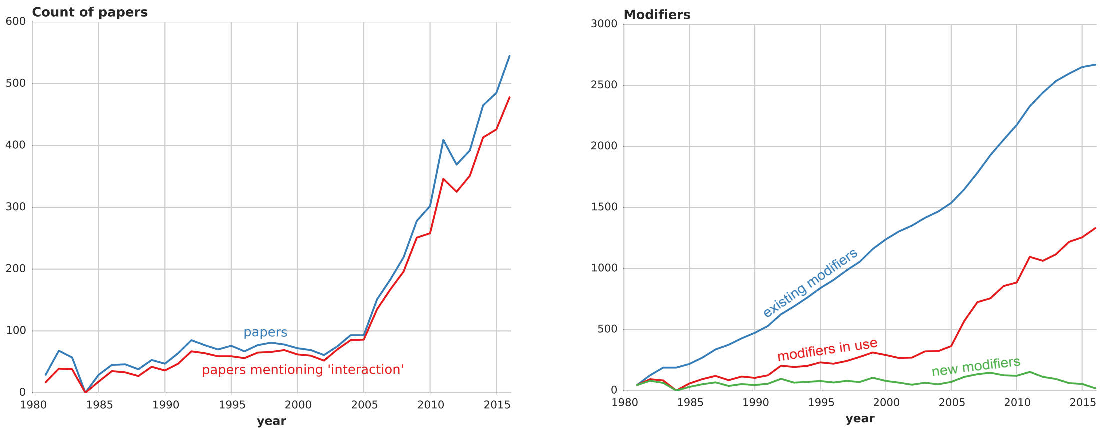
Fleiss' κ was computed to determine the inter-rater agreement between four raters. According to Landis and Koch there was a "substantial agreement" [123], $\kappa = .613$ (ranging from chance-level at 0 to perfect agreement at 1). All 705 disagreements (26.4%) were resolved through discussion.

Within these six interaction types, we further classified *themes* within each type of modifier. One author merged similar codes that had been handled separately, for instance because of differences in orthography (e.g., 'color' and 'colour') or because of abbreviations (e.g., 'f2f' or 'face-to-face'). Two authors then performed a thematic analysis of the two key uses of interaction, Style and Quality, spanning 515 and 371 modifiers respectively. We used the principles for thematic analysis described by Aronson [8] combined with the practices prescribed by affinity diagramming [19]. This chapter reports on the findings related to Quality only.

9.3 Findings

The use of ‘interaction’ is frequent and increasing over time. It represents many uses, of which Style and Quality account for the most prominent. In this section we provide a brief general overview, but then focus solely on affective and cognitive interaction, which represent an interaction quality.

The use of ‘Interaction’ at CHI



(a) **(b)**

Figure 9.2: Use of ‘interaction’ at CHI: (a) Absolute number of papers and of papers that mention ‘interaction’, (b) modifiers over years shown as the cumulative number of existing modifiers, the actual distinct modifiers in use in one year, and the absolute number of new modifiers introduced each year.

The word interaction has been used extensively over the course of CHI. Figure 9.2a shows that the absolute number of papers mentioning the word interaction increases, as does the total number of papers published. The ratio between those have changed too, so that the word interaction has increased from being mentioned in about 64% of the papers in the first five years of CHI to about 88% in 2016. Within papers, the word interaction is also used frequently. The median number of mentions is 6, increasing from 3-4 uses in the first five years of CHI, to approximately 9 over recent years.

Figure 9.2b shows the development in modifiers of ‘interaction’ over time. The figure shows that CHI authors use an increasing amount of modifiers. For instance, in 1981, only 44 distinct modifiers were used (e.g., ‘direct’, ‘menu-based’, ‘mode’). By 2016, the cumulative vocabulary of CHI had grown to 2,669 modifiers.

Figure 9.3 shows the distribution of the six types of modifiers over years, suggesting large-scale developments in how ‘interaction’ is used over the years. Modifiers related to Style appear to have increased in use over time more than other modifier types. Figure 9.3 also shows that Quality modifiers are a relatively constant fraction of the sentences containing interaction over time (around 10%). In contrast to Style, the increase in variants of Quality modifiers is smaller. But still, new qualities emerge (e.g., ‘extreme interaction‘ or ‘superior interaction’ from 2015).

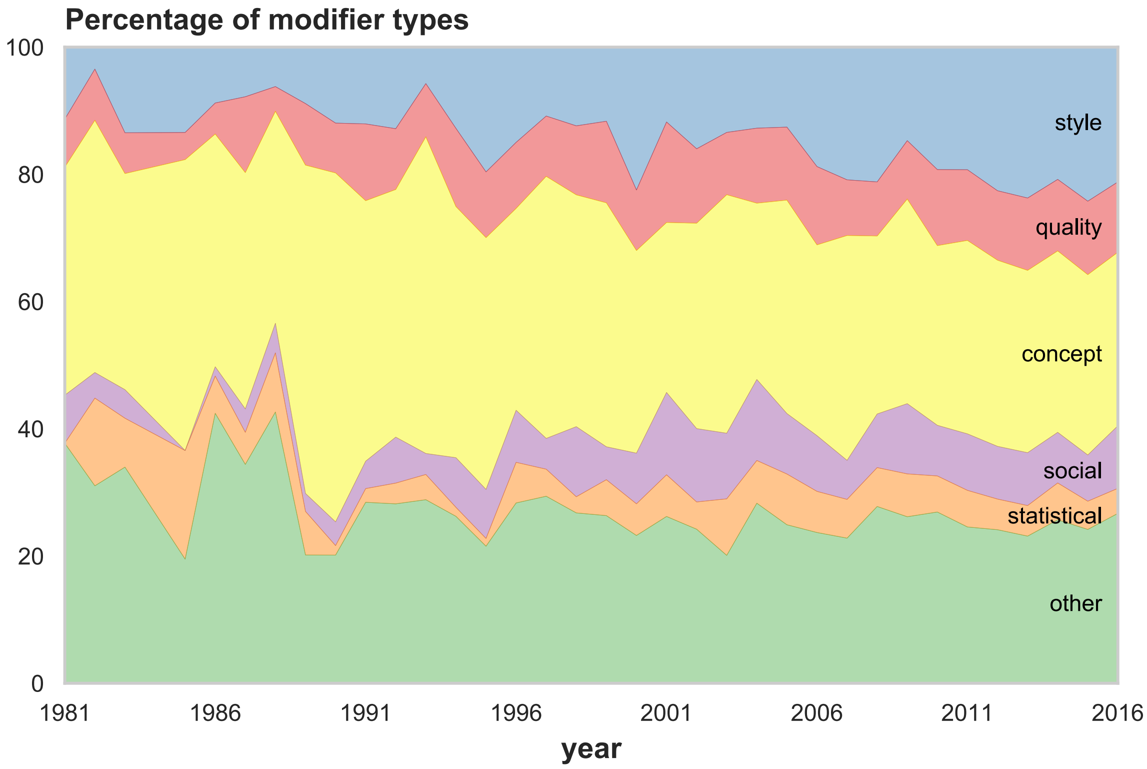


Figure 9.3: Percentage of modifier types used each year.

Qualities of Interaction

Quality modifiers capture valence, describing what is good or bad about the experience of interaction. Our examination of the 371 Quality modifiers suggests what kinds of things the CHI community values about interaction. We first define 12 themes of quality modifiers, listed in Table 9.2 with representative modifiers, and explore their development over over time, as shown in Figure 9.4. After, we discuss the specific Quality modifiers of particular interest to this thesis, namely Affective and Cognition.

Development of Quality Themes

The variability in vocabulary develops from early conferences to later ones: at CHI 1981 only the Quality modifiers 'logical', 'natural', 'unpleasant', and 'unpredictable' were used; at CHI 2016, 190 different modifiers of quality were used: the most frequent ones include 'rich', 'intuitive', 'traditional', 'seamless', 'spontaneous', and 'subtle'.

More generally, from CHI 2012–2016, through 2,256 papers, a total of 35 new quality modifiers emerged. In the same period the yearly use of distinct quality modifiers increased by 84, to 449. This suggests that new qualities establish themselves roughly with the same pace as old ones perish. Of all quality modifiers in our collective vocabulary, 150 did not occur in CHI 2012–2016. With the exception of 'conventional' which has been used in 13 different years at CHI, the other 149 modifiers have seldomly been used.

Theme	Examples	Papers
Feel	natural, complex, simple, intuitive, subtle	1,008
Comparison	new, novel, normal, common, traditional	919
Mode of Use	casual, continuous, frequent, explicit	517
Value Words	rich, active, difficult, easy	489
Resource Use	efficient, realtime, long, quick, rapid	396
Effectiveness	effective, appropriate, successful, precise	341
Affective	**positive, negative, intimate, emotional, intense, compelling, engaging, promising, uncomfortable, fun, intensive, comfortable, desirable, awkward, immersive, tedious, unexpected, unwanted**	**284**
Cognitive	**meaningful, interesting, expressive, realistic, challenging, serendipitous, imagined, purposeful, acceptable, ambiguity, meaning, acceptance, authentic, reflective, risky, secure, thoughtful, ambiguous**	**243**
Temporal	fluid, dynamic, sustained, concurrent	214
Adaptability	accessible, generic, adaptive, adaptable	59
Play	playful, serious, ludic, play-based	58
Look	aesthetic, sophisticated, nuanced	52

Table 9.2: The main themes of the 371 quality modifiers. The *Examples* column are most frequent modifiers across years, the *Papers* column is the number of papers in which modifiers from each theme occur.

Affective and Cognitive Interaction

The themes *Affective* and *Cognitive* both concern users' reactions and attitudes, and these themes sometimes overlap. *Affective* and *Cognitive* saw modest use in the first two decades of CHI (see Figure 9.4), and are still not the most prominent interaction qualities. Yet, they have seen increase in use since around year 2000 (also relative to other Quality themes), where they first experienced steady use. Together, affective and cognitive interaction account for between 10 and 20% percent of the use of interaction as a quality. The most used modifiers from within these categories are 'positive' and 'meaningful', respectively.

Affective Interaction

Affective is about users' emotional reactions to the interaction or, alternatively, ways of talking about interaction that may engender particular reactions. The modifier 'positive' is the most prominent example of an interaction quality in the Affective theme.

We find it interesting that positive words dominate *Affective* qualities. Not only is 'positive' (47) mentioned more frequently than 'negative' (26), but positive modifiers such as ('intimate', 'compelling', 'fun') are twice as frequent as negative ones (e.g., 'undesirable', 'awkward', 'frustrating'). The exception to this is 'uncomfortable', which is defined and discussed in one paper on the notion of uncomfortable interactions [16] and then used in 13 later papers.

