MOMENTS OF MEASUREMENT FOR AFFECT, STRESS, AND CHRONIC PAIN WITH MOBILE DEVICES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Philip James Adams

August 2015

MOMENTS OF MEASUREMENT FOR AFFECT, STRESS, AND CHRONIC PAIN WITH MOBILE DEVICES

Philip James Adams, Ph.D.

Cornell University 2015

With the goal of supporting daily wellbeing, this dissertation explores, unites, and extends several research domains: measurement and psychometrics, mobile HCI and user experiences, and the *in situ* assessment of affect, stress, and chronic pain. Inspired by the promises of mobile health (Richardson & M. C. Reid, 2013), this work contributes both artifacts (measures and software) and research findings that further the momentary measurement of these often unobservable and deeply human features.

Each of the three case studies involves investigating momentary measurement with mobile devices (Shiffman, Stone, & Hufford, 2008; Hektner, Schmidt, & Mihaly Csikszentmihalyi, 2007). In assessing affect, I report on the Photographic Affect Meter (Pollak, Adams, & Gay, 2011), a novel image-based single-item measure of emotion; underlying the PAM grid is the Russell's circumplex model of affect: a two-dimensional valence/arousal space within which each emotion can be placed (Russell, 1980). In assessing stress, I report on SESAME, a field trial comparing minimally invasive techniques for assessing stress in the wild (Adams et al., 2014); I am interested in supporting long-term engagement with one's own perceived stress levels through measurement by determining an effective balance between reliability and intrusiveness. In assessing pain intensity, I report on two projects intended to support the management of chronic pain by providing novel and effective self-report assessments of pain inten-

sity. Meter is a multi-stage user-centered research through design (Gay, 2004; Zimmerman, Forlizzi, & Evenson, 2007) investigation seeking optimal visual interfaces for the self-report of pain intensity on smartphone screens. Keppi is a novel pressure-based user input device for the self-report of scalar values; I show that users are able to consistently report pain intensity with four degrees of freedom, as well as continuously map pressure to visual cues.

I then reflect on the case studies holistically speaking to (1) the continued essentially value of self-report in a sensor-centric world, (2) the idea of a minimally viable moment of self-assessment, (3) the advantages and disadvantages of personalizing the measurement interfaces themselves, and (4) that there is not a one-size-fits all approach for developing novel self-reports for the EMA domain. I then contribute several design patterns useful to others continuing this work in new health spaces.

BIOGRAPHICAL SKETCH

Phil was born in the UK and lived in London and Colchester before moving to Philadelphia, where he attended Lower Merion High School and graduated in 2004. Following a gap year spent coaching soccer, traveling in Central America, and participating in service projects in East Africa, Phil moved to Ithaca to attend Cornell University where he studied Information Science. As a senior and then as a Master of Engineering (Computer Science) student, Phil began working as a research assistant in Dr. Geri Gay's Interaction Design Lab working in mobile technologies and museum navigation. He has since continued his work with Drs Gay, Choudhury, and Wethington exploring mobile and social computing with a focus on small and brief moments of interaction; his dissertation work is on the *in situ* momentary measurement of affect, stress, and chronic pain. Phil is also interested in how and why people participate in online communities (and why they don't!), in brief moments of socially supportive behavior, in terse visualizations affording glimpses of (personal) data, and in the capture and sharing of images for communication and self-assessment.

ACKNOWLEDGEMENTS

The work presented in this dissertation has only been possible because of the support of faculty, friends, and family members. Above all, I could not be more grateful to my advisor, Geri Gay, for the opportunity to begin a lifetime of scholarship. Geri's encouragement and advice, both in my professional work and personal life, have made the last five years a wonderful period of exploration, growth, challenge, and contribution. In her Interaction Design Lab, Geri fosters a mix of creativity, collaboration, work ethic, intellectual freedom, and scholarly rigor that I take with me as a pattern for all future efforts. I am also extremely grateful to my committee members Tanzeem Choudhury and Elaine Wethington who have been both fantastically available and deeply thoughtful in guiding my growth and this work. I consider myself beyond fortunate to have worked with this particular committee.

I also thank a range of collaborating researchers who have been an essential part of this work. Fellow graduate students JP Pollak and Alex Adams have worked most closely with me on a project each; learning from and with JP and Alex has been a rewarding and crucial aspect of the PAM and Keppi projects. I'm thankful for a series of fantastic post-docs who have been generous with their time and advice, most particularly Eric Baumer, Steve Voida, Jaime Snyder, and Mark Matthews. I would also like to thank the undergraduate and masters research assistants for their contributions, most especially Michael Elfenbein. A wide range of additional collaborators and scholars have additionally provided valuable insights and critiques: thank you to Deborah Estrin and the Small Data Lab at NYC Tech, the researchers and practitioners at TRIPLL, and clinicians and staff at Weill Medical.

I would like to thank the NSF for funding a large portion of this work, both through my Graduate Research Fellowship as well as the grant "Novel Techniques for Patient-centric Disease Management using Automatically Inferred Behavioral Biomarkers and Sensor-Supported Contextual Self-Report". Working toward an NIA-based TRIPLL pilot grant has also been extremely helpful in conducting this work.

I cannot imagine a more wonderful experience for graduate school than to be a part of the Information Science community at Cornell. My cohort (Shion Guha, Nitesh Goyal, Madeline Smith, and Karen Patzke) have kept me sane and provided encouragement throughout, as have honorary cohort members Liz Murnane and Victoria Schwanda Sosik and my desk-mates Saeed Abdullah and Eric Baumer. The wider IS graduate student body, and the incredibly generous and supportive IS faculty, have together created a safe, inspiring, and rewarding environment for scholarly activity and fun. I owe a special thank you to my close friends Shion, Liz, Scott O'Connor, and Annie Tomlinson for their encouragement and for sharing hour upon hour sitting across a table from me during writing sessions and study breaks.

Finally, I would like to thank my parents, grandparents, and siblings (Sarah and Tim) for their support and continued interest in my research and academic-world activities.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the goal of supporting daily wellbeing, this dissertation explores, unites, and extends several research domains: measurement and psychometrics, mobile HCI and user experiences, and the *in situ* assessment of affect, stress, and chronic pain. Inspired by the promises of mobile health (Richardson & M. C. Reid, 2013), this work contributes both artifacts (measures and software) and research findings that further the momentary measurement of these often unobservable and deeply human features.

How exactly does this work support wellbeing? Health is today understood to be the product of a massive society-behavior-biology nexus (Glass & McAtee, 2006). While some researchers are now attempting to leverage the entirety of such models in capturing whole-health data (e.g. UCSD's Delphi project[1]), many others use them as organizing frameworks that can position and enhance wellness research efforts (e.g. Wethington, 2005). In the context of such models it is clear that the self-report of affect, stress, and pain intensity are tiny subsets of human action at the micro-level (Figure 1.1) - and yet the effort among researchers and health practitioners to better capture such assessments is immense.

There is a great deal of value in finding more effective ways to assess affect, stress, and chronic pain in natural environments.

Humans are emotional beings. Indeed, as Lazarus and Lazarus have it, "Everything important that happens to us arouses emotion" (1994). More than be-

[1] http://delphi.ucsd.edu

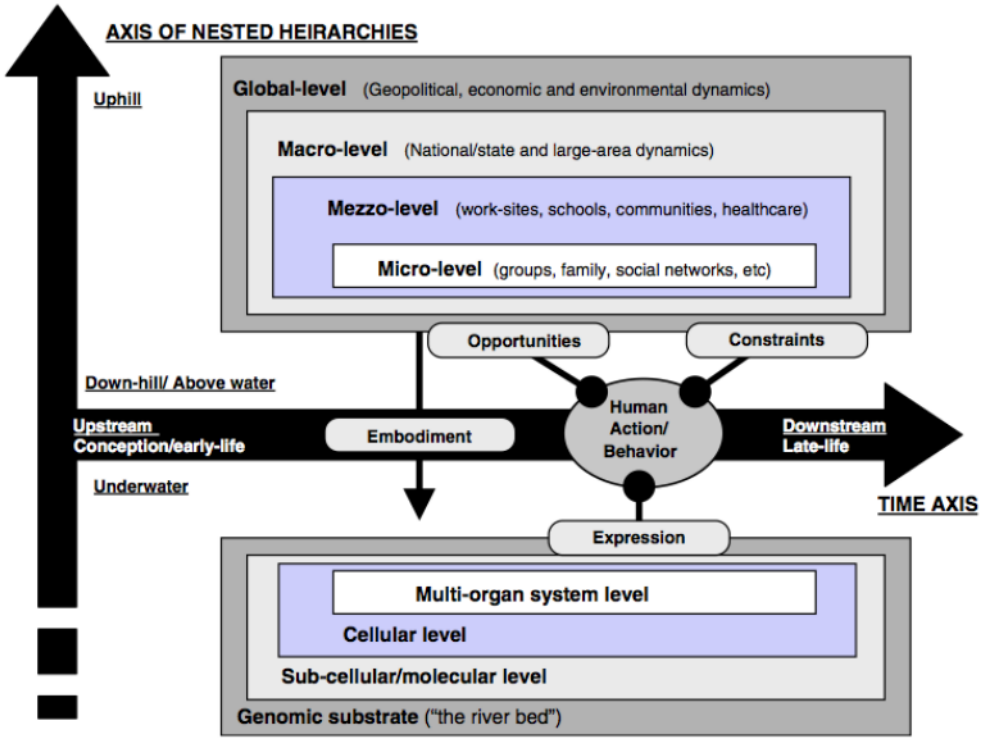


Figure 1.1: The Whole Health Model (Glass & McAtee, 2006).

ing one way we experience what happens to us, "the neurological evidence indicates emotions are not a luxury; they are essential for rational human performance" (R. W. Picard & J. Healey, 1997) and are "a vital tool for getting along in the world" (R. S. Lazarus & B. N. Lazarus, 1994). "We feel before we know, and in an important sense, feeling determines what we know" (Buck, 1986) - in a very real way, "emotions, motivations, and cognitive processes coexist and contribute to experience in every moment of our life" (Hektner et al., 2007). It is useful, therefore, to be able to capture and assess the emotions that people experience - perhaps particularly so as it becomes clear that positive affect is associated with positive health outcomes, including lower morbidity and decreased symptom and pain experiences, and is likely also a buffer to experiencing negative stress (Pressman & Cohen, 2005).

In the 2011 Stress in America survey, the American Psychological Association warned that stress is becoming a public health crisis (APA, 2011). Most Americans are suffering from moderate to high levels of stress, with nearly half reporting an increase in stress over the preceding five-year span. According to the APA, "job stress is estimated to cost U.S. industry $300 billion a year in absenteeism, diminished productivity, employee turnover and direct medical, legal and insurance fees" (APA, 2012).

Chronic pain, recurrent or long-lasting pain, affects an estimated 30.7% of US adults, and these numbers are considerably worse for the aging population: more than 50% of older adults and as many as 80% of older adults living in nursing homes (B. A. Ferrell, B. R. Ferrell, & Rivera, 1995; Helme & Gibson, 2001). Common chronic pain conditions include osteoarthritis (OA), rheumatoid arthritis (RA), lower back pain, and migraine/headache, as well as injury-related conditions, repetitive stress disorders, and other conditions. Chronic pain is common across a wide range of disease and demographic groups, but is more common in women than in men, and prevalence increases with age. Indicators of poor socioeconomic status, including lower household income and unemployment, are significantly correlated with chronic pain conditions (Johannes, Le, Zhou, Johnston, & Dworkin, 2010). Patients with chronic pain are frequently severely debilitated, with significant limitations in their ability to function or work. Chronic pain is associated with depression, sleep disturbance, fatigue and decreased cognitive and physical abilities (Ashburn & Staats, 1999).

Clearly, supporting high resolution and low burden measurement of affect, stress, and chronic pain, enabling improved self- and clinical-management, is a worthy goal. But, how best to measure them?

Observation (systematic and otherwise) of human traits and interactions are likely as old as humankind. The style of data collection used in this dissertation, Ecological Momentary Assessment (Shiffman et al., 2008), has roots in several self-monitoring traditions including diarying, action and interaction monitoring in behaviorism, and Experience Sampling (Hektner et al., 2007). In chapter 2, I discuss aspects of these traditions relevant to the research case studies and the resulting discussion. Features common to Ecological Momentary Assessment approaches are one modern representation Allport's 75 year old desire for "psychology to concern itself with life as it is lived, with significant total-processes of the sort revealed in consecutive and complete life documents" (Allport, 1942):

- Data are collected in natural settings, in the real world, as participants go about their daily lives.
- Assessments focus on participants' current, in-the-moment state, rather than relying on recall.
- Assessments tend to be as brief and as unobtrusive as possible, both to maintain ecological validity and so as not to overly burden participants.
- Moments for assessment are strategically selected (typically using one of time-, event-, or signal-contingent strategies).
- Subjects complete multiple assessments over time.

Each of the three case studies involves investigating momentary measurement with mobile devices, leveraging the microinteraction: a small piece of functionality that allows the user to complete a single task: setting phone volume to vibrate, logging into a service, or recording a piece of information (Saffer, 2013). Microinteractions represent the way users spend a much of their time on

mobile devices (Oulasvirta, Rattenbury, Ma, & Raita, 2012) - chapter 2 further describes the relevant (m)HCI literature around the mobile user experience and mobile usability, as well as the iterative user-centered research through design methods I draw on when conducting this work.

Chapters 3 through 5 each represents a case study of momentary measurement in a different domain: affect, stress, and chronic pain:

In chapter 3, I report on the design, development, and validation of PAM, the Photographic Affect Meter (Pollak et al., 2011). From a user's perspective, PAM is laid out in a grid (Figure 3.8). The user is presented with 16 photographs that represent a diversity of emotions (description to follow) arranged into a 4x4 grid, and is prompted to select a photo that best describes how they feel right now. Below the grid, a button to choose more photos reloads the grid with a different set of PAM images should the user not be able to comfortably express their state with one of the photos currently displayed.

Underlying the PAM grid is the Russell's circumplex model of affect: a two-dimensional valence/arousal space within which each emotion can be placed (Russell, 1980). The decision to use photographs is based in literature indicating that photographs and emotions are linked, often with universally shared meaning (Chalfen, 1989; Lang, 1995), but affording the user a sense of choice and control by way of interpretive flexibility (Sengers, Boehner, Mateas, & Gay, 2008).

In chapter 4, I report on SESAME, a field trial comparing minimally invasive techniques for assessing stress in the wild (Adams et al., 2014). In the 2011 Stress in America survey, the American Psychological Association warned that stress

is becoming a public health crisis (APA, 2011) - most Americans are suffering from moderate to high levels of stress, with nearly half reporting an increase in stress over the preceding five-year span (APA, 2012). In this case study, I am interested in supporting long-term engagement with one's own perceived stress levels through measurement. One of the central challenges in creating these types of systems is in determining what kind of stress-related data to collect in order to strike a balance between reliability - that is, how closely the data accurately and consistently track a person's perceived stress at the moment that it occurs, and intrusiveness - that is, how much effort is required on a participant or user's behalf to provide the data.

SESAME triangulates data from self-report, electrodermal-sensed arousal, and continuous inferencing of stress markers in human voice via the phone's microphone. Among these techniques for self-monitoring stress levels with a minimal impact on participants' daily lives, which track one another most closely? Under what conditions or in what contexts? When do these sensing modalities agree with one another, and when do they produce conflicting narratives about daily stress?

In chapter 5, I report on two projects intended to support the management of chronic pain by providing novel and effective self-report assessments of pain intensity. Chronic pain, recurrent or long-lasting pain, affects an estimated 30.7% of US adults (B. A. Ferrell et al., 1995). Those with chronic pain are frequently severely debilitated, with significant limitations in their ability to function or work. Chronic pain is associated with depression, sleep disturbance, fatigue and decreased cognitive and physical abilities. (Ashburn & Staats, 1999).

Meter is a multi-stage user-centered research through design (Gay, 2004; Zimmerman et al., 2007) investigation seeking optimal visual interfaces for the self-report of pain intensity on smartphone screens. Through design ideation, in-lab user studies, a three-week field trial, and an expert panel review, Meter results in two recommended self-report measures as well as rich insights regarding the ways those with chronic pain prefer to record and report their pain levels.

Keppi is motivated by observations of the ways that people in pain will sometimes grasp the arms of their chair or the hand of a loved one, and inspired by the uncomplicated action of squeezing a stress ball. These interactions are unobtrusive and can be very private. In seeking to integrate these types of interactions with intentional self-report, we developed Keppi, a novel pressure-based user input device for the self-report of scalar values. As the "intensity of pain is without a doubt the most salient dimension of pain" Turk and Melzack, 2011, we focus on the frequent *in situ* self-report of pain severity; to our knowledge, there has been no previous attempt to leverage a dedicated tangible user interface (TUI) for the EMA-style self-report of chronic pain.

In chapter 6, I reflect on the case studies primarily through four discussion points. First, from both empirically gathered data and the literature, I describe the continued value (the essential value!) of self-report in a world awash with increasingly capable passive sensors. Second, I explore the idea of a minimally viable moment of measurement. There has been a trend over the last 20 years of shrinking self-assessment measures, and in this and related work we have arrived at moments of self-measurement requiring only 1-3 seconds. Is this enough time for meaningful cognition to occur, for the data received to accu-

rately reflect some or all of the construct in question? Third, in the vein of personalization common in our commercial technology and reflecting modern medical trends toward 'N of 1' study designs and interventions, I discuss the advantages and disadvantages of tailoring the self-report interface to the individual. I discuss several mechanisms (including three leveraged in the case studies) through which such personalization can be achieved, and describe the impact of such tailoring on across-subjects assessments. Fourth, I discuss why through my work across momentary measurement in three domains, I do not believe there is a one-size-fits all approach for developing novel self-reports for the EMA domain, and I contribute several design patterns useful to others continuing this work in new health spaces.

Bringing all this together results in the contribution of this dissertation: to glimpse inside the "black box of daily life" (Myin-Germeys et al., 2009), to support daily wellbeing by capturing the "little experiences of everyday life that fill most of our working time and occupy the vast majority of our conscious attention" (Wheeler & Reis, 1991) in brief and effective moments of measurement.

CHAPTER 2

SELF-REPORT, PSYCHOMETRICS, AND THE MOBILE USER EXPERIENCE

This work sits at the intersection of the measurement and self-report literatures, the ways that mobile devices afford in-the-wild self-report, and the user experience of tiny, momentary interactions with one's mobile device. Together, these domains result in moments of measurement to support wellbeing.

I begin by describing the value of tracking and self-report, and discuss the various methods researchers have turned to when requesting participant self-reports. I particularly highlight the role of mobile devices in serving as a data collection platform. I then speak to elements of measurement theory with a focus on psychometrics and the cognitive processes undertaken by a self-report respondent. I am particularly interested in very brief moments of self-report: I discuss some literature on the relative effectiveness of very brief and single-item measures, including those presenting non-text icon and graphical form factors, before investigating what the academy has to say about the frequent and repeated completion of survey items and how single-item reporting has been taken up on the web and in mobile apps.

I close this chapter with a short description of the user experience and usability literatures in HCI, and how scholars conduct research through the design and evaluation of such experiences. My focus is primarily on how HCI scholars speak to mobile experience and usability.

Literature common to all of the above can be found in this chapter. Literature specific to the understanding and measurement of each wellbeing construct (af-

fect, stress, and chronic pain) sits within the appropriate following case study chapter.

2.1 The practice *in situ* self-report

2.1.1 Value of self-report

Know thyself. In this dissertation I am concerned with naturalistic self-monitoring of affect, stress, and pain levels. Why? Beyond the fact that accurate self-knowledge is a generally beneficial quality (T. D. Wilson & Dunn, 2004), self-monitoring in the context of self- or clinically-managed wellbeing is a powerful tool in assessment, treatment monitoring, target behavior selection, and behavior change (Korotitsch & Nelson-Gray, 1999).

Gaining self-knowledge is not easy - people do not have complete knowledge about themselves or about many things that affect their lives (T. D. Wilson & Dunn, 2004). Learning more about internal, unobservable states such as affect, stress, and pain levels is potentially even more difficult. They are largely unobservable - for these types of phenomena, subjective experience "is the most objective datum there is, and that reducing it to a more objective standard only decreases its objectivity" (Hektner et al., 2007). As we will see, for each of affect, stress, and pain, domain experts quite literally consider the subjective self-report the gold standard ground truth - for example "pain is what the patient tells us it is" (McCaffery, 1979). Self-report for understanding our moment-by-moment experiences enables us to look inside the "black box of daily life" (Myin-Germeys et al., 2009).

Self-report in the assessment and treatment of wellbeing

In the health domain, ESM-style self-report has been deployed in a wide range of assessment and treatment functions (Korotitsch & Nelson-Gray, 1999). Self-monitoring may be used to support diagnosis, for example in clarifying sources distress by asking subjects to self-report in moments of anxiety and worry. Signal-contingent sampling may aid in target behavior selection, for example making clear if a depressed client experiences few pleasant events, a large number of unpleasant events, and/or suffers from dysfunctional cognitions. ESM-style self-monitoring may also provide insights into scenarios and antecedents of certain states or behaviors, for example revealing that in the afternoon and evening hours, bulimic participants are at greater risk for bingeing and purging when at home than when not at home, while the alternative is the case for non-bulimic overweight participants (Schlundt, W. G. Johnson, & Jarrell, 1985). Event-contingent self-report is extremely useful for understanding and titrating treatment protocols for a range of conditions and interventions.

While in many cases used initially for assessment and treatment monitoring, self-report can have indirect positive reactive effects (Korotitsch & Nelson-Gray, 1999), for example greater self-awareness and reflection (Baumer et al., 2012). The data collected also need not remain private to the participant and their clinician: multiple studies report on positive outcomes associated with the simple sharing of health-related self-reports, leveraging mechanisms such as social pressure (Consolvo, Everitt, Smith, & Landay, 2006), social accountability (Newman, Lauterbach, Munson, Resnick, & Morris, 2011), and social support (Naaman, Boase, & Lai, 2010; Adams, Baumer, & Gay, 2014).

2.1.2 Traditions of naturalistic self-report

Ecological Momentary Assessment (Shiffman et al., 2008), the style of brief *in situ* self-monitoring on which I focus, draws on several self-reporting traditions. Each supports an idiographic perspective, attempting to identify patterns of behavior within an individual across a population of experiences or situations (Conner, Tennen, Fleeson, & Barrett, 2009).

Diaries and journals

Written diaries were deployed in clinical research in the 1940s (Verbrugge, 1980; Allport, 1942) and major diary studies since have focused on activity and time use (Hochschild & Machung, 1989; Robinson, 1977). Participants are typically equipped with folders or booklets containing questionnaires, one for each diary entry, and instructed on how and how often (typically once daily, or at meal-times) to complete the measures. Studies using pen-and-paper diaries for data capture have long been considered the easiest technology for participants to use (Bolger, Davis, & Rafaeli, 2003).

Pen-and-paper diarying methods do suffer from several specific limitations, however. First is a question of compliance: a given day's questionnaire may not be completed due to honest forgetfulness or due to the participant not having the booklet with them. This issue is made worse should the participant choose to complete missed entries relying on reconstruction or fabrication, introducing biases that defeat the purpose of diaries (Bolger et al., 2003). Second is an issue of data access: participants' reports are in pen-and-paper booklets, available to the researcher for analysis only after the booklet has been returned and the

data transcribed. Depending on the scale of the study, there may well be large amounts of data to manually handle and transform. This can result in missing data (booklets not returned) or invalid data (should data be transcribed incorrectly). Third is that metadata around the act of responding is not available without additionally burdening the respondent for manual entry. Such data might include accurate timestamps of when and for how long the response was made, or the physical or social contexts in which the report is being made.

A novel twist on diarying methods involves understanding users' activity (status updates, profile curation, etc.) on new media platforms such as Facebook or Twitter as diarying - for example, see (Humphreys, Gill, Krishnamurthy, & Newbury, 2013) for a historicizing of diarying in new media. Rather than prompting for self-report along specific dimensions, researchers have recently used content analysis (automatic or otherwise) to mine the entirety of these modern 'diaries' to infer personality traits (T. Ryan & Xenos, 2011) and attachment styles (Kang et al., 2015), and there is every indication that such inferences accurately reflect actual personality (Back et al., 2010).

Behavioral self-monitoring

The behavioral perspective has long history of self-monitoring, including both simple counts of relevant events and actions (Lindsley, 1968) as well as the self-report capture of the context in which the participant finds herself (Schlundt et al., 1985). Behaviors of interest varied widely, from daily hassles and uplifts (DeLongis, Folkman, & R. S. Lazarus, 1988) to pleasant activities' impact on mood and depression (Lewinsohn & Graf, 1973) to each and every of an individual's social interactions (Reis & Wheeler, 1991).

Extending the tradition of Taylor's activity sampling for the purposes of improving efficiency in the workplace (F. W. Taylor, 1911), Burns asked workers to use a pen-and-paper form to record information about each communication throughout the day, including the initiator and recipient, method of communication, and the location of any communication away from the recorder's desk - the forms were mailed to the Burns each day (Burns, 1954). Using a custom-built electronic random interval generator attached to a hearing-aid earphone that served as an early beeper, Hulbert prompted participants to report what they were thinking in the moment (Hurlburt, 1979); similarly, Klinger used signal-contingent sampling to access the types of thoughts and imagery respondents experienced throughout the day (Klinger, 1978).

Experience Sampling and Ecological Momentary Assessment

Perhaps the best known formalization of *in situ* repeated self-report is Experience Sampling (ESM) (Mihaly Csikszentmihalyi & Reed Larson, 1987), a method that collects information about the content and context of daily life. ESM "combines the ecological validity of naturalistic behavioral observation with the non-intrusive nature of diaries and the precision of scaled questionnaire measures" (Hektner et al., 2007), prompting self-report most commonly at pre-determined random intervals throughout the day. "One way to think about ESM is that it tries to provide psychology with a method and a theoretical perspective that makes it possible to study whole individuals functioning in their everyday environments, both as behaving organisms and as acting, conscious beings." (Hektner et al., 2007).

Three types of sampling exist (Reis & Gable, 2000; Wheeler & Reis,

1991). In *interval-contingent* sampling, participants complete self-reports at predetermined intervals for the duration of the study; this could be hourly, daily, or at some interval in-between. In *event-contingent* sampling, participants complete self-reports when a pre-designated event occurs - for example, after every change of location or after each social interaction. *Signal-contingent* sampling has participants complete self-reports when prompted by a randomly timed signal. According to (Wheeler & Reis, 1991), the first researcher to ask subjects to record at random times pre-selected throughout the day was Hinrichs in the 1960s. A computer pre-selected five random points in time for each of 11 workdays, and respondents were equipped with a daily schedule and an alarm watch; after each report about their communication activity, they set the watch alarm for the next scheduled time (Hinrichs, 1964).

Scollon presents a strong discussion of the five major strengths and four potential weaknesses of the ESM method (Scollon, Prieto, & Diener, 2009). The strengths are identified as:

- ESM allows us to understand the contingencies of behaviors/states.
- ESM increases the ecological validity of psychology by taking it out of the lab and into natural environments.
- Due to repeated measurement of each individual, ESM allows for the study of within-person processes.
- ESM avoids several of pitfalls of traditional self-report such as recall and global retrospective biases.
- ESM enables researchers to adopt multi-method strategies.

The potential weaknesses of the method include:

- Participant selection issues: given the relatively intensive nature of an ESM study, there are likely self-selection and attrition biases, large motivation-to-participate variances, and issues relating to familiarity with the reporting devices (decreasing over time as more ESM studies leverage the phones people are already carrying).
- Situation issues: response rates tend to decline after 2-4 weeks, and also tend to be lower in the evenings, when at home, and in places where signaling or response either cannot take place (e.g. when driving or swimming) or would be disruptive (e.g. in a lecture or church). An associated issue is the lag between signal and response, which can vary from several minutes to the case in which participants complete all responses at one time at the end of the study.
- Reactivity issues: repeated assessments can lead participants to become unusually aware of their behaviors or states, likely those of interest to the researcher, and may result in participants modifying their behaviors accordingly. It is also possible that repeated assessment of the same construct may influence the recall or self-assessment itself; not enough is known about the effects of ESM on participants' subjective experience.
- Data issues: ESM generates a large amount of noisy, incomplete, and irregularly spaced data that may be impacted by a variety of event- and time-dependencies (for example, day of week or circadian rhythms). Statisticians are yet to develop all the methods necessary for the rigorous analysis of these data (John Bunge, personal communication, 2013).

Marrying the diverse traditions of self-report with the value of the smartphone for self-monitoring is a modern collection of data collection methods

called Ecological Momentary Assessment, or EMA (Shiffman et al., 2008). It encompasses electronic device-based ESM as well as passively sensed data streams and other techniques under one common higher-order framework illuminating common features. Features common to EMA approaches are that:

- Data are collected in natural settings, in the real world, as participants go about their daily lives.
- Assessments focus on participants' current, in-the-moment state, rather than relying on recall.
- Assessments tend to be as brief and as unobtrusive as possible, both to maintain ecological validity and so as not to overly burden participants.
- Moments for assessment are strategically selected (typically using time-, event-, or signal-contingent strategies).
- Subjects complete multiple assessments over time.

2.1.3 The value of mobile devices for self-monitoring

Increasingly, researchers and clinicians are moving toward using the mobile phone as the data collection platform for a variety of tasks, including *in situ* self-report. Smartphones are particularly valuable in this role for four primary reasons (Klasnja & Pratt, 2012):

Powerful smartphones are increasingly pervasive

As of April 2015, according the Pew Internet Research Project, a full 90% of American adults have a cell phone, and 64% have a smartphone (Anderson,

2015). It is estimated that by 2020, 80% of the adults on earth will have a smartphone - essentially, everyone is or will be carrying a *de facto* pocket supercomputer (Evans, 2014). People use their phones for a wide variety of tasks all throughout the day. The Pew Research Center's surveys indicate that in 2014, 97% were texting, 75% were social networking, 47% played games, and 62% were looking for health and medical information online (Anderson, 2015). Bohmer et al tracked low level usage statistics of 4,100 smartphone users for four months, and confirms that mobile phones are still largely used for communication (voice and text) (Böhmer, Hecht, Schöning, Krüger, & Bauer, 2011). They find that some apps (for example music and social) show intense spikes in usage whereas other apps are used more broadly throughout the day. Bohmer further reports that the more someone is using their smartphone, she spends less time with each application - and that short sessions with only one app are much more frequent than longer sessions with two or more apps. All that time adds up, however, for while at a time each app is used for only 72 seconds (with great variability over different types of applications), mobile devices are used for almost an hour a day.

People carry their phones with them everywhere

44% of US adults sleep with their phone beside them so they do not miss calls, texts, or other updates, and 46% of cell phone owners describe their devices as "something they can't imagine living without" (Center, 2014; Anderson, 2015). Fogg tells us that we spend more time with our mobile devices than we do our partners or spouses (Fogg & Eckles, 2007), and indeed there is evidence that the overall adult population keeps their phones within arms' reach 50% of the time

(A. K. Dey et al., 2011). Falaki et al report that users interact with their phones 10-200 times a day (Falaki et al., 2010), and (Lookout, 2012) show that 58% of adult smartphone check their phones at least once every hour. Of note is that Oulasvirta et al suggest that 18-35% of these times users interact very briefly with the phone to check (for) dynamic content (Oulasvirta et al., 2012).

Smartphone users are very attached to their devices

As a function of the above habits, the ability to customize the interaction experience, and most importantly the ways the phone connects us with social others, people's relationships with their smartphones is often deeply personal. (Venta, Isomursu, Ahtinen, & Ramiah, 2008). Phones often contain personal information, intimate messages, pictures of family and friends, financial tracking, and calendaring tools. Such positive personal attachments should lower barriers to their use in personal health applications (Klasnja & Pratt, 2012).

Smartphones can provide context

The modern smartphone includes a large number of sensors and applications that can provide a great deal of data about a user's current situation. For example, GPS, WiFi fingerprinting, calendaring, and the accelerometer users' precise location and activities can be recorded. Increasingly, connections with wearable sensors and activity trackers are augmenting these inferences. Such context can be valuable in both diagnostic assessment and healthful intervention.

2.2 *In situ* self-report: the moment of measurement

As this work is about the measurement of unobservable human states, it is important to briefly mention here the relevant aspects of the literature on asking questions to measure subjective experiences and attitudes.

At a fundamental level, there are two primary concerns in measurement: representation and uniqueness (Suppes & Zinnes, 1963). The first, representation, concerns the philosophical problem of using numerical values to describe objects in the natural world, as well as proving that each given conceptual domain can be meaningfully cast in terms of numerical relations and operations. The second, uniqueness, deals with the problem of determining what type of measurement scale results from the measurement procedure. Suppes and Zinnes identify three broad categories of scales: absolute (counts), interval (values are ordered and equally spaced through the scale), and ratio (a scale with an absolute zero, thereby permitting ratio comparisons). We also tend to describe two further sorts of scales: nominal (a non-numerical set of category labels) and ordinal (an ordered set of labels, with a direction, such as low/medium/high or a 0-10 scale).