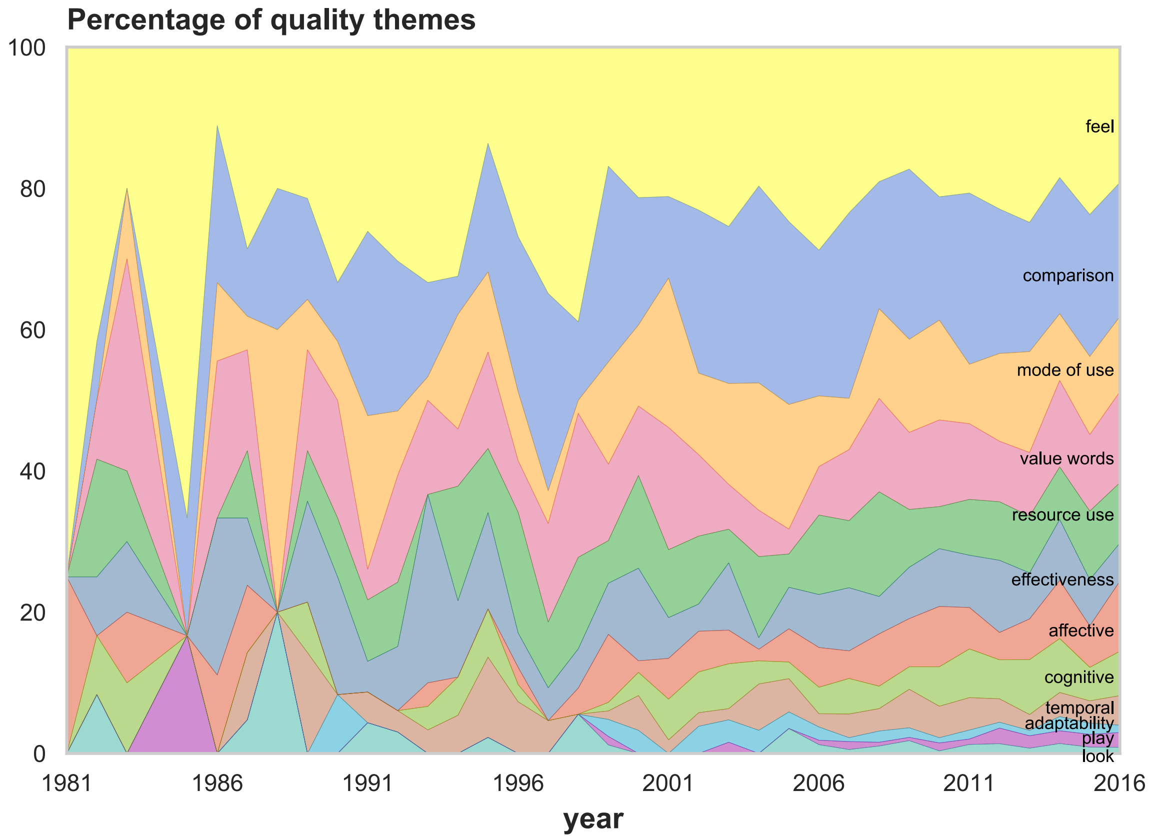


Figure 9.4: Twelve themes of quality over 35 years.

Cognitive Interaction

The theme *Cognitive* is about users' cognitive reactions to the interaction or, alternatively, ways of talking about interaction that may cause particular reactions.

The most used cognitive modifiers, 'meaningful', 'interesting', and 'expressive', saw limited use before the turn of the millennium (appearing in five papers combined), but are more frequent after (in 112 papers).

The cognitive modifiers are in particular qualitative, expressing the core valence about a particular interactional experience. This becomes clear, as formalizing definitions for these interaction ideals is remarkably difficult. While 'meaningful', 'expressive', and 'interesting' interactions are obviously all good properties, distilling exactly what makes interaction for instance 'expressive', is much harder to define. To this end, affective ideals appear somewhat less ambiguous, as seen by the lack of complexity by the most prominent affective modifiers ('positive', 'negative', 'intimate', 'emotional', and 'intense').

Co-occurences

To provide an analysis of cognitive and affective interaction, one level deeper, I here look at which words typically co-occur in sentences describing cognitive or affective interaction. I took each of the 2009 sentences containing a cognitive or affective interaction modifier. From these sentences I counted the occurences of all lemmatized words, exlcuding stopwords (such as 'the', 'me', or 'be'). The most frequently occurring words are therefore indicative of what domains, artifacts, or concepts CHI authors refer to when describing cognitive and affective interaction.

Co-word	Count	Co-word	Count	Co-word	Count	Co-word	Count
user	398	support	113	space	89	different	73
design	234	continuous	113	participant	88	effect	72
complex	208	number	112	technology	86	control	71
technique	187	display	111	dynamic	86	object	70
system	173	experience	110	gesture	86	people	68
social	160	information	110	appropriate	85	efficient	67
interface	151	casual	106	step	84	game	65
meaningful	145	effective	105	time	82	need	65
device	141	mobile	102	effort	82	approach	64
work	133	application	101	result	81	research	64
common	128	positive	98	model	81	version	64
study	124	touch	95	make	80		
task	114	expressive	91	new	73		

Table 9.3: Top 50 co-occuring words in sentences containing cognitive or affective interaction modifiers.

Table 9.3 shows the resulting top 50 co-occuring words from this analysis. The word ‘user’ tops the list, however the majority of the co-occuring words are interestingly not about human aspects (except ‘user’, ‘social’, ‘participant’, and ‘people’). Rather they describe interaction ideals, system artifacts, or activities.

Many frequently occurring co-words in cognitive or affective interaction sentences are about interaction qualities, such as ‘complex’ (cognitive), ‘meaningful’ (feel), ‘common’ (cognitive), ‘continuous’ (affective), ‘casual’ (mode of use), positive (affective), ‘expressive’ (comparison), or dynamic (affective). Many co-occurring words are also about system artifacts, as ‘system’, ‘interface’, ‘device’, ‘application’, ‘technology’, ‘model’, and ‘version’. Lastly, activities represent a large portion of the top 50 co-occuring words from affective/cognitive sentences: ‘work’, ‘study’, ‘task’, ‘support’, ‘time’, ‘game’, and ‘research’. This is indicative that cognitive and affective interaction is of interest to a large variety of tasks within HCI.

In summary, sentences describing affective and cognitive interaction are diverse, covering both system artifacts, ideals, and activities. Interestingly, these are to a relatively little degree about human aspects, but refer to more mainstream topics in HCI.

9.4 Discussion and Conclusion

It is clear that an analysis of an individual word and our relatively coarse analysis of the entire CHI paper history cannot paint a complete picture of the development and current state of cognitive or affective interaction, let alone human-computer interaction. Nevertheless, some general discussion points may be raised from the analysis.

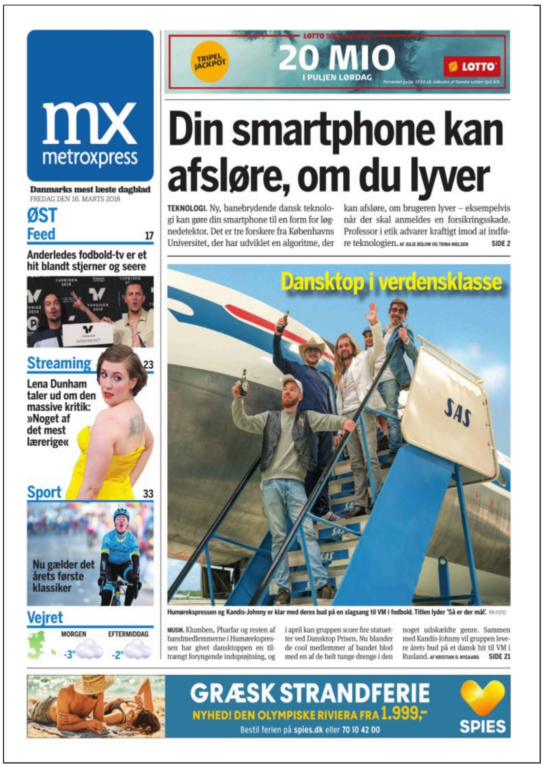
The notion of interaction is central to HCI, yet most accounts of it are conceptual. This chapter chartered its use over 35 years using quantitative and qualitative analyses of sentences containing the word interaction. The results show that 'interaction' is frequently used relative to other words, and that its use has increased over the years. The variation with which authors write about 'interaction', captured by modifiers, also develops. More than 2,600 different modifiers have been used, although new modifiers appear at a lower rate now than in 2006, when the number of papers at CHI grew.

Modifiers about the quality of interaction appear at a stable rate throughout the history of CHI, at between 8% and 11% over the years. Between quality modifiers, some changes can be observed. Since around year 2000, the themes Affective and Cognitive have experienced increased use.

While some interaction qualities can easier be explained (e.g., 'novel', 'positive, or 'aesthetic'), Cognitive ideals are mostly rather complex, experiential ideals. It seems however, that the modifiers from the Affective theme are somewhat simpler ideals, such as positive or intimate interaction. Despite this difference, the Affective and Cognitive themes are closely related, both describing users' reactions and attitudes towards interaction. These themes have both seen increasing use in the last years, with their combined use reaching more than 15% of all Quality modifiers in 2014 and 2016.

While the general trend at CHI shows an increasing use of the styles of interaction, an increase in descriptions of the experiential components of interaction, can also be observed. These themes are diversifying, with new modifiers emerging within the last few years such as 'effortless', 'coherent' and 'authentic'.

10 Ethics



mx metroxpress

Danmarks mest læste dagblad
FREDAG DEN 16. MARTS 2018

ØST

Din smartphone kan afsløre, om du lyver

TEKNOLOGI. Ny, banebrydende dansk teknologi kan gøre din smartphone til en form for løgnedetektor. Det er tre forskere fra Københavns Universitet, der har udviklet en algoritme, der kan afsløre, om brugeren lyver – eksempelvis når der skal anmeldes en forsikringsskade. Professor i etik advarer kraftigt imod at indføre teknologien. AF JULIE BÜLOW OG TRINA NIELSEN SIDE 2

Feed 17

Anderledes fodbold-tv er et hit blandt stjerner og seere

Streaming 23

Lena Dunham taler ud om den massive kritik: »Noget af det mest lærerige«

Sport 33

Nu gælder det årets første klassiker

Vejret

MORGEN	EFTERMIDDAG
-3°	-2°

Dansktop i verdensklasse

Humørekspressen og Kandis-Johnny er klar med deres bud på en slagsang til VM i fodbold. Titlen lyder 'Så er der mål'. PR-FOTO

MUSIK. Klumben, Pharfar og resten af bandmedlemmerne i Humørekspressen har givet dansktoppen en tiltrængt foryngende indsprøjtning, og i april kan gruppen score fire statuetter ved Dansktop Prisen. Nu blander de cool medlemmer af bandet blod med en af de helt tunge drenge i den noget udskældte genre. Sammen med Kandis-Johnny vil gruppen levere årets bud på et dansk hit til VM i Rusland. AF KRISTIAN D. NYGAARD. SIDE 21

Upon acceptance of the paper on truth estimation [154], I shared a press release with a couple of Danish and foreign news outlets. Metroxpress (a free Danish tabloid newspaper with more than 300,000 daily prints) brought a story March 16th 2018 on their front cover entitled "Your smartphone can reveal, whether you are lying" (shown left). Professor in Ethics at Aarhus University and prior member of the Ethical Council of Denmark, Thomas Ploug, warned in this article against the applications of the research, and was cited for saying that the "culture of trust could break down" with the deployment of this type of technology.

Besides being criticized by a professor in ethics on the front page of the most read newspaper in Denmark, I believe some ethical stance is generally fruitful as a creator of new disruptive technology, that has shown to change the lives of people, for better or worse. Because, as the American historian Melvin Kranzberg states it: "Technology is neither good nor bad; nor is it neutral". While the truth estimation received the most media attention of the work within this thesis, it is my opinion that studies involving mental processing in general should be ethically scrutinized. In this chapter I attempt at doing so.

The approach which I have chosen here is neither to unequivocally defend or reject the design and implementation of interactive technology that interfaces with mental processes, but instead discuss different takes on best practises. To do so, I present some critical reflections of my work in the light of (i) a recent blog post from the ACM Future of Computing Academy [43], and (ii) the ACM Code of Ethics and Professional Conduct [3].

10.1 Impacts, also the Negative

In a public appeal to computing researchers, the ACM Future of Computing Academy stresses that "The computing research community needs to work much harder to address the downsides of our innovations" [43]:

> *There clearly is a massive gap between the real-world impacts of computing research and the positivity with which we in the computing community tend to view our work. We believe that this gap represents a serious and embarrassing intellectual lapse. The scale of this lapse is truly tremendous: it is analogous to the medical community only writing about the benefits of a given treatment and completely ignoring the side effects, no matter how serious they are* [43].