The measurement of internal state through self-report has its basis in psychophysics, the subjective judgment of stimuli that can be measured objectively on physical scales. The earliest known work in this space is that of Fechner who studied 'just noticeable differences' of objectively different magnitudes of stimuli such as weight (Fechner, 1860). Applying psychophysics to human qualities such as attitudes or intelligence for which there is no physical scale is the domain of psychometrics, and required an acceptance among science researchers

of using subjective judgments as a valid approach to measurement.

While observation (systematic and otherwise) of human traits and interactions are likely as old as humankind, it was not until the rise of applicable statistical methods in the 19th century (thanks to scholars such as Galton, Pearson, and Spearman) that formal measurement in psychometrics became possible; one well-known early example is Binet's work on ability testing in the early 20th century. During World War II, screening of the mental and physical fitness of large numbers of recruits, at scale, further motivated the development of standard and rigorous methods and measures in psychometrics, which then formed the basis of psychological wellbeing measures in the postwar years (McDowell, 2006).

As psychometric scales became more prevalent, researchers developed robust descriptions of reliability and validity that could be used to assess the appropriateness and utility of any given scale. A scale is reliable if, assuming the underlying true value in question has not changed, the measure consistently provides the same result over and over.

There are multiple aspects of validity to which psychometrics attends. While it has not been widely adopted, I draw on a six item framework that explicitly considers the characteristics of pyschometric measures to be "social values that have meaning and force outside of measurement whenever evaluative judgments and decisions are made" (Messick, 1995). Validity, therefore, might best be considered an evaluative summary not just of the evidence for the score, but additionally for the actual and potential consequences of the score interpretation. The six aspects of construct validity addressed are:

- **content**: evidence of content relevance, representativeness, and technical quality: how well does the measure representatively cover and reflect the construct content domain?
- **substantive**: theoretical rationales for consistencies in the responses, including process models of performing the task - and "evidence that the theoretical processes are actually engaged by respondents in the assessment task"
- **structural**: the fidelity of the scoring structure to the structure of the construct
- **generalizability**: the extent to which a measure's properties and interpretations generalize across population groups, settings, and tasks.
- **external**: includes convergent and discriminant evidence as well as criterion relevance. Convergence suggests that the results of scales that measure similar constructs should correlate more closely than than scales measuring less related theoretical constructs. Discriminant validity is the absence of correlation between measures of unrelated constructs. Criterion relevance empirically assess whether the measure in question associates well with known 'gold standard' measures or other outcomes of the behavior/attitude/ability in question.
- **consequential**: the value implications of score interpretation as a basis for action.

Related to the idea of construct validity is face validity, an estimate of the extent to which a measure appears to evaluate the construct in question. Particularly in situations where one can assume a cooperative respondents and where the construct in question can be represented in plain language (or other ways

approachable by the respondent population), simply asking self-report questions with high face validity is recommended (Burisch, 1984). Times when scales with lower face validity, or indeed non self-report scales, should be used include those situations where respondents can see a clear mapping between their responses and the treatment they will likely receive - a example of Messick's consequential validity.

2.2.1 Question structure and wording

So far I have considered, for the types of self-report measures in which I'm most interested, what the measures attempt to quantify (unobservable human state) and a description of how we can assess how well they do their work (hopefully, reliably and validly). In this section, I look at the questions within the self-report measures themselves. While we often view self-report scales solely as measurement devices to elicit information from participants, we forget that respondents draw on a great deal of information from various aspects of the measure's questions to determine their task and answer the questions - to a great extent, which questions and the ways we ask them determine the answers we receive (Schwarz, 1999; Krosnick & Presser, 2010).

There is a huge variety of possible question types. Questions can assess constructs directly or indirectly, be open- or closed-ended, be required or forced-choice, and be influenced by the questions that appear before or after them in the measure (Schwarz, 1999; Fabrigar, Krosnick, & MacDougall, 2005; Schaeffer & Presser, 2003). Within the realm of rating scales, measures can be unipolar or bipolar; take the form of likert, equal-appearing intervals, or semantic differ-

entials; present endpoints and/or midpoint labels and/or interval descriptors; be single- or multi-item; use verbal, numerical, or graphical features (Krosnick, Judd, & Wittenbrink, 2005; Schaeffer & Presser, 2003). In general, researchers who study the science of asking questions lean toward more straightforward, direct, and concise measures over a more axiomatic, representational, even elegant approach (Krosnick et al., 2005); survey research has found that decompositional approaches, which ask direct questions about the smallest possible units, are superior to open-ended, global questions, which invite heuristic processing (Menon, 1997; Reis & Gable, 2000).

Questions can have a variety of reference periods. In this work, I attend to in-the-moment responses, but even given this more narrow constraint, ESM and EMA studies have considered such a reference period to include everything from 'right now' through 'in the last hour' to 'over the day'.

Measures can include or require a great deal of participant training, or be designed to (hopefully) be intuitive and straightforward to pick up and begin answering.

2.2.2 Cognitive process of responding to self-report items

When a respondent answers a question, the accuracy of the obtained data is dependent at least partly on how well the respondent performs the required cognitive tasks (Vannette & Krosnick, 2014). Researchers have settled on four basic component steps involved in cognitively responding to a self-report question item (Tourangeau, Rips, & Rasinski, 2000):

1. comprehension: the respondent interprets the intended meaning of the question, representing the logical form, identifying the question focus, and linking key terms to internal concepts;
2. retrieval: the respondent searches their memory for all related information, accessing specific and generic memories, and fills in missing details;
3. judgment: the respondent summarizes and integrates the retrieved information, drawing inferences and making assessments;
4. response: the respondent maps the judgment(s) onto a response category or value, editing the response as necessary.

Proceeding carefully through these stages is known as optimizing; optimizing is influenced by a variety of respondent motivations and biases that may impact the resulting reported value(s).

Krosnick and others contribute a processing framework that describes attitudinal evaluation and response consisting of three steps (Krosnick et al., 2005).

1. Spontaneous evaluation phase: in this phase, an attitude object or its symbolic representation may provoke evaluations automatically, potentially without conscious awareness. This phase is the result of a passive process that likely triggers differentially for different constructs depending on attitude strength or experience frequency and recency.
2. Deliberative evaluation phase: a controlled memory search for stored evaluations and related associations followed by a deliberation and reflection on the construct and the respondent's evaluation of it. Respondents are likely to deliberate more carefully depending on their motivation (e.g. an internal need for accuracy or perceived outcome on the respondent)

and opportunity: the construct in question can be consciously evaluated, and the cognitive resources required are available.

3. Response phase: The evaluations generated automatically or deliberately shape the response either implicitly (the respondent may be unaware of the evaluative processing) or explicitly following deliberate consideration of the two previous processing stages. The respondent may also engage in a variety of metacognitions at this point, shaping or correcting the eventual response through perceived biases, self-presentation motivations, or resulting impact or outcome to the respondent.

The two evaluative phases presented by Krosnick have a close association with systems 1 and 2 of dual-process theory. System 1 is a quick and intuitive reasoning with close ties to emotion, formed by habits and hard to change, while system 2, perhaps incorporating parts of the last two of Krosnick's phases and certainly the third of Tourangeau's, is a slower reasoning process that is 'rational', analytical, and subject to conscious judgments and attitudes (Kahneman, 2011).

2.2.3 Brief or single-item questions and their repeated use

Of particular interest to this work are very brief measures that fit into tiny moments form measurement. This interest reflects an ongoing trend in psychometrics to the point where what once seemed radically short can now seem tediously long - for example, the PANAS is a 20 item measure that in 1988 was introduced as "a brief measure of positive and negative affect" (Watson, L. A. Clark, & Tellegen, 1988). "The demand for super-short measures is growing...

examples of this trend toward minimal measurement are the single-item self-esteem scale (Robins, Hendin, & Trzesniewski, 2001), single-item ability ratings (Rammstedt & Rammsayer, 2002), and even a 10-item measure of the Big Five (Gosling, Rentfrow, & Swann, 2003)" (Rammstedt & John, 2007).

Of primary concern when using such abbreviated measures must be, are such measures still effective? One of the first to test the assumption that the trade-off between complex measures providing full coverage and depth of understanding and brief measures that are less burdensome is decreased effectiveness was Burisch. He first found that shorter scales are at least as valid as longer scales on average, even when the shorter scales were simply subsets of the longer scales, and that simple scales were as valid as more sophisticated scales reflecting "psychodynamic theorizing or elaborate multivariate approaches to scale construction" (Burisch, 1984). Returning to this theme later, Burisch found that not only did extremely short scales retain respectable psychometric characteristics, but that lengthening a scale can in fact weaken its validity (Burisch, 1997). Short-form versions of traditional self-report scales continue to reflect these findings, for example a 10-item version of the Big Five Inventory that is completed in under a minute (Rammstedt & John, 2007).

Brief self-report measures may be deployed in EMA-style studies, and therefore are completed by each participant many (even into the hundreds) of times. One positive outcome of frequent and repeated measurement is that there is no need to reintroduce participants to the measure - the participant quickly learns to complete it and (at least with electronic device-based reporting) therefore exhibits a decreasing completion time (Stone & Broderick, 2007).

On the other hand, frequent and repeated use of even the shortest scales may

result in negative outcomes too. One example would be respondent fatigue, which occurs when "individuals respond to survey questions but do not provide truthful or consistent responses in order to reduce the burden of answering questions" (Egleston, S. M. Miller, & Meropol, 2011). Similarly, some EMA studies report response fatigue (a trailing off of completion rates over time) - and importantly, the presence or absence of response fatigue is not commonly reported (S. C. Reid et al., 2008).

2.2.4 Single-item assessments in modern media

There has been some attention to very brief self-reports in the world of recommendation systems, where user ratings of collection objects (such as movies, books, etc.) improve not just the systems' ability to recommend better content for the reporting user, but to learn about the objects themselves at an aggregate level. Persuading users to perform many repeated ratings without them getting bored is essential (Swearingen & Sinha, 2002), and of great value to companies such as Amazon, Netflix, Yelp, and more. Harper and others have modeled user motivations to rate as a cost/benefit trade-off where the benefit is improvement in ratings and fun, and the cost is time (Harper, Li, Chen, & Konstan, 2005).

Cosley and others investigated users' satisfaction and rating consistency when rating movies with three different types of scales: a binary scale, a plus/minus 3 scale with no zero, and a five star rating scale (with half-star increments). Users preferred the star rating scale, and expressed a preference for greater granularity as it allowed them to be more accurate in their ratings. Extending Harper's work, Sparling and Sen looked at users rating with four dif-

ferent rating scales (unary, binary, five-star, and a 100-point slider), comparing time taken to rate, cognitive effort, and user satisfaction with each scale (Sparling & Sen, 2011). The found that in general, users have to work harder with more granular rating scales, although these effects are moderated by the rating domain (e.g. movies or product reviews). Furthermore, rating time varied significantly depending on whether the item in question was liked or disliked, and users spent more time assigning ratings at the middle of the scales compared to the extreme ends of the scales. As with Cosley, users preferred the five star rating scale although the binary thumbs up/down scale was a close second.

Other examples of very brief self-report signals include interactions such as 'likes' on Facebook[1], 'hearts' on Instagram[2], and 'favorites' on Twitter[3].

In the mobile environment, researchers are seeking to leverage microinteractions users are already taking in order to solicit either self-reports (Truong, Shihipar, & Wigdor, 2014) or crowdsourced micro-tasks (Vaish, Wyngarden, Chen, Cheung, & Bernstein, 2014). In both these works, the standard 'slide to unlocks' paradigm of the Android lockscreen is replaced by a 'report to unlock' model, where answering a question unlocks the device. The interaction is either tap-and-submit, or even more clever, the 'slide to unlock' interaction is replaced by a slider widget (a visual analog scale in the language of psychometrics) such that you 'slide to report *and* unlock'.

Other examples of leveraging small moments of self-report, not in the EMA tradition but increasingly common in both research and 'in the wild' are the single or short surveys in Google Consumer Surveys[4] and the Amazon Mechanical

[1]http://facebook.com
[2]http://instagram.com
[3]http://twitter.com
[4]http://google.com/insights/consumersurveys

Turk[5] systems.

2.3 (Designing for) the mobile user experience

2.3.1 User experience

The HCI field commonly distinguishes between 'usability', a narrow definition typically referring to the ease of use of the design features, especially around a task, and 'user experience' which includes the user's thoughts and feelings about their interactions with the device (J. Heo, Ham, Park, Song, & Yoon, 2009), as well as involving (potentially complex) interactions among social, locational, task, and user state contexts. Defining user experience is not easy; reporting on a survey of 275 researchers and practitioners, Law and others show agreement around the idea that user experience is dynamic, context-dependent, and subjective (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009).

2.3.2 User-centered research through design

Modern HCI researchers and designers take a user-centered approach. There are a (sometimes bewildering) range of user-centric iterative design and evaluation strategies in HCI and related domains: participatory design, user-centered design, activity-centered design, co-creation, co-design, bodystorming, and more (Muller & Kuhn, 1993; Oulasvirta, Kurvinen, & Kankainen, 2003; Gay, 2004; Sanders & Stappers, 2008; Zimmerman et al., 2011). The common thread

[5] http://mturk.com

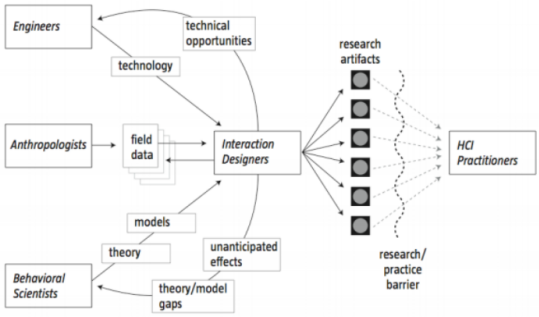


Figure 2.1: An illustration of the pathways and deliverables between and among Interaction Design Researchers and HCI Researchers (Zimmerman, Forlizzi, & Evenson, 2007).

is the idea that in some important way, the end-user of the designed object has a role in the creation of the use of the object, implicitly or explicitly, and that this role should be at least recognized, if not intentionally foregrounded and brought ever earlier into the design process. Typically this involves an iterative process (Gay, 2004).

Extending Frayling's work, Zimmerman and others describe what it means to conduct *research through design*: how researchers can engage 'wicked problems' and generate design artifacts that transform the world from its current to a preferred state (Frayling, 1993; Zimmerman et al., 2007). This design-based process begins with *analysis* which concerns how things currently are- the truth; moves through *projection*, concerning how things could be- the ideal; and results in *synthesis*, speaking to how things will be- the real (Jonas, 2006).

Zimmerman presents both a framework outlining the collaborative pathways between researchers, designers, and practitioners (Figure 2.1) as well as four lenses for the evaluation of an interaction design research contribution. In presenting these lenses, Zimmerman both contrasts the style and value of research through design contributions to those in engineering or behavioral sci-

ences, and exhorts interaction design researchers to articulate with greater and more clear rigor their contributions to the research and practitioner communities:

1. Process: while in design there is no expectation that reproducing the process will reproduce the results (in sharp contrast to base assumptions in the scientific method!), rigor in reporting the process and the rationale for selecting enables reproduction of the design process.
2. Invention: research through design should result in significant contributions not just to the HCI research community, but the research articulation should give rise to technical and situated opportunities in the domain(s) of interest.
3. Relevance: In contrast to behavioral science or engineering where validity can often be a function of proving hypothesis or increasing performance, research through design moves the conversation from what is true to what is real. Outcomes should be articulated such that, in an almost anthropological way, it is clear why and how the design in the real world is relevant and moves the world into a preferred state.
4. Extensibility: a discussion as the extent that a design and/or its process can be built upon, either employing the process in approaching future problems or leveraging and understanding the knowledge created by the resulting artifacts.

In research through design, "research is guided through design process logic and design is supported and driven by phases of scientific research and inquiry" (Jonas, 2007).

Torsi et al have presented findings on the early stages of interdisciplinary user-centered design research as applied to the self-management of chronic illnesses (Torsi, Nasr, Wright, Mawson, & Mountain, 2009). Drawing on both Activity Theory and Goals in Health and Social Care, the authors contribute an understanding of patient experience using the metaphor of illness as a journey. Among other findings, they highlight for future work the high level of heterogeneity among potential users with chronic illness as a challenging design implication.

2.3.3 Usability of mobile devices

An evaluation of the usability of a system running on a mobile device is typically conducted just like the traditional evaluations of desktop software: in the laboratory setting. Increasingly, however, attention is being paid to evaluating the usability of systems in the context-of-use. This is a paradigmatic shift, from artificial settings toward conducting field trials in natural, every day settings - requiring new and hybrid methods (Alshehri & Freeman, 2012). The standard elements of usability evaluation remain, but are applied in the natural settings; researchers are finding this process well worth the additional hassle (Mack & J. Nielsen, 1994; C. M. Nielsen, Overgaard, Pedersen, Stage, & Stenild, 2006).

Mobile devices also introduce novel types of interactions. Increasingly, we're seeing a growth in gestural interactions (swipes, taps, pinches, and more) that are often OS-specific (G. Inc, 2014; A. Inc, 2014). While experienced users who have developed their skills through routine use, media following, and ad hoc problem solving situations are comfortable with these interactions (Oulasvirta,

Wahlström, & Anders Ericsson, 2011), there is some concern among usability researchers that gestures represent a backwards step in usability (Norman & J. Nielsen, 2010).

Usability of mobile devices in health data capture

The fact that people have their phones with them all the time, and are using them with great regularity makes them the perfect vehicle for real-time in-situ data capture. The implications of the mobile phone as a research platform are immense (G. Miller, 2012). To leverage this ability, and to compete with all the other app usage on a typical smartphone, any self-report data capture interface must be brief, usable, intuitive, and ideally, in some way fun - in addition to being a valid measure of the physiological feature in question. Stinson et al discuss four features of usability for self-reporting on an mobile device, in this around chronic pain in teenagers (Stinson, Petroz, et al., 2006). First, and especially because the self-report will be taking place outside a lab or clinical setting, the interface must be learnable and easy to start using. Second, and perhaps of ultimate importance for a long-running data collection, reporting must be and feel efficient. It cannot feel that it gets in the way. Third, the measures should be tested for all manner of errors - several of Stinson et al's participants failed to move a VAS-P slider all the way to the left end when attempting to report 'no pain'. Lastly, as much as possible the interface should be satisfying, even pleasurable to use.

2.3.4 The microinteraction

Many of these behaviors are small interactions or a sequence of such interactions. A microinteraction is a small piece of functionality that allows you to complete a single task: setting your phone volume to vibrate or logging in to a service are examples. Microinteractions can be forgettable or delightful. Effective microinteractions can provide great utility, such as intelligent autocomplete when searching with Google or typing on a mobile device, and also delight, as with the playful and deeply satisfying double-tap to 'heart' an Instagram photo. The tiny signal reaching the recipient of such a 'heart' is quite powerful in providing social connection, support, and validation. A microinteraction consists of four components (Saffer, 2013):

1. A trigger that initiates the microinteraction. Triggers can be manual (a user flipping a light switch) or automatic based on events or context (an email arrives, or the user walks into a scheduled meeting).
2. Rules that determine how the microinteraction functions. Given the trigger, what happens next? After a user has flipped a light switch, the light will stay on and fully lit until the switch is turned off. Rules can become more complex, such as dimming the light over time if no motion is detected in the room.
3. Feedback generated by the rules. Anything the user sees, hears, or feels that helps them understand the system state and, in particular, the rules, count as feedback. The light turns on and the switch stays in the 'on' position, or an 'unread' counter increments on the email application's icon.
4. Loops and modes making up the microinteraction's meta rules. Rules may be updated over time with repeated use (such as a 'buy now' but-

ton become 'buy another' if the user has previously purchased the item in question), or there may be special modes (such as 'settings' on a weather widget).

These components sit well within Norman's seven fundamental design principles for interactions: what do I want to accomplish, what are the alternatives, what action can I do now, how do I do it, what happened, what does it mean, have I accomplished my goal (Norman, 2013).

The effective design of microinteractions for ESM-style self-report on the mobile phone should do more than simply collect self-report data. They should assist the user/participant in condensing a complex psycho-physical experience into an actionable (clinical) form. In the space of mobile text-entry, an inviscid entry rate is one at which the bottleneck is not the text-input interface, but the user's creation of what to type (Kristensson & Vertanen, 2014). Certainly any ESM measures should be at least inviscid - the process of actually entering self-report should not take any longer than the cognitive process of self-understanding - and indeed, ESM measures can and should be designed to support this cognitive work. This ensures such measures are truly as brief and non-intrusive as possible.

We can also see the points at which the psychometric and EMA methods map onto the microinteraction and mobile usability literature and where there is opportunity for exploration. For example, initiating triggers map well onto time-, event-, and signal-contingent sampling schedules, while the meta rules speak to a tailored and responsive reporting interface that at first glance may conflict with measurement values such as consistency and standardization. These issues are unpacked in chapter 6.

CHAPTER 3

CASE STUDY: THE MEASUREMENT OF AFFECT

Humans are emotional beings. Indeed, as Lazarus and Lazarus have it, "Everything important that happens to us arouses emotion" 1994. More than being one way we experience what happens to us, "the neurological evidence indicates emotions are not a luxury; they are essential for rational human performance" (R. W. Picard, 2000) and are "a vital tool for getting along in the world" (R. S. Lazarus & B. N. Lazarus, 1994). "We feel before we know, and in an important sense, feeling determines what we know" (Buck, 1986) - in a very real way, "emotions, motivations, and cognitive processes coexist and contribute to experience in every moment of our life" (Hektner et al., 2007).

It is useful, therefore, to be able to capture and assess the emotions that people experience - perhaps particularly so as it becomes clear that positive affect is associated with positive health outcomes, including lower morbidity and decreased symptom and pain experiences, and is likely also a buffer to experiencing negative stress (Pressman & Cohen, 2005).

Affect is traditionally assessed with fairly long questionnaires (e.g. the Positive and Negative Affect Schedule (Watson et al., 1988)) that are typically administered only at the start and/or end of a study or protocol. However, emotion varies over time (Hektner et al., 2007; Scherer, 2005) and beyond its own biases, recall in such settings may well be closer to experience reconstruction than memory retrieval and subject to various other biases at the moment of recall (Stone & Shiffman, 2002; Schwarz, 1999). We therefore want to assess affect *in situ* using an Experience Sampling or Ecological Momentary Assessment method. "The only way to capture emotions... is by asking people to describe

them in the moment they occur" (Hektner et al., 2007).

In this case study, I describe the development and evaluation of the Photographic Affect Meter (PAM), designed to provide a solution to this problem. Much of this work have been published in the Proceedings of the Human-Computer Interaction Conference (Pollak et al., 2011) and in the Journal of Diabetes and Technology (Gay, Pollak, Adams, & Leonard, 2011).

3.1 How affect is understood

Affect is important, then. But what exactly is affect? What is an emotion? Researchers have taken a myriad of approaches in searching for an answer - psychological, ethological, sociological, developmental, neurobiological, anthropological, evolutionary, and more - and for many, the question remains open (Scherer & Ekman, 2014). Indeed, "the number of scientific definitions proposed has grown to the point where counting seems quite hopeless" (Scherer, 2005).

The issue is further complicated by different understandings of the relative definitions of 'emotion', 'affect', 'feeling', 'mood', and other affective states or traits. For Scherer, emotion is the comprehensive construct with subjective feeling as one part; for many other researchers, Scherer's subjective feeling is what they mean when they speak to emotion or feeling. 'Mood' is a more muddy situation; for some moods and emotions hold equivalent meaning; for others, a mood is an affective state without a clear cause, or that is unrelated to an event, and as such is more difficulty to fit within their definition of emotion and emotional response. Following both Ekman and Scherer, many emotion researchers distinguish these constructs as separate from preferences (relative stable evalu-

ative judgments) and attitudes (enduring beliefs and predispositions) (Ekman, 1992; Scherer, 2005). In this work, and as is common in the HCI and affective computing literature, I use affect, emotion, mood, and feeling all to mean the subjective and experienced quality of an affective state.

In this section, I draw largely on affect as defined in psychology and report on three main schools of thought. First, a component process model that attempts to capture the most comprehensive picture of the biological and psychological aspects of emotion. Second, the idea that there exist a relatively small subset of discrete emotions, common to all humans, that (in different ways among different scholars) can be composed. And third, dimensional models of affect in which any emotion can be described as a functional combination of two or more dimensions such as valence or control. Finally, I discuss the *state* and *trait* aspects of emotion.

3.1.1 A component process model of affect

Scherer in 2005 provides great detail on a component process definition of emotion and feeling. He sees an emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus as relevant to major concerns of the organism" (Scherer, 2005). The five components, and their associated biological subsystems and functions in emotion, are as follows:

1. Cognitive component: taking place in information processing centers, this component is about the appraisal of a stimulus concerning its relevance to the organism.

2. Neurophysiological component: taking place in the nervous and neuroendocrine systems, this component concerns bodily systems and system regulation.
3. Motivational component: taking place in the central nervous system, this executive component concerns the preparation and direction of action.
4. Motor expression: the somatic nervous system controls the communication of reaction and behavioral intentions, for example through facial and vocal expression.
5. Subjective feeling component: a monitoring of internal state and interaction with the environment taking place in the central nervous system, this component is that of the emotional experience.

This comprehensive multi-modal component process perhaps most completely describes an emotion event. The fifth component, the subjective feeling component, is that which other (psychological and sociological) models attend to most closely. Particularly in self-report, it is this subjective quality that is described by the words 'affect', 'emotion', and 'feeling', rather than speaking to some of the more objective and biological responses characteristics in the other components of the model.

3.1.2 Discrete, basic emotions

When conceptualizing the space of the human emotional experience, various researchers hold that, as a function of evolutionary processes, humans have a discrete and finite set of basic emotions (Tomkins & Carter, 1964; Ekman,

1992). These basic emotions amplify and drive states of the organism, optimizing for survival, and each is generated from a separate neural pathway. Evidence for the basic emotion position includes the universality of the basic emotions across cultures, as well as strong mappings from basic emotions to specific facial expressions (Ekman, 1992; Posner, Russell, & Peterson, 2005). While Ekman considers these basic emotions to be fully independent, other basic emotion theorists have arranged the basic emotions into bipolar relationships, such as joy/sadness or disgust/trust (Plutchik, 1980). The set of basic emotions does slowly change over time, typically growing longer.

How do basic emotion theorists account for those emotional experiences that are not perfectly described by one of the basic emotions - for example, fondness or hate? First is the idea that the basic emotions are really 'prototypes' or good descriptors of 'emotion families': other feelings can therefore be described as variations on a theme or nested within a tree structure (Ekman, 1992; Shaver, Schwartz, Kirson, & O'connor, 1987). Others of these feelings can be rejected as instead understood as moods (such as euphoria or irritation), attitudes (such as hatred), or traits (such as hostility, melancholy, or Pollyanna-ism) rather than poorly described emotions (Ekman, 1992). Or, some of what we consider 'emotions' are complex experiences consisting of actors, basic emotions, and contexts called plots: grief for example is not just a lot of sadness, but is an experience made up of the individual and the object(s) of their loss, and the state of loss (Ekman, 1992). It is also possible that some emotional experience are complex blends of two or more prototypical emotions (Shaver et al., 1987).

Bridging into the next section (dimensional models of affect) is Plutchik's three dimensional wheel of emotions (Plutchik, 1980). Eight basic emotions are

arranged into the bipolar relationships described above, but then each of the eight can be experienced at different intensities (a little joy is serenity, much joy is ecstasy).

3.1.3 Dimensional models of affect

Increasing numbers of researchers reject the idea of basic, discrete, prototypical emotions. Noting that individuals have difficulty in assessing, identifying, and describing their own emotions (Saarni, 1999), these researchers "rather recognize emotions as ambiguous and overlapping experiences" and embrace, rather than obscure, the intercorrelations among emotions (Posner et al., 2005).

The question of what these emotional dimensions are, and how many there may be, continues to be debated. In the two-dimensional arena, researchers have described emotions along the continua of positive and negative affect (Watson et al., 1988), approach and withdrawal (Lang, Bradley, & Cuthbert, 1998), valence and arousal (Russell, 1980), and others (Posner et al., 2005). A two-dimensional structure is common (good models are simple models), although other dimensional models, including the first (Wundt, 1874), present a third dimension, most often describing either a sense of intensity or dominance, or the degree of control over the experienced emotion (Mehrabian & Russell, 1974).

With a dimensional model, any given emotion can be placed within the two- or three-dimensional space described by the axes. For example, refer to Figure 3.1 to see 28 emotions arranged along the dimensions of valence (left/right) and arousal (up/down) in a circumplex model (Russell, 1980).

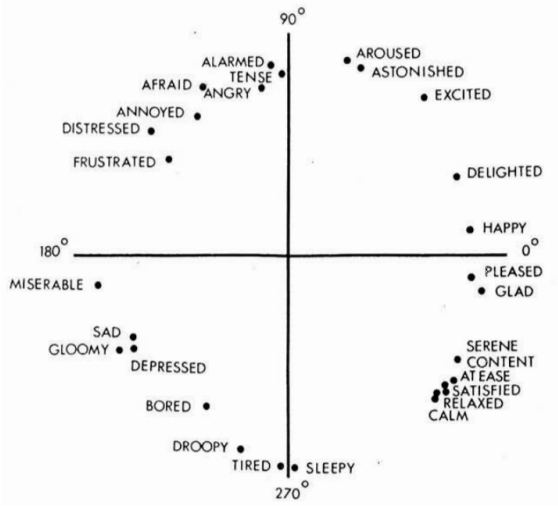


Figure 3.1: Coordinates for 28 words of affect from (Russell, 1980).

3.1.4 Trait and state affect

However they operationalize affect, psychologists distinguish between *trait* affect and *state* affect. Trait affect is one's 'current mood' or 'mood in the last day'; when self-reported affect is about a 'general mood' or a mood 'over the last few weeks', this is trait affect (Pressman & Cohen, 2005). Trait affect is relatively stable within individuals, while state affect can vary frequently and widely and is more associated with short-term events.

3.2 How affect is measured

The classic, gold standard measure of affect is the Positive And Negative Affect Schedule (PANAS), essentially combines dimensions of arousal and valence into two measures, one for positive affect (PA) and one for negative affect (NA) (Watson et al., 1988). PANAS consists of 20 items: single emotion or feeling words that represent positively and negatively valenced feelings as well as arousal and

activation. For the PA scale, higher arousal and more pleasurable selections result in higher scores; low arousal and less pleasure result in a low score. The NA scale functions the same with respects to arousal but features negatively valenced items. More recently, PANAS has been suggested to be more reflective of positive activation than of pleasure, as items such as happiness are not directly assessed (Crawford & Henry, 2004; D. H. Epstein et al., 2009), which could lead to misleading interpretation if PA is assumed to be pleasure-driven. Examination of PANAS scores in subjects experiencing extreme state anger has highlighted this issue, as the high level of activation from anger leads to confusingly high PANAS Positive scores (Harmon-Jones, Harmon-Jones, Abramson, & Peterson, 2009). Still, the validity and reliability of the measure is difficult to question; the relative simplicity combined with repeatability across thousands of studies has given PANAS a place among the more widely used measures in all of science.

3.2.1 Brief self-report measures

Researchers have sought to streamline the process of measuring affect for some time. Ironically, even Watson et al originally referred to the 20-item PANAS as "a brief measure of positive and negative affect" (Watson et al., 1988). PANAS may be brief in comparison to prior work, but certainly not to the point that it would be suitable for use in the context of Ecological Momentary Assessment or frequent measurement. In fairness, PANAS was not developed to be used as a frequent, in situ measure, but there have been several efforts to develop more suitable single-item measures of affect meeting with varying degrees of success.

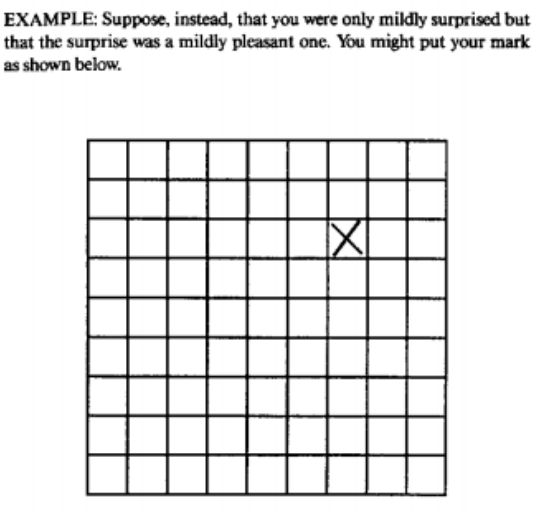


Figure 3.2: One of multiple examples from the instructions for Russell's Affect Grid (Russell, Anna Weiss, & Mendelsohn, 1989).

The first work that is relevant to this research is Russell's Affect Grid (Russell, Anna Weiss, & Mendelsohn, 1989). In short, the Affective Grid is a pen and paper based instrument in which subjects are presented a 10x10 empty grid representing two-dimensional affect space with valence in the x-axis and arousal in the y-axis (Figure 3.2). Subjects check the cell of the grid that represents how they currently feel, and that location can be mapped to a score that correlates strongly with the Semantic Differential Scale and PANAS Positive. The most significant issue with the Affect Grid is the difficulty that it presents subjects; the instructions are lengthy (two full pages!) in order to explain the less-than-clear concepts described above and even then, subjects must be able to cognitively process their current emotional state and quantify it on a box in a grid.

A second important single-item measure of affect is Bradley and Lang's Self Assessment Manikin, or SAM (Bradley & Lang, 1994). SAM presents users three sets of five drawings of a simple character, each set representing the range of states in the pleasure, arousal, and dominance dimensions. Subjects identify the

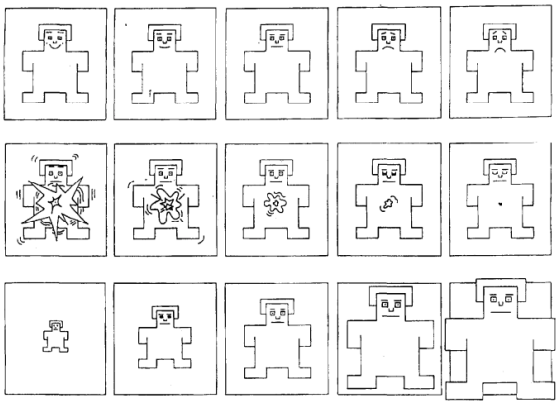


Figure 3.3: The Self-Assessment Manikin (SAM) used to rate valence (top), arousal (middle) and dominance (bottom) (Bradley & Lang, 1994).

character in each set that represents their current state in that particular dimension (Figure 3.3). SAM scores were not reported as validated against PANAS. SAM attempts to simplify the cognitively difficult task of locating ones affective state in two-dimensional space by instead leveraging the human response to imagery, a subject in which Lang is expert (Lang, 1995).

A facial-expression single-item measure that lays drawings of either a female, male, or gender neutral cartoon character around Russell's affective space is the Pictorial Mood Reporting Instrument (Figure 3.4); while promising, the authors report several issues with the measure largely around handling the overlap of several pairs of moods, including tension/irritation and calm/relaxed (Vastenburg, Romero Herrera, Van Bel, & Desmet, 2011).

To a certain extent, the two measures described above could be considered as the conceptual predecessors to most of the more recent work in this area. Perhaps the best example of how modern mobile devices can be combined with these classic theoretically grounded measures of affect is the assessment of affect on mobile phones with a colored two-dimensional Mood Map (Morris et al., 2010). The system is essentially an abstract version of the original Affect

Figure 3.4: Mood map for the female character arranged in the circumplex model of affect (Vastenburg, Romero Herrera, Van Bel, & Desmet, 2011)

Grid that has been designed to run on the MyExperience experience sampling application (Froehlich, Chen, Consolvo, Harrison, & Landay, 2007). The simplistic but seemingly effective Mood Map is perfectly suited for such applications; however it has not been validated against reliable measures of affect, and given the level of abstraction of the grid space and arbitrary color use, it may prove difficult to validate (Figure **??**). Other examples of mobile phone-based assessments include Isomursu's Feedback (Isomursu, Tähti, Väinämö, & Kuutti, 2007), although this is more focused at emotional response to a mobile app for use in usability testing and Meschtscherjakov's emoticon-based work, which is promising, but also not yet validated (Meschtscherjakov, Astrid Weiss, & Scherndl, 2009).

3.3 Design and development of the Photographic Affect Meter

As described above, in this work I am seeking a self-report system that (1) reliably measures affect, (2) is unobtrusive and pleasant enough to administer at

Figure 3.5: The color-based Mood Map (Morris et al., 2010)

least daily, and (3) is able to be administered *in situ*. Looking at previously existing systems, there is no single instrument fully satisfies all three criteria. Russell and Lang's work certainly satisfy the reliability criterion. The Affect Grid could easily be implemented on a mobile phone, but due to the above-discussed issues would remain a cumbersome instrument. Lang's SAM is quicker and more pleasant, but requires three distinct decisions to be made regarding one's emotional state. The Mood Map and other similar works are ideally suited for frequent, in context measurement but have in general not been validated as reliable measures of affect. Designing a mobile phone based tool for frequently assessing a user's emotions in context sounds decidedly like a problem for the HCI research community, yet there has been little work done along these lines. Taking aim at the problem with what research in HCI has taught us about mobile technology, context, usability, design, and affective computing seems likely to yield significant contributions.

Research in affective computing suggests that computers are inherently bad

at detecting users' emotional states and that designers should steer clear of falling into this trap (Boehner, DePaula, Dourish, & Sengers, 2007). The resulting representation of emotion therefore needed to both offer opportunity for interpretation and personal meaning, as well as allow users (rather than the designers or the system) to determine the meaning of each representation of emotion. Therefore, ambiguity of emotion was incorporated into the design of the system (Sengers et al., 2008).

Along these lines, the first important consideration was choosing a medium for representing emotion. Text-based instruments are obtrusive in the amount of time they take to complete, and even simplified measures like Russell's Affect Grid can be unwieldy. Others, such as SAM and Mood Map, instead make use of graphical representations, taking a successful step towards an abbreviated scale. Along these lines, color has been a popular medium for representing emotion in the HCI community (Sundstrom, Stahl, & Hook, 2007). A body of research connects color with emotion (Mayer, DiPaolo, & Salovey, 1990; Naz & Helen, 2004) but suggests that interpretation of color would prove to be too equivocal to be useful for assessment purposes.

3.3.1 Generating a corpus of affect-laden photographs

Photographs offer a richer and more engaging representation of affect. The link between photographs and emotion is well researched, with evidence suggesting that photos themselves can be emotionally charged and can have universal emotional legibility (Lang, 1995), but also can have very private meaning based on prior or shared experiences (Chalfen, 1989). We hoped this tension

between emotional legibility and private meaning would offer users a range of interpretive flexibility from images with more specific emotional content (such as an expressive human face) to something more ambiguous (a drop of water rippling in a glass). Certainly this had appeared to be the case in an early system designed to support cancer outpatients by encouraging them to share their emotions using photos (Gay et al., 2011).

In selecting a photo set to use for PAM, we first considered using Lang's International Affective Picture System (IAPS), an archive of photos that make up a validated instrument for evoking a variety of emotional responses (Lang, 1995). However, in order for these photos to consistently evoke such responses, the subject matter they depict falls at extreme ends of the spectrum (including violence, sexuality, etc.), and our pilot testers indicated that most images in IAPS would not be photos they would choose for themselves. Further, the IAPS photos have been chosen to provoke emotional response, whereas the goal of PAM is assessment.

Instead, we turned to the online photo-sharing service Flickr. Using the Flickr API, we downloaded roughly 9,000 Creative Commons-licensed images that had been tagged by the community of Flickr users with one of Russell's 28 words of affect (Russell, 1980) in hopes of finding photos that PAM users would find representative of a variety of emotions. After removing images with inappropriate content, 7,714 photos remained. To prototype the concept and illuminate the subset of these images with which people would be most likely to identify, testers (N=70) were encouraged to use an early version of PAM to select photos to represent their current emotional state repeatedly over a period of one to two weeks. From these data, we identified the most frequently user-

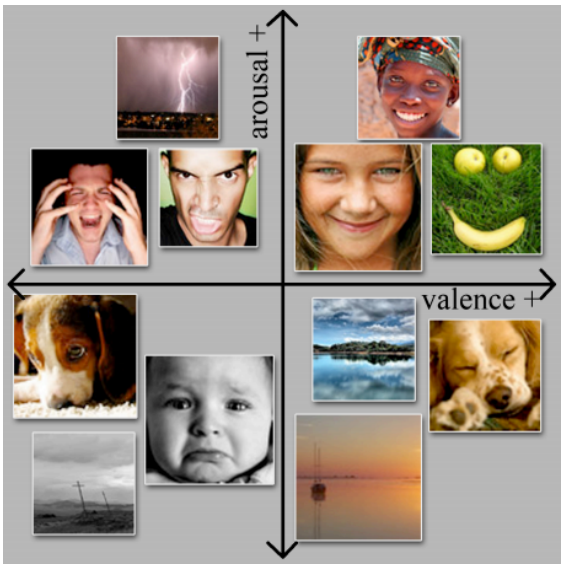


Figure 3.6: Several PAM images laid atop Russell's affective space (Russell, 1980)

selected photos. It should be noted that the communal tagging of a photo with an emotion word on Flickr has the potential to be somewhat arbitrary, loaded, or unrelated to the image. As such, we used the photos' tag words in the initial sorting and handling of the images only—our assumptions moving forward about the relationship of a given image to affect are always based on empirical evidence rather than the emotion-word tags. The code is available on GitHub.[1]

3.3.2 Arranging the photos in a grid

We decided to arrange the photos in a grid, from low arousal and negative valence in the bottom left corner, to high arousal and positive valence in the top right, imitating Russell's Affect Grid (Figure 3.6. Brief testing on a variety of smart phones such as Android, iPhone, and Blackberry indicated that a 4x4 grid was an appropriate balance of photo size given limited screen size. To deal

[1]https://github.com/philadams/flickr-images-grab

with the openness for interpretation of photos described above, multiple images would be assigned to each grid cell, and those images would rotate randomly each time PAM loads. Under this solution, hopefully with use of the 'more photos' button and repeated use, users will select an appropriate photo the majority of the time.

Finally, we chose our list of most frequently selected photos and arranged them in the grid based on the dimensions of valence and arousal and Russell's itemization of emotion words (Russell, 1980). Starting with a list of 100 or so photos, the prototype system was piloted with 70 more testers who also completed PANAS after selecting a PAM image, and these results determined the final photoset (3 images per grid cell) as well as the location each would appear in the grid.

3.3.3 PAM scoring system

PAM produces four separate scores upon completion. The first, just referred to as PAM, is a 1-16 score that maps to PANAS PA. In keeping with the conceptualization of PANAS PA as ranging from low pleasure and low arousal to high pleasure and high arousal at the high end, the PAM Score is derived by starting a counter at the negative valence, low arousal corner of the grid and working increasingly in arousal then valence toward the positive valence, high arousal corner of the grid (Figure 3.7).

The other three PAM values are also derivatives of position within the grid (Figure 3). The PAM Quadrant is the valence and arousal quadrant of the grid from which the subjects' selection was made, and is represented by a score of 1

High Arousal

Negative Valence	6	8	14	16
	5	7	13	15
	2	4	10	12
	1	3	9	11

Positive Valence

Low Arousal

Figure 3.7: PAM scoring in the grid layout

(Negative/ Low), 2 (Negative/ High), 3 (Positive/ Low), or 4 (Positive/ High). PAM Valence is a score of -2 to 2 (excluding 0) that represents the absolute position in the grid in absolute terms on the x-axis, and is a measure of the extent of displeasure to pleasure the subject is feeling. PAM Arousal is also a -2 to 2 (excluding 0) score, instead derived from the position on the y-axis. PAM Arousal represents the subject's state of arousal or activation ranging from low to high.

3.3.4 The PAM interface

From a user's perspective, PAM, like its predecessors, is laid out in a grid (Figure 3.8). The subject is presented with 16 photographs that represent a diversity of emotions (description to follow) arranged into a 4x4 grid, and is prompted to select a photo that best describes how they feel right now. Below the grid, a button to choose more photos reloads the grid with a different set of PAM images should the user not be able to comfortably express their state with one of the photos currently displayed. Upon selection of a representative photo, the other images are removed, allowing the subject to review his/her selection before touching the report mood button.

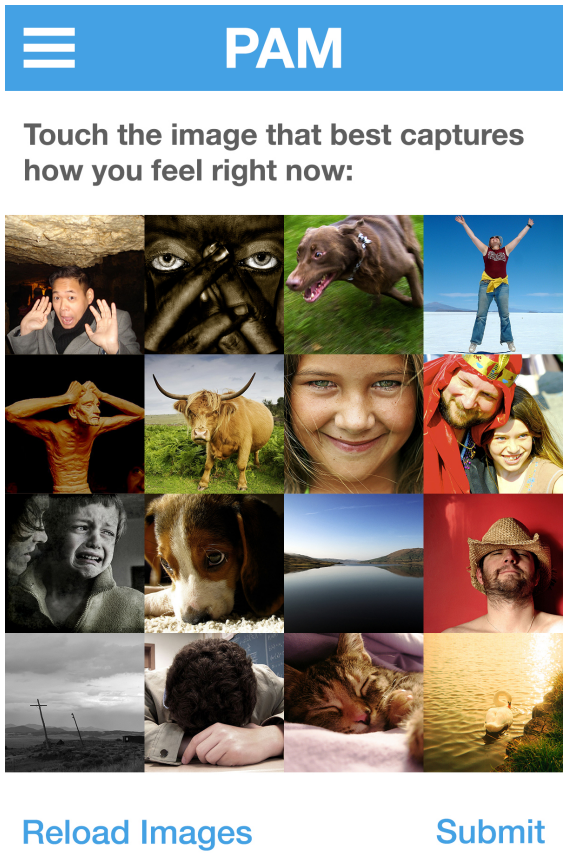


Figure 3.8: An instance of PAM as it appears on an iOS device

3.4 Validation of PAM by comparison to PANAS

3.4.1 Protocol and participants

Subjects were 315 individuals (45% male; Asian 21.3%, Black or African American 7.8%, Hispanic or Latino 8.5%, White 54.6%, Other 7.8%) recruited through a combination of university departmental and student list serves, snowball sampling, flyers, and postings on a variety of websites. Subjects who participated were given the option of providing an email address to be entered into a drawing to win an iPod.

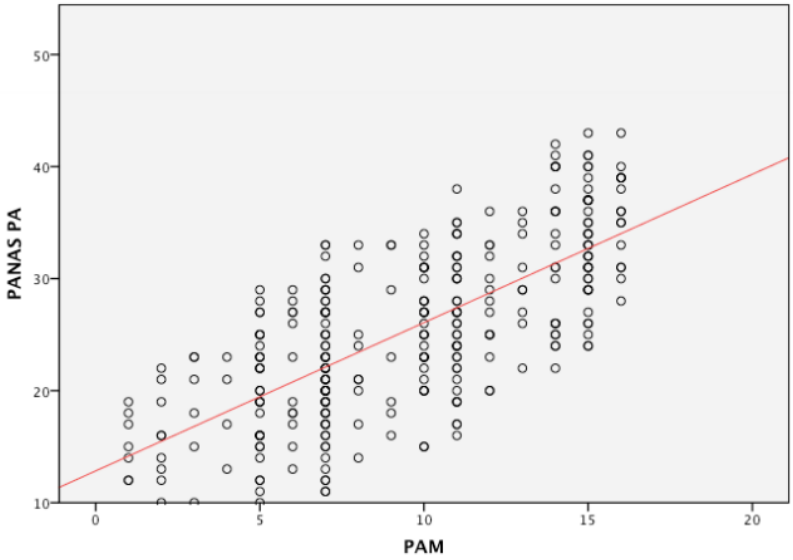


Figure 3.9: Scatterplot of PAM vs PANAS PA

Subjects logged into the study website and completed the PAM assessment with the prompt "Select the image that best captures how you feel right now." Next, subjects were taken to a web-based version of PANAS, with the time scale of the PANAS prompt also changed to be "Indicate to what extent you feel this way right now, that is, at the present moment". Finally, subjects were asked to provide basic demographic information. This provided a data set with which to examine the relationship between each subject's PAM and PANAS results, both theoretically representing measures of state affect.

3.4.2 Results

The primary point of comparison of this study is subjects' PAM scores and PANAS scores, both representing state affect. In this sample (N=315), mean scores for each measure were PAM, M=9.72, SD=4.06; PANAS PA, M=25.75, SD=7.64; and PANAS NA, M=15.65, SD=6.20. Figure 3.9 depicts a scatterplot of

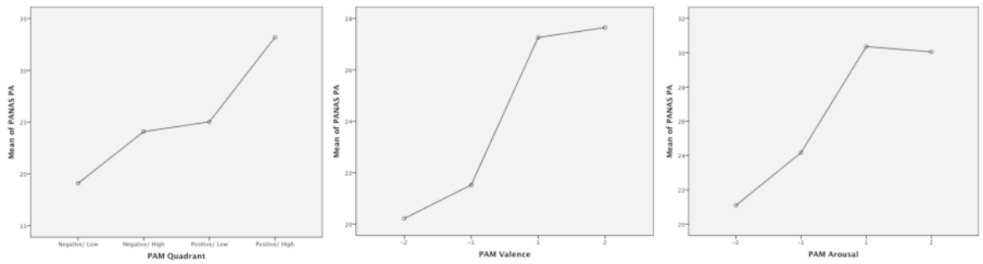


Figure 3.10: Plot of mean PANAS scores by PAM Quadrant, Valence, and Arousal. ANOVA for each significant at p¡0.001. Difference in means (t-test) are all significant at p ¡ 0.001 except for Negative/High vs. Positive/Low, Valence = -2 vs Valence = -1. Valence = 2 vs. Valence = 1, Arousal = 2 vs. Arousal = 1.

PANAS PA vs PAM. PAM strongly correlates with PANAS Positive (Table 3.1. PAM scores are in fact a good indicator of positive affect. Note that there is a weak negative correlation (-.37) between PAM and PANAS Negative, which is consistent both with theoretical expectations (Watson et al., 1988) and with the correlation between PANAS Positive and PANAS Negative (-.30). While it is difficult to compare correlations between studies, it is worth noting that Russell's Affect Grid produced a correlation of only .62 with PANAS PA.

Given the dimensionality of PAM, it is also important to examine the component data (arousal and valence) as well as just the compiled PAM score. Remember that PANAS Positive is in reality a combination of valence and arousal, with lower arousal and obviously negative valence indicative of lower scores and higher arousal and positive valence indicating higher scores. The graphs in Figure 3.10 clearly demonstrate the same pattern, providing strong evidence of the validity of the PAM valence and arousal components. It is interesting to note that PANAS Positive scores are slightly higher at Arousal=1 than Arousal=2 (although not significantly so), as this suggests that the scores are somewhat valence (pleasure) dominated.

	PANAS Positive	PANAS Negative
PAM	.71	-.37
PANAS Positive		-.30

Table 3.1: Correlations between PAM scores and PANAS scores, N=315 (p¡0.001 for all values).

3.5 Discussion and conclusions

These results clearly demonstrate that PAM can be utilized as a brief, reliable measure of affect. Establishing the validity of any measure can be an arduous task involving the exhaustive assessment of various forms of validity (Campbell & Fiske, 1959); the process has been described as iterative and nearly unending (Cronbach, 1988). At this stage, our approach has been to establish construct validity for PAM - that is, that PAM is in fact measuring affect and doing so in a way that is meaningful and fits within expectations. (Campbell & Fiske, 1959) classically identifies convergent validity the extent to which the measure converges with other similar or theoretically correlated measures) as a requisite component to establish construct validity in a new measure. This argument is very simple: PAM scores strongly correlate with PANAS Positive state affect scores, and analysis of the valence and arousal components of PAM leads to expected results when compared to PANAS. This is strong evidence of convergent validity considering that PANAS is the measure we have selected as most representative of our desired construct of affect.

3.5.1 Supporting the process of self-assessment

Users of PAM need not be given instructions other than a brief prompt, in contrast to Russell's Affect Grid (Russell et al., 1989), for example, which requires

lengthy instruction. Design and implementation undoubtedly play a role in the effectiveness of PAM; a touch screen interface with one step that is built into existing apps is far from obtrusive. Further, the grid layout of photos, whose underlying structure is not spelled out explicitly to users, may still become familiar over time, increasing the ease of response. Another possibility is that the emotional representation of images allows for a very rapid response. Lang suggests that images elicit a visceral response; perhaps users are forming almost precognitive decisions about their current emotional state in response to the imagery. Whatever the mechanism, it appears that PAM is supporting the self-assessment of affect, something that people tend to have difficulty doing (Saarni, 1999). All the above is certainly subject matter for future experimentation, and my work with PAM invites the suggestion that photos may be of value in developing brief self-report measures for other constructs (such as stress, or pain).

3.5.2 How else can PAM be used?

PAM has already been used in a range of EMA-style experiments and field trials around health decision making (Baumer et al., 2012; Adams et al., 2014; Adams et al., 2014). Equally as important as establishing construct validity, PAM's successful deployment in such studies demonstrates the viability and effectiveness of PAM for use as a tool for Ecological Momentary Assessment in field studies.

Beyond EMA-style studies, PAM is may be suitably deployed into any experience sampling scenario, for example to examine the variability of affect over time, the affective response to a system or interface, or to gain understanding

of emotional responses to scheduled events such as political addresses or sports tournaments. Indeed, PAM was used to understand the relative affective responses of users to the use of several social networking services during my internship on the Social Research Team at Google, and has also been deployed as part of the Weill Cornell Medical College system log-in and -out for medical residency interns as part of an investigation into the impact on quality of life for interns as policies about their work hours changed.

PAM has also been deployed as part of an effort not to capture the affective state of one individual over time, but to capture and visualize the mood of an entire building as part of the mood.cloud sculpture (Kim, Gay, Reynolds, & Hong, 2015). PAM is presented to visitors and 'inhabitants' of Gates Hall in an iPad-based kiosk, providing MoodCloud its affective input stream.

Here is a list of institutions and companies that have or are currently using PAM: Cornell University, National University of Singapore, NY Hospital for Special Surgery, Stanford University, University of California San Francisco, University of Rochester, Vanderbilt University Medical Center, Weill Cornell Medical College, bLife, Google, Intel, MyStrength, Otsuka America.

CHAPTER 4

CASE STUDY: THE MEASUREMENT OF MOMENTARY STRESS

In the 2011 Stress in America survey, the American Psychological Association warned that stress is becoming a public health crisis (APA, 2011). Most Americans are suffering from moderate to high levels of stress, with nearly half reporting an increase in stress over the preceding five-year span. According to the APA, "job stress is estimated to cost U.S. industry $300 billion a year in absenteeism, diminished productivity, employee turnover and direct medical, legal and insurance fees" (APA, 2012).

In our research, we are interested in supporting long-term engagement with one's own perceived stress levels. One of the central challenges in creating these types of systems is in determining what kind of stress-related data to collect in order to strike a balance between reliability - that is, how closely the data accurately and consistently track a person's perceived stress at the moment that it occurs, and intrusiveness - that is, how much effort is required on a participant or user's behalf to provide the data. Previous research has focused on developing techniques for monitoring stress levels, often by triangulating among multiple data sources (Carbonaro et al., 2011; Chang, Fisher, Canny, & Hartmann, 2011; Plarre et al., 2011).

In this case study, I present the results of a study designed to compare different kinds of data understand how we might collect these data with the minimal cost to the users. Among the kinds of sensing technologies that continuously monitor stress levels with a minimal impact on participants' daily lives, which track one another and participants' self report most closely? Under what conditions or in what contexts? When do these sensing modalities agree with one

another, and when do they produce conflicting narratives about daily stress?

Along with a group of collaborators in the Interaction Design and People-Aware Computing Labs, I designed and ran a study to answer these questions. Over the course of 10 days, we collected a variety of stress measures from a small group of participants during their everyday activities. We then compared the different data streams produced by traditional self-report measures and minimally invasive sensing devices (electrodermal and voice-based stress recognition). In addition, we conducted post-study interviews, asking the participants, themselves, to reflect on the accuracy and completeness of the data that had been collected. This case study presents several outcomes that represent specific research contributions:

1. I present evidence that voice-based stress sensing tracks with variations in EDA and self-reported stress measures in real-world environments
2. I describe the range of participant experiences - positive and negative - as they reported their stress levels
3. I reflect on some of the limitations of the various sensing approaches and the ways that our participants' experiences can help to inform the design of future personal stress informatics systems.

Much of this work has been been published in Pervasive Health (Adams et al., 2014).

4.1 How stress is understood

In general terms, stress is the reaction of an organism to a change in its equilibrium. In more practical terms, stress is the tension that a person experiences in response to an external stimulus or threat, and stress may have positive or negative outcomes, depending on whether or not a person is able to effectively cope with stress.

4.1.1 Stressors and mediating resources

Stress is commonly understood to be a feeling of strain and pressure: (Cohen, R. C. Kessler, & Gordon, 1995) describes stress as people's subjective evaluations of their resources and ability to cope with the demands presented to them by certain situations and experiences. Pearlin describes the stress process as the coming together of self-concepts, sources of stress, and stress-mediating resources - specifically, life events affect role strains which erode positive self-concepts, which resources such as social support and other coping mechanisms mediate the experience at various points in the process (Pearlin, Menaghan, Lieberman, & Mullan, 1981). Folkman discusses in detail the cognitive appraisal of stressors, available mediating resources, and likely event outcomes (Folkman, Lazarus, Dunkel-Schetter, DeLongis, & Gruen, 1986).

4.1.2 Three forms of stressors

'Stressors' can be classified into three major forms: life events, chronic strains, and daily hassles (Thoits, 1995). Life events are acute changes that necessitate major behavioral change within a short period of time - examples include divorce or the birth of one's first child. Chronic strains include experiences such as marital problems or a disabling injury, and are recurring demands that require continued readjustment over long(er) periods of time. Daily hassles are smaller events that occur within day-to-day life, such as getting back test results or sitting in traffic.

4.1.3 Mediators: social support as an example

The resources and processes that mediate the stress experience impact the qualitative, subjective perception of stress levels. Coping resources, coping strategies, stress resilience, and social support are all examples of moderating factors, each with a large and growing literature discussing complex mechanisms. For example, consider that social support, those processes "through which social relationships might promote health and wellbeing" (Cohen, Underwood, & Gottlieb, 2000), may well mediate stress by attenuating the appraisal of stress, by lessening deleterious outcomes, or both. And social support appears to both protect people under stress from the influence of stress - the buffering model - as well as be beneficial whether an individual is under stress or not - the main-effect model (Cohen & Wills, 1985). Receiving social support might not be the most salient mechanism in mediating stress: the perception of support available is at least as important as actually receiving said support in adjusting to stressful

events (Wethington & R. C. Kessler, 1986) and providing social support may be more beneficial than receiving it (S. L. Brown, Nesse, Vinokur, & Smith, 2003).

4.2 How stress is measured

Because of the highly subjective nature of perceived stress levels, researchers traditionally have relied upon self-report measures to gather data about people's experiences of stress. Researchers have conducted both diary studies of stress (Almeida, 2005; Ferreira, Sanches, Höök, & Jaensson, 2008; Sanches et al., 2010) and *in situ* and Experience Sampling studies (R. Larson & M. Csikszentmihalyi, 1983).

The gold standard for the self-report of stress are the Perceived Stress Scales (Cohen, Kamarck, & Mermelstein, 1983), and stress inducers can be assessed by using the Daily Inventory of Stressful Events (Almeida, Wethington, & R. C. Kessler, 2002). Measurement becomes more difficult in the world of momentary self-report where we are reduced to short versions of the Perceived Stress Scale (Cohen & Williamson, 1988) or single-item (S. E. Taylor, Welch, Kim, & Sherman, 2007) scales with single-number output.

Increasingly, stress researchers note that self-report approaches, while suitable for short-term research studies, present challenges when incorporated as part of a personal informatics system that is intended to provide benefits to its users over the long term. Diary studies are prone to memory effects and reduced compliance over time (Almeida, 2005; R. Larson & M. Csikszentmihalyi, 1983), and experience sampling can become highly interruptive (Intille, Rondoni, Kukla, Ancona, & Bao, 2003; Scollon et al., 2009) - which may itself

become a source of stress for a study participant or a system user. Although the HCI community has developed adaptations to the ESM method, including delivery of surveys electronically and based on sensed contextual information (Intille et al., 2003; Klasnja et al., 2008; Meschtscherjakov, Reitberger, & Tscheligi, 2010), these approaches still require a considerable investment in time and effort, in order to provide insights about everyday stress and stressors over time. Although this style of data collection might be well suited to helping individuals to discover sources of stress within a particular time period, it would clearly not be as helpful when reflecting over an arbitrary window of time or when aiming to maintain an intended or desired response to stressors (Li, A. K. Dey, & Forlizzi, 2011).

The pervasive sensing capabilities of modern computational devices (G. Miller, 2012) provide a valuable opportunity for continuously and non-intrusively measuring stress levels (Almeida, 2005; Ertin et al., 2011; Ferreira et al., 2008; Gaggioli et al., 2013; R. W. Picard & Liu, 2007; Plarre et al., 2011; Sanches et al., 2010). These devices are also being adopted by medical professionals and incorporated into long-term clinical treatments and behavioral interventions that are designed to improve healthcare outcomes (Chatterjee & A. Price, 2009). However, because stress is a complex and multifaceted health issue, there are a variety of methodologies for automatically collecting data about people's levels of stress. Some approaches, such as heart rate or heart rate variability, provide relatively direct and accurate measurements of stress, but come with undesired trade-offs in terms of the intrusiveness of measurement (Carbonaro et al., 2011). Other approaches rely on secondary physiological signals, such as skin conductivity (referred to as electrodermal analysis or EDA), to detect changes in arousal that may be linked to stress (Ayzenberg, Hernan-

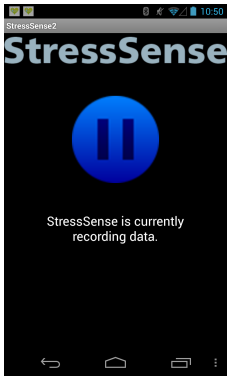
dez Rivera, & R. Picard, 2012; Ertin et al., 2011; R. W. Picard & Liu, 2007; Poh, Swenson, & R. W. Picard, 2010; Sanches et al., 2010) at the cost of some degree of fidelity (e.g., affective valence). Monitoring changes in vocal production (Chang et al., 2011; Lu et al., 2012; Scherer, 1986) is a far less invasive approach but may be limited in the accuracy that can be achieved or the diversity of environments in which it is effective.

Because different sensing modalities capture different representations, granularities, and quantities of stress data, most of the previous systems that have been proposed or developed for collecting stress data triangulate among different data collection methods, depending on whether they have aimed to create the most robust possible user model (Carbonaro et al., 2011; Chang et al., 2011; Plarre et al., 2011) or to explore issues related to sensor deployment and integration, as in (Ertin et al., 2011) or to design visual representations of stress to reflect back to end users, as with (Ferreira et al., 2008; McDuff, Karlson, Kapoor, Roseway, & Czerwinski, 2012; Sanches et al., 2010).

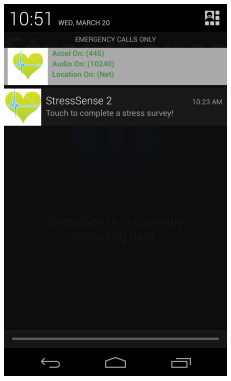
4.3 Stress Experience Sampling And Measurement Experiment

4.3.1 The SESAME system

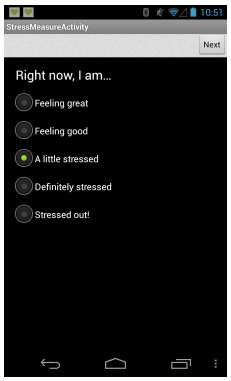
The smartphone app, SESAME (Stress Experience Sampling And Measurement Experiment), was designed to collect data about individuals' stress levels and the environmental contexts within which this stress is experienced. It runs on the Android mobile operating system. Data is collected in the following three ways:



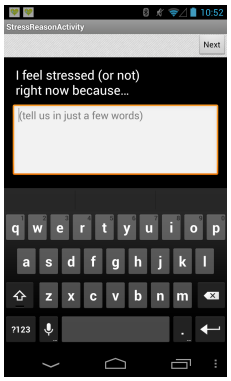
(a) Control panel for starting and pausing passive sensing.



(b) Notifications appear when it is time to self-report.



(c) Self-report of stress level with (S. E. Taylor, Welch, Kim, & Sherman, 2007).



(d) Free-text rationale for previously-selected stress level.

Figure 4.1: SESAME user interface

1. passively via sensors on the mobile phone,
2. via self-report measures also on the mobile device, and
3. via an Affectiva Q Sensor, worn on the wrist.

The data collected on the smartphone device were cached locally on the device and pushed to the server in batches when the device is both plugged in to a power source and connected to a Wi-Fi network, most commonly during each night.

At short time intervals, SESAME infers an audio profile (silence, non–human-voice noise, stressed voice, not stressed voice) from microphone data. As recording audio can raise particular privacy concerns, the audio feature extraction and profile labeling takes place on the device—see previous work for more details (Lu et al., 2012). Due to limitations of the operating system, audio profile sensing is suspended when the participant makes or receives a phone call.

Figure 4.2: The Affectiva Q electrodermal activity sensor.

SESAME's user interface (Figure 4.1) is minimal, consisting of two icons in the system notification bar (Figure 4.1b). Tapping the first icon presents the option to pause passive data collection (Figure 4.1a). Tapping the second launches the self-report panel, which includes single-item measures assessing momentary stress and momentary affect, as well as an optional short free-text response to the prompt, "I feel stressed (or not) right now because..." (Figure 4.1d). To capture momentary stress we used Taylor's 5-item measure [37] that prompts "right now, I am (1) feeling great! (2) feeling good (3) a little stressed (4) definitely stressed (5) stressed out!" (Figure 4.1c).

We used the Affectiva device (Poh et al., 2010) to gather data about participants' electrodermal activity, which provides an indication of physiological arousal; as described above, this measure is associated with momentary stress. The Affectiva is worn on the inside of the wrist and is about the size of a wrist-

watch. Data gathered by an Affectiva was cached on that device and retrieved by the re-searchers at the conclusion of the study.