The suggestion put forward in this blog post is a change to the peer-reviewing process: "Peer reviewers should require that papers and proposals rigorously consider all reasonable broader impacts, both positive and negative". The authors suggest that researchers and computing practitioners should consider also the possible negative side effects of their new technological advances. Authors should be expected to address each identified negative outcome of their new technology, making an argument of how the positive outcomes outweigh the negative, and how to mitigate negative outcomes.

The intention of the appeal is for computing researchers to list and discuss the disadvantages of technological advances within their publications or grant proposals, such that it can become a open part of the peer-reviewing process. Here, I will instead exercise this ethics task post publication (in addition to some of the ethical reflections presented throughout the individual chapters).

Negative Impacts

As the individual chapters contained in this thesis have plenty of examples of how they can positively influence HCI research and practice, this section will solely focus on the possible negative impacts of the work. Each negative implication is backed by actionable suggestions for mitigating unethical use of the research.

Negative Impacts of Sensing Mental Processes

A fair open question is whether interactive systems should be aware of users' mental states at all. Conversely one could argue that if the system knows your emotional state, or had the ability of knowing it, and yet does nothing about it; it would be like passing a crying person without interfering. Here I list some of the problematic views of interfacing mental processes using HCI.

Respecting privacy

Software should typically function without the need of users' personal information. As many details of interaction convey intimate details of the user, respecting privacy is of even higher concern. It is not hard to imagine a situation where a user unknowingly is passing on interactional information that discloses e.g. gender, age, handedness, or more delicate information about the user's thoughts or emotions, while being under the assumption of anonymity.

Section 1.6 of the ACM Code of Ethics and Professional Conduct states that "Only the minimum amount of personal information necessary should be collected in a system" [3]. From this follows that any sensing algorithm should not store or operate on data that not directly follows from its stated purpose.

I believe there are few examples of interactive systems that could not function without sensing-led classification, and the utilization of this class of predictive algorithms should generally be on an opt-in basis, for improved usability, for instance. In that regards, it also follows that any user prediction should be disclosed to the user. This obviously also entails that any sensitive information should not be disclosed to 3rd parties.

Implicit sensing without awareness

The work on sensing presented in this thesis makes inferences on users cognitive or affective states based on explicit interaction. That is, data generated directly based on the user's aware interaction, such as swiping details, acceleration, or the inter-tap details of number entry.

Many sensing approaches utilize data from implicit interaction, such as battery state, signal strength, time of day, which arguably still derives from interaction (the battery level is for instance a function of device usage). This not only allows for sensing when the device is not activated (for instance idling in your pocket), but also allows for sensing without you awareness, or at times where privacy is of even higher concern. These sensing approaches can have legitimate purposes (e.g., monitoring walking or running), but requires additional ethical consideration. The ACM Code of Ethics suggests in section 1.3 to "Be honest and trustworthy" [3]; from this follows that sensing without awareness should be avoided, thus requiring pre-sensing consent from any user.

Marginalizing users

The modest history of machine learning-led user classification is abundant with examples of marginalization[1], from inferring crime risks from face analysis [242] to using unknowingly sexist AI recruiting tools [101]. Regardless of the accuracy of interaction-based sensing techniques, there is similarly a range of potential unethical use-cases of such technology.

A first blatant risk of marginalizing users comes from inevitable misclassifications; situations where the system identifies you as belonging to some class when that is in fact not the case. In the work presented with truth estimation we handle this issue by not disclosing negative classifications; that is, users are not told if the algorithms find it plausible that an interaction is untruthful, but instead report the absence of a truthful classification (the same happens when confidence is low for positive classifications). This alleviates some of the issues related to being misclassified as a liar, yet requires some communication for users to understand that absence of a truth classification does not equal a lie classification. The ACM Code of Ethics, advocates in section 1.4, to "be fair and take action not to discriminate" [3]. Sensing mental processes, and especially identifying users' affective states poses a critical ethical issue. This is the case if the purpose is to identify and discriminate specific vulnerable users, or monitoring users to identify peaks and lows of for instance valence levels; be it for advertisement or political coercion. While affect detection techniques can sensibly be utilized, for instance, for improved UX or self-monitoring, consent should always be given prior to processing, and a vast range of applications of affect should be avoided in general.

[1] See for instance `https://github.com/daviddao/awful-ai` for recent examples

Discrimination based on physiology

In a previous chapter I reported how positive affect was correlated with slower interactions. Interaction speed is obviously not only determined by our affective state, and many traits could easily influence the speed with which an interaction is performed (e.g., age, weight, or finger size).

For machine learning modeling, the feature engineer will look for patterns in data that show a correlation to the modeling domain at hand. It is therefore common to include features, not based on causation, but correlation to optimize the model's accuracy. While the features included in the work presented in this thesis are to some degree based on hypotheses, and links retrieved from affect literature, I have not carried out an analysis that determines how, for instance, physiology influence the model's accuracy.

I call upon both empirical research on common sensing approaches' sensitivity to user discrimination, and for thorough testing of classification pipelines before production. One obvious way to somewhat alleviate these concerns (which has been the methodology taken in this thesis), is to source the model's data from a large and heterogeneous user pool.

Negative Impacts of Influencing Mental Processes

If sensing mental processes poses ethical issues, influencing does it even more so. Commonly employed affect manipulation techniques, such as watching movie clips or affective imagery, are first of all mostly conducted in laboratories with strict ethical procedures, but are also explicit in their purpose. Velten, a very common procedure to manipulate affect, asks participants to relive an emotional story from their past [226], making it obvious to the participants what is happening. This is, however, not the case when manipulating cognition or affect using more subtle techniques, such as those based on computer interaction.

During 2012, Facebook conducted an emotion manipulation experiment on 689,003 users without their knowledge [117]. The content users were presented with was selected to manipulate their emotions, and to study subsequent emotional contagion (i.e., by posting behavior on the social media). The study showed that "longer-lasting moods (e.g., depression, happiness) can be transferred through networks" [117].

Facebook, and the two universities affiliated with the research (Cornell University and the University of California at San Francisco) faced strong criticism as a result of the study. Manipulating affect (without consent or knowledge) can have widespread consequences for the individuals ranging from decreased mood to depression. The authors responded to this critique, noting that "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." [117]. The Facebook study poses a potent example of the potentials of subtle affect manipulation, showing significant changes to the vocabulary use of the participants who unknowingly participated.

In general I believe it is objectionable to induce negative affect outside of a controlled laboratory. As the whereabouts or psychological vulnerability of the participants is unknown, I believe this practice should in general be avoided. This was also why we chose only to induce positive affect (and neutral) for the affect detection study [150].

Negative Impacts of Crowdsourcing

While the practice of crowdsourcing, including micro task markets such as Amazon Mechanical Turk (AMT) and Figure Eight, have shown to diversify participant recruiting, offering higher external validity to studies, it does pose some critical questions, especially about labor rights. Workers on platforms such as AMT have few rights or means to object decisions made by HIT creators: whoever creates the job decides the compensation, the maximum amount of time allowed, and the acceptance criteria. AMT is frequently used (around 100K users; 2K active users at any given time [55]), why some ethical critique of its workers rights have been discussed (e.g, [68]). It is usually needed to include what Kittur refers to as 'verifiable questions' [112] to verify the integrity of a worker. These could be outright easy questions that anyone should be able to answer correct, just to make sure that the worker is paying attention. A labor market where the employer dictates all the rules, with few to no means of employees to organize or object decisions is problematic. It undermines the minimum salary, and advocates precarious working conditions. The problematic relationship is underlined by Amazons wording, calling Mechanical Turk a "24x7 workforce", yet refer to employers as "requesters", and employees as "turkers" [47].

While the use of AMT remains problematic, I believe a few guidelines may help mitigate some of these issues. We have, first of all, never paid less that the nationwide minimum US wage ($7.25 per December 2018). Next, we have had our "exclusion tests" prior to the actual study participation, thus minimizing unnecessary time spent by participants who did not follow the rules. These tests have been meta questions about the study (only requiring reading the study introduction), rather than "common knowledge questions". Lastly, we have not rejected any workers once they have finalized a study; in cases where we have deemed the participation as unfit for the analysis, we have simply removed the data instead of reporting the worker.

10.2 Summarizing Suggestions

Then, how to avoid that the "culture of trust could break down"? In addition to exercising the task of imagining potential negative outcomes of computing research put forward by ACM Future of Computing Academy [43], and the general advice from ACM Code of Ethics and Professional Conduct [3], I have presented a few practical suggestions for mitigating ethical issues in computing research that relates to mental processes, as summarized below.

Opt-in. Applications should ask before interfacing mental processes.

Consent. Ask participants prior to partaking experiments.

Privacy. Be open about the data processed.

Applications. Most interactive systems will not need users' affective states to function properly.

Avoid black/white classifications. If a specific classification can be detrimental to users, consider reworking the discourse.

Avoid elicitation of negative affect, especially outside of the laboratory.

Pay crowdworkers. Avoid declining pay-out, instead consider reworking your recruitment design. Report the actual hourly wage in the participants section.

11 Discussion

I will here present a brief overview of the methodological and validity considerations related to the design and presentation of the content of the thesis. In conclusion, I will argue for the ideal of human-computer interfaces to mental functions.

11.1 Methodological Considerations

Overview

Table 11.1 shows an overview of the study types, designs, and analyses used throughout the papers presented in the thesis. A predominantly empirical and quantitative approach has been adopted; all four papers use experiments, and subsequent analysis via statistical, mathematical, or computational techniques. The work presented in Chapter 9, while predominantly quantitative, did employ some qualitative methods, such as affinity diagramming and and thematic analysis.

	Paper 1	**Paper 2**	**Paper 3**	**Paper 4**
Study				
Crowdsourcing	×	×		
Out-of-lab			×	×
Laboratory			×	×
Design				
Between-subjects	×		×	×
Within-subjects	×	×		
Analysis				
Machine learning	×	×		
Statistical hypothesis test			×	×
Objective	Model	Model	Method	Theory

Table 11.1: Methodology overview of the thesis' included papers.

Place of study and Recruitment

In Table 11.1 I have separated 'crowdsourcing' and 'out-of-lab', to make the distinction between between outsourced recruitment (e.g., micro markets), and conventional recruitment with unsupervised study participation.

An argument I have used at several occasions throughout the thesis is that conducting studies out of the laboratory can increase heterogeneity and power of studies, thus increasing external validity and generalizability. It has been a primary goal of the presented work to shift study participation out of the laboratory; the laboratory studies I have conducted, have mainly been to compare results from unsupervised studies.

Power

It has been a priority in the design of the studies to have adequate power to make the conclusions drawn from the data.

	Paper 1	**Paper 2**	**Paper 3**	**Paper 4**
Study I	*55*	*87*	*31*	*168*
Study II	*127*	*41*	*57*	*41*
Study III		*51*		

Table 11.2: Number of participants included in the final analysis for each study.

Table 11.2 shows the power of the individual studies presented in the thesis. An adequate amount of participants is an ongoing scholarly discussion, becoming ever more relevant with recent replication crisis in science.

As the temporal integrity of human affect and cognition can be hard to enforce in a within-subjects design, the studies presented in this thesis have mainly employed a between-subjects design. This both reduces the risk of the participant guessing the purpose, and the integrity of, for instance emotion elicitation. This however comes with a cost of participant recruitment, as the study will generally require a participant count coefficient equal to the number of conditions. Caine [29] surveyed studies at CHI, and found that 70% of CHI studies have less than 30 participants. For between-subjects experiments the mean total subject count is 26, considerably lower than what I have presented.