4.3.2 Study design

We recruited a small cohort of participants in person by convenience and snowball sampling. In an effort to hold stress profiles as constant as possible, I recruited only graduate students and postdoctoral researchers from within a single academic department at Cornell. Participants received no compensation for participating in the study.

Participants

Of the original 11 participants recruited, n=7 completed the study; two participants chose not to continue beyond the first two days, and two others did not respond consistently to the self-reporting prompts during the data collection phase. Six of the seven participants who completed the study were male and one was female; six were aged 26–35 and one was aged 18–25. All but one participant owned and regularly used a smartphone (six Android, one iOS), and while six participants reported tracking personal information (such as sleep, exercise, spending, or mood) using websites or apps, no participant reported regularly using or wearing external sensors like the Nike FuelBand[1].

[1] http://www.nike.com/cdp/fuelband

Experimental protocol

Prior to the study period, each participant completed a short questionnaire. In addition to basic demographic information and questions about prior smartphone usage experience (including use of personal informatics (Almeida, 2005; Li, A. Dey, & Forlizzi, 2010; Li et al., 2011)), participants reported traits for affect (PANAS) (Watson et al., 1988) and stress (PSS 14) (Cohen et al., 1983), as well as a measure of mindful attention awareness correlated with a variety of wellbeing constructs (MAAS) (K. W. Brown & R. M. Ryan, 2003).

During the day preceding the study, each participant was introduced to the SESAME app. After being trained on pausing/restarting sensing and responding to ESM prompts, participants were encouraged to make a handful of sample reports using the app and receive answers to any questions they might have. Because the continuous sensing and sense-making components of the system are computationally intensive and can drain a smartphone battery before the end of a full day, six participants used the system on a secondary, loaned phone (an LG Nexus 4), which they carried with them for the duration of the study. Over the next ten days, participants were asked to run the SESAME app between the hours of 8:00am and 11:00pm (at a minimum) and to make self-reports in response to notifications (issued every half hour with a small random variation) as they were able. Participants were also free to volunteer additional self-reports at any additional time. Because we were most interested in collecting ground truth data from our participants using ESM over a relatively short window of time, we opted to increase the prompt frequency to the upper end of what is typically considered acceptable practice (Scollon et al., 2009) and to forego contextual suppression of stress reporting prompts (Intille et al., 2003; Klasnja et al.,

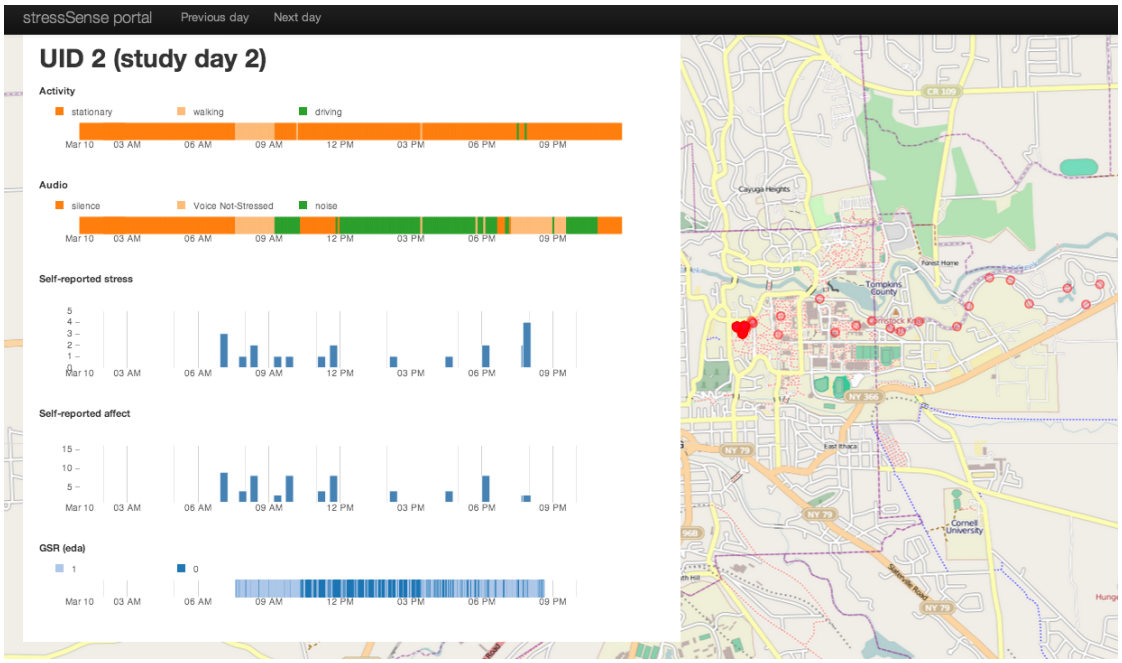


Figure 4.3: Visualization of one day of one participant's data, showing sensed locations to the right, and to the left a series of time-aligned charts showing sensed activity, audio-inferred stress state, self-reported stress level, self-reported affect level, and EDA-inferred stress state.

2008). Participants were informed that our goal was to collect as much stress data from them as was practical, but that they should feel free to ignore these frequent prompts if they occurred during attention-sensitive tasks like driving or if responding would be socially inappropriate (e.g., during a family dinner or while on a date).

Due to the limited number of Affectiva Q Sensor devices available for deployment during the study, we randomly divided our participants into two groups; participants in Group 1 collected EDA data on study days 1–5 and participants in Group 2 collected EDA data on days 6–10.

At the conclusion of data collection, we conducted a 20- to 30-minute semi-structured interview with each participant to gather qualitative information about their experiences using the SESAME app. To aid participants in recalling

the scenarios in which they were reporting and in which they found themselves, and to help provide common ground between participant and researcher, we developed a per-participant, per-day visualization tool (Figure 4.3). Within one day, the participant and researcher could see a map of places the participant traveled to the right, and to the left a series of time-aligned charts showing sensed activity, audio-inferred stress state, self-reported stress level, self-reported affect level, and EDA-inferred stress state. The participant could easily navigate across the 10 days of their displayed data to make between-day comparisons.

4.3.3 Analyses

Over the ten-day deployment, and with a pre-test questionnaire and a semi-structured interview, we have collected many types of data often at different temporal resolutions. Here, we describe how each form of data was preprocessed and then analyzed.

Data preparation

Data from the pre-test questionnaire was prepared according to each instrument's directions (K. W. Brown & R. M. Ryan, 2003; Cohen et al., 1983; Watson et al., 1988).

For both passively collected and self-reported data, we discarded samples outside of the time range dictated by the study parameters. To maintain consistency, participants were asked to use the system between the hours 8:00am and

11:00pm daily, but some used the system outside this time range (e.g., making self-reports early in the morning or wearing the Affectiva late at night).

We inferred physiological arousal based on the EDA data provided by the Affectiva Q Sensor. In stressful situations, the sympathetic nervous system activates the sweat glands. EDA devices like the Q Sensor estimate the amount of sweat secreted by measuring changes in the electrical conductance of the skin (Poh et al., 2010). Once we retrieved the raw EDA data from the wrist-worn devices, we normalized the time-series data using a Z-normalization technique, which centers the data distribution about a zero-mean and scales it, resulting in unit variance. We used a 20-second long window with 10-second shift to extract high level features from the normalized EDA data. EDA has a fast-changing response (startle response) when stressors are present, so some of the features that we extracted include the mean crossing rate, the energy associated with low-frequency filter bank, the slope of the linear regression, and minimum and maximum values, all of which have been shown to capture aspects of this startle response (J. A. Healey & R. W. Picard, 2005). We trained two single-component Gaussian mixture models (one for modeling the aroused condition and another for modeling neutral or non-aroused conditions) with a full covariance matrix. We then used a GMM-based model to classify the EDA data into a binary value—aroused or not aroused—every 10 seconds over 20-second windows, using a threshold of 0.85. (On the pre-existing dataset, this GSR-based stress model has a performance of 74.3%, 76.5% and 78.4% in terms of accuracy, recall and precision, respectively.)

Continuously sensed audio yielded inferences every 1.2 seconds. To aid in comparison across measures, we elected to smooth all passively sensed infer-

ences into 5-minute windows using a simple majority rule; smoothing to tighter (1-minute) windows did not appreciably change the results.

Comparison of measurement modalities

One of the central questions that this research seeks to address is the reliability of various widely deployable stress-sensing techniques in a range of real-world scenarios. It is clear that some of these approaches will be more suitable in detecting stress in certain situations. For example, recognizing stress from voice will necessarily be more accurate when a person is engaged in a conversation than when they are working alone; the EDA signal will change in different ways during physical exercise than when a person is experiencing emotional or cognitive stresses (Poh et al., 2010). Furthermore, while subjective self-assessment of stress levels (e.g., with the PSS-14 instrument) has been shown to have high internal consistency and predictability, many of these types of instruments have been designed to examine stress as a trait, framing stress in the context of life events that take place over the course of weeks or months. Since there is no clear-cut and established gold standard for globally measuring stress levels in an ecologically valid, non-intrusive way, we set out to determine the circumstances in which a variety of established stress measuring mechanisms, including EDA, continuous voice-based stress recognition, and self-reported stress levels, do and do not align with each other.

In comparing self-report measures, captured at a resolution of 30 minutes, inferences from continuously sensed sources were smoothed over 1 hour windows prior to the self-report. This is because the self-report stress literature indicates that momentary psychological stress is a function of the current stress

trait, and recent daily stressors; further, we confirmed that this is how our participants made self-report assessments from the semistructured interviews. Because data most closely co-occurring with the self-report will have a greater effect on experienced stress, we computed a weighted mean over the data in the window, giving preference to the most recent data points; again, feature selection was by simple majority. This happens for each self-reported value for each user; we then normalize and assess the relationship for each self-report value, one to five, with both electrodermal and voice-stress inferences.

Participant narratives

We referred to the semi-structured interview data to help make sense of discontinuities in the sensor data streams and to inform our understanding of the participants' experiences using the system, including how and when participants provided data over the various modalities that we used.

4.3.4 Results and discussion

General participant experience

Over the course of the study, 15 hours per day for 10 days, SESAME collected a significant amount of data about stress and stress contexts experienced by our 7 participants. The system recorded some 17,415,310 audio profile inferences, 884 self-reports, 56,837 location measurements, and 9,400,139 EDA measurements. Smoothing and binning the data into manageable windows (as described above) yielded, on average, 1,192 location measurements, 1,066 audio

profile inferences, 126 self-reports, and 15,368 aroused or not aroused inferences from the EDA data per participant.

The number of stress self-reports completed over the course of the study (a $\tilde{4}$0% response rate) had dramatic variance (1,982), with several participants [P4, P6, P7] each providing more than 150 responses (Table 1). Most of the self-report survey submissions appear to have been direct responses to SESAME's experience sampling prompts—appearance of an auxiliary status bar icon, a short vibration sequence, and the notification LED set to pulse a purple color. However, most participants voluntarily submitted at least a few instances of un-prompted self-reports, with P2 submitting the largest number (11).

During the study, the smartphone application was programmed to issue the participants a reminder to complete the experience sampling questionnaire approximately once every 30 minutes. In practice, many experience-sampling responses were delayed due to constraints of the threading model used by the operating system or an app crash, or the participants simply did not respond to the notifications. There were a number of reasons why this may have been the case: the vibration pattern associated with the notification was somewhat subtle on some of our participants' phones (particularly on the Nexus 4s), and participants told us during our post-study interviews (see also below) that they would voluntarily ignore the notifications if they were engaged in an activity that demanded their attention (e.g., a conversation or driving) or if their hands were otherwise occupied (e.g., cooking, playing with children). The algorithm that generated the Experience Sampling notifications was also linked directly to the system timer, rather than being driven by the last time that a self-report survey was submitted. This resulted in a number of occasions in which a par-

ticipant would notice that they had neglected to complete a survey in response to a prior reminder, submit the survey, and immediately be notified that it was time to complete another; many of these subsequent self-report reminders were ignored.

On average, our participants completed surveys a little less than 1/3 of the time that an Experience Sampling notification was issued. Compliance ranged from 9% to nearly 50% across our group of participants. For those self-reported surveys that were completed in response to an experience sampling notification, the average delay between the system issuing an experience sampling reminder and the participant invoking the self-report survey mechanism was 8 minutes, 54 seconds (excluding delays longer than 30 minutes, which indicated an error or crash in the application). There were a fairly large number of very quick responses in our dataset (as little as 3 seconds elapsed from notification to invocation of the survey), but much of the time, the participants simply did not or could not respond until tens of minutes had elapsed. The only real drawback to this response rate is that our self-report data is scattered somewhat unevenly over our data collection window, as our analysis examined the passively sensed stress rates at whatever time the self-report surveys were completed. Interviews suggest that periods of time associated with very high levels of stress are under-reported in the self-report data, leading an artificial downward skew of the self-reported stress levels.

Although participating in the study did result in frequent interruptions due to our use of experience sampling to collect self-report data, participants were able to complete these three-question surveys quickly, with an average start-to-finish time of 20 seconds and a median of 17 seconds; this excludes 5 outliers

over 5 minutes where the survey activity appears to have been interrupted by another smartphone function.

We also ran a number of paired t-tests in order to determine whether participants' responses to the pre-test surveys (MAAS, I-PANAS-SF, or PSS-14) were effective predictors for the average stress levels reported over the course of the study. We found no significant differences that would suggest a relationship, although we did observe a very weak trend (p=.112) suggesting a correlation between participants' score on the PSS-14 and their average self-reported stress levels across the 10 days of the study (r=.562). A study with a larger sample size would be needed to more rigorously assess the predictive power of PSS-14 in empirically determining a per-person baseline stress level.

Scenarios and stress

We anticipated that each capture modality would, for individuals and in aggregate, reflect similar daily stress rhythms, and this bore out in our data. Voice-based stress measures were most pronounced on weekdays in the early and mid afternoon. There is a second, smaller peak in the late morning; detection of voice stress before mid-morning or after dinnertime is rare. Voice stress is prevalent for several users on one or two particular days (e.g., P4's day 10 and P6's day 3). EDA data also peaks in the early- to mid-afternoon, overall, and dramatically so for participant P3. Overall, self-reported stress remains relatively constant throughout the day, although there are fewer reports of "A little stressed" once the evening begins. Of note is that not a single participant self-reported the most stressed value "Stressed out!" over the course of the study. This unexpected gap in the data highlights one of the key shortcomings of the

traditional self-reporting approach to tracking stress levels: in situations where a participant is experiencing the highest levels of stress, they are extremely unlikely to stop what they're doing and attend to an experience sampling prompt. This reporting bias has been noted previously (Scollon et al., 2009), but is of particular interest when the construct being investigated relates so closely to a major factor in survey response compliance.

The sound profile that appeared in conjunction with experiences of stress is also interesting. The audio classification corresponding to the times that self-reports were given was most commonly noise or silence. P2 and P4 additionally reported feeling 'good' when there was unstressed voice present on multiple occasions. On four devices, the audio capture silently failed occasionally, and so self reports, particularly for P1 and P7, were also made at times for which we have no noise profile information; we have no reason to believe this occurred in periods of particularly high or low perceived stress and should not therefore bias our findings.

Open-ended free-text rationales

Participants provided open-ended rationales for 264 of the 358 completed experience-sampling prompts. These responses shed light on the specific kinds of stress that participants were experiencing during the study, which is significant since there are many different facets of stress: cognitive, social, physical, physiological, and so on.

Participants most frequently provided rationales categorized as working (n=54); these responses were associated with diverse stress ratings ranging from

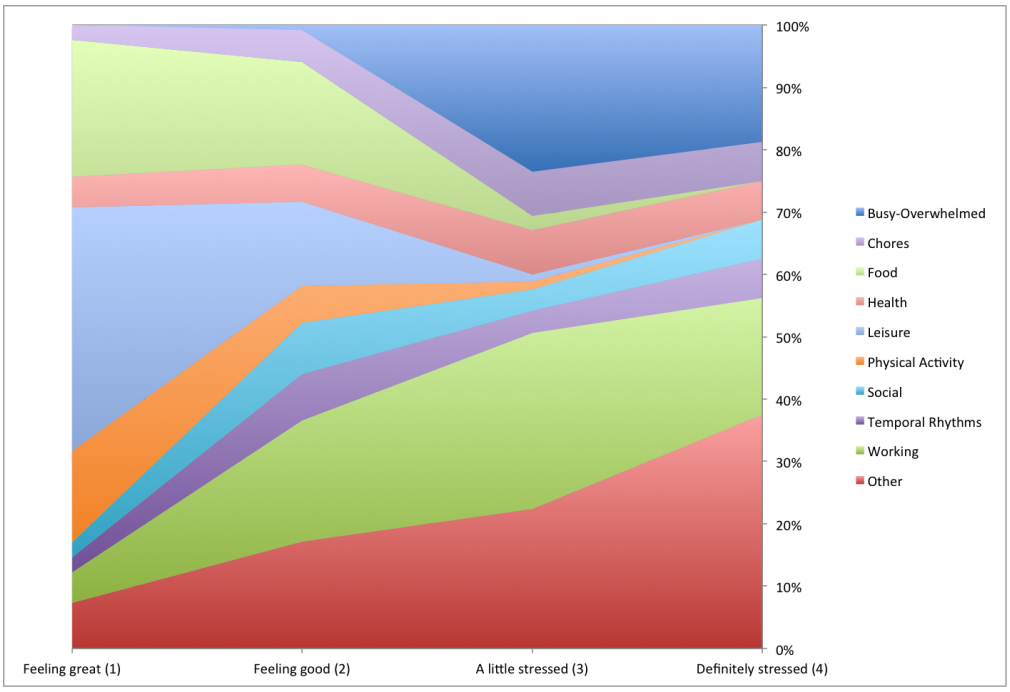


Figure 4.4: Distribution of the various categories of rationales provided for self reported stress levels, displayed as a percentage of the total responses at each stress level.

"feeling great" to "definitely stressed" (1 to 4 on a 5-point Likert item). For example, participants reported "feeling great" (1) when giving working rationales such as "cranking things out" [P2], and "definitely stressed" (4) when giving working rationales such as "interview go time" [P1].

The rationales most frequently associated with the higher levels of stress included (see also Figure4.4):

- a general sense of being overwhelmed (e.g., "so. much to do." [P2])
- technology issues (e.g., "confused database" [P7])
- rationales that suggested that a trend was emerging reflecting a loss of control (e.g., "nice slow morning but increasingly more to do today" [P1])

The rationales most frequently associated with lower levels of stress in-

cluded:

- environmental surroundings (e.g., "The weather is really nice" [P5])
- experiences with food (e.g., "capuchin oh!" [P4])
- activities associated with leisure (e.g., "relaxing" [P3])
- physical activity (e.g., "brisk walk in the rain" [P2])
- social interactions (e.g., "chatting over dinner" [P1])
- adherence to a temporal rhythm (e.g., "almost bedtime" [P7])
- rationales that suggested that a trend was emerging reflecting increased control over the pace of the day (e.g., "feeling better. trying breakfast" [P2]).

Some categories of rationales were distributed more evenly across high- and low-stress conditions, such as chores, health, travel, and working, suggesting that these rationales are associated both positive and negative stressors.

Associating stress from the three modalities

In investigating whether the minimally invasive passive measures (voice-based stress detection and EDA) could be reliably used to monitor stress levels in the wild, we report the voice-stress and EDA data collected simultaneously with self-reported stress, as well as with one another. Not all self-report points have associated EDA or voice-stress data, because (1) each participant wore the Q Sensor for only for half of the study days; (2) as a result of the occasional audio capture crash, as described above, there are some windows of time lacking raw audio data upon which to draw voice-based stress inferences; and (3) there were

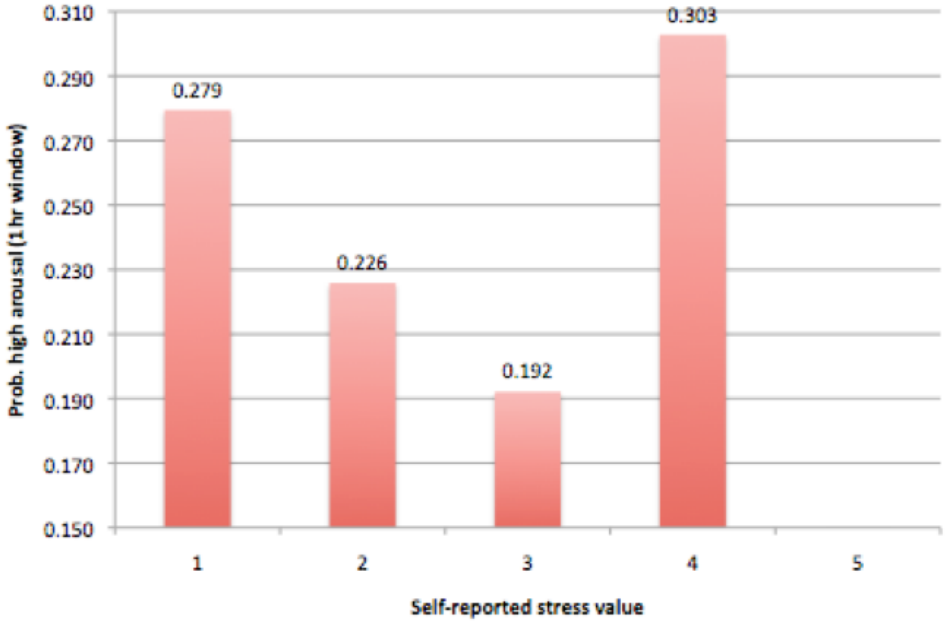


Figure 4.5: The probability of aroused EDA in a 1-hour window associated with self-reported stress.

occasions when participants chose to disable the passive sensors, such as when swimming, washing dishes, or to conserve the phone battery.

Self-report with electrodermal activity Over the self-report values 1–4, we see a cup-shaped curve, as responses 1 ("feeling great!") and 4 ("definitely stressed") correspond to higher levels of arousal than responses 2 ("feeling good") and 3 ("a little stressed"). This distinction emerges most clearly at higher EDA classification thresholds; here we report with a threshold of 0.8, selected by experimentation (Figure 4.5).

These data confirm that EDA provides an indication of the intensity of perceived stress responses, that is, the more strongly a participant agrees or disagrees that they are under stress, the stronger the EDA signal recorded by the system. However, the main drawback of the EDA approach is also highlighted

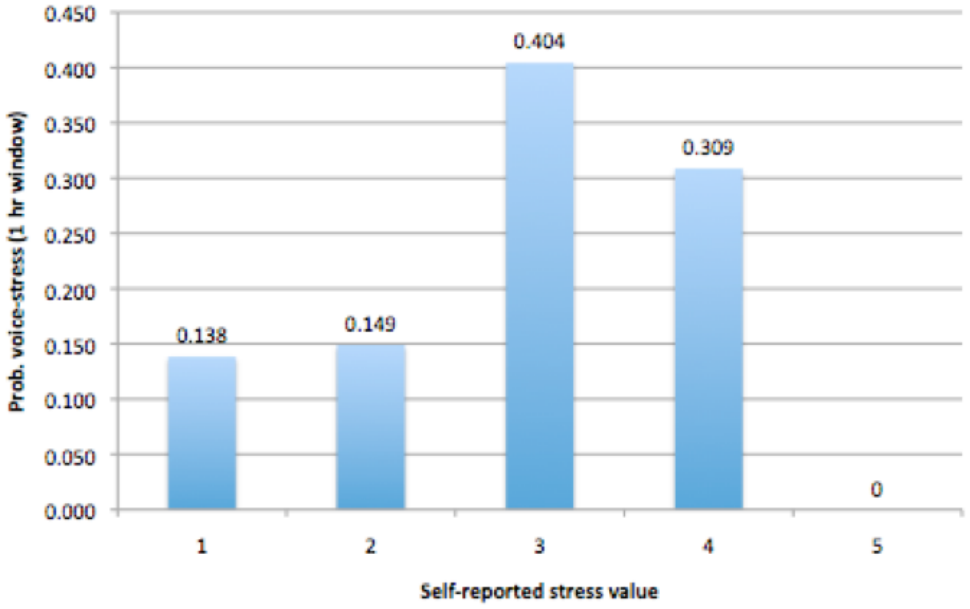


Figure 4.6: The probability of voice-stress presence in a 1-hour window associated with self-reported stress.

here: it is statistically impossible to detect from EDA data alone the valence of the perceived stress response, that is, whether increased arousal is associated with pleasant experiences or the negative experiences that we typically associate with being under duress.

Self-report with voice-stress In associating self-reported stress values with voice-stress, we report a weighted mean of voice-stress (1) and voice-no-stress (0), also computed over a 1-hour window (Figure 4.7). There is a positive correlation of r=0.59 over the self-report values 1–4, which we anticipated. However, the relationship peaks not at 4 ("definitely stressed") but at 3 ("a little stressed"); we believe this is because participants self-reported much less frequently when experiencing higher levels of stress.

This positive correlation revealed in our data suggests that analysis of pas-

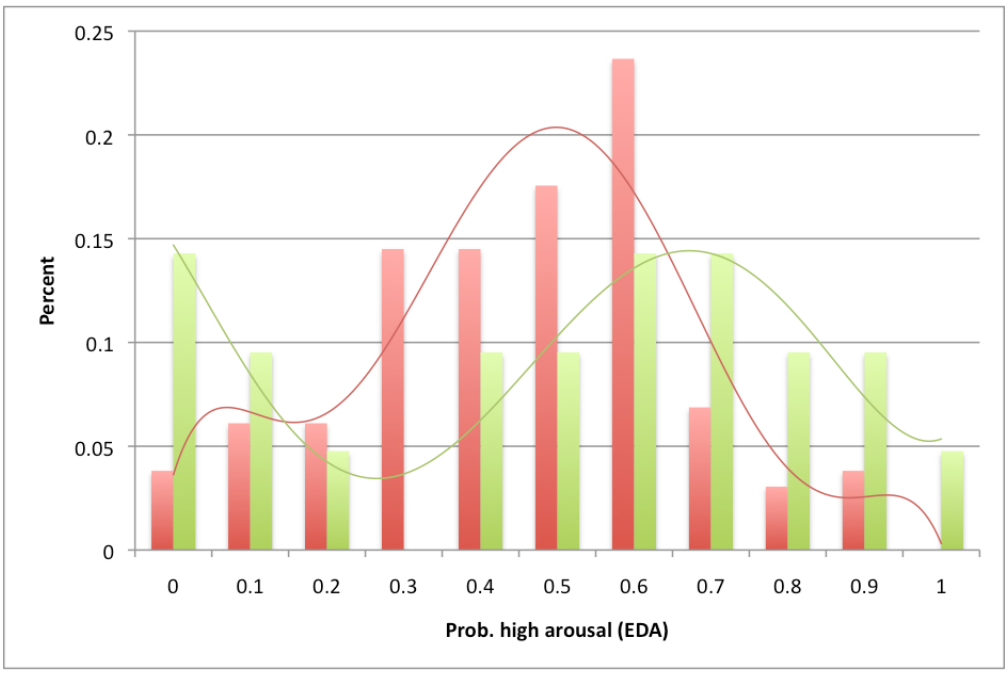


Figure 4.7: Distribution of aroused EDA, both overall (in red) and when voice-stress is detected (in green).

sively collected voice signals can result in reasonably accurate detection of stress episodes without requiring any participant/user intervention at all. However, the success of this approach is clearly dependent on the presence of a clear voice signal; stressful situations in which vocalizations are not present cannot be inferred at all.

Voice-stress with electrodermal activity In Figure 4.5, we examine the distribution of EDA aroused/non-aroused responses (threshold=0.75, selected by experimentation) both independent of, and in the presence of, positive voice-stress recognition. The red bars show the distribution of arousal inferred from EDA in one-hour windows over the entire study. For example, in about one-fifth (y-axis) of the windows, we detected a state of high arousal for 40% (x-axis) of the reports collected during that window. Similar to a histogram, the relatively

normal distribution of the red bars illustrate that it is relatively common to observe a relatively even mix of aroused and non-aroused readings from the EDA sensor during any given hour; it is more unusual for a one-hour window to be dominated by aroused or non-aroused inferences.

The green bars also show the distribution of EDA arousal in one hour windows, but in only those windows in which we also inferred the presence of voice stress for the user in question. The probability mass shifts rightward when we examine only these EDA reports that correspond to sensed stress from the voice channel. This shows that EDA and voice-based stress recognition generally track one another in the positive case; that is, over a one hour window, when we observe consistent stressed responses using the voice-based recognizer, we are also more likely to see aroused inferences from the EDA sensor during this time.

The small rise at the left end of the green bar distribution is worthy of note, however. This artifact reveals that there were a small (but noticeable) number of one-hour windows in which the EDA sensor was resulting in a much greater number of non-aroused inferences than aroused inferences, but the voice-based stress recognizer was detecting instances of stressed voice. When triangulating raw sensor and location data with interviews, it appears that this disagreement is a result of P6, a teaching assistant, calmly listening to several hours of student presentations, and P7 relaxing at home while eating breakfast or dinner while watching action movies on TV. In both cases, the system was (correctly) detecting stressful vocalizations, just not ones generated by the participants themselves. Adding a speaker ID filter to the system would mitigate these instances of false positives.

Semi-structured interview feedback

Some of the frustrations that participants reported included the troublesome drain on the phone battery caused by our implementation of continuous passive sensing and a concern that being repeatedly asked to consider and report on one's stress could itself become a stressor—P7 even went so far as to "mentally rename the app Keep Calm" so that he could continue participating without feeling overwhelmed. Although, P7 also felt that "sometimes the device felt like a box I could put my stress in, and move on". Several participants agreed with P4 that even through the self-report prompts occurred very frequently, the interaction required to acknowledge the prompt and complete the associated survey was "very light".

The self-report stress measure was criticized in two ways. First, participants felt that there was need for a "not stressed and not unstressed" option between "feeling good" (2 on a 5-point Likert item) and "a little stressed" (3 on a 5-point Likert item). Second, two participants agreed with P6 who wanted greater resolution of the scale:

> I would compare it to previous times I had answered, and I would feel a little more stressed, but not enough to take it to the next level, so I'd put the same as last time. [P6]

The experience of wearing the Q Sensor during the study was widely reported to be "easy" although one participant found the strap too tight and another commented that the device kept slipping out of contact with her skin. The Affectiva was the only visible aspect of participation in the study and it did

spark some conversation about the device, the study, and the stressors that had been experienced (and noted) by the participants during the study.

Participants did not widely report privacy concerns as a result of participation in the study. P6 explained that any privacy concerns were mitigated by the potential reflective benefit of such a tool: "The question [of whether there are privacy issues implicated in use of the system] is, do you believe in the values of the app? Here, I feel like the app is helping me be aware of and control my stress." P5 noted that while she had few concerns since she was informed about and understood what the system was doing, her friends became uncomfortable when they learned that audio sensing was taking place via her smartphone's microphone; friends of P2 expressed similar concerns. This finding illustrates one of the more complex trade-offs in the design of systems like SESAME: even when approaches can be developed to sense stress without directly burdening the user (e.g., voice-based stress recognition), there may be tacit social costs implicated in these decisions.

The use of the microphone in a smart phone as a sensor for capturing voice, while effective in the lab (Lu et al., 2010), had various consequences in the wild. First, multiple participants reported unexpected holes in the audio profile data with respect to voice-stress. Because SESAME could not access the microphone during phone calls, phone conversations with remote friends and family, not unusually an opportunity to talk through experiences of stress, were not captured. Secondly, even though the device is not actually recording audio files, there is a sense among a few participants, and others around them, that this might be a privacy concern. Allowing users to censure sensed data after the fact has been effective in other physiological sensing systems, such as (Kay et al.,

2012); more control over personal sensed data could also help alleviate privacy concerns in SESAME. Third, the voice-stress classifier used in SESAME did not learn, so using an adaptive voice stress recognition algorithm could improve the classification.

Although our data collection app also recorded location and activity during the study, we do not report on the relationship between location or activity and the other sensed/reported data here. Future analysis will consider the automated recognition of physical regions of interest and their association with stress levels, as well as the potential for using accelerometer-based activity recognition for identifying stress or increasing the robustness of voice- or EDA-based approaches.

While we attempted to control for stress profiles and daily stressors for this study, it is possible that characteristics of our participant population impacted the results. For example, much of their work takes place collaboratively in spaces relatively free of background noise. Our results should be confirmed in larger, more diverse populations.

4.3.5 Conclusions

We examined minimally intrusive mechanisms for measuring, inferring, and eliciting characterizations of a user's stress. We demonstrated that voice- and EDA-based classifiers produce representations of stress that correlate with self-reported measures of perceived stress and with one another in real-world environments. As a result of this research, we identified contexts in which the various sensing approaches are more or less effective (Table 4.1).