Validity

Shadish et al. [191] describe a validity taxonomy, helpful for inspecting threats to validity using four categories: *Internal Validity*, *External Validity*, *Construct Validity*, and *Statistical Conclusion Validity*. These categories describe ways to determine and analyze threats to the validity of experimentation.

Internal

Internal validity concerns whether findings can be attributed experimental manipulation of the independent variable [191]. Internal validity often comes as a trade-off with external validity.

Because the studies presented in this thesis were mainly conducted without the presence of any experimenter, and that the studies tap into the participants' regular daily lives, there are several internal validity threats that cannot be controlled. Events that occur simultaneously, or naturally occurring changes to the settings are not controlled. Specifically, results may vary if participants instance are on a train while conducting a crowdsourced study. Differences might also arise if a study is conducted during night, if a participant is intoxicated, or if the participants is not paying sufficient attention, to name a few.

Obvious threats to internal validity are handled, such as multiple participation, avoiding multiple or varied affect manipulation per participant, and requiring comparable equipment.

External

External validity concerns whether causal relations between measures holds over variations in persons, settings, treatments, and outcomes [191].

A primary concern for external validity is the extent to which findings generalize to other populations, beyond the recruited sample. While the participants available for studies using crowdsourcing systems are not in perfect alignment with the general population (slightly younger and lower household income on average [55]), they have been found to be more diverse, than samples typically recruited for research [140, 166, 181].

As Table11.2 shows, each paper presented 2-3 studies in an attempt to verify relations in variation to study design, context, settings, persons, etc.

Construct

Construct validity concerns making higher-order inferences based on what is claimed to be measured [191].

Many constructs are used when measuring mental processes, notably the the distinct emotions, affect and arousal, PANAS, and the use of Likert scales. The Self-Assessment Manikin [24] tries to mitigate some of this cross participant understanding of affect scales, yet it is no silver bullet to obtaining labels of affect.

Both studies in Paper 1 and Paper 2 showed an RBF kernel SVM as superior in classifying mental aspects. This could be a testament to the construct validity issues related to labeling user activity around mental functions; the analyses showed different clusters (which gaussian functions often model effectively), even for same labelled data, which could indicate different interpretations of the same affect constructs.

Statistical Conclusion

Statistical conclusion validity refers to whether the suspected cause and effect covary [191].

The most fundamental threat to statistical conclusion validity is a low powered study (not considering misuse of data analysis, p-hacking, etc). I have tried to mitigate this by recruiting more participants than what is usually seen at HCI venues (Paper 4 for instance, to my knowledge, employed more participants than any single VR experiment previously). It is however difficult to guarantee that statistical conclusion validity is maintained. Tools like power analysis may help give an indication of subject counts needed. However, for small effect sizes this usually renders participant counts much higher than what is practically possible. The studies in this thesis took a pragmatic midway; recruiting as many as possible and aiming for at least 20 subjects per condition.

11.2 Computer-Cognition Interfaces

Throughout this thesis I have referred to the ideal of computer-cognition interfaces, broadly understood as the capability of interactive systems to understand and influence humans' mental processes.

An essential question arises when designing interactive systems that interface with mental processes: *why* advocate for this ideal? As noted in the ethical discussion of such technology, many interactive systems should not enforce, let alone should entail, such interactions. I believe, however, there are multiple reasons to why this is both an appealing quality for many interactive systems, and an important leap for human-computer interaction research.

Benefits of sensing mental processes

Giving interactive systems the capabilities of understanding users' mental processes, enables a range of possibilities ranging from improved UX to clinical applications, and as described in this thesis, enhanced veracity and thus trust of digital communication.

While many systems already employ machine learning models to generate content based on individual browsing history, country of origin, time of day, etc.; this is largely unused for the fundamentals of the UX of the systems. Sensing mental functions could open up for personalized notifications, flows, or menu layout, focused features, based on the user's current mood, stress level, or alertness. In addition, one could foresee the importance of sensing mental functions for humanoid robotics, that adapt and personalize their behavior in correspondence to their human peers.

Benefits of influencing mental processes

Arguably more pressing than questioning the practice of *sensing* mental functions, is arguing for *influencing* mental functions.

The fundamental understanding of human mental processes often requires researchers to manipulate certain mental functions within their subjects. If we are to understand how, for instance, happiness influences creativity, a researcher would need a happy person to begin with. Computer systems that employ interfaces to human mental functions allow researchers to employ more subtle, and less cumbersome study designs, in turn allowing fascinating findings linked to human cognition.

Inducing certain cognitive aspects of users could also be beneficial for certain tasks. It is easy to imagine how a digital systems that increased focus, and decreased the urge to procrastinate would be compelling, for instance. This also has an application for safety-critical systems, where human lives are potentially at stake, such as systems deployed for the hospital or aviation sectors.

There is consensus about the dominating factor affective aspects have on human life, influencing decisions, behavior, creativity, and other most fundamental parts of everyday life [31]. Affective computing as a research paradigm was initiated in 1997, yet, nowadays computer systems generally do not process affect, let alone the multitude of mental processes that are increasingly conducted sensing research on. I believe there are fundamental work left to be explored, besides perfecting existing directions which is evidently already happening. Many mental aspects of humans are open for digital interfacing, such as decision- or judgement-making, attention (and not diversion), imagination, sense-making, and memory, to name a few.

12 Conclusion

Research on user sensing is increasingly focusing on cognitive domains. Where sensing techniques have previously primarily targeted physical activity and device position, these domains have in recent years been extended to users' complex mental functions such as depression, stress, and emotions. Conversely, interactive systems have shown capabilities in altering mental aspects of users, and increasingly so with the proliferation of immersive technology such as virtual reality.

Through four papers I have shown advances to models, methods, applications, and theory for mediating human thinking in digital systems. Paper 1 and Paper 2 showed how these complex mental processes can be sensed with encouraging accuracy using only the details of touch interaction, disregarding actual data input.

Paper 3 showed the feasibility of conducting unsupervised VR experiments, paving the way for increased power and diversity in studies of complex VR phenomena, including body ownership and immersion. Paper 4 adopted this methodology, and showed a link between visuo-synchronous virtual avatars, and human affect. I have presented evidence for a link between positive affect and body ownership, showing that humans exhibiting more positive affect, may easier accept a first person avatar as their body.

Additionally, I have provided some perspective: a chapter bridging quantitative and qualitative analyses of the contents of all CHI papers tried distilling a notion of what cognitive and affective interaction might entail.

Together, this thesis has contributed to the ideal of computer-cognition interfaces, that aspire towards interactions between humans and computers that process thoughts and emotions. I have shown advances to this ideal through the chapters about affect and truth estimation using mobile interaction, crowdsourced VR studies, affective avatar manipulations, and disfluent UIs.

Bibliography

[1] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, Matthew Kay, Julie A. Kientz, Geri Gay, and Tanzeem Choudhury. Cognitive Rhythms: Unobtrusive and Continuous Sensing of Alertness Using a Mobile Phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 178–189, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: 10.1145/2971648.2971712.

[2] Johnny Accot and Shumin Zhai. Beyond Fitts' Law: Models for Trajectory-based HCI Tasks. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, pages 295–302, Atlanta, Georgia, USA. ACM, 1997. ISBN: 0-89791-802-9. DOI: 10.1145/258549.258760.

[3] ACM. ACM Code of Ethics and Professional Conduct. https://ethics.acm.org/, 2018. [Online; accessed 01-December-2018].

[4] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468, 2008. ISSN: 0036-8075. DOI: 10.1126/science.1160379.

[5] Fahd Albinali, Stephen Intille, William Haskell, and Mary Rosenberger. Using Wearable Activity Type Detection to Improve Physical Activity Energy Expenditure Estimation. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 311–320, Copenhagen, Denmark. ACM, 2010. ISBN: 978-1-60558-843-8. DOI: 10.1145/1864349.1864396.

[6] Adam L. Alter, Daniel M. Oppenheimer, Nicholas Epley, and Rebecca N. Eyre. Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of experimental Psychology*, 136(4):569–576, November 2007. ISSN: 0096-3445. DOI: 10.1037/0096-3445.136.4.569.

[7] Luigi Anolli and Rita Ciceri. The Voice of Deception: Vocal Strategies of Naive and Able Liars. en. *Journal of Nonverbal Behavior*, 21(4):259–284, December 1997. ISSN: 0191-5886, 1573-3653. DOI: 10.1023/A:1024916214403.

[8] Jodi Aronson. A pragmatic view of thematic analysis. *The qualitative report*, 2(1):1–3, 1995.

[9] Kevin W. Arthur, Kellogg S. Booth, and Colin Ware. Evaluating 3D Task Performance for Fish Tank Virtual Worlds. *ACM Transactions on Information Systems*, 11(3):239–265, July 1993. ISSN: 1046-8188.

[10] Cedric Bach and Dominique L Scapin. Comparing inspections and user testing for the evaluation of virtual environments. *International Journal of Human-Computer Interaction*, 26(8):786–824, 2010.

[11] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. Detecting Drinking Episodes in Young Adults Using Smartphone-based Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):5:1–5:36, June 2017. ISSN: 2474-9567. DOI: 10.1145/3090051.

[12] Domna Banakou, Raphaela Groten, and Mel Slater. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *PNAS*, 110(31):12846–12851, 2013. ISSN: 0027-8424. DOI: `10.1073/pnas.1306779110`.

[13] R. M. Baños, C. Botella, M. Alcañiz, V. Liaño, B. Guerrero, and B. Rey. Immersion and emotion: Their impact on the sense of presence. *CyberPsychology & Behavior*, 7(6):734–741, 2004. DOI: `10.1089/cpb.2004.7.734`.

[14] Rosa María Baños, Víctor Liaño, Cristina Botella, Mariano Alcañiz, Belén Guerrero, and Beatriz Rey. *Changing induced moods via virtual reality*. In Springer, Berlin, 2006, pages 7–15. ISBN: 978-3-540-34293-9. DOI: `10.1007/11755494_3`.

[15] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):37:1–37:20, September 2017. ISSN: 2474-9567. DOI: `10.1145/3130902`.

[16] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. Uncomfortable Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2005–2014, Austin, Texas, USA. ACM, 2012. ISBN: 978-1-4503-1015-4. DOI: `10.1145/2207676.2208347`.

[17] Joanna Bergstrom-Lehtovirta, Aske Mottelson, Andreea-Anamaria Muresan, and Kasper Hornbæk. Tool Extension in Human–Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, UK. ACM, To appear in 2019.

[18] Damien Besancenot, Delphine Dubart, and Radu Vranceanu. The value of lies in an ultimatum game with imperfect information. *xxx*, 93:239–247, 2013. ISSN: 0167-2681. DOI: `http://doi.org/10.1016/j.jebo.2013.03.029`.

[19] Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[20] Kristopher J. Blom, Jorge Arroyo-Palacios, and Mel Slater. The effects of rotating the self out of the body in the full virtual body ownership illusion. *Perception*, 43(4):275–294, 2014. DOI: `10.1068/p7618`.

[21] Benoît Bolmont. *III: Role and Influence of Moods Including Anxiety on Motor Control*. In *Causes, Role, and Influence of Mood States*. Nova Biomedical Books, 2005, pages 57–75. ISBN: 9791594542502.

[22] John Bolton, Mike Lambert, Denis Lirette, and Ben Unsworth. PaperDude: A Virtual Reality Cycling Exergame. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 475–478, Toronto, Ontario, Canada. ACM, 2014. ISBN: 978-1-4503-2474-8.

[23] C. F. Bond and B. M. DePaulo. Accuracy of deception judgments. *Pers Soc Psychol Rev*, 10(3):214–234, 2006.