Data collection method	Measures	Effective contexts	Less-effective contexts
Experience sampling-driven self-report	Collection of Taylor single-item stress measure, an open-ended rational for stress level, and affect	Scenarios in which the user's subjective perception of stress is valuable; provides (potentially) finer descriptive resolution	When interruptions are not desirable (e.g. work, driving, social engagement); interruptions may adversely influence the user's stress level
Electrodermal analysis (Affective Q Sensor)	Physiological arousal via skin conductivity	Most day-to-day contexts; most valuable when contextual valence already known	Physical discomfort of or preferences against wearing device; expense limits scaling of participant pool
Voice-based stress analysis	Variation in vocal characteristics (pitch, speaking speed, vocal energy)	Many day-to-day contexts in which user will be regularly speaking; where on-device feature extraction possible	Ineffective in quiet or noisy spaces; currently only provides coarse metrics

Table 4.1: An overview of stress detection methods explored in this study and their contexts of use.

We found that self-report remains the mechanism through which the most accurate representations of low and moderate levels of stress can be collected from participants, as well as the only mechanism that can be easily augmented to understand the source of stress; however, contextual augmentation could be provided in the personal informatics reviewing interface. Our results indicate that self-report is also useful for validating or correcting stress models con-

structed based on automatically sensed data. The main drawbacks of self-report are its intrusiveness, which might be mitigated through the use of context-aware prompting, such as (Intille et al., 2003; Klasnja et al., 2008), and the fact that the method is still unlikely to reveal episodes of intense stress, simply because users can choose to ignore experience sampling prompts during those experiences.

Based on our empirical data, EDA- and voice-based stress recognition both provide less invasive yet still reasonably robust representations of stress in real-world environments; certainly when these two channels agree we can be confident the user is experiencing stress. Further research is needed to develop robust sensing of EDA or the intensity of stress using only those sensors users already carry with them. And future work will also need to address a number of weaknesses of voice-based stress sensing that we have identified, such as a higher incidence of false positives and the potential for raising privacy concerns. Several of these false positives can be mitigated by adding a speaker ID filter to the system, while affording user censure of sensed data, for example (D. A. Epstein, Borning, & Fogarty, 2013), should continue to alleviate privacy concerns already partially addressed by the system design.

Stress is a factor in so many facets of health and wellbeing. Our study provides an encouraging starting point for informing the design of minimally invasive UbiComp systems for sensing stress.

CHAPTER 5

CASE STUDY: THE MEASUREMENT OF CHRONIC PAIN

Chronic pain, recurrent or long-lasting pain, affects an estimated 30.7% of US adults, and these numbers are considerably worse for the aging population: more than 50% of older adults and as many as 80% of older adults living in nursing homes (B. A. Ferrell et al., 1995; Helme & Gibson, 2001).

Common chronic pain conditions include osteoarthritis (OA), rheumatoid arthritis (RA), lower back pain, and migraine/headache, as well as injury-related conditions, repetitive stress disorders, and other conditions. Chronic pain is common across a wide range of disease and demographic groups, but is more common in women than in men, and prevalence increases with age. Indicators of poor socioeconomic status, including lower household income and unemployment, are significantly correlated with chronic pain conditions (Johannes et al., 2010). Patients with chronic pain are frequently severely debilitated, with significant limitations in their ability to function or work. Chronic pain is associated with depression, sleep disturbance, fatigue and decreased cognitive and physical abilities. (Ashburn & Staats, 1999).

Chronic pain is best managed with an interdisciplinary model of care that addresses patients' physical, psychological, and social wellbeing. Management techniques include medication, physical therapy, behavioral and psychological interventions, and ergonomic evaluation. (Ashburn & Staats, 1999; Woolf, 2004); high-quality pain management can require physicians to understand their patients' activities, mental condition, subjective experience of pain, and mood, as well as more traditional medical indicators. For example, in the clinical setting a practitioner will take something like the PQRSTU approach: P (provoca-

tion, palliation, and past), Q (quality of pain), R (region of pain, radiation and spread), S (severity), T (timing, onset, time of day, duration), and U (interference, how the pain affects 'u').

User-provided input in this population is essential given that many of the factors and symptoms of interest can only be captured through user input. Indeed, there is evidence that recalling pain and coping strategies can lead to positive outcomes (Haythornthwaite, Menefee, Heinberg, & Clark, 1998), although there is some evidence that drawing the patient's attention to the pain itself has potential negative effects (Younger, McCue, & Mackey, 2009); this is also indicated by the importance of distraction on pain intensity (Kohl, Rief, & Glombiewski, 2013). This motivates a minimally intrusive method of obtaining self-report data about pain.

In order to provide better care (for example, around times of treatment titration), there is a desire for a far higher data resolution for pain reporting than simply once every clinic visit. And indeed, self-management practices indicate value in regular, frequent measurement. Furthermore, a growing body of literature suggests that pain reporting as it is typically performed is affected by recall bias (e.g. Eich, Reeves, Jaeger, and Graff-Radford, 1985; Niven and Brodie, 1996). As in the domains of affect and stress, pain researchers are turning to Experience Sampling and Ecological Momentary Assessment.

Stinson et al and de la Vega et al appear to be leaders in considering both device- and user-specific usability issues as well as rigor in validity Stinson et al., 2013; de la Vega et al., 2014. However, there does not appear to be a standard way either of translating and deploying self-report measures to the smartphone platform, and this is true in the context of pain reporting. As a result, there is

a broad practice of simply deploying variations of various pen-and-paper measures in novel technologies and with alternate interaction and interface functionality without overly attending to issues of usability and feasibility that are compounded by age-related differences in ability and familiarity with mobile devices.

In this chapter, then, I take a comprehensive user-centered approach to investigate the ways the those with chronic pain self-report chronic pain, the reporting modalities they prefer and dislike in the mobile context, and how repeated and frequent reporting changes when using one's own mobile device. I report on two design projects: first, iteratively refined smartphone interfaces for the momentary self-report of chronic pain intensity, and second a novel tangible user interface that supports the self-report of scalar values (in this case, pain intensity) without even reaching for the phone.

Parts of this work are in submission at a special issue of JAMIA (Meter) and at the ASSETS conference (Keppi).

5.1 How chronic pain is understood

In the self-report of pain, it has long been agreed that multiple dimensions are necessary for an adequate pain representation (Charles S. Cleeland et al., 1996). But the dimensionality has (and continues to be) a moving target in the research. Melzack and Casey suggested the three dimensions of discriminative, motivational-affective, and cognitive-evaluative based on the neurophysiology of pain mechanisms (Melzack & Casey, 1968) and then found three typical response types in the words patients used to describe pain (Melzack & Torgerson,

1971). Commonly, two dimensions variously called pain and reaction to pain (Beecher, 1959), sensory-discriminative and attitudinal (W. C. Clark & Yang, 1983), and sensory and reactive (C. S. Cleeland, 1989).

Today, these two dimensions are commonly labeled pain intensity (how much it hurts) and pain interference (what the pain prevents me from doing); (Charles S. Cleeland et al., 1996) uses multidimensional scaling to show that pain interference is likely a combination of (a) interference with mood, life enjoyment, and relationships with others, and (b) interference with walking, work, sleep, and activity. As the "intensity of pain is without a doubt the most salient dimension of pain" Turk and Melzack, 2011, in this work I attend to the frequent *in situ* self-report of pain intensity.

5.2 How chronic pain is measured

5.2.1 Pen-and-paper self-report

Chronic pain has been measured through self-report (and reported this way in the literature) since the late 1940s with the simple descriptive pain scale (Keele, 1948): pain intensity is self-reported on a 5-point scale: none, slight, moderate, severe, and agonizing. I include here several of the most widely used self-report assessment tools: the VAS-P, the NRS-11, the BPI, the SF-MPQ, and the FPS-R.

The Visual Analog Scale for Pain, or VAS-P, is a unidimensional measure of pain intensity that has been widely used in diverse adult populations (Woodforde & Merskey, 1971; E. C Huskisson, 1974; Scott & E. C. Huskisson, 1976). It

is a continuous scale represented as a horizontal or vertical line (typically 10 centimeters in length), with text descriptors at each end the describe the extremes of the scale. These descriptors, as well as instructions for use, vary widely in the literature, although the endpoints "no pain" and "pain as bad as it could be" are not uncommon. Intermediate text labels are not recommend, mostly to avoid clustering reports around these positions. The VAS-P takes less than a minute to complete, and is easy to learn and administer. Scoring the pen-and-paper version can be burdensome. It cannot be administered by telephone.

Participants are asked to place a line on the scale at the point that represents their pain intensity; scoring the pen-and-paper version involves using a ruler to measure in millimeters (0-100) how far along the line the participant has responded. Higher scores mean more pain. Variations intended to improve both usability and ease of scoring include the Mechanical Visual Analog (D. D. Price, Bush, Long, & Harkins, 1994) and Pain-O-Meter (Gaston-Johansson, 1996) which have a built-in slider the patient moves to indicate their pain level; on the reverse of the measure the researcher can read off a numerical measure from the slider position.

Test-retest reliability is high, but lower among illiterate patients (Ferraz et al., 1990). It correlates well with the simple descriptive pain scale and the NRS (Downie et al., 1978). The VAS-P is considered highly sensitive to change (Hawker, Mian, Kendzerska, & French, 2011, S11).

There is some disagreement over the relative efficacy of the horizontal and vertical versions of the VAS-P. (Scott & E. C. Huskisson, 1979) indicates that the correlation between the two is 0.99 but finds a more uniform distribution of scores with the use of the horizontal VAS-P; (Stephenson & Herman, 2000)

suggests that the vertical version is easier to see and correlates better with the SF-MPQ's PPI and also notes that the vertical version may result in higher reported values.

The NRS is a unidimensional measure of pain intensity, intended for use with adults. There are many versions (using different integers and number of segments), but the 11-item (Farrar, Young Jr, LaMoreaux, Werth, & Poole, 2001) has become standard. The common format is a horizontal line anchored by text describing intensity extremes (C. Johnson, 2005), just like the VAS-P. The NRS takes less than 1 minute to complete, and is easy to administer and score. Chronic pain patients find the NRS easy to complete, but it may be inadequate for capturing complex changes (Williams, Davies, & Chadury, 2000).

Can be administered using pen-and-paper, computers, or verbally. The participant is asked to choose the numeric value on the scale that best describes their pain intensity. Higher scores mean more pain. The NRS-11 has high test-retest reliability, and is highly correlated with the VAS. On average, a reduction of 2 points or 30% indicates a clinically important change; percent change is consistent over baseline levels while higher baseline scores require an accordingly larger reduction in raw change (Farrar et al., 2001).

The decision to use the NRS-11 or the VAS-P for the self-report of pain is a complex one. Simple differences are discussed in (C. Johnson, 2005), while (Serlin, Mendoza, Nakamura, Edwards, & Cleeland, 1995) offers a detailed set of reasons why the NRS-11 should be preferred. Their relative usability have been compared (Williams et al., 2000), and the conclusion is often that the VAS-P is more sensitive and mathematically useful (D. D. Price, McGrath, Rafii, & Buckingham, 1983), while the NRS-11 might be easier to administer and interpret

(Bolognese, Schnitzer, & Ehrich, 2003).

The Faces Pain Scale Revised (FPS-R) is also a unidimensional measure for the assessment of pain intensity (Hicks, von Baeyer, Spafford, van Korlaar, & Goodenough, 2001). Instead of a continuous range or sequence of numerical values, the FPS-R asks the participant to select from a range of cartoon faces that face that best represents how much pain she is experiencing. Like the original FPS (Bieri, Reeve, Champion, Addicoat, & Ziegler, 1990), the scale was intended for use with children who might struggle with the abstractions inherent to the VAS-P or NRS-11. The FPS-R improves upon the original (a) by providing an easy scaling on a 0-10 range, and (b) by ensuring equal intervals in pain intensity between each of the 6 faces. The FPS-R has since been shown to be valid for use with adults (Stuppy, 1998), and correlates well with the VAS-P (Hicks et al., 2001) and other VAS-style measures (Miró, Huguet, Nieto, Paredes, & Baos, 2005).

The next two measures are multidimensional measures, attending to both pain intensity and other pain qualities.

The Short-form McGill Pain Questionnaire (SF-MPQ) (Melzack, 1987) was developed because the original MPQ (Melzack, 1975) can take 5-20 minutes to administer, and because the VAS-P and Present Pain Intensity (a 5-item measure from the original: mild, discomforting, distressing, horrible, excruciating) only consider intensity. The measure does correlate well in all dimensions with the original.

The measure consists of three items. First, a set of 15 descriptors (11 sensory, 4 affective, such as ‘throbbing’ and ‘fearful’) that are rated on a 4-item scale

(none, mild, moderate, severe). Then the VAS-P, and lastly the PPI described above. Completing the MPQ can require a sophisticated vocabulary and may not be appropriate for low literacy respondents. Sex and ethnic differences may exist in both children and adults in the selection of pain descriptors.

The Brief Pain Inventory Short-form (BPI-SF) is a comprehensive pain self-report tool (C. S. Cleeland, 1989) that includes four pain severity items (an NRS-11 for each of pain intensity 'right now', 'at worst', 'at least', and 'on average' anchored with 'no pain' and 'pain as bad as you can imagine'), seven pain interference items (an NRS-11 for each of general activity, mood, walking ability, normal work, relations with other people, sleep, and enjoyment, anchored with 'does not interfere' and 'completely interferes'), a body map for localizing bodily pain, and information about current treatments. The measure has been used in over 400 studies worldwide.

Lastly, the SF-36 Bodily Pain Score (BPS) is one of eight subscales of the Medical Outcomes Study, and assesses bodily pain as a dimension of health status (McHorney, WARE JOHNE, & ANASTASIAE, 1993). It contains two items: a 6-point NRS for pain intensity anchored with 'none' and 'very severe', and a 5-point NRS for the extent to which pain has interfered with work anchored with 'not at all' to 'extremely'. The two values are combined (using the simple algebraic sum) and then scaled to a range 0-100; population normative data can be included such that a score of 50 is the 'average' for the population and a single standard deviation is 10 points; population normative data is available for the US and the UK. The SF-36 BPS has been shown to be both reliable and valid in multiple populations, has been shown to detect change well, and takes less than two minutes to complete.

5.2.2 Self-report of pain electronically

The self-report of pain using electronic reporting systems has kept track with experience sampling and ecological momentary assessment in other domains. For each self-report measure (including those described above), there are confirmatory studies ensuring the measure's validity and applicability when deployed electronically, with an additional note that the electronic version tends to be preferred by subjects; it is not clear if this is a function of novelty, utility, user experience, or something else - although all the normal wins for EMA appear to also apply to pain self-report (Gaertner, Elsner, Pollmann-Dahmen, Radbruch, & Sabatowski, 2004). Nearly all occur in a clinical or laboratory setting, and not in natural environments. Such confirmatory studies have been performed for the FPS-R (Wood & von Baeyer, 2011), the VAS-P (Jamison et al., 2002; Reips & Funke, 2008), and the SF-MPQ (Cook et al., 2004).

There does not appear to be a standard way either of translating and deploying all (or even each) measure, although Reips et al have made an attempt to do so for HTML-based VAS-P with their VAS generator (Reips & Funke, 2008). As a result, there is a broad practice of simply deploying variations of the above measures in novel technologies and with alternate interaction and interface functionality without any of the above confirmations. Stinson et al appear to be leaders in considering both device- and user-specific usability issues as well as rigor in validity (Stinson et al., 2013), while Aaron et al report the important finding that although participants view daily electronic assessments of pain to have both positive and negative effects, pain-related measures do not show reactive effects (Aaron, Turner, Mancl, Brister, & Sawchuk, 2005).

Building on earlier work by (Stinson, Petroz, et al., 2006), Jibb outlines a set

of four system requirements for a comprehensive pain management electronic application (Lindsay A Jibb et al., 2014). These are (1) overall endorsement of the system, (2) the need for a clinical expert, (3) the need to individualize the system, and (4) changes over time to improve clinical effectiveness.

We are in an interesting time when a lot of the development of mHealth applications, including self-report measures, are being performed outside the research community by large companies through single-person start-ups. Again, many of the pain self-report measures are creative variations of the VAS-P or simple descriptive scale - and the VAS-P and NRS-11 in particular have been shown to be influenced by simple changes such as the measure description or the text anchoring each end. Examples include Catch My Pain[1] and the Chronic Pain Client[2].

5.3 Effective mobile measures for pain (Meter)

In the Meter project, I take an iterative user-centered research through design approach. The project consists of four stages: design ideation and review, an in-lab user study with 10 participants and 9 candidate interfaces, a 3-week 12 participant field trial with two iteratively refined measures, and an expert review with a panel of researchers and practitioners in chronic pain.

[1]https://play.google.com/store/apps/details?id=com.sanovation.catchmypain.phone
[2]https://play.google.com/store/apps/details?id=com.program.chronicpainclient

5.3.1 Design ideation and review

We drew on the self-report and usability literatures, as well as our experience developing and deploying ESM systems, to ideate a large number of potential interfaces. We were holding several design constraints in tension:

- Design to be highly usable: particularly as one target population was older adults, and recognizing age-related changes in psychomotor and perceptual capabilities, any reporting interface needed to use large touch-target regions, readable fonts and font sizes, simple interaction(s), and given potential hand or wrist pain, low manipulability. We knew touch-based interactions tend to be well-received by older adult populations, but how did this change for those with certain types of physical pain?
- Design for the smartphone medium: working with the smartphone as our starting point meant considering affordances not available in (e.g.) traditional pen-and-paper instruments. For example, touch, responsive/reactive displays, interfaces with memory of past reports, animations, and more. What interactions might be possible, useful, or effective?
- Design for repeated use: participants and then end-users would be using these interfaces repeatedly, frequently, perhaps over periods of years. What elements of playfulness or delight might we embed in such interfaces to motivate continued use? How might/should the interfaces change over time, with repeated use? A measure that is okay to complete once or infrequently can become incredibly frustrating when engaged with more often.
- Design to support cognitive translation processes: as in our work with

affect or stress reporting, can we support the effort of translating a complex subjective experience into what may well become a single numerical value? Perhaps by way of visual/audio aids that have 'pain legibility', or by drawing on (e.g.) a framework describing attitude report processing such as (Krosnick et al., 2005)?

- Design to elicit reports on sensitive content: the interfaces would be asking end users to focus on their pain experience, potentially making pain more salient or increasing anxiety, among other potential negative effects (Younger et al., 2009). Could we design interfaces that abstracted away from or ameliorated this process?
- Design for *in situ* use: end users would be using these measures at home, but also potentially at work, in social situations, or in public. The interfaces should be usable on-the-go. And, combined with the sensitive/private nature of pain, *in situ* use informs interaction modality selection - perhaps voice-based reporting would make less sense.

The resulting set of sketched interfaces each attempted to consider these constraints. We were additionally informed by the design thinking around the microinteraction: a small piece of functionality that allows you to complete a single task. Examples include setting your phone volume to vibrate or logging in to a service, but also perfectly describe a momentary self-report measure. A microinteraction consists of four components: a trigger that initiates it, rules that determine how it functions, feedback generated by the rules, and loops and modes making up its meta rules (Saffer, 2013). These aspects sit will within HCI fundamental design principles for interaction: (as described by Norman) 'what do I want to accomplish', 'what are the alternatives', 'what action can I

do now', 'how do I do it', 'what happened', 'what does it mean', and 'have I accomplished my goal?' (Norman, 2013).

By way of discussion and (HCI) expert review, we selected seven novel designs. We then directly translated the NRS-11 using native Android widgets and included a usability-improved version of the VAS-P. These nine measures were then prototyped for the Android OS as custom Views. We describe the resulting nine candidate measures:

NRS-11 a standard NRS-11, vertical orientation, using native radio buttons. Prompt *Please indicate your current pain level with the following NRS*, anchored at bottom with *No pain at all* and at top with *Worst pain imaginable*.

SuperVAS a standard VAS-P, vertical orientation, using a custom seek bar widget that dramatically increases the target touch region. Prompt *Please indicate your current pain level*, anchored at bottom with *No pain at all* and at top with *Worst pain imaginable*.

SuperVAS+ a non-standard VAS-P, vertical orientation, using a custom seek bar widget taking the form of a screen-wide rectangular color block. Prompt *Please indicate your current pain level by moving this bar*, anchored at bottom with *No pain at all* and at top with *Worst pain imaginable*.

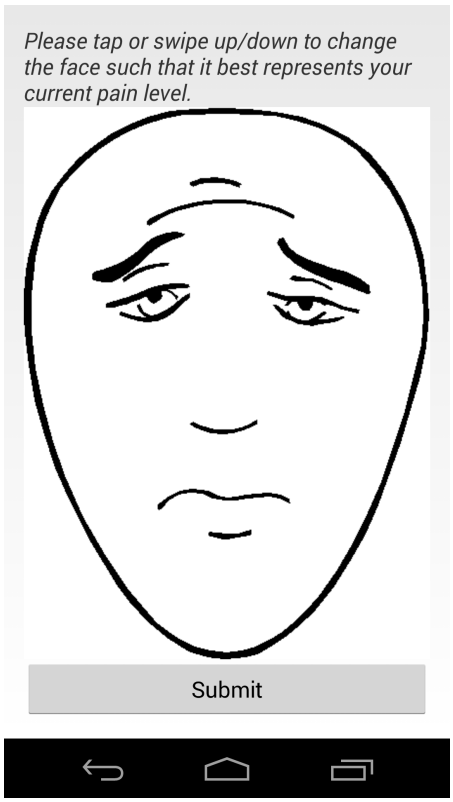
Suureta An empty circle that slowly becomes filled with color when any part of the interface is touched and held. Prompt *Please touch and hold the screen to report your current pain level from empty circle (no pain) to full circle (worst pain possible)*. Suureta can be completed even by resting a knuckle, nose, or chin on the screen. Figures 5.4a and 5.4b.

Many Fingers respondents report their level of pain by touching the interface with a certain number of fingers simultaeneously. Prompted 0...5, but on most hardware capable of sensing up to ten simultaeneous touch events. Prompt *Please touch the screen with 1 to 5 fingers to report your current pain intensity level (5 = worst possible)*. Figures 5.3a and 5.3b.

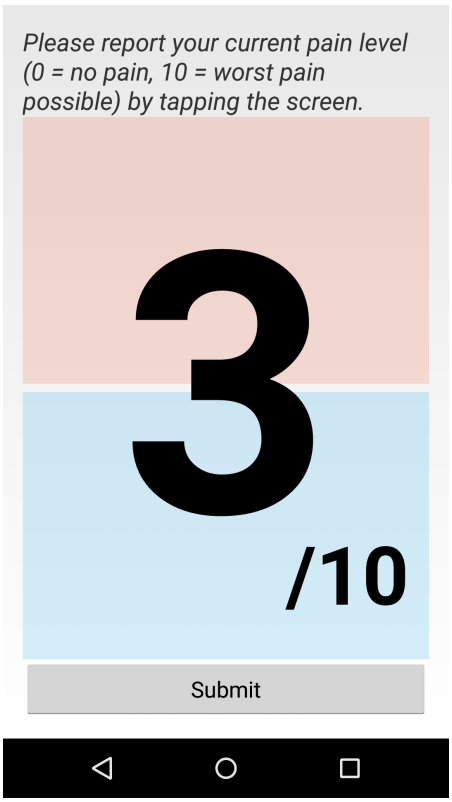
Tap Tap essentially an oversized number picker widget for reporting on a numerical 0...10 range. The currently selected value is displayed in a very large font on the screen; tapping anywhere on the top half of the screen increments the value; tapping anywhere on the bottom half of the screen decrements the value. Visual feedback is provided by upping the saturation of the tapped region. Prompt *Please report your current pain level (0 = no pain, 10 = worst pan possible) by tapping the screen*. Figure 5.1b.

PhotosPeople essentially an NRS-6 using photographs of people's faces instead of numbers to display each level of pain. To generate a set of photos portraying varying levels of pain, we followed a similar protocol as we did in generating affect-laden images for PAM (Pollak et al., 2011): starting with a list of 18 descriptive words describing degrees and qualities of pain we downloaded a 500 images from the Flickr API for each word [3] [4]. A large number of these images, particularly at the 'high pain' end were really quite intense - reminding us of IAPS (Lang, 1995). For our prototypes, we selected faces and photographs that conveyed varying pain levels spread over the NRS range, but that also we believed would not provoke a strong negative response in our participants. We arranged the photos in an NRS to have a transparent underlying logic to the instrument's arrangement; in this prototype we were particularly contrasting

[3] https://www.flickr.com/services/api/
[4] https://github.com/philadams/flickr-images-grab



(a) Meter: 'SAFE'. Tapping on the screen between the two anchors *No pain at all* and *Worst pain imaginable*, or sliding a finger up or down the screen, smoothly transitions across the faces images.



(b) Meter: 'tap tap'. Essentially a supersized number picker. Tapping in the red section incremented the reportable value; tapping in the blue section decremented the reportable value.

Figure 5.1: Two of the candidate measures resulting from the design ideation stage.

the two design tensions of *supporting cognitive translation processes* and *reporting sensitive content*.

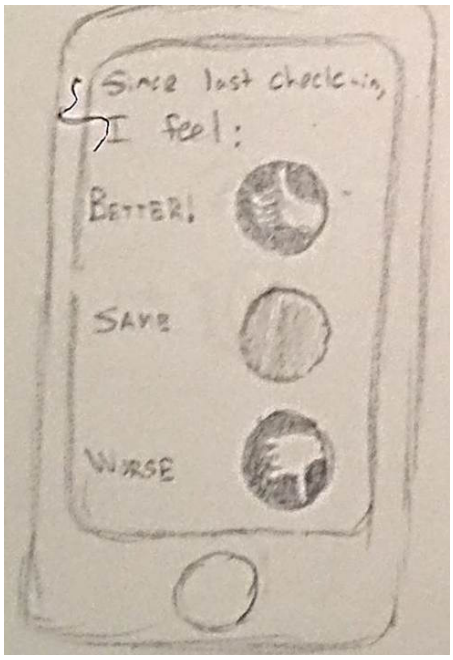
PhotosLandscape Analogous to the above measure, but with abstract and landscape images representing pain levels rather than faces.

SAFE We read with interest that to create the FPS-R, (Hicks et al., 2001) used a computer animated cartoon face created by Champion et al (Champion, von Baeyer, Trieu, & Goodenough, 1997) to select four faces representing pain levels

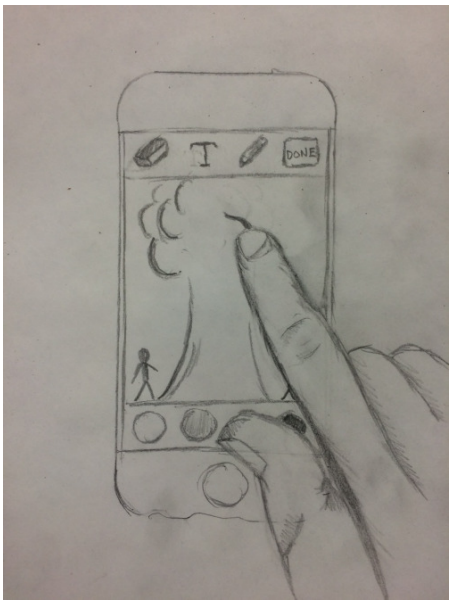
at equal intervals between the two anchor faces. The animated face is made up of 101 separate face-in-pain images, and all 101 images have been deployed in a laptop-based self-report scale for children, such that the face pain level can be incremented or decremented by pressing the left/right arrow keys (Good-enough et al., 2005). With permission from the authors, we prototyped SAFE such that tapping on the screen between the two anchors *No pain at all* and *Worst pain imaginable*, or sliding a finger up or down the screen, smoothly transitions across the faces images. It seemed a much more natural interaction for the smartphone, and side-stepped the issue of displaying 6 face images on such a small screen.

Several intriguing design directions were not selected, but are worth mentioning here:

Several ideated self-report approaches pushed further on the idea of reporting relative to recent pain experiences, either having the smartphone measure draw on its memory of recent responses, or by prompting the participant to do so (Figure 5.2a). This design direction was informed by both an understanding that the data stream resulting from repeated use would offer local and global up/down trends - so why not simply ask the end user directly to provide local trends? Further evidence came from the literature: Joyce et al report that patient preference for seeing their previous scores or not when reporting is 3:1 (Joyce, Zutshi, Hrubes, & Mason, 1975). However, meaningful reporting in this style requires the user to have memory of their previous n reports, or to make sense of a reminder, and opens up a series of pyschometric bias issues. We are exploring this direction further, however, in a project on what we're calling 'ecological momentary feedback'.



(a) Meter sketch: 'vs last time'. On-device instruments can have memory of the previous n reports, and/or ask respondents to report relative to the last n reports.



(b) Meter sketch: 'doodles'. Prompt for a doodle or sketch or phrase, giving great freedom to the respondent to report as they choose.

Figure 5.2: Two promising design directions that were not pursued.

We have had success in the past having participants capture photographs with the smartphone as part of an ESM self-report data point, and reported on outcomes of such systems (Baumer et al., 2012; Adams et al., 2014). There are also encouraging (although early stage) efforts in the automatic sensing of pain levels from facial expressions (S. Wang et al., 2010; Tian, Kanade, & Cohn, 2005). A couple ideated designs therefore involved the use of one of the device cameras, and prompting the user to take a picture of their face (or other photographic representation of their pain level). Issues resulting in this direction not being pursued further include differential comfort with mobile photography in the older population, a limited (or at least, under-reported) selfie tradition in the older population, the difficulty in using the front- or rear-facing camera to

capture a pain expression, and lower confidence in the system's likely ability to consistently and meaningfully translate from expression to a pain score.

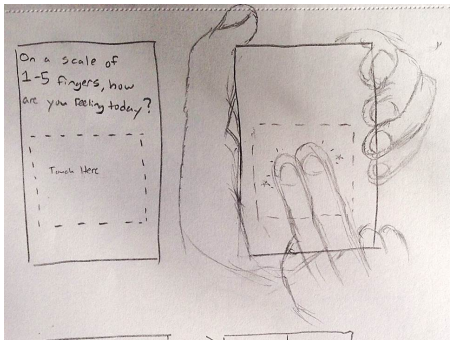
Finally, a promising design direction was in drawing and doodling interfaces (Figure 5.2b). Drawing with one's finger on the smartphone is quick and easy, and we had experience effectively using image-based ESM data. Furthermore, there is a good sized body of research indicating the positive therapeutic value of painting and drawing across chronic conditions, including pain (Nainis et al., 2006), and we were inspired and moved by community galleries[5] which give voice to chronic sufferers and express the various facets of their lived experiences. As with the photo-based interface, we concluded that translating from doodle to a utilitarian value (e.g. for use by clinicians) would be inconsistent. However, both these design directions suggest new and potentially rich avenues for pain journaling and self-reflection. We hope to explore them further.

5.3.2 In-lab user study

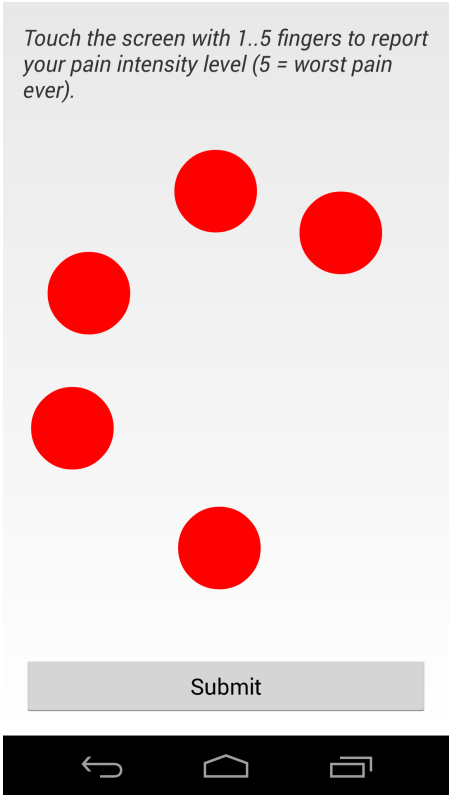
We built out interactive prototypes of each novel measure, as well as directly translating two widely used pen-and-paper measures (NRS-11 and VAS-P) to the smartphone medium using the native platform standard widgets, as has been commonly done in others' work. We then iterated on the VAS-P, custom building a visually similar version that behind the scenes considered some of the above-described tensions.

We then recruited N=10 lab study participants, each of whom experienced some form of waxing and waning chronic pain. Recruitment was variously by

[5]http://painexhibit.org



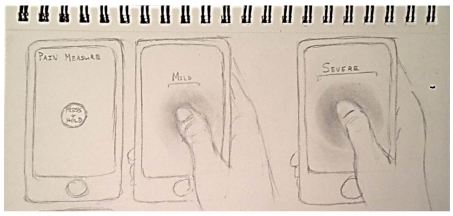
(a) Meter sketch: 'many fingers'. Taking advantage of multi-touch affordances, the user touches the screen with 1 to 5 fingers simultaneously.



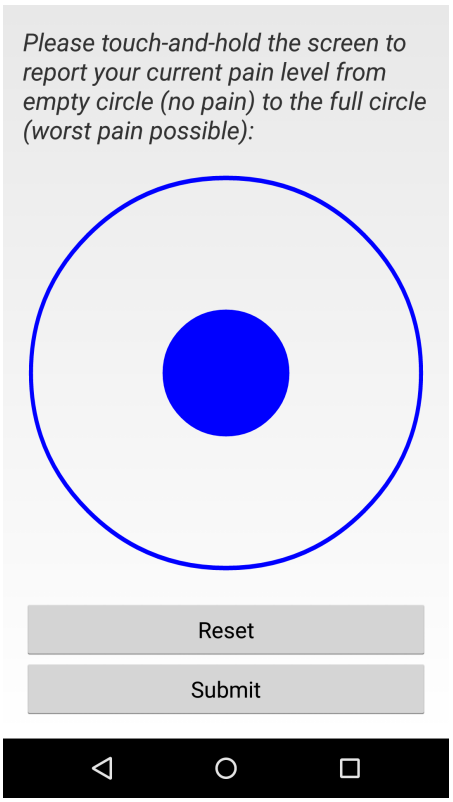
(b) Meter: 'many fingers'. Taking advantage of multi-touch affordances, the user touches the screen with 1 to 5 fingers simultaneously.

Figure 5.3: Sketch and resulting interface of Meter 'many fingers'.

way of the Wellness, Retirees, and Elder Care campus mailing lists, flyers at the Physical Therapy and across campus, and through snowball sampling. Of the ten participants, nine were female, five were 55 or older, and 9 were reasonably to very comfortable with smartphones (the last owned one, but used it only as an emergency device and map). Gender skew is not uncommon in smaller-scale pain studies, for example 51/78 participants were female in (Guillory et al., 2015). Participants' recent experiences with pain ranged from low-level recurring joint pain or injury recovery through severe arthritis or back and neck pain from a recent car accident.



(a) Meter sketch: 'suureta'.



(b) Meter: 'suureta'.

Figure 5.4: Meter sketch: 'suureta'. The user simply touches and holds anywhere on the screen, and circle slowly grows over time. We anticipated users being able to complete this meter with a fingertip, knuckle, or even their nose or chin.

Each lab session had three sections. In the first, we asked questions about the participant, their recent experience(s) with pain, how they had reported pain in any setting in the past, their self-tracking practices (if any) and familiarity with smartphones. In the second section, the participant interacted with each of the 9 candidate measures, used it to report their current pain level, and provided feedback (we also took observational notes). The final section was reflective, compare/contrast (both qualitatively and via ordinal rankings across four usability dimensions), and offered an opportunity for further free response.

Participants were compensated $10 for the 40-60 minute in-lab user study.

Cornell's Institutional Review Board approved this procedure.

5.3.3 *In situ* field trial

Drawing on the findings from the in-lab study, we retained a 'numbers people' interface and a more abstract interface. The two measures used in the field trial were *SAFESlider* (Figure 5.5) and *SuperVASNumbered* (Figure 5.6).

The two measures have the same dual interaction modes: tap anywhere on the screen (as if underneath the screen lay a standard NRS-11 or VAS-P with low pain at the bottom through high pain at the top), or swipe anywhere on the screen to control the slider on the side.

These two candidate ESM measures, then, were deployed in an Experience Sampling style three-week field trial. Participants, again each experiencing some form of waxing and waning chronic pain, were recruited variously by way of the Wellness, Retirees, and Elder Care campus mailing lists, flyers at the Physical Therapy and across campus, and through snowball sampling.

Of the N=12 participants, nine were female, four were 50 or older, and all reported be at least comfortable with their Android OS smartphones. Again, participants' current experiences with pain ranged from low-level recurring joint pain or injury recovery through severe arthritis, fibromyalgia, or compartment syndrome.

Participants installed two applications onto their own Android phone. The first is Meter (https://github.com/philadams/Meter), an app we developed to house prototypes of several of the ideated reporting interfaces described above.

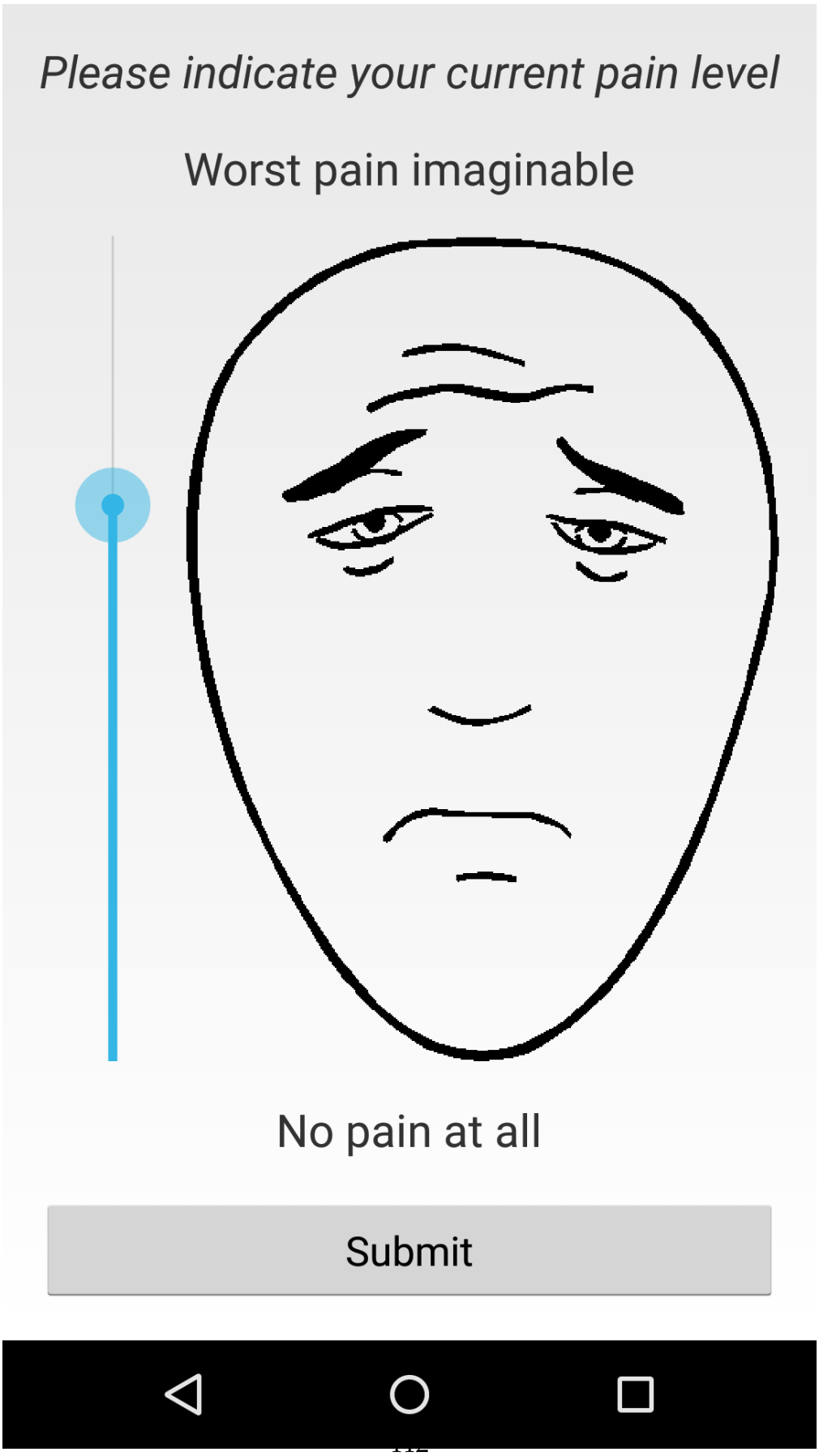


Figure 5.5: Meter: ‘SAFE slider’. The user reports a more qualitative pain level using cartoon faces by either touching anywhere on the screen, or by sliding a finger up or down anywhere on the screen.

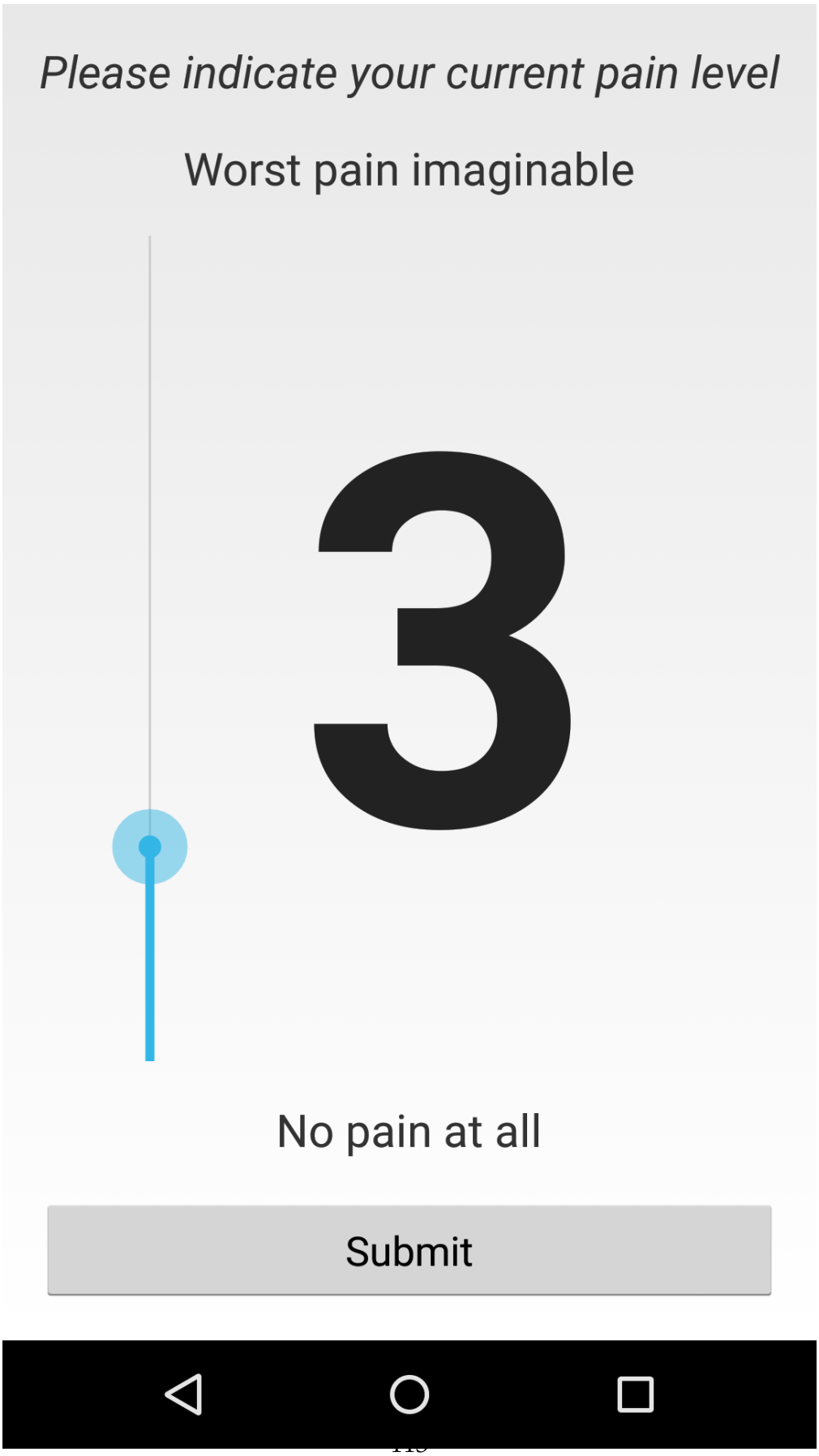


Figure 5.6: Meter: 'super VAS numbered'. The user reports a pain level from 0 to 10 by either touching anywhere on the screen, or by sliding a finger up or down anywhere on the screen.

The second is Ohmage (http://ohmage.org), an open-source participatory sensing platform; one of the authors is a co-PI responsible for the technology behind Ohmage. Among its various capabilities, the Ohmage platform allows researchers to configure and deploy Experience Sampling Method-style electronic diary studies with a variety of question types; one question type is *remote_activity*, which will launch any screen from any application on the device. This allowed us to have the Ohmage survey to launch a particular candidate reporting interface from within Meter as a survey question, and to then persist data about the participant's interaction with the measure (e.g. the selected pain value or the time taken to complete the item) back to the Ohmage server.

Twice a day, once in the morning and once in the evening, participants were prompted via a notification from Ohmage to complete a short 1-3 minute survey. Each survey consisted of four questions:

1. pain level self-report using one of the candidate reporting interfaces
2. subjective usability rating of the above interface using a 5-point likert scale
3. unstructured feedback about the above interface (optional)
4. pain level self-report using Ohmage's version of an NRS

At the end of the three weeks, all but two participants completed a 30-40 minute follow-up semi-structured interview about the candidate measures and their experience in the field trial. Participants were compensated $10 for each week of ESM survey responses and $10 for the follow-up interview for a total of $40. Cornell's Institutional Review Board approved this procedure.

You can review the participant enrollment instructions at http://cornellhci.org/meter.

5.3.4 Expert panel

This work was presented at a Works-in-Progress meeting of the Translational Research for Pain on Later Life group[6]. TRIPLL represents a collaboration among investigators at Weill Cornell Medical College, Cornell University (Ithaca campus) and the Hebrew Home at Riverdale. TRIPLL also maintains ongoing partnerships with Columbia University's Mailman School of Public Health, Hospital for Special Surgery, Memorial Sloan Kettering Cancer Center, Visiting Nurse Service of New York and the Council of Senior Centers & Services of NYC, Inc.[7].

A work-in-progress paper was disseminated one week ahead of the review. Each expert shared initial thoughts, and then an unstructured panel conversation followed.

5.3.5 Results and discussion

I first discuss qualitative feedback from the in-lab study and the field trial. I then provide Meter use statistics from the field trial, before concluding with feedback from the expert panel.

It became clear quite quickly that our participants were diverse in the ways that they thought about their pain and the ways they wanted to report their pain. Over the themes below, there are not strong correlations between such preferences and demographics, or level or type of pain experienced. In these sections I include a large number of quotes, reporting with the participants'

[6]http://tripll.org
[7]http://tripll.org/about-us/background/

voices.

Reporting pain with faces

Participants had strong feelings either for or against the use of faces for reporting pain levels. Those who valued reporting with faces could connect with the people and expressions in SAFE and PhotosPeople: "[SAFE] was the most fun, and the photos with people best describe my pain most times." P1.8; "It's a little bit different as to what I'm expressing compared to the number based ones... there's a cold hard number vs communicating what I'm feeling. I can kind of imagine him [SAFE] feeling what I feel, feeling more than just the sterile measure." P1.3; "I can definitely imagine feeling the way this person looks. [SAFE]." P1.2; "With the numbers my brain needs to actually think more about the body part and the pain sensation explicitly and somehow try to map that onto a number; face is a different channel, goes through emotion and feeling" P1.3; "I really think [SAFE] helped me to capture what my true level is." P1.10.

The majority of participants felt that mapping pain to facial expressions just could not work for them. General responses included: "I don't like coordinating pain to facial expressions, because I just can't believe it's accurate" P1.4; "I can't relate to this [SAFE], I don't think I've ever looked like that and I've been in a lot of pain." P1.7; "I couldn't use an app based on just faces. Like when they have the faces [FPS in a clinical setting] that have the numbers too, I don't even look at the faces [waves them away with her hands] I just use the numbers." P1.7; "Because the people's faces it's hard to tell how much or how little pain they express, but if you've got color or a storm, it's easier to gauge how severe [the pain] is. P1.9"; "I don't know what pain looks like on a face, my face doesn't

look like this [SAFE]." P1.7; "It's hard having to rely on the accuracy of these [face pictures]... I don't think this is a good way to quantify pain, for me." P1.5; "The face itself can make me confused" P1.1.

Several participants had more visceral responses to PhotosPeople or even SAFE, and reported feeling uncomfortable with images of humans (or representations of humans) in pain: "Okay this I really don't like. I don't like seeing people in pain" P1.5; "I'm way too distracted by the actual faces of these people to figure out what's going on, to report my pain [PhotosPeople]." P1.7; "With [SAFE] I could disassociate from the emotional impact [in PhotosPeople] of seeing a human in pain" P1.5; "I don't have a problem reporting pain with people's faces, but I feel sad for the sadness of these people's faces" P1.6.

Several participants rejected using PhotosPeople and even SAFE because the human representations did not look like or connect with their own faces or self-image: "I wonder if the male/female thing is a problem, you know if I'm a female I won't choose the male pictures." P1.4; "The face also looks ageless, sexless, it's almost like a parody of what a human face should be [SAFE]." P1.1; "I am in pain, but I don't look like any of these people" P1.4; "I don't know if it would make a difference if, someone was a woman, if all the pictures were women. You have to be able to see yourself there." P1.2; "I think the [SAFE] person was enough of a blank canvas that I could express myself through it but these are people [PhotosPeople] who are different from me, so that's another layer of disconnect... when the person in the picture is different from me, anything about the picture that I can't connect with myself individually, it's harder to pick that one, to connect with myself. even if it's the right degree in the scale." P1.3

It is quite possible that beyond individual preferences, or indeed perhaps correlated with preference, there is are individual differences regarding the ability to make sense of the levels of pain expressed by facial expression. Certainly there does appear to be growing and recent evidence that in non-clinical populations there is a range of capabilities in facial expression processing (Palermo, O'Connor, Davis, Irons, & McKone, 2013). In addition to the visceral responses some participants have in reporting pain with facial expressions, it could be that having to report with one large face in Meter, vs optionally using one of several faces (among an array of other non-face-based images) in PAM could explain some of the negative reactions to face-based reporting.

Conflating pain and affect through faces

Some participants questioned whether SAFE was really about pain levels, or about affective state, for example: "This [SAFE] is of course a more qualitative representation... but does it represent my current feeling about pain? You're asking more about my current mood." P1.1. At the very least, there was a sense that the faces supported reporting mood through a different 'channel' than did numerical scales: "With the numbers, my brain needs to actually think more about the body part and the pain sensation and somehow try to map that onto a number... with the faces it is a different translation channel that goes through emotion and feeling, thought the underlying component is still my pain." P1.3

'Numbers people'

It also became quite clear that about half the participants had a strong preference for reporting with numbers, while the others rejected numbers and strongly preferred something more abstract or qualitative, or at least something they could play with and 'dial in'. The use of words like 'qualitative', 'precise', 'accurate', 'scale', 'subjective', 'objective', and more were all used to mean various things and to apply almost equally to more numbers-based interfaces and more abstract interfaces. Lastly, there's a long history of using 0..10 and 1..5 scales both in pain reporting but across domains such as restaurant and movie ratings. Mathematical representations are culturally dependent and learned over time so we cannot discount familiarity. "Everywhere you're taught to think in, 1..10 what's your pain level" P1.4; "I'm very familiar with this one [NRS-11], I've this numbers of times before." P1.5; "The number system [NRS-11] is very familiar, they use this in the hospitals and doctors' offices" P1.2.

Numbers people said things like: "I tend to be a numbers person, so I like numbers. I relate to them" P1.5; "I don't like this one, I'm too analytical so I can't handle it because I can't quantify 20% just by looking at it. I don't like that, I'm done [closes Suureta]" P1.7; "Even though it's hard to put a feeling like pain into percentages I have an easier time doing that than by doing visual aids and trying to think about pain... I want a scale! I want numbers on everything!"; "I guess we're used to dealing with numbers rather than a scale." P1.2; "I always think of a number out of 0..10 initially. So I mentally compute '3' and then try to use the interface to map into that measure" P1.3.

Participants who were not numbers people said things like: "It's hard to put a numerical value on something like discomfort" P1.9; "With [SuperVAS] you

don't have to be analytical and pin it down... the task-oriented ask very specific questions, analytic questions, which can be uncomfortable for me." P1.9; "Well it's vague [NRS-11]. It really isn't 'pain' at all. It's common, we rate a lot of things in life, so I can see why people might like to rate things this way but, I'm more of a qualitative person." P1.6; "I like the fact that there are no numbers involved. i think you can really be more accurate." P1.4; "Once again we're back to the numbers. numbers might make more sense for a health professional, but I'm not sure they do for me. No numbers is somehow more descriptive. I think it's just got to be easier for a patient to do that, than try to put a number on it" P1.4.

Many participants valued being able to adjust their reported score, to dial it in: "You've got the ability to move up and down while you think about it, until it kind of feels right [SuperVASPlus]. That's kind of what i did even with the first one [SuperVAS]." P1.4; "It's pretty cool that you can slide it up and down" P1.10; "I can change my mind because I can go up and down." P1.2; "It's just nice to be able to move things up and down without having to hit a whole bunch of different buttons" P1.10; "I can still go up and down a bit, that's a good thing [SAFE]" P1.8. That said, a single-tap into a range was still considered important for others: "It's [NRS-11] much clearer than [SuperVAS], because you can see numbers. I like filling in dots." P1.8; "[NRS-11] with 10 [11] options is a relief, vs having to finagle with an unlimited number [in SuperVAS]" P1.3.

Reporting range and resolution

This last quote also speaks to preferences around reporting resolution and ranges. For most participants, 0..5 was too small a range and anything bigger

(0..100) was too big. "You can't really define your pain as well, it only went from 0..5 not 0..10, and because you can't slide you can't be as precise" P1.10; "But I would want to report 4.5/5. There's such a gap between 4 and 5." P1.8; "1..10 is really good, really anything in my life would fit on that scale." P1.8; "For me this [0..10] was good enough for a ballpark assessment of my pain" P1.7; "There's not a lot of variance from 1 to 5" P1.7; "This [SuperVAS+] is more fine-grained than I need. Jeez, I don't know. I don't know. I will just go to a whole number if I can" P1.3

The low-end of any scale, but particularly the NRS-11 and SAFE, was often called for lacking resolution. With SAFE this often demonstrated the conflation of pain with affective response to pain level, but also revealed end-user attention to never being in no-pain, but wanting to parse out greater degrees of lower-pain days (sometimes due to medicine): "It is fair to say that the lowest face doesn't describe a comfortable enough level" P1.9; "Is that the best smile I'm going to get? I might actually give this person [SAFE] an opportunity to have more of a smile if you have no or low pain... the lowest pain looks content, not happy." P1.6; "There's a lot of being in pain faces, but there could be a wider low pain range." P1.2; "I want it to go happier too! If I had no pain, I'd be grinning ear to ear!" P1.2; "[NRS-11] should also have better resolution at the low end of the scale, like not 0 or 1 but .5 and 1.5 too." P1.7; "Most of the photographic faces were pretty extreme [too few lower-pain options]" P1.10.

Findings particular to older adults

Our use of smartphones as the reporting medium, confirming trends reported in the literature, was validated by several older adult participant statements such

as “I like being able to use a smartphone for these sorts of things. It’s portable. I’m moving away from paper. We live with our phones now. I would be much more inclined to pop it onto my phone real quick, vs find that piece of paper.” P1.2

Furthermore, three older adult participants reported caring for an elderly relative. Without our prompting, they reflected on their family members’ likely reactions to these interfaces: “An example is my mom (93) who is always in the hospital, they always ask 1..10 and she always says ‘I hate that’, but she might really respond to visually sliding her finger [SuperVAS] or in fact really the the big bar [SuperVAS+].” P1.5; “I’ve just gone through this extended illness with my mother, and they were always asking what’s your pain level. She was never able to put things into words, especially as she got dementia. I really wonder if she had a scale to move her finger and said tell us how you’re feeling, and push the little ball up and down [SuperVAS], that that might really be helpful with older patients. She [mother] would have been able to do that and be pretty accurate.” P1.4

The role of playfulness

Elements of playfulness and aesthetics were drawn out and valued, particularly when the participant considered using each interface over time. “I like doing things that are new, it makes things more fun, more interesting... when I started with WeightWatchers I wrote it all down (pen-and-paper) and then I needed to spice it up... [With the WeightWatchers app] they change [the reporting interface] up once in a while, so people don’t get bored.” P1.8; “It’s not just about giving me good information, or me giving it good information - it should be fun

to do" P1.9; "If it's something you'd use every day, it has to be a little fun, not boring" P1.2; "I just like to play! [Many Fingers]" P1.10; "The visual is more attractive to me, I guess, rather than [NRS-11/SuperVAS]" P1.10.

Participants also provided a series of helpful feedback on more traditional usability issues, ranging for the size and color of text and images to their expectation that one should be able to tap on the scale anchors to report extreme values. It also became clear through observation that many participants do not read scale instructions or anchor texts, even when explicitly prompted to do so. "I just assumed that the lower end... did it say no pain?" P1.2. In the field trial, we would need to ensure participants were all introduced to the scale effectively.

No evidence of negative reactivity

The literature makes conflicting claims as to whether (frequent) self-report of pain intensity might result in negative, positive, or neutral outcomes (Haythornthwaite et al., 1998; Aaron et al., 2005; Hektner et al., 2007). Of particular concern is that repeated self-assessment of such a difficult and potentially traumatic experience draws more attention to, and foregrounds, the negative experience of being in pain, and thereby makes the subjective lived experience worse.

Among our participants, not a single participant reported negative consequences of repeatedly reporting pain levels. Those experiencing a subjectively lower degree of pain throughout the field trial suggested that perhaps others in more pain might experience negative such effects, those participants actually living with high levels of pain did not find this to be so. In fact, encouragingly,

several participants suggested that simply reporting their pain levels helped them in their day-to-day experience. P2.3 speaks to this:

> On the '7' days, it almost felt like the scales were an outlet for me, in a way. Like, I'm feeling bad, but, it almost felt like I could report it in the thing [phone or measure] and maybe compartmentalize it. Like, I've reported it and externalized the pain a little more? Rather than just having it on my mind. In a sort of a sense, it's maybe like the way I might vent if I'm upset. Somehow reporting it, externalizing it, actually eased it. When I'm at 7 or 8 it's not like the system is going to remind me I'm in pain, I'm not going to forget that! But the reporting is a venting.

Similarly, P2.9 found that while Meter did not make her think about her pain more, or be more aware of it, it did help her think about her pain differently, for example being more aware of contextual effects such as the weather (there was an extremely cold and windy week during the field trial). P2.7 found that the regular cycle of reporting helped him maintain his treatment habits:

> I don't think my pain increased, but I was more aware of it. I've been doing PT [physical therapy, at home, daily], but when I'm doing the PT [for a few days in a row], I feel good so I don't do the PT. But I *have* to do the PT. This reminder and filling out the measure didn't make my perception of the pain different, but it kept it on my mind so I would do things about it when I went home.

pID	response rate	response time	pain level
p2.03	57%	2.97s	3.0
p2.04	71%	2.34s	1.0
p2.06	95%	2.74s	2.0
p2.07	81%	2.45s	2.0
p2.09	95%	5.92s	4.0
p2.10	81%	3.45s	1.0
p2.13	81%	5.60s	4.0
p2.14	86%	3.67s	1.0
p2.18	105%	2.90s	3.5
p2.19	76%	4.26s	1.0
p2.20	90%	4.81s	2.0
p2.22	86%	4.64s	1.5

Table 5.1: Per-participant metrics when completing SuperVASNumbered.

Use statistics from the field trial

Overall, field trial participants completed 453 surveys over the course of the three weeks, a 90% response rate. Participant details are presented per-measure.

SuperVASNumbered This survey was completed in the mornings. The N=12 participants completed it 211 times, an 84% response rate. Overall, the SuperVASNumbered measure took a median of 3.6 seconds to complete, and participants used it to report a mean pain intensity value of 2.171 on a range of 0 (no pain) to 10 (worst pain imaginable). Pain intensity scores reported with SuperVASNumbered correlated very well with the standard NRS in the survey (Pearson's $r = 0.98$, $p = 2.2 \times 10^{-16}$). Over the 211 usability assessments, SuperVASNumbered was described as 'Very easy' 130 times, 'Somewhat easy' 63 times, 'Neither difficult nor easy' 17 times, 'Somewhat difficult' 1 time, and 'Very difficult' 0 times.

Table 5.1 shows per-participant metrics when reporting with the Super-

pID	response rate	response time	pain level
p2.03	95%	3.95	4.0
p2.04	90%	3.03	1.0
p2.06	95%	4.81	2.0
p2.07	105%	3.23	2.0
p2.09	100%	5.27	4.0
p2.10	86%	8.93	2.0
p2.13	110%	9.86	6.0
p2.14	95%	3.30	2.0
p2.18	105%	5.05	4.5
p2.19	71%	5.62	1.0
p2.20	110%	7.23	1.0
p2.22	90%	9.62	1.0

Table 5.2: Per-participant metrics when completing SAFESlider.

VASNumbered measure; response time and pain level are both median values within each participant.

SAFESlider This survey was completed in the evenings. The N=12 participants completed it 242 times, a 96% response rate. Overall, the SuperVASNumbered measure took a median of 5.0 seconds to complete, and participants used it to report a mean pain intensity value of 2.74 on a range of 0 (no pain) to 10 (worst pain imaginable). Pain intensity scores reported with SAFESlider correlated very well with the standard NRS in the survey (Pearson's $r = 0.93$, $p = 2.2 \times 10^{-16}$). Over the 242 usability assessments, SuperVASNumbered was described as 'Very easy' 95 times, 'Somewhat easy' 89 times, 'Neither difficult nor easy' 29 times, 'Somewhat difficult' 29 times, and 'Very difficult' 0 times.

Table 5.2 shows per-participant metrics when reporting with the SAFESlider measure; response time and pain level are both median values within each participant.

5.4 A Tangible User Interface for self-report (Keppi)

Motivated to better support those struggling with chronic pain, and looking to sidestep the frequent use of a mobile phone screen, we observed the way that in moments of pain, people will sometimes grasp the arms of their chair or the hand of a loved one. We were inspired by uncomplicated action of squeezing a stress ball. These interactions are unobtrusive and can be very private. In seeking to integrate these types of interactions with intentional self-report, we contribute a novel pressure-based user input device for the self-report of scalar values. As the "intensity of pain is without a doubt the most salient dimension of pain" (Turk & Melzack, 2011), we focus on the frequent *in situ* self-report of pain severity. While there is precedent in the older adult population for dedicated health devices - for example, the emergency alert systems worn as necklaces or bracelets - and a long history of dedicated self-report devices in the behavioral literature (e.g. Lindsley, 1968), to our knowledge, there has been no previous attempt to leverage a dedicated tangible user interface (TUI) for the EMA-style self-report of chronic pain.

The Keppi consists of a conductive foam based, force-resistive sensor (FSR) covered in a soft rubber with embedded signal conditioning, an ARM Cortex-M0 microprocessor, and BLE. We conducted in-lab usability and feasibility studies with 28 participants and find participants were able to consistently report with four degrees of freedom, as well as continuously map pressure to visual cues.

Much of this work is being submitted to ASSET 2015.

5.4.1 Pressure-based input and TUIs for self-report

Numerous studies have investigated pressure-based input actions. Srinivasan and Chen asked subjects to interact with a force sensor with their finger pads and attempt to match several types of force targets both with and without visual feedback of the pressure there were exerting (Srinivasan & Chen, 1993). When attempting to match constant pressure targets without visual feedback, they report absolute average error increasing with target force magnitude; with visual feedback it held constant. In a linear ramp target task, force rates had no effect on subjects' performance. In a sinusoidal target task, longer tasks showed worse performance than shorter tracking tasks, and performance was worse with higher average target pressure. However, they report performance only for the positively sloped parts of the sinusoidal curves.

Ramos et al, investigating pressure-based input with a stylus, report that subjects having reasonable control of pressure levels depends on (1) having a fixed number (at most, six) of discrete pressure levels, (2) providing good visual feedback, and (3) the type of selection task in question (Ramos, Boulos, & Balakrishnan, 2004). The first finding is confirmed by others, including (Cechanowicz, Irani, & Subramanian, 2007) and Mizobuchi et al. who report that three to seven discrete pressure levels make sense for the accurate control of input values (Mizobuchi et al., 2005). Mizobuchi et al. further report that forces of greater than 3N decrease the input experience from both performance and comfort perspectives, and that when providing visual feedback of the pressure being applied, analog feedback such as sliders are recommended over digital feedback.

Clarkson and others placed a simple continuous pressure sensor within mobile phone casing, supporting (for example) variable speed scrolling or navi-

gation based on pressure level (Clarkson, Patel, Pierce, & Abowd, 2006). They speak to the value of pressure-based interaction in mobile settings: namely, that pressure sensors are inexpensive, do not replace or obscure existing interactions and affordances, and therefore leverage and extend users' familiarity with their existing devices. Unlike tilt or motion sensors, pressure-based input does not require additional gross physical motions.

Heo and Lee present a novel touchscreen device that senses normal and tangential forces, and evaluate it in two contexts: web-browsing and an e-reader (S. Heo & Lee, 2011). For example, a greater force when dragging in the e-reader turned more pages at once. Chu and others report on 'haptic conviction widgets', widgets that allow (and at times require) users to convey their degree of conviction in performing an action, such as buttons that take different degrees of force to click or considerable force being required to permanently delete files from a trash can (Chu, Moscovich, & Balakrishnan, 2009).

Blaskó and Feiner developed a single-handed multi-strip device made of "a set of physically independent, pressure-sensitive multi-functional strips, each approximately the width of a finger" (Blaskó & Feiner, 2004), and report on a variety of interaction techniques possible with linear strips, including controlling an on-screen slider or spring wheel. They report that participants learned very quickly to exert the proper amounts of pressure.