[24] M. M. Bradley and P. J. Lang. Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy & Experimental Psychiatry*, 25:49–59, 1994. ISSN: 0005-7916. DOI: `10.1016/0005-7916(94)90063-9`.

[25] Agata Brajdic and Robert Harle. Walk Detection and Step Counting on Unconstrained Smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiqui-*

tous Computing, UbiComp '13, pages 225–234, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493449.

[26] Barry Brown, Stuart Reeves, and Scott Sherwood. Into the Wild: Challenges and Opportunities for Field Trial Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1657–1666, Vancouver, BC, Canada. ACM, 2011. ISBN: 978-1-4503-0228-9.

[27] Michael Buettner, Richa Prasad, Matthai Philipose, and David Wetherall. Recognizing Daily Activities with RFID-based Sensors. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 51–60, Orlando, Florida, USA. ACM, 2009. ISBN: 978-1-60558-431-7. DOI: 10.1145/1620545.1620553.

[28] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's Mechanical Turk a new Source of Inexpensive, yet High-quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[29] Kelly Caine. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 981–992, San Jose, California, USA. ACM, 2016. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858498.

[30] Paul Cairns, Pratyush Pandab, and Christopher Power. The Influence of Emotion on Number Entry Errors. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2293–2296, Toronto, Ontario, Canada. ACM, 2014. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557065.

[31] Rafel A. Calvo, Sidney K. D'Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford Handbook of Affective Computing*. Oxford University Press, 1st edition, 2015. ISBN: 978-0199942237.

[32] Andrew Campbell and Tanzeem Choudhury. From Smart to Cognitive Phones. *IEEE Pervasive Computing*, 11(3):7–11, July 2012. ISSN: 1536-1268. DOI: 10.1109/MPRV.2012.41.

[33] Luca Canzian and Mirco Musolesi. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1293–1304, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2805845.

[34] Scott Carter, Jennifer Mankoff, and Jeffrey Heer. Momento: Support for Situated Ubicomp Experimentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 125–134, San Jose, California, USA. ACM, 2007. ISBN: 978-1-59593-593-9.

[35] Gayathri Chandrasekaran, Tam Vu, Alexander Varshavsky, Marco Gruteser, Richard P. Martin, Jie Yang, and Yingying Chen. Vehicular Speed Estimation Using Received Signal Strength from Mobile Phones. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 237–240, Copenhagen, Denmark. ACM, 2010. ISBN: 978-1-60558-843-8. DOI: 10.1145/1864349.1864386.

[36] Keng-hao Chang, Drew Fisher, John Canny, and Björn Hartmann. How's My Mood and Stress?: An Efficient Speech Analysis Library for Unobtrusive Monitoring on Mobile Phones. In *Proceedings of the 6th International Conference on Body Area Networks*, BodyNets '11, pages 71–77, Beijing, China. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011. ISBN: 978-1-936968-29-9. DOI: 10.4108/icst.bodynets.2011.247079.

[37] Olivier Chapuis, Renaud Blanch, and Michel Beaudouin-Lafon. Fitts' Law in the Wild: A Field Study of Aimed Movements. Technical report, CNRS-Université Paris Sud, December 2007. LRI Technical Repport Number 1480, Univ. Paris-Sud, 11 pages.

[38] Driss Choujaa and Naranker Dulay. Predicting Human Behaviour from Selected Mobile Phone Data Points. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 105–108, Copenhagen, Denmark. ACM, 2010. ISBN: 978-1-60558-843-8. DOI: `10.1145/1864349.1864368`.

[39] Keum San Chun, Sarnab Bhattacharya, and Edison Thomaz. Detecting Eating Episodes by Tracking Jawbone Movements with a Non-Contact Wearable Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):4:1–4:21, March 2018. ISSN: 2474-9567. DOI: `10.1145/3191736`.

[40] CIA. The World Factbook, 2016. `https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html`.

[41] J.A. Coan and J.J.B. Allen. *Handbook of Emotion Elicitation and Assessment*. Series in Affective Science. Oxford University Press, USA, 2007. ISBN: 9780195169157.

[42] Andy Cockburn, Per Ola Kristensson, Jason Alexander, and Shumin Zhai. Hard Lessons: Effort-inducing Interfaces Benefit Spatial Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1571–1580, San Jose, California, USA. ACM, 2007. ISBN: 978-1-59593-593-9. DOI: `10.1145/1240624.1240863`.

[43] ACM Future of Computing Academy. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. `https://acm-fca.org/2018/03/29/negativeimpacts/`, 2018. [Online; accessed 01-December-2018].

[44] Stephen A. Coombes, Christopher M. Janelle, and Aaron R. Duley. Emotion and Motor Control: Movement Attributes Following Affective Picture Processing. *Journal of Motor Behavior*, 37(6):425–436, 2005. DOI: `10.3200/JMBR.37.6.425-436`.

[45] National Research Council. *The Polygraph and Lie Detection*. The National Academies Press, January 2003. DOI: `10.17226/10420`.

[46] Céline Coutrix and Nadine Mandran. Identifying Emotions Expressed by Mobile Users Through 2D Surface and 3D Motion Gestures. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 311–320, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370265`.

[47] Geoff Cox and Allan MacLean. *Speaking Code - Coding as Aesthetic and Political Expression*. Software Studies. MIT Press, 2012.

[48] John R. Crawford and Julie D. Henry. The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43(3):245–265, 2004. DOI: `10.1348/0144665031752934`.

[49] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3):1–18, March 2013. DOI: `10.1371/journal.pone.0057410`.

[50] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *Public Library of Science ONE*, 8(3):1–18, March 2013.

[51] Charles Darwin. *The Expression of the Emotions in Man and Animals*. New York: Oxford University Press. (Original work published 1872), 3rd edition, 1998. ISBN: 9780899460048.

[52] L.C. De Silva and Suen Chun Hui. Real-time facial feature extraction and emotion recognition. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, 1310–1314 vol.3, December 2003. DOI: `10.1109/ICICS.2003.1292676`.

[53] C. Diemand-Yauman, D. M. Oppenheimer, and E. B. Vaughan. Fortune favors the bold (and the Italicized): effects of disfluency on educational outcomes. *Cognition*, 118(1):111–115, January 2011.

[54] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Muhlberger. The impact of perception and presence on emotional reactions: A review of research in virtual reality. *Frontiers in Psychology*, 6:26, 2015.

[55] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 135–143, Marina Del Rey, CA, USA. ACM, 2018. ISBN: 978-1-4503-5581-0. DOI: `10.1145/3159652.3159661`.

[56] Trinh Minh Tri Do and Daniel Gatica-Perez. Contextual Conditional Models for Smartphone-based Human Mobility Prediction. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 163–172, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370242`.

[57] Panteleimon Ekkekakis. *The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research*. Cambridge University Press, 2013. ISBN: 9781107011007.

[58] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*:169–200, 1992.

[59] Clayton Epp, Michael Lippold, and Regan L. Mandryk. Identifying Emotional States Using Keystroke Dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 715–724, Vancouver, BC, Canada. ACM, 2011. ISBN: 978-1-4503-0228-9. DOI: `10.1145/1978942.1979046`.

[60] Jonathan St.B.T. Evans. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459, 2003. ISSN: 1364-6613. DOI: `http://dx.doi.org/10.1016/j.tics.2003.08.012`.

[61] J.S.B.T. Evans and D.E. Over. *Rationality and Reasoning*. Essays in cognitive psychology. Psychology Press, 1996. ISBN: 9780863774386.

[62] University of Copenhagen Faculty of Science. General rules and guidelines for the PhD programme. `https://www.science.ku.dk/english/research/phd/student/filer/regelsaet/SCIENCE_regels_t_2015_FINAL.pdf`, 2016. [Online; accessed 01-December-2018].

[63] Sebastian Feese, Bert Arnrich, Gerhard Troster, Michael Burtscher, Bertolt Meyer, and Klaus Jonas. CoenoFire: Monitoring Performance Indicators of Firefighters in Real-world Missions Using Smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 83–92, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: `10.1145/2493432.2493450`.

[64] Anna Felnhofer, Oswald D. Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human–Computer Studies*, 82(C):48–56, October 2015. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2015.05.004.

[65] Raul Fernandez. *A Computational Model for the Automatic Recognition of Affect in Speech*. PhD thesis, Massachusetts Institute of Technology, 2004.

[66] Rebecca A. Ferrer, Emily G. Grenen, and Jennifer M. Taber. Effectiveness of Internet-Based Affect Induction Procedures: A Systematic Review and Meta-Analysis. *Emotion*, May 2015, 2015. DOI: 10.1037/emo0000035.

[67] Paul M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381–391, 1954.

[68] Karën Fort, Gilles Adda, and K. Bretonnel Cohen. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37:413–420, 2011.

[69] Shane Frederick. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42, December 2005. DOI: 10.1257/089533005775196732.

[70] Rui Fukui, Masahiko Watanabe, Tomoaki Gyota, Masamichi Shimosaka, and Tomomasa Sato. Hand Shape Classification with a Wrist Contour Sensor: Development of a Prototype Device. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 311–314, Beijing, China. ACM, 2011. ISBN: 978-1-4503-0630-0. DOI: 10.1145/2030112.2030154.

[71] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(4):31:1–31:30, December 2012. ISSN: 1073-0516. DOI: 10.1145/2395131.2395138.

[72] Azucena Garcia-Palacios, Hunter G. Hoffman, Albert Carlin, Thomas A. Furness, and Christina Botella. Virtual Reality in the Treatment of Spider Phobia: a Controlled Study. *Behaviour Research and Therapy*, 40(9):983–993, 2002. ISSN: 0005-7967.

[73] Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. Is the Web as Good as the Lab? Comparable Performance from Web and Lab in Cognitive/Perceptual Experiments. *Psychonomic Bulletin & Review*, 19(5):847–857, 2012.

[74] Francesca Gino and Scott S. Wiltermuth. Evil Genius? How Dishonesty Can Lead to Greater Creativity. en. *Psychological Science*, 25(4):973–981, April 2014. ISSN: 0956-7976. DOI: 10.1177/0956797614520714.

[75] Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data Collection in a Flat World: the Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[76] Sandy J. J. Gould, Anna L. Cox, and Duncan P. Brumby. Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(3):19:1–19:29, June 2016. ISSN: 1073-0516.

[77] Lilian de Greef, Mayank Goel, Min Joon Seo, Eric C. Larson, James W. Stout, James A. Taylor, and Shwetak N. Patel. Bilicam: Using Mobile Phones to Monitor Newborn Jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 331–342, Seattle, Washington. ACM, 2014. ISBN: 978-1-4503-2968-2. DOI: `10.1145/2632048.2632076`.

[78] H. P. Grice. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA, 1975.

[79] Leonard H. Grokop, Anthony Sarah, Chris Brunner, Vidya Narayanan, and Sanjiv Nanda. Activity and Device Position Recognition in Mobile Devices. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 591–592, Beijing, China. ACM, 2011. ISBN: 978-1-4503-0630-0. DOI: `10.1145/2030112.2030228`.

[80] Weixi Gu, Zheng Yang, Longfei Shangguan, Wei Sun, Kun Jin, and Yunhao Liu. Intelligent Sleep Stage Mining Service with Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 649–660, Seattle, Washington. ACM, 2014. ISBN: 978-1-4503-2968-2. DOI: `10.1145/2632048.2632084`.

[81] Weixi Gu, Yuxun Zhou, Zimu Zhou, Xi Liu, Han Zou, Pei Zhang, Costas J. Spanos, and Lin Zhang. SugarMate: Non-intrusive Blood Glucose Monitoring with Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):54:1–54:27, September 2017. ISSN: 2474-9567. DOI: `10.1145/3130919`.

[82] Yu Guan and Thomas Plötz. Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):11:1–11:28, June 2017. ISSN: 2474-9567. DOI: `10.1145/3090076`.

[83] Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel. ElectriSense: Single-point Sensing Using EMI for Electrical Event Detection and Classification in the Home. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 139–148, Copenhagen, Denmark. ACM, 2010. ISBN: 978-1-60558-843-8. DOI: `10.1145/1864349.1864375`.

[84] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3:1157–1182, March 2003. ISSN: 1532-4435.

[85] Nils Y. Hammerla and Thomas Plötz. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1041–1051, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: `10.1145/2750858.2807551`.

[86] Jeff Hancock, Jeremy Birnholtz, Natalya Bazarova, Jamie Guillory, Josh Perlin, and Barrett Amos. Butler Lies: Awareness, Deception and Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 517–526, Boston, MA, USA. ACM, 2009. ISBN: 978-1-60558-246-7. DOI: `10.1145/1518701.1518782`.

[87] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, 45(1):1–23, 2007. DOI: `10.1080/01638530701739181`.

[88] Tian Hao, Guoliang Xing, and Gang Zhou. RunBuddy: A Smartphone System for Running Rhythm Monitoring. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive*

and Ubiquitous Computing, UbiComp '15, pages 133–144, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: `10.1145/2750858.2804293`.

[89] Holger Harms, Oliver Amft, Gerhard Tröster, Mirjam Appert, Roland Müller, and Andreas Meyer-Heim. Wearable Therapist: Sensing Garments for Supporting Children Improve Posture. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 85–88, Orlando, Florida, USA. ACM, 2009. ISBN: 978-1-60558-431-7. DOI: `10.1145/1620545.1620558`.

[90] Jeffrey Heer and Michael Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212, Atlanta, Georgia, USA. ACM, 2010. ISBN: 978-1-60558-929-9.

[91] Niels Henze, Martin Pielot, Benjamin Poppinga, Torben Schinke, and Susanne Boll. My App is an Experiment: Experience from User Studies in Mobile App Stores. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3(4):71–91, October 2011. ISSN: 1942-390X. DOI: `10.4018/jmhci.2011100105`.

[92] Benjamin E. Hilbig and Corinna M. Hessler. What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, 49(2):263–266, 2013. ISSN: 0022-1031. DOI: `http://dx.doi.org/10.1016/j.jesp.2012.11.010`.

[93] Christian Holz and Patrick Baudisch. Understanding Touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2501–2510, Vancouver, BC, Canada. ACM, 2011. ISBN: 978-1-4503-0228-9. DOI: `10.1145/1978942.1979308`.

[94] Christian Holz and Edward J. Wang. Glabella: Continuously Sensing Blood Pressure Behavior Using an Unobtrusive Wearable Device. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):58:1–58:23, September 2017. ISSN: 2474-9567. DOI: `10.1145/3132024`.

[95] Kasper Hornbæk. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends in Human-Computer Interaction* , 5(4):299–373, June 2013. ISSN: 1551-3955.

[96] Kasper Hornbæk, Aske Mottelson, Jarrod Knibbe, and Dan Vogel. What Do We Mean by 'Interaction'? An Empirical Analysis of 35 Years of CHI. *ACM Transactions on Computer-Human Interaction.* TOCHI, In review.

[97] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. Is Once Enough?: On the Extent and Content of Replications in Human-computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3523–3532, Toronto, Ontario, Canada. ACM, 2014. ISBN: 978-1-4503-2473-1.

[98] Anette Hulth. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003. DOI: `10.3115/1119355.1119383`.

[99] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):13:1–13:27, March 2018. ISSN: 2474-9567. DOI: `10.1145/3191745`.

[100] Sozo Inoue, Naonori Ueda, Yasunobu Nohara, and Naoki Nakashima. Mobile Activity Recognition for a Whole Day: Recognizing Real Nursing Activities with Big Dataset. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1269–1280, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2807533.

[101] Technology Intelligence. Amazon scraps 'sexist AI' recruiting tool that showed bias against women. https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/, 2018. [Online; accessed 01-December-2018].

[102] A. M. Isen, K. A. Daubman, and G. P. Nowicki. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52(6):1122–1131, June 1987. ISSN: 0022-3514. DOI: 10.1037/0022-3514.52.6.1122.

[103] ISO. 9241-400:2007, Ergonomics of human–system interaction – Part 400: Principles and requirements for physical input devices, 2007. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=38896.

[104] Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the Fisher kernel method to detect remote protein homologies. In *In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158. AAAI Press, 1999.

[105] Joohee Jun, Myeongul Jung, So-Yeon Kim, and Kwanguk (Kenny) Kim. Full-body ownership illusion can change our emotion. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 601:1–601:11, Montreal QC, Canada. ACM, 2018. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174175.

[106] D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237–251, July 1973.

[107] Spencer Kaiser, Ashley Parks, Patrick Leopard, Charlie Albright, Jake Carlson, Mayank Goel, Damoun Nassehi, and Eric C. Larson. Design and Learnability of Vortex Whistles for Managing Chronic Lung Function via Smartphones. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 569–580, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: 10.1145/2971648.2971726.

[108] Hsin-Liu (Cindy) Kao, Bo-Jhang Ho, Allan C. Lin, and Hao-Hua Chu. Phone-based Gait Analysis to Detect Alcohol Usage. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 661–662, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370354.

[109] IftikharAhmed Khan, Willem-Paul Brinkman, and Robert Hierons. Towards estimating computer users' mood from interaction behaviour with keyboard and mouse. English. *Frontiers of Computer Science*, 7(6):943–954, 2013. ISSN: 2095-2228. DOI: 10.1007/s11704-013-2331-z.

[110] Konstantina Kilteni, Antonella Maselli, Konrad P. Kording, and Mel Slater. Over my Fake Body: Body Ownership Illusions for Studying the Multisensory Basis of Own-body Perception. *Frontiers in Human Neuroscience*, 9:141, 2015.

[111] Konstantina Kilteni, Jean-Marie Normand, Maria V. Sanchez-Vives, and Mel Slater. Extending Body Space in Immersive Virtual Reality: A Very Long Arm Illusion. *Public Library of Science ONE*, 7(7):1–15, July 2012.

[112] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, Florence, Italy. ACM, 2008. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357127.

[113] Jesper Kjeldskov and Mikael B. Skov. Was It Worth the Hassle?: Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, MobileHCI '14, pages 43–52, Toronto, ON, Canada. ACM, 2014. ISBN: 978-1-4503-3004-6.

[114] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic Recognition of Non-Acted Affective Postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038, August 2011. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2010.2103557.

[115] Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. Evaluating Tooth Brushing Performance with Smartphone Sound Data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 109–120, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2804259.

[116] Yvonne A. W. de Kort, Wijnand A. Ijsselsteijn, Jolien Kooijman, and Yvon Schuurmans. Virtual Laboratories: Comparability of Real and Virtual Environments for Environmental Psychology. *Presence: Teleoper. Virtual Environ.*, 12(4):360–373, August 2003. ISSN: 1054-7460.

[117] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014. ISSN: 0027-8424. DOI: 10.1073/pnas.1320040111.

[118] Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004. ISSN: 1532-0464. DOI: http://dx.doi.org/10.1016/j.jbi.2004.08.004. Named Entity Recognition in Biomedicine.

[119] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[120] Tim Kühl and Alexander Eitel. Effects of disfluency on cognitive and metacognitive processes and outcomes. *Metacognition and Learning*, 11(1):1–13, April 2016. ISSN: 1556-1631. DOI: 10.1007/s11409-016-9154-x.

[121] Cassim Ladha, Nils Hammerla, Emma Hughes, Patrick Olivier, and Thomas Ploetz. Dog's Life: Wearable Activity Recognition for Dogs. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 415–418, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493519.

[122] Benjamin Lafreniere, Tovi Grossman, Fraser Anderson, Justin Matejka, Heather Kerrick, Danil Nagy, Lauren Vasey, Evan Atherton, Nicholas Beirne, Marcelo H. Coelho, Nicholas Cote, Steven Li, Andy Nogueira, Long Nguyen, Tobias Schwinn, James Stoddart, David Thomasson, Ray Wang, Thomas White, David Benjamin, Maurice Conti, Achim Menges, and George Fitzmaurice. Crowdsourced Fabrication. In *Proceedings of the 29th Annual Symposium on User*

Interface Software and Technology, UIST '16, pages 15–28, Tokyo, Japan. ACM, 2016. ISBN: 978-1-4503-4189-9. DOI: 10.1145/2984511.2984553.

[123] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 1977.

[124] Eric C. Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N. Patel. SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 280–289, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370261.

[125] Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N. Patel. Accurate and Privacy Preserving Cough Sensing Using a Low-cost Microphone. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 375–384, Beijing, China. ACM, 2011. ISBN: 978-1-4503-0630-0. DOI: 10.1145/2030112.2030163.

[126] Kevin Larson and Mary Czerwinski. Web Page Design: Implications of Memory, Structure and Scent for Information Retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, pages 25–32, Los Angeles, California, USA. ACM Press/Addison-Wesley Publishing Co., 1998. ISBN: 0-201-30987-4. DOI: 10.1145/274644.274649.

[127] Janina Lehmann, Christina Goussios, and Tina Seufert. Working memory capacity and disfluency effect: an aptitude-treatment-interaction study. *Metacognition and Learning*, 11(1):89–105, April 2016. ISSN: 1556-1631. DOI: 10.1007/s11409-015-9149-z.

[128] Timothy R. Levine, Rachel K. Kim, and Lauren M. Hamel. People Lie for a Reason: Three Experiments Documenting the Principle of Veracity. *Communication Research Reports*, 27(4):271–285, 2010. DOI: 10.1080/08824096.2010.496334.

[129] Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett. *Handbook of Emotions*. Guilford Press, New York, 3rd. Edition, 2008. ISBN: 978-1-59385-650-2.

[130] Sugang Li, Xiaoran Fan, Yanyong Zhang, Wade Trappe, Janne Lindqvist, and Richard E. Howard. Auto++: Detecting Cars Using Embedded Microphones in Real-Time. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):70:1–70:20, September 2017. ISSN: 2474-9567. DOI: 10.1145/3130938.

[131] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 389–402, Taipei, Taiwan. ACM, 2013. ISBN: 978-1-4503-1672-9. DOI: 10.1145/2462456.2464449.

[132] Sally A. Linkenauger, Markus Leyrer, Heinrich H. Bülthoff, and Betty J. Mohler. Welcome to Wonderland: The Influence of the Size and Shape of a Virtual Hand On the Perceived Size and Shape of Virtual Objects. *Public Library of Science ONE*, 8(7), July 2013.

[133] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. CHI 1994-2013: Mapping Two Decades of Intellectual Progress Through Co-word Analysis. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3553–3562, Toronto, Ontario, Canada. ACM, 2014. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2556969.

[134] Beth Logan, Jennifer Healey, Matthai Philipose, Emmanuel Munguia Tapia, and Stephen Intille. A Long-term Evaluation of Sensing Modalities for Activity Recognition. In *Proceedings of the 9th International Conference on Ubiquitous Computing*, UbiComp '07, pages 483–500, Innsbruck, Austria. Springer-Verlag, 2007. ISBN: 978-3-540-74852-6.

[135] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 351–360, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370270`.

[136] Yuki Maekawa, Akira Uchiyama, Hirozumi Yamaguchi, and Teruo Higashino. Car-level Congestion and Position Estimation for Railway Trips Using Mobile Phones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 939–950, Seattle, Washington. ACM, 2014. ISBN: 978-1-4503-2968-2. DOI: `10.1145/2632048.2636062`.