The above systems are based around a single-sided 'push' or 'touch'-style interaction. Analogous to our inspiration of squeezing loved ones' hands or chair arms, or playing with a stress ball, there has been recent attention to two-sided pressure-based input: a 'grasp' or 'squeeze'-style interaction. Hoggan et al. experimentally confirm that squeezing is a viable input technique for mobile

device interaction, over other recent mobile-centric interactions such as device tilting (Hoggan, Trendafilov, Ahmaniemi, & Raisamo, 2011), and Stewart et al. report that grasping or squeezing outperforms single-sided pressure-based input, and therefore that pressure-based input for mobile devices is better delivered through a grasp or squeeze (Stewart, Rohs, Kratz, & Essl, 2010).

Hoggan et al. also report that in squeeze-based interactions, input errors increase when higher target forces (4N) are used over longer time periods - and here, 'longer' means more than about three seconds (Hoggan et al., 2011).

There have been several custom-built devices for the self-report of physiological states. One of the closest to the current work is that of Boormans and others' work on the Continuous Pain Score Meter (Boormans, Van Kesteren, Perez, Brölmann, & Zuurmond, 2009). The CPSM is a dedicated physical slider intended for and evaluated in a clinical setting; subjects move the slider to indicate pain levels throughout the clinical session. The CPSM is not very portable, and was neither designed for nor evaluated in natural settings. Laurans et al. have developed a similar physical slider for the continuous assessment of affect (Laurans, Desmet, & Hekkert, 2009).

The closest previous work to our pressure-based pain self-report TUI is a reference in Huskisson *et al* to a project by Armstrong *et al* who "used a mechanical method of assessing pain as an alternative to verbal statements, asking the subjects to squeeze a bag in proportion to the severity of their pain"; the citation is incomplete and we have not been able to recover the original work (E. C Huskisson, 1974). Huskisson *et al* continue by observing that most subjects are able to report verbally.

Finally, several researchers report on passively sensed pressure values in user interfaces as a way of inferring physiological state, such as stress or affect. Hernandez and others, using pressure sensing for physiological assessment in a passive context, show that higher typing pressure on a per-keystroke basis, as well as greater contact area with mouse (i.e., more pressure on mouse) is associated with higher self-reported and EDA-sensed stress levels (Hernandez, Paredes, Roseway, & Czerwinski, 2014). Their work uses a pressure-sensitive keyboard by (Dietz, Eidelson, Westhues, & Bathiche, 2009) who report through informal user testing that users are able to very quickly learn to control their input pressure. Both (Dietz et al., 2009) and (Clarkson et al., 2006) suggest pressure input is valuable for the sharing of affective state in computer-mediated communication, and (Chu et al., 2009) reports requiring greater force to communicate stronger emotions effective in instant messaging.

Continued contributions to pressure-based user input literature is especially timely given the likely widespread release of commercial devices with variable pressure inputs - for example, Apple's Force Touch technology in the Apple Watch[8].

5.4.2 The Keppi system

Design Considerations

The initial idea to design a pressure-sensitive TUI for self-report was inspired by a stress ball. After briefly exploring the idea of a pressure sensitive stress

[8]https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/OSXHIGuidelines/ControlsAll.html

ball, we came up with the idea of creating a compressible stick, which has many of the same affordances as a ball and not nearly as many physical design challenges. The design of Keppi was an iterative, experimental process. During this process we explored a variety of commodity pressure, flex, piezo-electric, and force sensors, all with different attributes such as flexibility, pressure thresholds, shape, and size among many others.

The primary issue encountered with the commodity sensors was form factor. The two major issues with the form factor were that the cylindrical shape of the stick did not allow the sensor to sit flush against a rigid surface (which causes the signal to be very unpredictable and noisy) and that in order to cover enough surface area, many sensors would have to be used. Another issue we found with commodity FSRs is that they generally are far too sensitive to withstand the force of a person's grip. We did have some success with Flex sensors and plan to investigate them further. We also saw promise in some custom piezo-electric ceramic rings, however they are quite fragile and subject to fracture which could be harmful to the users (could cut them while squeezing). After exploring all of these possibilities, we decided to design and experiment with custom FSRs.

In order to develop a robust FSR that could handle high thresholds of pressure while also having good resolution and sensitivity to lower pressures, we designed a series of sensors to test different force-sensitive resistive materials, different amounts of the material, different types and designs of electrodes, as well as different housings and mechanical designs. These prototypes were developed and benchmarked concurrently through a series of tests such as applying constant force with weights and clamps, testing the recovery time for varying the amount of impact and surface compression, as well as exploring

the effect of different types of leads: both how they affect the change in resistance during compression and how they hold up physically since they are subject to bending. In the next section we will go over the hardware design and the compare two versions of the Keppi TUI.

Hardware Design

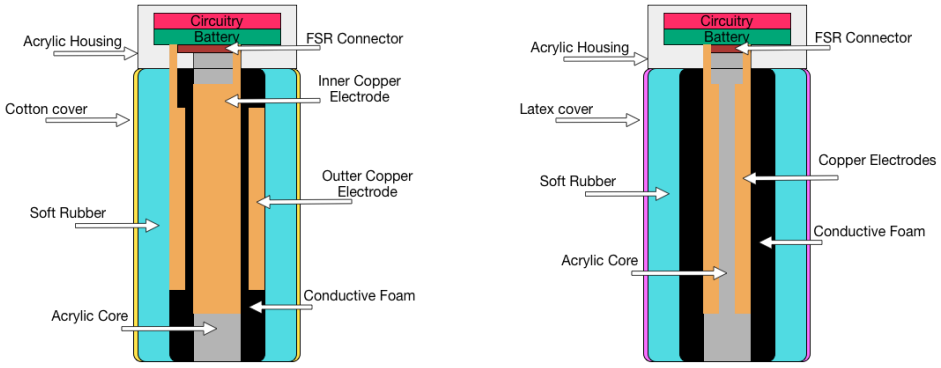
Keppi's Force Resistive Sensor After running several pilot studies to determine the best basic structure of the hardware (weight, height, diameter, texture) and establishing the majority of the materials that we were going to use, we developed two fully functional prototypes (as seen in Figures 5.7a and 5.7b). While the two prototypes are very similar, there are two major differences: the FSR design and the compressibility of the device.

(a) Keppi version 1 (b) Keppi version 2

Figure 5.7: Two versions of the Keppi

Keppi FSR v1 The FSR design in Keppi v1 consisted of a core electrode (copper tape) that covered the core shaft inside of Keppi, which is covered in medium density conductive foam. On top of this foam is an outer electrode which wraps all the way around the foam (Figure 5.8a). The electrodes have

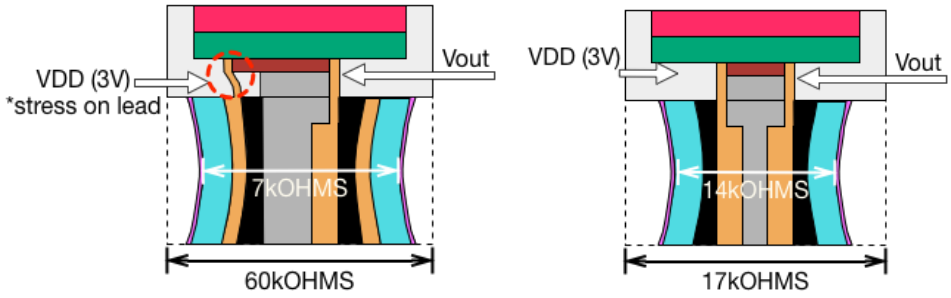
industrial grade aluminum foil leads that are soldered into a board mounted at the top of the core. This entire system is cover in electrical tape to keep it isolated and stable. The outer electrode has a 3V charge pass through it, which then travels through the conductive foam to the core electrode. The output of the core electrode is passed into a signal conditioning circuit, which we will discuss in the electronics design section. When compressed, the foam is compressed changing the density of the material, which inherently changes the resistance (from approximately 60k Ohms to 7k Ohms (see Figure 5.9a)) and the voltage/current.



(a) Keppi v1: sensor diagram (b) Keppi v2: sensor diagram

Figure 5.8: Sensor diagrams for the two versions of the Keppi

This design provided good resolution ($\tilde{3}$0 points) and could handle a great deal of force. There are two major drawbacks to this system: the outer electrode's lead encountered a great deal of movement, and the outer electrode would cause the conductive foam to temporarily deform in the case that a large change in pressure occurred. This resulted in the output signal to sometimes get stuck, or behave erratically. When heavy or rapid compression would occur, the outer electrode would crinkle, deforming its shape and the foam's resistive qualities.



(a) Keppi v1: compression (b) Keppi v2: compression

Figure 5.9: The change in resistance and stress on the sensor while undergoing compression for the two FSR's

Keppi FSR v2 Before retiring Keppi v1, we already began developing Keppi v2. Taking into account some of the previous design issues, we primarily focused on creating a more stable FSR and increasing the resolution. The FSR in Keppi v2 was designed using two core electrode plates, each covering approximately half of the core shaft (Figure 5.8b). This design immediately resolved the issue of mechanical wear on the FSR's leads. The two electrodes were wrapped with the same conductive foam as v1, however instead of covering it in tape, it is gently held in place with thread that is evenly wrapped around it. This is then covered in thin latex sleeve in order to prevent the outer layers, which are somewhat sticky, from damaging the foam. This FSR also has a change in resistance from 17k Ohms when not compressed, to 9.5k Ohms when fully compressed (see Figure 5.9b). While this is a much smaller range than demonstrated in v1, it has a much higher resolution, affording the user more control over the output signal.

Casing An important consideration when designing technologies that leverage force sensitive resistive materials is the design of the enclosure. This is particularly important when dealing with materials that are constantly being

physically manipulated. It is important to design the system in a way that affords easy manipulation on the user end but easily and quickly returns to a nominal state.

In Keppi v2, we achieved this through balancing the distribution of physical constraints with the ability to manipulate the device in meaningful way. The solution we arrived at was using a flexible, thin latex that was firm enough to hold everything together, but also allowed the sensor to return to a nominal state rapidly, even after elongated periods of intense manipulation. The only part of the system that demonstrates physical constraint is the top of the device. This is in order to prevent the sensor from being pulled of or damaging the electrodes.

Making it squishy We explored many different materials in order the find the balance between a squishy, stress ball-like texture and affording users control over pressure. The material we ended up using is a soft, polyurethane rubber (Durometer 40A). This combined with the soft, compressible conductive foam underneath provided a good balance in tactile sensation and control. This material demonstrated rapid recovery so the sensor could easily normalize, and was extremely resilient to tearing or misshaping.

Electronics

The electronic core of Keppi consists of three main components: analogue signal conditioning, Microprocessor, and BLE (see Figure 5.10. For the signal conditioning circuit, we make use of a transimpedance (voltage to current) op amp design using the Texas Instruments TL082 JFET input opamp (see Figure 5.12c).

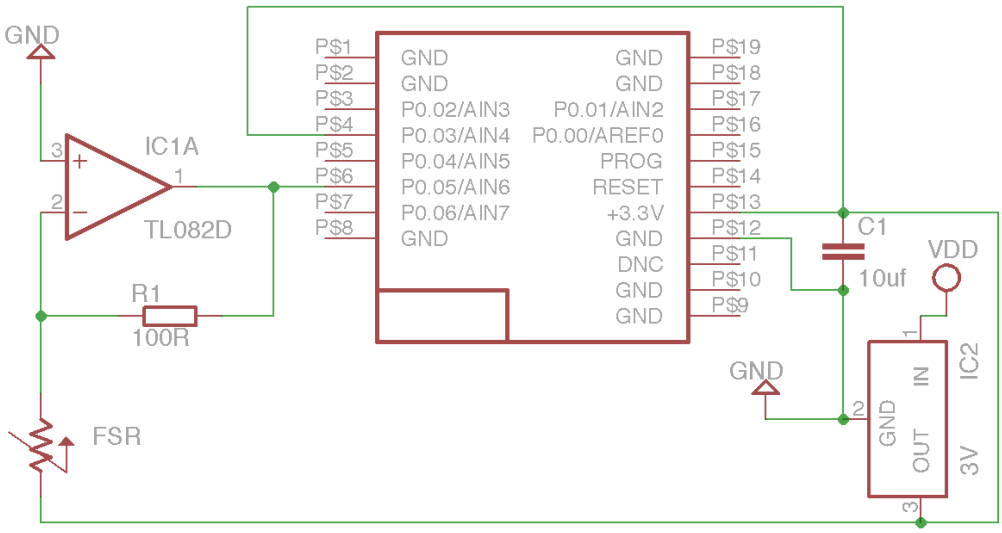


Figure 5.10: Electronic schematic for Keppi

We use a combined microprocessor/BLE chip (RFD22301) to read the signal with its onboard 10-bit ADC and transmit it to a mobile device with BLE (see Figure 5.12a).

The circuit also has a 3V, 0.25mA voltage regulator (See Figure 5.12b), break-out cable for USB programming, and connects directly to a 110mAh Polymer Lithium Ion Battery (which is attached to a micro-USB charging circuit) (see Figure 5.11). All of these components fit in an acrylic housing on the device, in which they connect directly to the FSR. In order to manage voltage irregularity in the ADC, a signal is sent directly from the voltage regulator to another ADC, which can then be used as a coefficient to get rid of noise on the signal.

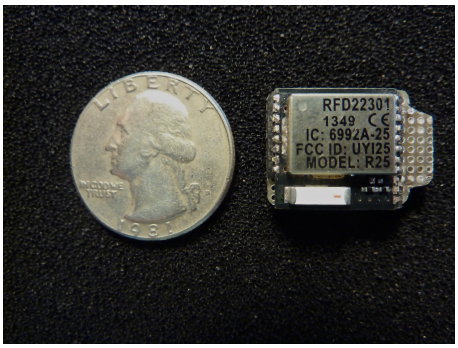
5.4.3 System evaluation

We conducted in-lab usability and feasibility studies with a convenience and snowball sample of 28 participants recruited via in-person intercept and email.

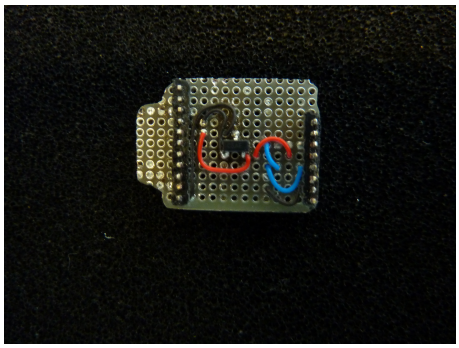
Figure 5.11: Electronics outside of the housing connected to the FSR

Their average age was 24.9 (min 19, max 38); 10 participants were female, 18 male; 10 of the 28 experience some form of chronic pain (for example, injury-related ankle or neck pain, frequent migraines, or lower back pain). Participants were compensated $10. Cornell's IRB approved this protocol.

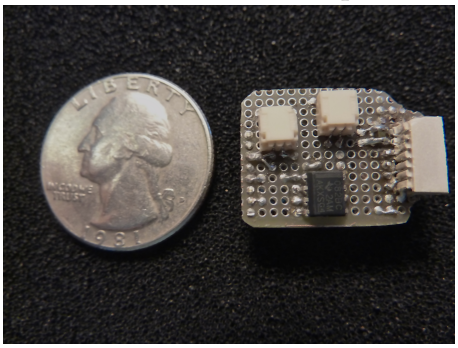
Participants were given the Keppi to hold and asked for their initial impressions regarding the object and how it might be used to report a value, for example pain intensity. After explaining (or confirming) the mapping from squeeze intensity to value range, each participant familiarized her/himself with this mapping by squeezing the device while being provided real-time visual feedback of squeeze intensity on a slider widget (no numbers were displayed). When familiarized, participants then completed three tasks while provided visual feedback of the pressure they were using:



(a) BLEand micro-processing unit (RFD22301)



(b) Voltage regulator lies underneath micro-processing BLE chip



(c) Analog signal conditioning circuit for Keppi's FSR

Figure 5.12: Different layers of the Keppi circuit

1. Report the highest possible value (hardest squeeze)
2. Report a medium value and release, a high value and release, a low value and release.
3. Watch an animation of a red circle tracing a series of sinusoidal curves and a step function, and continuously report the values traced using the Keppi (Figure 5.13)

Participants then repeated these three tasks, this time without visual feedback of the pressure they were using. In closing, participants were asked some open-ended questions about their interaction with the Keppi.

The first ten participants used Keppi v1 and were provided their visual feed-

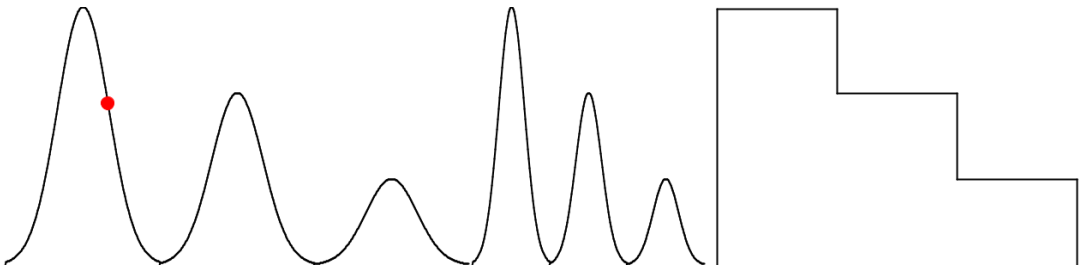


Figure 5.13: Sinusoidal and step function curves for the continuous tracking tasks.

back on a smartphone, the Keppi transmitting data via Bluetooth LE. The next 18 participants used Keppi v2 and were provided with the same visual feedback but this time on a laptop screen, the Keppi transmitting data via a serial port. While qualitative findings come from all 28 participants (although primarily the first ten), for the reasons described above we report quantitative findings only from the 18 participants using Keppi v2.

5.4.4 Results and discussion

Qualitative responses

When initially handed the Keppi and asked how they might use Keppi to report values or pain levels, many respondents assumed squeezing to be the reporting modality. The squishy stick form-factor invites such a grip. Our design inspirations may also have been evident as multiple participants suggested Keppi was like a stress ball. Two thought it might be a microphone, and one (P9) thought Keppi might "test my physiological status while holding it" - perhaps thinking of the handles on gym exercise machines that infer heartrate while you hold them.

After the mapping from squeeze intensity to pain level was confirmed (or revealed), participants on the whole found this made sense - "it is pretty intuitive, I naturally relate pain to more squeezing" P5. Several participants, including those reporting a natural mapping, did identify several scenarios in which squeezing to report pain would make less sense. For example, "if I have a headache I don't want to do anything to force more pressure" P7, or not being sure about moments of accidental reporting, for example by sitting on the Keppi.

In daily life, most participants imagine carrying the Keppi in trouser pockets, jacket pockets, and bags: "in the side pocket of my backpack - its the exact same size as my mace which I keep there too" P6. P10 "would wear it around my neck, but then I don't care about how things look to other people". We had not anticipated that some participants would view Keppi as something you could report with if it were still in your bag or in a pocket, without having to take it out.

Although we did not ask about a clinical environment either, several participants observed that Keppi could be deployed here, too - perhaps inspired by the continuous tracking tasks. For example, P7 observed "if I am having my teeth removed, I cannot communicate with doctor. This could be used to gauge scales to doctor based on force I use on the device."

Toward the end of the session we invited participants to describe Keppi in three words. Keppi was viewed as convenient, useful, intuitive, easy-to-use, and practical. For some Keppi was portable, compact, light, and mobile, but for others it seemed bulky, heavy, and jerky. While it seemed advanced, innovative, technological, and interesting, Keppi for some others came across as medical,

confusing, and weird. Similarly, some users found Keppi to be fun, squishy, and to feel good to hold, while others did not like the (v1) texture.

One aspect we probed was whether, without the on-screen visual feedback, participants felt that in addition to the fact that they could feel themselves squeezing the Keppi we should add another signal to Keppi to indicate that a squeeze was being monitored, or a confirmation signal that a squeezed report had been received. We got mixed responses from participants - just over half felt that additional feedback either was not necessary or that at least, after you had seen the on-screen visual feedback you did not need to see it all the time. The others reported interest in either a light or vibration-based indicator of use or report of capture.

Based on this feedback, and while developing v2, we both removed the cloth material cover from the Keppi as well as made it easier to squeeze because Keppi v1 was widely considered not squishy enough: "I would like it be more elastic because when I want to squeeze with more strength, I feel too much resistance with this" P2. This issue of squishiness was one of both hedonics as well as perceived reporting range and resolution - with a squishier Keppi, it would be easier to report more values more accurately.

Quantitative findings

We separated out the data on low/medium/high reporting from the continuous tracking data; for each participant there is data from the visual feedback condition (VF) and the no visual feedback condition (NFV).

In the low/medium/high reporting task (Figure 5.14), we find via a one-way

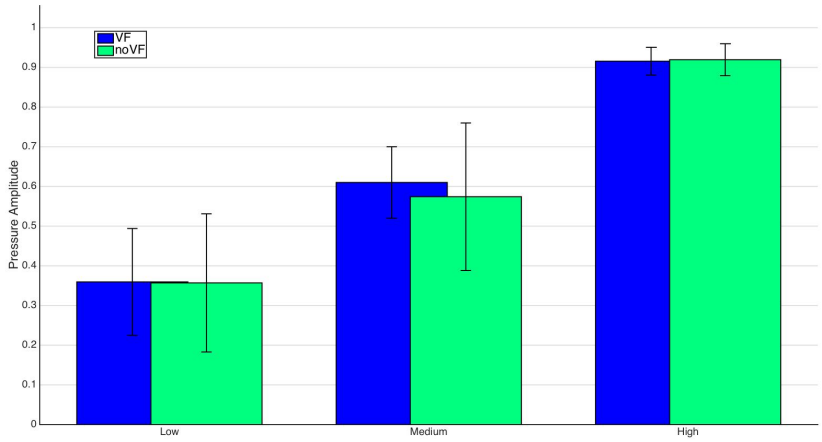


Figure 5.14: Participants are able to report intensities with four degrees of freedom (no pain is 0) with visual feedback or no visual feedback.

ANOVA that the means of none/low/medium/high reporting levels are significantly different in both the VF ($p = 5.13226\times10^{-21}$) and NVF ($p = 5.37399\times10^{-14}$) conditions, as well as when we ignore visual condition and average them together ($p = 2.87841 \times 10^{-20}$). A two-way ANOVA between the VF and NVF conditions overall found no significant difference ($p = 0.78$), meaning that participants were able to report intensities with four degrees of freedom both with and without visual feedback.

In the continuous tracking task, we first prepared the data by normalizing both the baseline (Figure 5.13) curve and the user-tracking data. We did not have to normalize or fit the data in the x dimension as it is all equally spaced 1D data at the same sampling resolutions. Our next question was, how well were participants able to track the baseline curve? Inspecting the data visually (Figure 5.15) indicated that participants on average were tracking the baseline curve very well. Performing a cross-correlation of baseline and VF, baseline and NVF, and VF and NVF (Figure 5.16) confirms that each of these series are pairwise

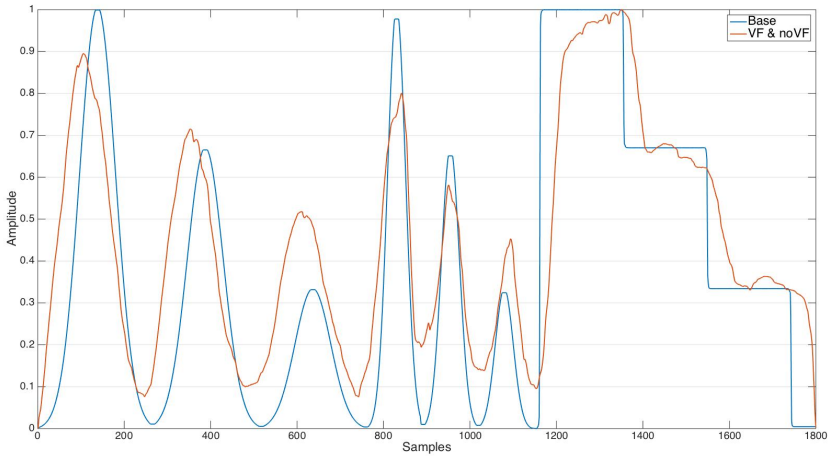


Figure 5.15: Visual comparison of the normalized and averaged continuous tracking data, over all data (orange) as compared to the baseline curve (blue).

highly cross-correlated (0.98, 0.93, and 0.92). Cross-correlation is a measure of the similarity of two series as a function of the lag of one relative to the other; as in our data we neither anticipated nor found a lag (participants tracked the baseline curve in real-time), the peak in the cross-correlation plot is at lag=0. This means that in both with and without visual feedback, participants were able to very closely track the baseline curve in the continuous tracking task.

Further discussion

Participants report a perception of less pressure control at lower pressure levels. While contrasting with the (Srinivasan & Chen, 1993) who find that tracking error does not vary with target force magnitude, there is evidence from (Ramos et al., 2004) that there may be less pressure control at lower pressure levels. The squeezable material wrapped about the Keppi does not deform linearly - it deforms more easily at lower pressures when it is relatively less dense - and this may contribute to a perception of less control at lower pressure levels. While

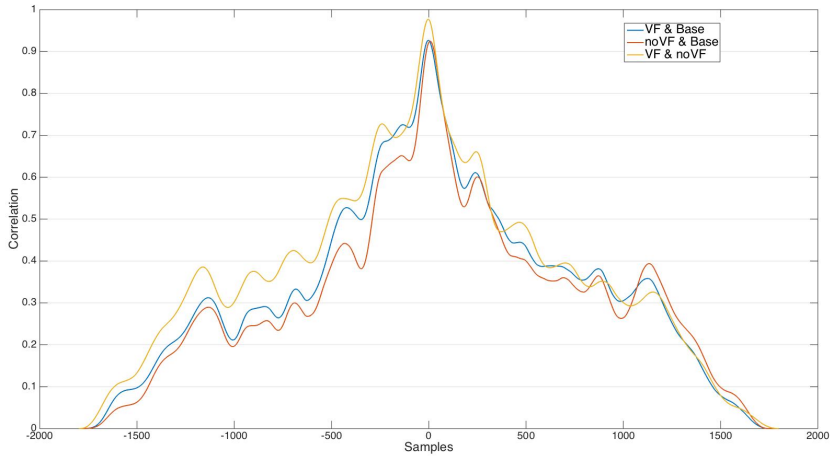


Figure 5.16: Cross-correlation of Keppi continuous tracking task data: VF vs baseline (blue), NVF vs baseline (red), and VF vs NVF (yellow). The series are all highly correlated.

quantitatively we were able to distinguish between no squeeze, low squeeze, and medium squeeze, this perception of differential control should be investigated further, both in refining the user experience and to rule out its impact as a possible reporting bias.

Much prior work described above reports that providing users visual feedback regarding the pressure they are using is necessary for accurate pressure-based input (Stewart et al., 2010); (Shi, Irani, Gustafson, & Subramanian, 2008) even calls visual feedback essential (although not sufficient) for enhancing pressure control. Our findings suggest that, at least after training, visual feedback may not be necessary for accurate reporting at four degrees of freedom. While learning happens very quickly, we should confirm that participants can continue to accurately report day after day. Given one intended use case for Keppi is EMA-style day-to-day report, the repeated use should serve as a sort of continued training and familiarity with the input interaction. In the only previous work we are aware of dealing with pressure-based inputs in natural settings -

as opposed to in laboratory settings - (G. Wilson, Brewster, Halvey, Crossan, & Stewart, 2011) suggest an alternative: audio feedback while walking is both highly accurate and useful.

Limitations and future work

This squeeze-based self-report TUI is not intended for use by individuals experiencing some form of hand or wrist pain or movement limitation. Diminished strength or grip ability otherwise is not such an issue; the input sensitivity range can be adjusted and the output normalized. It has also become clear that for some non-hand or wrist pain conditions, such as migraines, pressure-based reporting may not be ideal.

Particularly in a naturalistic setting, we need to investigate the necessity of on-device feedback either during the reporting action (mid-squeeze) or to indicate that the report has been captured (post-squeeze), or both. Our intuition is that, like a button, simply squeezing the Keppi would provide the necessary (haptic) feedback that a squeeze is taking place; it may be that some vibration on-device or a notification of some sort on the connected smartphone to indicate the recording of a report would be useful. It is also possible that as the user gets used to the Keppi, and can see accumulated reports or end-of-day summaries on the phone, post-squeeze feedback would not be perceived to be useful. Indeed, over time or in more public settings, such feedback might become intrusive or a privacy concern.

It may be that, at higher pressure levels and for longer target tasks, pressure-based input can result in some participant muscle fatigue (S. Heo & Lee, 2011;

Mizobuchi et al., 2005). Empirically, the only participants to report fatigue when using the Keppi were those seeking to maximize the high pressure value - to make the slider widget go all the way to the top - and, in three cases, to report some fatigue after the continuous tracking task (48 seconds, performed twice). We believe participant fatigue can be ameliorated in three ways. First, by increasing the squishiness of the device such that reporting across all pressure levels requires less exertion. Second, by tailoring the reporting range to the individual, such that for any strength capability, reporting a high value requires (say) 70% of a max squeeze. Third, by recognizing that in EMA-style reporting there will rarely be target tasks longer than a few seconds, and in a clinical continuous reporting setting, clinicians can invite subjects to report a low baseline level of pain and indicate only spikes in pain using Keppi, where appropriate.

We next intend to conduct field trials of the Keppi to evaluate it in natural settings. To our knowledge, there is far less work on pressure-based interactions as they occur outside controlled laboratory settings, and this can impact the real-world applicability of our findings. For example, Wilson et al. suggest that mobility (i.e., walking) negatively affects pressure-based linear targeting, but that using a rate-based control mitigates this negative influence (G. Wilson et al., 2011).

We also hope to further shrink the fabricated TUI such that it can comfortably fit within a pocket, on a keychain, or when worn around the neck.

These findings report on the development of Keppi, a novel user input device for the self-report of scalar values - in this case, pain intensity. Keppi consists of a conductive foam based, force-resistive sensor (FSR) covered in a soft rubber with embedded signal conditioning, an ARM Cortex-M0 micropro-

cessor, and BLE. We conducted in-lab usability and feasibility studies with 28 participants and found participants were able to consistently report with four degrees of freedom, as well as continuously map pressure to visual cues and scenarios.

CHAPTER 6

GENERAL DISCUSSION

In this chapter, I will reflect on the case studies primarily through four discussion points. First, from both empirically gathered data and the literature, I describe the continued value (the essential value!) of self-report in a world awash with increasingly capable passive sensors. Second, I explore the idea of a minimally viable moment of measurement. There has been a trend over the last 20 years of shrinking self-assessment measures, and in this and related work we have arrived at moments of self-measurement requiring only 1-3 seconds. Is this enough time for meaningful cognition to occur, for the data received to accurately reflect some or all of the construct in question? Third, in the vein of personalization common in our commercial technology and reflecting modern medical trends toward 'N of 1' study designs and interventions, I discuss the advantages and disadvantages of tailoring the self-report interface to the individual. I discuss several mechanisms (including three leveraged in the case studies) through which such personalization can be achieved, and describe the impact of such tailoring on across-subjects assessments. Fourth, I discuss why through my work across momentary measurement in three domains, I do not believe there is a one-size-fits all approach for developing novel self-reports for the EMA domain, and I contribute several design patterns useful to others continuing this work in new health spaces.

This discussion is based in the literature and in my experiences working on the three 'moments of measurement' case studies:

In chapter 3, I report on the design, development, and validation of PAM, the Photographic Affect Meter (Pollak et al., 2011). From a user's perspective,

PAM is laid out in a grid (Figure 3.8). The user is presented with 16 photographs that represent a diversity of emotions (description to follow) arranged into a 4x4 grid, and is prompted to select a photo that best describes how they feel right now. Below the grid, a button to choose more photos reloads the grid with a different set of PAM images should the user not be able to comfortably express their state with one of the photos currently displayed.

Underlying the PAM grid is the Russell's circumplex model of affect: a two-dimensional valence/arousal space within which each emotion can be placed (Russell, 1980). The decision to use photographs is based in literature indicating that photographs and emotions are linked, often with universally shared meaning (Chalfen, 1989; Lang, 1995), but affording the user a sense of choice and control by way of interpretive flexibility (Sengers et al., 2008).

In chapter 4, I report on SESAME, a field trial comparing minimally invasive techniques for assessing stress in the wild (Adams et al., 2014). In the 2011 Stress in America survey, the American Psychological Association warned that stress is becoming a public health crisis (APA, 2011) - most Americans are suffering from moderate to high levels of stress, with nearly half reporting an increase in stress over the preceding five-year span (APA, 2012). In this case study, I am interested in supporting long-term engagement with one's own perceived stress levels through measurement. One of the central challenges in creating these types of systems is in determining what kind of stress-related data to collect in order to strike a balance between reliability - that is, how closely the data accurately and consistently track a person's perceived stress at the moment that it occurs, and intrusiveness - that is, how much effort is required on a participant or user's behalf to provide the data.

SESAME triangulates data from self-report, electrodermal-sensed arousal, and continuous inferencing of stress markers in human voice via the phone's microphone. Among these techniques for self-monitoring stress levels with a minimal impact on participants' daily lives, which track one another most closely? Under what conditions or in what contexts? When do these sensing modalities agree with one another, and when do they produce conflicting narratives about daily stress?