[137] Lara Maister, Natalie Sebanz, Günther Knoblich, and Manos Tsakiris. Experiencing Ownership Over a Dark-skinned Body Reduces Implicit Racial Bias. *International Journal of Cognitive Science*, 128(2):170–178, 2013. ISSN: 0010-0277.

[138] Alex Mariakakis, Megan A. Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N. Patel. BiliScreen: Smartphone-Based Scleral Jaundice Monitoring for Liver and Pancreatic Disorders. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):20:1–20:26, June 2017. ISSN: 2474-9567. DOI: `10.1145/3090085`.

[139] Tim Marsh. Evaluation of Virtual Reality Systems for Usability. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, pages 61–62, Pittsburgh, Pennsylvania. ACM, 1999. ISBN: 1-58113-158-5.

[140] Winter Mason and Siddharth Suri. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012. ISSN: 1554-3528.

[141] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. A Dose of Reality: Overcoming Usability Challenges in VR Head-Mounted Displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2143–2152, Seoul, Republic of Korea. ACM, 2015. ISBN: 978-1-4503-3145-6.

[142] Joseph E. McGrath. *Human-computer Interaction*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 1995. Part Methodology Matters: Doing Research in the Behavioral and Social Sciences, pages 152–169. ISBN: 1-55860-246-1.

[143] A. Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Social Environmental and Developmental Studies. Oelgeschlager, Gunn & Hain, 1980. ISBN: 9780899460048.

[144] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. MyTraces: Investigating Correlation and Causation Between Users&Rsquo; Emotional States and Mobile Phone Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):83:1–83:21, September 2017. ISSN: 2474-9567. DOI: `10.1145/3130948`.

[145] A. Meyer, S. Frederick, T. C. Burnham, J. D. Guevara Pinto, T. W. Boyer, L. J. Ball, G. Pennycook, R. Ackerman, V. A. Thompson, and J. P. Schuldt. Disfluent fonts don't help people solve math problems. *J Exp Psychol Gen*, 144(2):16–30, April 2015.

[146] Rada Mihalcea and Carlo Strapparava. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 309–312, Suntec, Singapore. Association for Computational Linguistics, 2009.

[147] Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):85:1–85:20, September 2017. ISSN: 2474-9567. DOI: `10.1145/3131894`.

[148] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. Automated Estimation of Food Type and Amount Consumed from Body-worn Audio and Motion Sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 451–462, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971677`.

[149] Reham Mohamed and Moustafa Youssef. HeartSense: Ubiquitous Accurate Multi-Modal Fusion-based Heart Rate Estimation Using Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):97:1–97:18, September 2017. ISSN: 2474-9567. DOI: `10.1145/3132028`.

[150] Aske Mottelson and Kasper Hornbæk. An Affect Detection Technique Using Mobile Commodity Sensors in the Wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 781–792, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971654`.

[151] Aske Mottelson and Kasper Hornbæk. An Affect Detection Technique Using Mobile Commodity Sensors in the Wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 781–792, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971654`.

[152] Aske Mottelson and Kasper Hornbæk. Emotional Avatars: The Interplay between Affect and Ownership of a Virtual Body. In *Manuscript*, 2018.

[153] Aske Mottelson and Kasper Hornbæk. Virtual Reality Studies Outside the Laboratory. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17, 9:1–9:10, Gothenburg, Sweden. ACM, 2017. ISBN: 978-1-4503-5548-3. DOI: `10.1145/3139131.3139141`.

[154] Aske Mottelson, Jarrod Knibbe, and Kasper Hornbæk. Veritaps: Truth Estimation from Mobile Interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 561:1–561:12, Montreal QC, Canada. ACM, 2018. ISBN: 978-1-4503-5620-6. DOI: `10.1145/3173574.3174135`.

[155] Aske Mottelson, Christoffer Larsen, Mikkel Lyderik, Paul Strohmeier, and Jarrod Knibbe. Invisiboard: Maximizing Display and Input Space with a Full Screen Text Entry Method for Smartwatches. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '16, pages 53–59, Florence, Italy. ACM, 2016. ISBN: 978-1-4503-4408-1. DOI: `10.1145/2935334.2935360`.

[156] Annamalai Natarajan, Abhinav Parate, Edward Gaiser, Gustavo Angarita, Robert Malison, Benjamin Marlin, and Deepak Ganesan. Detecting Cocaine Use with Wearable Electrocardiogram Sensors. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 123–132, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: 10.1145/2493432.2493496.

[157] Paula M. Niedenthal. Embodying emotion. *Science*, 316(5827):1002–1005, 2007. ISSN: 0036-8075. DOI: 10.1126/science.1136930.

[158] Paula M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-Gruber, and F. Ric. Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, 9(3):184–211, 2005. DOI: 10.1207/s15327957pspr0903_1.

[159] Jean-Marie Normand, Elias Giannopoulos, Bernhard Spanlang, and Mel Slater. Multisensory stimulation can induce an illusion of larger belly size in immersive virtual reality. *PLOS One*, 6(1):1–11, 2011. DOI: 10.1371/journal.pone.0016128.

[160] Hessel Oosterbeek, Randolph Sloof, and Gijs van de Kuilen. Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics*, 7(2):171–188, June 2004. ISSN: 1573-6938. DOI: 10.1023/B:EXEC.0000026978.14316.74.

[161] Daniel M. Oppenheimer. Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, 20(2):139–156, 2006. ISSN: 1099-0720. DOI: 10.1002/acp.1178.

[162] Daniel M. Oppenheimer and Michael C. Frank. A rose in any other font would not smell as sweet: Effects of perceptual fluency on categorization. *Cognition*, 106(3):1178–1194, 2008. ISSN: 0010-0277. DOI: http://dx.doi.org/10.1016/j.cognition.2007.05.010.

[163] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Portland, Oregon. Association for Computational Linguistics, 2011. ISBN: 978-1-932432-87-9.

[164] Kazushige Ouchi and Miwako Doi. Indoor-outdoor Activity Recognition by a Smartphone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 537–537, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370297.

[165] Xueni Pan, Mel Slater, Alejandro Beacco, Xavi Navarro, Anna I. Bellido Rivas, David Swapp, Joanna Hale, Paul Alexander George Forbes, Catrina Denvir, Antonia F. de C. Hamilton, and Sylvie Delacroix. The Responses of Medical General Practitioners to Unreasonable Patient Demand for Antibiotics - A Study of Medical Ethics Using Immersive Virtual Reality. *Public Library of Science ONE*, 11(2), February 2016.

[166] Gabriele Paolacci and Jesse Chandler. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[167] Jun-geun Park, Ami Patel, Dorothy Curtis, Seth Teller, and Jonathan Ledlie. Online Pose Classification and Walking Speed Estimation Using Handheld Devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 113–122, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370235.

[168] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, and Mel Slater. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition*, 22(3):779–787, 2013. ISSN: 1053-8100. DOI: http://dx.doi.org/10.1016/j.concog.2013.04.016.

[169] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. AROMA: A Deep Multi-Task Learning Based Simple and Complex Human Activity Recognition Method Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2):74:1–74:16, July 2018. ISSN: 2474-9567. DOI: 10.1145/3214277.

[170] Valeria I. Petkova and H. Henrik Ehrsson. If I were you: Perceptual illusion of body swapping. *PLOS One*, 3(12):1–9, 2008. DOI: 10.1371/journal.pone.0003832.

[171] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997. ISBN: 0-262-16170-2.

[172] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):91:1–91:25, September 2017. ISSN: 2474-9567. DOI: 10.1145/3130956.

[173] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 825–836, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2804252.

[174] R. Plutchik. The Nature of Emotions. *American Scientist*, 89:344, July 2001. DOI: 10.1511/2001.4.344.

[175] Henning Pohl and Aske Mottelson. How we Guide, Write, and Cite at CHI. In *Proceedings of the 2019 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '19, Glasgow, UK. ACM, In review.

[176] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. A survey on mobile affective computing. *Computer Science Review*, 25:79–100, 2017. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2017.07.002.

[177] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. Passive and In-Situ Assessment of Mental and Physical Well-being Using Mobile Sensors. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 385–394, Beijing, China. ACM, 2011. ISBN: 978-1-4503-0630-0. DOI: 10.1145/2030112.2030164.

[178] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. EmotionSense: A Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 281–290, Copenhagen, Denmark. ACM, 2010. ISBN: 978-1-60558-843-8. DOI: 10.1145/1864349.1864393.

[179] Katharina Reinecke and Krzysztof Z. Gajos. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1364–1378, Vancouver, BC, Canada. ACM, 2015. ISBN: 978-1-4503-2922-4.

[180] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. Affective

interactions using virtual reality: The link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56, February 2007. ISSN: 1094-9313. DOI: 10.1089/cpb.2006.9993.

[181] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, Atlanta, Georgia, USA. ACM, 2010. ISBN: 978-1-60558-930-5.

[182] Steven V. Rouse. A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, 43:304–307, 2015. ISSN: 0747-5632. DOI: http://dx.doi.org/10.1016/j.chb.2014.11.004.

[183] Jonathan Rubin, Hoda Eldardiry, Rui Abreu, Shane Ahern, Honglu Du, Ashish Pattekar, and Daniel G. Bobrow. Towards a Mobile and Wearable System for Predicting Panic Attacks. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 529–533, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2805834.

[184] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980. ISSN: 0022-3514.

[185] Maria V. Sanchez-Vives, Bernhard Spanlang, Antonio Frisoli, Massimo Bergamasco, and Mel Slater. Virtual Hand Illusion Induced by Visuomotor Correlations. *Public Library of Science ONE*, 5(4):1–6, April 2010.

[186] Leonard Saxe, Denise Dougherty, and Theodore Cross. The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist*, 40(3):355–366, 1985. ISSN: 1935-990X 0003-066X. DOI: 10.1037/0003-066X.40.3.355.

[187] Joey Scarr, Andy Cockburn, Carl Gutwin, and Sylvain Malacria. Testing the Robustness and Performance of Spatially Consistent Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3139–3148, Paris, France. ACM, 2013. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2466430.

[188] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010. DOI: 10.1080/02699930903274322.

[189] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):32:1–32:20, March 2018. ISSN: 2474-9567. DOI: 10.1145/3191764.

[190] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W Picard. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14(2):93–118, 2002. DOI: 10.1016/S0953-5438(01)00059-5.

[191] W. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and Quasi Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company., Boston, 2002, pages 33–102.

[192] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston, MA,

USA, 2002. Part Statistical Conclusion Validity and Internal Validity, pages 33–102. ISBN: 0395615569.

[193] Rahul C. Shah, Chieh-yih Wan, Hong Lu, and Lama Nachman. Classifying the Mode of Transportation on Mobile Phones Using GIS Information. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 225–229, Seattle, Washington. ACM, 2014. ISBN: 978-1-4503-2968-2. DOI: `10.1145/2632048.2632109`.

[194] Masamichi Shimosaka, Shinya Masuda, Kazunari Takeichi, Rui Fukui, and Tomomasa Sato. Health Score Prediction Using Low-invasive Sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 1044–1048, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370439`.

[195] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V. Sanchez-Vives. A Virtual Reprise of the Stanley Milgram Obedience Experiments. *Public Library of Science ONE*, 1(1):1–10, December 2006.

[196] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria V. Sanchez-Vives. Towards a Digital Body: The Virtual Arm Illusion. *Frontiers in Human Neuroscience*, 2(6), 2008. ISSN: 1662-5161.

[197] Mel Slater, Daniel Perez-Marcos, H. Henrik Ehrsson, and Maria V. Sanchez-Vives. Inducing Illusory Ownership of a Virtual Body. *Frontiers in Human Neuroscience*, 3(2):214–220, September 2009. ISSN: 1662-4548.

[198] Mel Slater, Aitor Rovira, Richard Southern, David Swapp, Jian J. Zhang, Claire Campbell, and Mark Levine. Bystander Responses to a Violent Incident in an Immersive Virtual Environment. *Public Library of Science ONE*, 8(1):1–13, January 2013.

[199] Mel Slater, Bernhard Spanlang, Maria V. Sanchez-Vives, and Olaf Blanke. First Person Experience of Body Transfer in Virtual Reality. *Public Library of Science ONE*, 5(5):1–9, May 2010.

[200] Mel Slater, Martin Usoh, and Anthony Steed. Depth of Presence in Virtual Environments. *Presence Teleoperators and Virtual Environments*, 3(2):130–144, January 1994. ISSN: 1054-7460.

[201] Mel Slater, Martin Usoh, and Anthony Steed. Taking Steps: The Influence of a Walking Technique on Presence in Virtual Reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2(3):201–219, September 1995. ISSN: 1073-0516.

[202] Mel Slater and Sylvia Wilbur. A Framework for Immersive Virtual Environments Five: Speculations on the Role of Presence in Virtual Environments. *Presence Teleoperators and Virtual Environments*, 6(6):603–616, December 1997. ISSN: 1054-7460.

[203] Steven A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119:3–22, 1996.

[204] Frank Soboczenski, Paul Cairns, and Anna L. Cox. *Increasing Accuracy by Decreasing Presentation Quality in Transcription Tasks*. In *Human-Computer Interaction – INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part II*. Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler, editors. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pages 380–394. ISBN: 978-3-642-40480-1. DOI: `10.1007/978-3-642-40480-1_25`.

[205] H. Song and N. Schwarz. If it's hard to read, it's hard to do: processing fluency affects effort prediction and motivation. *Psychol Sci*, 19(10):986–988, October 2008.

[206] R. William Soukoreff and I. Scott MacKenzie. Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *International Journal of Human-Computer Studies*, 61(6):751–789, December 2004. ISSN: 1071-5819. DOI: `10.1016/j.ijhcs.2004.09.001`.

[207] S. A. Spence, T. F. Farrow, A. E. Herford, I. D. Wilkinson, Y. Zheng, and P. W. Woodruff. Behavioural and functional anatomical correlates of deception in humans. eng. *Neuroreport*, 12(13):2849–2853, September 2001. ISSN: 0959-4965.

[208] Misha Sra, Aske Mottelson, and Pattie Maes. Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 85–97, Hong Kong, China. ACM, 2018. ISBN: 978-1-4503-5198-0. DOI: `10.1145/3196709.3196788`.

[209] Misha Sra, Xuhai Xu, Aske Mottelson, and Pattie Maes. VMotion: Designing a Seamless Walking Experience in VR. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, pages 59–70, Hong Kong, China. ACM, 2018. ISBN: 978-1-4503-5198-0. DOI: `10.1145/3196709.3196792`.

[210] STEAM. Steam Hardware & Software Survey: September 2017. `http://store.steampowered.com/hwsurvey`, 2017. [Accessed 01-September-2017].

[211] Anthony Steed, Sebastian Frlston, Maria M. López, Jason Drummond, Ye Pan, and David Swapp. An 'In the Wild' Experiment on Presence and Embodiment using Consumer Virtual Reality Equipment. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1406–1414, April 2016. ISSN: 1077-2626.

[212] Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers. *Judgment and Decision Making*, 10(5):479–491, 2015.

[213] F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5):768–777, May 1988. DOI: `10.1037/0022-3514.54.5.768`.

[214] Mayu Sumida, Teruhiro Mizumoto, and Keiichi Yasumoto. Estimating Heart Rate Variation During Walking with Smartphone. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 245–254, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: `10.1145/2493432.2493491`.

[215] David Sun, Pablo Paredes, and John Canny. MouStress: Detecting Stress from Mouse Motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 61–70, Toronto, Ontario, Canada. ACM, 2014. ISBN: 978-1-4503-2473-1. DOI: `10.1145/2556288.2557243`.

[216] Alistair Sutcliffe and Brian Gault. Heuristic Evaluation of Virtual Reality Applications. *Interacting with Computers*, 16(4):831–849, 2004.

[217] M. Suwa, N. Sugie, and Fujimora. A preliminary note on pattern recognition of human emotional expression. In *The 4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

[218] Brandon Taylor, Anind Dey, Daniel Siewiorek, and Asim Smailagic. Using Physiological Sensors to Detect Levels of User Frustration Induced by System Delays. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 517–528, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2805847.

[219] Ann E. Tenbrunsel and David M. Messick. Ethical Fading: The Role of Self-Deception in Unethical Behavior. en. *Social Justice Research*, 17(2):223–236, June 2004. ISSN: 0885-7466, 1573-6725. DOI: 10.1023/B:SORE.0000027411.35832.53.

[220] Edison Thomaz, Irfan Essa, and Gregory D. Abowd. A Practical Approach for Recognizing Eating Moments with Wrist-mounted Inertial Sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1029–1040, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2807545.

[221] A. Toet, M. van Welie, and J. Houtkamp. Is a dark virtual environment scary? *CyberPsychology & Behavior*, 12(4):363–371, August 2009. DOI: 10.1089/cpb.2008.0293.

[222] Jessica J. Tran, Shari Trewin, Calvin Swart, Bonnie E. John, and John C. Thomas. Exploring Pinch and Spread Gestures on Mobile Devices. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 151–160, Munich, Germany. ACM, 2013. ISBN: 978-1-4503-2273-7. DOI: 10.1145/2493190.2493221.

[223] Martin Usoh, Kevin Arthur, Mary C. Whitton, Rui Bastos, Anthony Steed, Mel Slater, and Frederick P. Brooks Jr. Walking >Walking-in-place >Flying, in Virtual Environments. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 359–364, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co., 1999. ISBN: 0-201-48560-5.

[224] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):168:1–168:22, January 2018. ISSN: 2474-9567. DOI: 10.1145/3161192.

[225] Lyn M. Van Swol, Deepak Malhotra, and Michael T. Braun. Deception and Its Detection: Effects of Monetary Incentives and Personal Relationship History. en. *Communication Research*, 39(2):217–238, April 2012. ISSN: 0093-6502. DOI: 10.1177/0093650210396868.

[226] Emmett Velten. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4):473–482, 1968. ISSN: 0005-7967. DOI: https://doi.org/10.1016/0005-7967(68)90028-4.

[227] J. M. Vendemia, R. F. Buzan, and S. L. Simon-Dack. Reaction time of motor responses in two-stimulus paradigms involving deception and congruity with varying levels of difficulty. *Behav Neurol*, 16(1):25–36, 2005.

[228] Aldert Vrij. *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley Series in Psychology of Crime, Policing and Law. Wiley, 2011. ISBN: 9781119965763.

[229] Harald G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998. ISSN: 1099-0992. DOI: 10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W.

[230] Edward Jay Wang, William Li, Doug Hawkins, Terry Gernsheimer, Colette Norby-Slycord, and Shwetak N. Patel. HemaApp: Noninvasive Blood Screening of Hemoglobin Using Smartphone Cameras. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 593–604, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971653`.

[231] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. CrossCheck: Toward Passive Sensing and Detection of Mental Health Changes in People with Schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 886–897, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971740`.

[232] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 295–306, Osaka, Japan. ACM, 2015. ISBN: 978-1-4503-3574-4. DOI: `10.1145/2750858.2804251`.

[233] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):110:1–110:24, September 2017. ISSN: 2474-9567. DOI: `10.1145/3130976`.

[234] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):43:1–43:26, March 2018. ISSN: 2474-9567. DOI: `10.1145/3191775`.

[235] Weichen Wang, Gabriella M. Harari, Rui Wang, Sandrine R. Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T. Campbell. Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):141:1–141:21, September 2018. ISSN: 2474-9567. DOI: `10.1145/3264951`.

[236] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54:1063–1070, 1988.

[237] Mark Weiser. The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11, July 1999. ISSN: 1559-1662. DOI: `10.1145/329124.329126`.

[238] Rainer Westermann, Kordelia Spies, Gunter Stahl, and Friedrich W. Hesse. Relative Effectiveness and Validity of Mood Induction Procedures: A Meta-Analysis. *European Journal of Social Psychology*, 26:557–580, 1996. DOI: `10.1002/(SICI)1099-0992(199607)26:4<557::AID-EJSP769>3.0.CO;2-4`.

[239] C. Williams and K. Stevens. Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America*, 52(4):1238–1250, 1972.

[240] Emma J. Williams, Lewis A. Bott, John Patrick, and Michael B. Lewis. Telling Lies: The Irrepressible Truth? *PLOS ONE*, 8(4):1–14, April 2013. DOI: `10.1371/journal.pone.0060713`.

[241] Margaret Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, 2002. ISSN: 1531-5320. DOI: `10.3758/BF03196322`.

[242] Xiaolin Wu and Xi Zhang. Automated Inference on Criminality using Face Images. *CoRR*, abs/1611.04135, 2016.

[243] Haoyi Xiong, Yu Huang, Laura E. Barnes, and Matthew S. Gerber. Sensus: A Cross-platform, General-purpose System for Mobile Crowdsensing in Human-subject Studies. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 415–426, Heidelberg, Germany. ACM, 2016. ISBN: 978-1-4503-4461-6. DOI: `10.1145/2971648.2971711`.

[244] Chenren Xu, Sugang Li, Gang Liu, Yanyong Zhang, Emiliano Miluzzo, Yih-Farn Chen, Jun Li, and Bernhard Firner. Crowd++: Unsupervised Speaker Count with Smartphones. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 43–52, Zurich, Switzerland. ACM, 2013. ISBN: 978-1-4503-1770-2. DOI: `10.1145/2493432.2493435`.

[245] Koji Yatani and Khai N. Truong. BodyScope: A Wearable Acoustic Sensor for Activity Recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 341–350, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370269`.

[246] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D. Abowd, and James M. Rehg. Detecting Eye Contact Using Wearable Eye-tracking Glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 699–704, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370368`.

[247] Chuang-Wen You, Martha Montes-de-Oca, Thomas J. Bao, Nicholas D. Lane, Hong Lu, Giuseppe Cardone, Lorenzo Torresani, and Andrew T. Campbell. CarSafe: A Driver Safety App That Detects Dangerous Driving Behavior Using Dual-cameras on Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 671–672, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: `10.1145/2370216.2370360`.

[248] Donghan Yu, Yong Li, Fengli Xu, Pengyu Zhang, and Vassilis Kostakos. Smartphone App Usage Prediction Using Points of Interest. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):174:1–174:21, January 2018. ISSN: 2474-9567. DOI: `10.1145/3161413`.

[249] Ulrich von Zadow, Sandra Buron, Tina Harms, Florian Behringer, Kai Sostmann, and Raimund Dachselt. SimMed: Combining Simulation and Interactive Tabletops for Medical Education. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1469–1478, Paris, France. ACM, 2013. ISBN: 978-1-4503-1899-0.

[250] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2008. DOI: `10.1109/TPAMI.2008.52`.

[251] Mi Zhang and Alexander A. Sawchuk. A Preliminary Study of Sensing Appliance Usage for Human Activity Recognition Using Mobile Magnetometer. In *Proceedings of the 2012 ACM Confer-*

ence on Ubiquitous Computing, UbiComp ’12, pages 745–748, Pittsburgh, Pennsylvania. ACM, 2012. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370380.

[252] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):176:1–176:30, January 2018. ISSN: 2474-9567. DOI: 10.1145/3161414.

[253] Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. Verbal and Nonverbal Communication of Deception. *Advances in Experimental Social Psychology*, 14:1–59, January 1981. ISSN: 0065-2601. DOI: 10.1016/S0065-2601(08)60369-X.