In chapter 5 I report on two projects intended to support the management of chronic pain by providing novel and effective self-report assessments of pain intensity. Chronic pain, recurrent or long-lasting pain, affects an estimated 30.7% of US adults (B. A. Ferrell et al., 1995). Those with chronic pain are frequently severely debilitated, with significant limitations in their ability to function or work. Chronic pain is associated with depression, sleep disturbance, fatigue and decreased cognitive and physical abilities. (Ashburn & Staats, 1999).

Meter is a multi-stage user-centered research through design (Gay, 2004; Zimmerman et al., 2007) investigation seeking optimal visual interfaces for the self-report of pain intensity on smartphone screens. Through design ideation, in-lab user studies, a three-week field trial, and an expert panel review, Meter results in two recommended self-report measures as well as rich insights regarding the ways those with chronic pain prefer to record and report their pain levels.

Keppi is motivated by observations of the ways that people in pain will sometimes grasp the arms of their chair or the hand of a loved one, and inspired by the uncomplicated action of squeezing a stress ball. These interactions are unobtrusive and can be very private. In seeking to integrate these types of

interactions with intentional self-report, we developed Keppi, a novel pressure-based user input device for the self-report of scalar values. As the "intensity of pain is without a doubt the most salient dimension of pain" Turk and Melzack, 2011, we focus on the frequent *in situ* self-report of pain severity; to our knowledge, there has been no previous attempt to leverage a dedicated tangible user interface (TUI) for the EMA-style self-report of chronic pain.

6.1 Self-report remains essential in a sensor-centric world

There is an exciting trend in both the research and commercial worlds of increasingly sensitive, accurate, and affordable on-body passive sensing of activities and states. Such sensors can be always on, and many are nearly completely uninstrusive. Several have experienced wide adoption already - for example, FitBit's activity trackers[1]. In this section, drawing on both my own research and the literature, I make the argument that even in an increasingly sensor-centric world, self-report maintains an essential role - and not only for those constructs that we cannot (yet?) effectively passively measure. As in Boehner *et al*, I am definitely not arguing against passive sensing - an approach one I've taken in my own work, including the second case study presented here! - but I *am* suggesting "a recognition of the limits and liabilities of both objective and subjective accounts" (Boehner et al., 2007), in addition to stepping back and appreciating the positive side-effects of active self-monitoring.

Each of the case studies in this dissertation concern the assessment of internal states for which the gold standard ground truth assessment is inherently

[1] http://fitbit.com

and definitionally subjective and experienced. For example, "pain is what the patient tells us it is" (McCaffery, 1979). Any passive sensing of affect, stress, or chronic pain is done by proxy - for example, inferring pain intensity from facial expressions (Kaltwang, Rudovic, & Pantic, 2012), inferring stress levels from voice features (Lu et al., 2012), or inferring affective state from electro-dermal activity and cameras (R. W. Picard & J. Healey, 1997) - and for many scholars it is simply the case that a computer cannot meaningfully quantify stress, affect, or pain in their comprehensive complexities (Boehner et al., 2007). The accuracy of such systems is increasing, although as discussed above there are a variety of scenarios and contexts in which the system is unavailable or unaware that it is providing inaccurate data (Adams et al., 2014).

As we saw in the SESAME project, passive monitoring can provoke an uneasy sense of being surveilled, and even some participants' friends felt somewhat uncomfortable sharing physical space with a participant whose phone was 'constantly listening' to them. And this was in a project where the data was all entirely private, and would live just long enough for the duration of the study. In an intriguing and complementary manner, for other users the set-and-forget model assumed by many passive sensing systems means individuals will not remember or realize that their behaviors are being monitored.

While costs continue to decline, cutting-edge wearables and passive sensing systems are likely to remain accessible only by wealthier citizens or by those living in regions with good internet access. This is somewhat ameliorated by the explosive penetration of smartphones - 80% of adults worldwide are predicted to own a smartphone by 2020 (Evans, 2014) - and the work of researchers focused on comprehensive monitoring systems using only those sensors com-

monly included in an average smartphone (Lane et al., 2011).

The complexity and intelligence of such inference systems engender their own user-facing issues. Passively collected data at the scale we are starting to see, particular those concerning something as personal and private as one's health, requires a new form of literacy for the end-user, and a new way of presenting and communicating data management on the part of system developers. Simply looking at the degree of difficulty users have in understanding the cloud back-up models presented them regarding the photos taken on smartphones suggests that this is a complex and perhaps underestimated engineering and user experience problem. The steps both Google and Apple, for example, have taken with Fit [2] and Healthkit[3] serving as trusted repositories and brokers of such data are encouraging, but it remains to be seen how effective they will be in the long term.

The very mindfulness and foregrounding of even brief self-report measures avoid many (if not all) of the above issues, and further, have been associated with various positive side-effects. In the Meter project, I reported on participants sharing experiences such as finding the self-report practice helping maintain at-home physical therapy exercises, or finding that performing frequent self-report afforded new or greater awareness of contextual pain correlates, such as weather. These findings agree with those in the related literature on medication reminders for supporting patient compliance, as well as the obesity literature where the manual self-monitoring practices, beyond simply being a data collection method, may in fact be necessary for successful weight control (Baker & Kirschenbaum, 1993).

[2]https://developers.google.com/fit
[3]https://developer.apple.com/healthkit

One concern raised with the self-report of sensitive or difficult experiences, such as chronic pain, is that the self-report can foreground and make more salient the negatives - that through reactivity effects, self-report can make the subjective experience of living with pain worse. Encouragingly, I reported on empirical evidence that for no single participant was this the case - as in (Cruise, Broderick, Porter, Kaell, & Stone, 1996; Stone et al., 2003) - and in fact, the practice of brief moments of self-report for several participants had therapeutic benefit. For example, one participant spoke to feeling that she could put her pain into the measure on the phone and let it go - she felt that as the 'box' was handling the pain she did not therefore have to handle it herself. These types of findings fit with with others reporting positive reactive outcomes associated with the self-report of subjective experiences (Haythornthwaite et al., 1998; Aaron et al., 2005; Hektner et al., 2007). Indeed, early behavioral self-monitoring was primarily and explicitly considered an intervention for behavioral modification, the idea being that the mere act of recording a behavior changes its frequency in a desirable direction: increasing the frequency of positively valued behavior and decreasing the frequency of negative behavior (Wheeler & Reis, 1991; Lindsley, 1968).

While there are many domains in which self-report is not a good solution (step counting, for example, is something almost impossible to manually count or to estimate by way of even frequent recall), for the vast majority of domains, the best monitoring system is likely one that incorporates both passive sensing and self-report. Keeping 'the human in the loop' enables a continual refinement and accuracy checking of the algorithms behind the passive system (Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Adams et al., 2014) as well as the ability to effectively protect privacy in nuanced settings (Kay et al., 2012). A dual ap-

proach, passive sensing and experiential assessment through self-report, allows the system and the researcher to leverage the advantages of both approaches while simultaneously ameliorating many of the disadvantages of each.

6.2 Is there a minimum viable momentary self-report?

There is a nearly 20 year trend in self-report measurements becoming shorter and shorter. This trend began with measures being abbreviated to what now can seem tediously long - for example, the PANAS is a 20 item measure that in 1988 was introduced as "a brief measure of positive and negative affect" (Watson et al., 1988).

Of primary concern when using such abbreviated measures must be, are such measures still effective? While (as discussed in Chapter 2) there is every indication that short-form measures retain respectable psychometric characteristics (Burisch, 1984, 1997; Rammstedt & John, 2007; Gosling et al., 2003; Pollak et al., 2011), this question has not been explicitly addressed in the context of repeated and extremely brief self-reports via smartphones. As researchers, assessment experts, and interaction designers, we are dealing with a tension between meaningful and accurate data on the one hand, and participant burden and compliance on the other. How brief can the measure become before inviting the participant not to cognitively engage with it? How often can a participant be assessed with the same measure before respondent fatigue (Egleston et al., 2011) sets in? How minimal can EMA measures become before compliance or assessment quality significantly degrades?

As the measure becomes extremely brief (on the order of a few seconds in

each of the case studies) or is presented too repeatedly, is it possible that respondents may no longer put forth the cognitive effort required to answer meaningfully? Recall that in answering a survey item, respondents conduct a four-stage cognitive process. First, respondents interpret the question its intent. Next, they search their memories for relevant information and then integrate that information into a single judgment. Finally, they must translate the judgment into a response by selecting one of the alternatives offered - quite possibly shaping the response by taking into account concerns of self-presentation and consequential validity at the same time (Messick, 1995; Krosnick et al., 2005). Each of these steps can be quite complex and involve a great deal of cognitive work, conscious and unconscious, to answer even a single question (Krosnick, 1999; Krosnick et al., 2005).

In order to get accurate data, researchers indicate that it is important that the respondent go through each of these cognitive steps (Krosnick et al., 2005) - known as optimizing the response. But for a variety of reasons, participants are known to shift their response strategy to 'satisficing': compromise their standards and exert less energy in responding (Krosnick, 1991). Satisficing can be weak - all cognitive steps are executed, but each one less diligently than when optimizing - or strong - the retrieval and judgment steps are skipped altogether, resulting in a response that is either easily or even arbitrarily selected.

It may be the case that agreeing to complete a survey may be a relatively automatic compliance process (Cialdini & James, 2009), and this may be particularly true in EMA-style reporting where the cognitive and interactional effort expected is so small. This risk is compounded by the fact that modern EMA commonly prompts for each response by leveraging smartphone notification sys-

tems. Signal-contingent sampling was once a clear strength of ESM and EMA self-report; now, the signal can become lost in the noise of a bewildering number of smartphone notifications that overwhelmed users triage as quickly and as effortlessly as possible. Apple, in seeking to support notification-overwhelmed Watch users, have resorted to developing a new Taptic interaction model that varies both the strength, length, and pattern of notification vibrations to differentiate notification categories.

It may not be, however, that very quick responses to brief measures are the result of satisficing behaviors. Another way of understanding how respondents are answering such brief questions draws on the two systems of Dual Process Theory: system 1 operations are fast, automatic, associative, and governed by habit, while system 2 operations are slower, more effortful, and more likely to be consciously monitored and deliberately controlled (Kahneman, 2003). While at first it seems that system 1 'gut reactions' might always provide less accurate and meaningful responses, Kahneman indicates that the impressions and intuition generated by system 1 can also be powerful and accurate. Importantly, powerful and accurate intuition is often the result of repeated practice. It seems likely that for the repeated use of a given measure prompting a participant for input on a subjective experience, especially when we know that response time decreases due to familiarity (Stone & Broderick, 2007), may well begin to rely on skilled intuition.

The final cognitive step, mapping from the self-assessment to selecting a response to the measure itself, may become intuitive, even habitual. Certainly anecdotal evidence from users of PAM in deployments such as (Baumer et al., 2012) suggests that, over time, respondents develop a sense of the underlying

two dimensional model as well as begin to memorize where in that two dimensional space specific images appear. They report, for example, feeling that they are more 'up here' (top right) at the current moment, or knowing that 'the cowboy' perfectly captures how they feel right now and so look out for and immediately select that image.

Certainly in looking again at the four cognitive steps involved in answering a self-report question (question interpretation, memory/self exploration for current experience, judgment making, and finally translating the judgment into an available response), it seems clear that in repeatedly responding to a brief measure not all steps are necessary each and every time. The question's intent and meaning need not be understood and reinterpreted, and skill and intuition become involved in the translation of a self-assessment into an item response. The middle two steps are still necessary and at risk of satisficing, but, as I argue in the next two sections of this chapter, an effectively designed single-item EMA measure intentionally supports the cognitive processes of generating a self-assessment.

I therefore conclude that there is no momentary measurement too minimal to be effective. In fact, in many ways it is surprising that the self-report community has arrived at such brief, momentary measures after decades developing and deploying such comprehensive and complex multi-section surveys, instead of beginning with the briefest possible measure and growing if necessary. In the health intervention literature, Glasgow *et al* have recently made analogous observations. They call for wider use of minimal interventions needed for change (MINC): interventions that are just sufficient for change, require few resources of settings and participants, make use of common intervention strate-

gies, and are only as complex as needed for change (Glasgow et al., 2014). Furthermore, the authors claim that any intervention larger, more expensive, or more complex than the equivalent MINC should first be demonstrated to intervene equivalently more effectively and produce equivalently greater impact. Similar research in small interventions is reported effective by (Phillips-Caesar et al., 2015) leveraging 'small changes and lasting effects' (SCALE).

6.3 Tailoring/personalization of self-reporting experience

People are individuals. We have different personalities, learning styles, and tastes, varying abilities in literacy, numeracy, and self-expression. Self-report researchers need to recognize this and deploy self-report measures that consider, even embrace, their participants individuality. Meeting people where they are is what naturalistic methods are all about, and due to targeted ads, customizable mobile phone experiences, and increasingly intelligent digital concierge agents[4], participants live in and are coming to expect experiences tailored to and for them. Indeed, rejecting the idea that biological responses are standard, Kessler *et al* have called for a moratorium on randomized controlled trials (R. Kessler & Glasgow, 2011). One-size-fits-all is not the right approach (Torsi et al., 2009).

In a self-report measure, I see personalization as a means of implementing Vannette's recommendations for optimizing response quality: tailoring the experience to both enhance respondent motivation as well as minimizing the difficulty of meaningfully answering the question (Vannette & Krosnick, 2014).

[4]For example, Siri, Cortana, or Google Now

Minimizing task difficulty involves supporting each of the four cognitive stages involved in answering questions: make it easier to interpret the question, to retrieve information from memory or about one's state, to integrate that information into a self-assessment, and to report that self-assessment using the measure in question. Beyond external motivations (such as a sense of accountability or an understanding of why participation is important), respondent motivation can be enhanced by ensuring that completing the measure, potentially tens or hundreds of times, is not only painless but potentially enjoyable or even delightful.

One method of tailoring the measure to each individual is to allow each respondent to approach and engage with the measure's response options differently. Meter does this in two ways:

- Both measures contain a VAS-style slider widget which can be appropriated for precision or both spatial thinkers.
- Meter affords both a tap-to-report interaction as well as responding to swipes and drags, allowing users either to report with one tap or to 'dial in' their pain level, adjusting the value in realtime until it 'feels right'.

PAM does in three different ways:

- The use of ambiguity and interpretive flexibility in the association of affect with each photo gives users more control over the representation and reporting of their own emotions.
- Not every image can or will speak to every user. One of three images will randomly appear in each cell, and a 'load more images' button allows

users to generate one of 3^{16-1} other possible PAM views - and indeed, some users reporting knowing that they feel like a previously seen/used photo, and refreshing PAM until they can report with that photo.

- On the surface and in the directions, PAM is entirely photo-based. But the images are laid out in the two dimensional valence/arousal space that describes emotions, and over time there is some evidence that users appreciate, explicitly or even intuitively, these dimensions. This gives them two modalities to report with, by their preference.

Another way to tailor the measure to the user is to provide more than one measure, and let the user choose. This is also an underutilized although promising direction: "Not enough is known about personality or symptom effects on response compliance or styles... individual differences in various dimensions may lead to selective biases in the ability to respond to diary questionnaires" (Bolger et al., 2003); "[For children,] no single pain intensity measure is appropriate across ages or types of pain" (Stinson, Kavanagh, Yamada, Gill, & Stevens, 2006). One of the two resulting Meter measures very explicitly speaks more to 'numbers people', while the other is based on facial expressions, affording a more subjective, qualitative response. In within-subject studies this is not a problem assuming each measure demonstrates construct and content validity. If user ratings correlate well between scales, then between subject comparisons (or even allowing the user to change which scale they use for each momentary measurement) becomes possible: "in fact... a designer might choose to allow users to rate on any scale they wish, computing recommendations using normalized scores" (Cosley, Lam, Albert, Konstan, & Riedl, 2003).

Both of these tailoring techniques make the question more interpretable, bet-

ter fit with the respondent's cognitive process for recall/self-assessment, and support the mapping of self-assessment to the reported photo, face, value, or slider position. Each of the novel measures in PAM and Meter are further designed to optimize for meaningful responses by enhancing motivation to report. The measures are visually appealing and are designed for the mobile screen environment, and each measure presents as playful, even delightful - for example, Meter participants commonly reported very much enjoying interacting with the SAFESlider faces. These playful elements are intended to reduce respondent burden and fatigue, somewhat as (Dai, Rzeszotarski, Paritosh, & Chi, 2015) report that providing micro-distractions to micro-task crowd workers retains work quality and results in each worker continuing executing more tasks over time.

Based on this work, I anticipate tailoring measures to the participant to have several positive outcomes. First, that as response burden is down, motivation to report is up, and that the reporting interface is not exactly the same all the time, participants will maintain their reporting habits longer. Second, that there will be an improved sense that the system (or care provider) both recognizes the participant's uniqueness and preferences and tries to consider them. And finally, that satisficing in reporting will be far less common.

6.4 Patterns for developing novel EMA measures

It was my initial desire to present a strict recipe for developing novel EMA measures. Specifically, I began the Meter project by applying the same photo-based scale development as had been successful with PAM, hoping to both develop

a photo-based single-item measure for the self-report of pain intensity and to gather evidence that this development method generalized to all experiential assessment domains. The sub-method of generating a very large set of candidate photos and then winnowing it down through a hybrid process of expert review and aggregating crowd participant reports did seem effective. However, it very quickly became clear that this method would not work for pain.

The primary issue, as described in Chapter 5, was that the vast majority of both participants and informal testers really do not want to see photographs of other people in pain. Removing photographs containing people did not help, as the best abstract representations of pain intensity I was able to generate proved far too interpretable and yielded widely inconsistent ranking among testers. And then even many of those individuals who were okay with, or even enjoyed, reporting pain intensity with photos of people reported not wanting to report on their pain using photographs of people who did not look like them.

Pain appeared to be both deeply provocative and upsetting when seen in others, as well as an experience so personal, even private or intimate, that it attempting to use photos of (dissimilar) others to identify with was not possible. Only by using abstractions of faces (cartoons) was I able to both ameliorate the intensity of upset at seeing another in pain and present a representation of a human simple enough that there were really only commonalities between the cartoon face and the respondents (essentially, a head with eyes, a nose, a mouth, and facial expressions ranging from reasonably happy to being in severe pain). Even then, some participants found the high 'in pain' end of the scale a little upsetting, and others found the cartoon face so abstract that they could not see themselves in it.

Therefore, I have not uncovered a recipe anyone can follow for the creation of novel experiential assessment measures. But I do recommend a series of development patterns that I believe do hold over all domains and measure types, and will therefore be useful to developers of future novel self-report measures. I have already discussed the importance of recognizing and embracing the variability and individuality present in any reporting population, and of having your measures allow for individual differences in cognition and interaction styles and preferences. And, having described why very brief measures remain effective and accurate, I have recommended that EMA-style measures therefore be as momentary and as simple as possible. In this section, I share further development patterns for the design of novel EMA measures.

The domain matters Understanding the underlying construct is essential. How the construct is understood and how it is measured is the first and potentially most essential task. Along with this is the ways that people want to share their experiences with the construct - or not. For example, contrary to my expectations, many people living with chronic pain very much want to share their experiences with others, often in great detail. And in the field trials, while people did not report high '10/10' values, it became clear in follow-up interviews that this was more a function of self-presentation and condition management than unwillingness to report great pain. At first I thought this was a question of sampling bias (only those who wanted to talk about their pain would sign up for my study), but clinical practitioners report that this is not at all uncommon. Conversely, it was surprising to me at the time that not a single SESAME participant ever self-reported a single maximum '5/5' value for their own stress. In follow-up interviews, participants reported that they did experience a great

deal of stress, but simply would not ever choose to self-report stress levels in those moments.

Practice user-centered design throughout For an HCI researcher this point is not novel. But learning how to practice user-centered design with at-risk populations, including how to react when faced with participants in semi-structured interviews sharing very personal and upsetting narratives is essential. Without my training in peer counseling and active listening, and being able to lean on my lab-mates for advice, the interviews would have been extremely difficult for me to complete.

Develop with a multi-disciplinary team Developing EMA measures to assess subjective states for the support of daily wellbeing is a big goal requiring expertise in measurement and repeated assessments, interface design, usability and user experience, app development, and a lot of domain-specific knowledge. Effective assessment measures cannot be constructed without access to researchers and practitioners in each of these fields who will share their knowledge and criticism (Torsi et al., 2009).

Support meaningful resolution across scale ranges Regardless of what the recipients of the data may request, each participant has a range of the scale within which they desire greater or lesser reporting resolution. Two participants in Meter reported wanting a 0..5 point scale, but also with a 0.5 and 1.5 - some days their pain was not zero, but did not merit a 1 or a 2. Some regions of a scale will be utterly useless to some participants. For example, some participants reported that they could never imagine using the top half of the faces measure

in Meter; similarly, in passing judgments about movies online for the purposes of receiving recommendations, (Cosley et al., 2003) points out that users may not feel the need to distinguish between degrees of badness. This is not unlike placing a fish-eye lens on top of one part of a ruler, and simply covering up another region with blue tape and giving the blue region all the same value.

Integrate passive sensing with self-report measures As hardware costs decrease and we all carry or wear a growing number of (physiological) sensors at all times, as sensing algorithms become more refined and capable, and as widespread and pervasive self-tracking becomes fully a part of our culture (Lupton, 2014), in very few situations should naturalistic measurement be conducted without at least one form of passive sensing. Beyond providing fantastic data resolution, passive sensing of the user's environment and attentional and social behaviors can provide a rich picture of the contexts in which the measured construct states appear. While for each construct, and indeed for each user, the balance of passive sensing to self-report will vary, I recommend in general leveraging the findings from the SESAME project, summarised in Table 4.1.

In general, and this will become increasingly true for the above reasons, lean toward passive sensing whenever possible. There are three exceptions and one caveat. First, when the sensing subsystem is uncertain about its inferences, it is incredibly valuable to invite the user to provide self-report data that is then used to refine and improve the inference engine. Second, when the user is new to the system or protocol, inviting more frequent self-reports both helps the user understand what the resulting data means as well as improves user confidence in the inferences the system is making. Third, if the construct in question is particularly susceptible to positive reactivity effects then foregrounding the

construct through more frequent self-report is a very good idea - for example, in the self-management of diet (Baker & Kirschenbaum, 1993).

Perhaps the most exciting integration of passive sensing and self-report is in affording after-the-fact self-report and data annotation. In contextual recall, the system does not prompt for self-report in the moment as in true EMA-style measurement, but rather at the end of the day or the week. To ameliorate recall biases and other memory issues, this method captures meaningful contextual cues in the moment a self-report is desired (location, time-of-day, audio profile, or last scheduled activity for example) and then presents these cues to the user at a later moment in time when prompting for self-report.

Leverage novel wearables for naturalistic self-report The Keppi project indicates strong potential for novel devices and wearables for more naturalistic self-report. There are many scenarios in which using a phone for self-report is impossible, impractical, or socially inappropriate, but within those same scenarios tremendous value could be leveraged from meaningful self-report data. We imagined Keppi being worn around the neck or stored in a pocket and being used unobtrusively, without even removing it from a pocket. A similar pressure-based sensor could be wrapped around a car's steering wheel, or deployed within the arms of a dentist's chair. In group therapy, unobtrusively providing continuous or momentary self-report about particular feelings or comfort would provide tremendous value to the therapist, not just during the session but as part of their continued training. In settings like athletics, grasping the Keppi to capture moments of interest for later review, or to continuously record subjective performance indices, could be extremely valuable.

As commercial wearables become increasingly available, it will make sense to leverage them (or existing forms such as the Keppi) in research settings - for example, using the watch form factor. However, as in each of the case studies described above, carefully and iteratively designing for specific use-cases and settings can result in more effective and tailored measurement interfaces.

Choose system names carefully The presentation of a measure or system can have (negative) influences on the end-user before they have even begun reading the instructions. SESAME was originally titled 'StressSense'; one participant in particular reported that simply the system monicker appearing in his notification tray caused him additional stress! 'SickKids', 'FatBelt', and 'e-Ouch' all appear similarly problematic (Stinson et al., 2013; Pels, Kao, & Goel, 2014; Stinson, Petroz, et al., 2006). I suggest that the naming of app systems is at least as important as the text used to provide anchor or midpoint labels on a scale, or that to provide instructions on measure use.

6.5 Limitations and future work

None of the measures presented in this dissertation have been deployed outside the United States, and even within the United States all but the crowd workers involved in winnowing PAM, participants have not been fully representative of the demographic diversity in the nation. While there is some evidence that visual methods may be more generalizable across languages and cultures, culture plays a role in the acceptance of even simplistic non-text-based interfaces (Evers & Day, 1997). Certainly PAM's image set contains some images whose

content is US-centric (for example, Michelle Obama or the cowboy hat), and research indicates that meaning-making of photos and colors is culturally mediated (Peesapati, H.-C. Wang, & Cosley, 2010), although it is not clear whether this extends to parsing out affective content - something researchers are beginning to do automatically (Machajdik & Hanbury, 2010). There is also strong evidence the perception of pain varies across cultures (Free, 2002). Before deploying PAM or Meter outside the US, researchers should validate it against the locality's gold standard emotion or pain intensity measure.

Each user study was conducted only with self-selected participants who elected to participate in a study involving self-report and self-monitoring. But not everyone would want to self-track, or having self-tracked, enjoys it. Kaiton Williams has conducted a multi-year autoethnographic project, self-tracking weight-loss, diet, and exercise. While he appreciated the support the systems gave him in losing weight, Williams became aware of and uncomfortable with different ways of thinking about his body and lived experience:

> I do not enjoy contemplating my self as blood and sinews and electrical signals... I might have preferred to accomplish my self-transformation within broader measures, and I still long for that; to comprehend my body in longer and longer scales: seasons instead of hours, some other, coarser, property than calories.

There is evidence in the literature that self-tracking may not always be either beneficial or something a participant wants to do. For example, it may well be that inaccurate but positive beliefs about oneself or one's behaviors are actually advantageous (J. D. Brown & Dutton, 1995), and in the context of self-tracking cessation behaviors being reminded of positive experiences associated with bad

habits can be counterproductive (Ploderer, Smith, Howard, Pearce, & Borland, 2012). It is not clear how PAM, SESAME, Meter, Keppi, or other self-report measures might provide too much self-knowledge in these or related scenarios.

In a fantastic and timely article, Lupton provides the first sociological positioning of cultures increasingly engaged in self-tracking. She presents self-tracking as we know it today as "a profoundly social practice, both in terms of the enculturated meanings with which it is invested and the social encounters and social institutions that are part of the self-tracking phenomenon" (Lupton, 2014).

BIBLIOGRAPHY

Aaron, L. A., Turner, J. A., Mancl, L., Brister, H., & Sawchuk, C. N. (2005). Electronic diary assessment of pain-related variables: is reactivity a problem? *The Journal of Pain*, *6*(2), 107–115.

Adams, P., Baumer, E. P., & Gay, G. (2014). Staccato social support in mobile health applications. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems* (pp. 653–662). ACM.

Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voida, A., Gay, G., … Voida, S. (2014). Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild.

Allport, G. W. (1942). The use of personal documents in psychological science. *Social Science Research Council Bulletin*.

Almeida, D. M. (2005). Resilience and vulnerability to daily stressors assessed via diary methods. *Current Directions in Psychological Science*, *14*(2), 64–68.

Almeida, D. M., Wethington, E., & Kessler, R. C. (2002, March 1). The daily inventory of stressful events an interview-based approach for measuring daily stressors. *Assessment*, *9*(1), 41–55.

Alshehri, F. & Freeman, M. (2012). Methods for usability evaluations of mobile devices.

Anderson, M. (2015). 6 facts about americans and their smartphones [Pew research center]. Retrieved April 28, 2015, from http://www.pewresearch.org/fact-tank/2015/04/01/6-facts-about-americans-and-their-smartphones/

APA. (2011). Stress in america: our health at risk 2011. Retrieved May 12, 2014, from http://www.apa.org/news/press/releases/stress/2011/health-risk.aspx

APA. (2012). Creating a psychologically healthy workplace [APA center for organizational excellence]. Retrieved May 12, 2014, from http://www.apaexcellence.org/resources/creatingahealthyworkplace/

Ashburn, M. A. & Staats, P. S. (1999). Management of chronic pain. *The Lancet*, *353*(9167), 1865–1869.

Ayzenberg, Y., Hernandez Rivera, J., & Picard, R. (2012). FEEL: frequent EDA and event logging–a mobile social interaction stress monitoring system. In *CHI'12 extended abstracts on human factors in computing systems* (pp. 2357–2362). ACM.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010, March 1). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*(3), 372–374.

Baker, R. C. & Kirschenbaum, D. S. (1993). Self-monitoring may be necessary for successful weight control. *Behavior Therapy*, *24*(3), 377–394.

Baumer, E. P., Katz, S. J., Freeman, J. E., Adams, P., Gonzales, A. L., Pollak, J., ... Gay, G. K. (2012). Prescriptive persuasion and open-ended social awareness: expanding the design space of mobile health. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 475–484). ACM.

Beecher, H. K. (1959). *Measurement of subjective responses: quantitative effects of drugs.* New York: Oxford University Press.

Bieri, D., Reeve, R. A., Champion, G. D., Addicoat, L., & Ziegler, J. B. (1990). The faces pain scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. *Pain*, *41*(2), 139–150.

Blaskó, G. & Feiner, S. (2004). Single-handed interaction techniques for multiple pressure-sensitive strips. In *CHI'04 extended abstracts on human factors in computing systems* (pp. 1461–1464). ACM.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291.

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., & Bauer, G. (2011). Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 47–56). ACM.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual review of psychology*, *54*(1), 579–616.

Bolognese, J. A., Schnitzer, T. J., & Ehrich, E. W. (2003, July). Response relationship of VAS and likert scales in osteoarthritis efficacy measurement. *Osteoarthritis and Cartilage*, *11*(7), 499–507.

Boormans, E. M., Van Kesteren, P. J., Perez, R. S., Brölmann, H. A., & Zuurmond, W. W. (2009). Reliability of a continuous pain score meter: real time pain measurement. *Pain Practice*, *9*(2), 100–104.

Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59.

Brown, J. D. & Dutton, K. A. (1995). Truth and consequences: the costs and benefits of accurate self-knowledge. *Personality and Social Psychology Bulletin*, *21*(12), 1288–1296.

Brown, K. W. & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*, *84*(4), 822.

Brown, S. L., Nesse, R. M., Vinokur, A. D., & Smith, D. M. (2003). Providing social support may be more beneficial than receiving it results from a prospective study of mortality. *Psychological Science*, *14*(4), 320–327.

Buck, R. (1986). The psychology of emotion. In J. LeDoux & W. Hirst (Eds.), *Mind and brain: dialogues in cognitive neuroscience* (pp. 275–300). Cambridge, UK: Cambridge University Press.

Burisch, M. (1984). You don't always get what you pay for: measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality*, *18*(1), 81–98.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, *11*(4), 303–315.

Burns, T. (1954). The directions of activity and communication in a departmental executive group: a quantitative study in a british engineering factory with a self-recording technique. *Human relations*.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Carbonaro, N., Anania, G., Mura, G. D., Tesconi, M., Tognetti, A., Zupone, G., & De Rossi, D. (2011). Wearable biomonitoring system for stress management: a preliminary study on robust ECG signal processing. In *World of wireless, mobile and multimedia networks (WoWMoM), 2011 IEEE international symposium on a* (pp. 1–6). IEEE.

Cechanowicz, J., Irani, P., & Subramanian, S. (2007). Augmenting the mouse with pressure sensitive input. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1385–1394). ACM.

Center, P. R. (2014). Mobile technology fact sheet (pew research center) [Pew research center's internet & american life project]. Retrieved May 13, 2014, from http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/

Chalfen, R. (1989). Snapshot versions of life.

Champion, G. D., von Baeyer, C. L., Trieu, J. D. H., & Goodenough, B. (1997). Sydney animated facial expressions (SAFE). Randwick, NSW, Australia: Pain Research Unit, Sydney Children's Hospital.

Chang, K.-h., Fisher, D., Canny, J., & Hartmann, B. (2011). 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* (pp. 71–77). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Chatterjee, S. & Price, A. (2009). Healthy living with persuasive technologies: framework, issues, and challenges. *Journal of the American Medical Informatics Association*, *16*(2), 171–178.

Chu, G., Moscovich, T., & Balakrishnan, R. (2009). Haptic conviction widgets. In *Proceedings of graphics interface 2009* (pp. 207–210). GI '09. Toronto, Ont., Canada, Canada: Canadian Information Processing Society.

Cialdini, R. B. & James, L. (2009). *Influence: science and practice*. Pearson education Boston, MA.

Clark, W. C. & Yang, J. C. (1983). Applications of sensory decision theory to problems in laboratory and clinical pain. In *Pain measurement and assessment* (pp. 15–25). New York: Raven Press.

Clarkson, E. C., Patel, S. N., Pierce, J. S., & Abowd, G. D. (2006). Exploring continuous pressure input for mobile phones. In *UIST*.

Cleeland, C. S. [C. S.]. (1989). Measurement of pain by subjective report. *Advances in pain research and therapy*, *12*, 391–403.

Cleeland, C. S. [Charles S.], Nakamura, Y., Mendoza, T. R., Edwards, K. R., Douglas, J., & Serlin, R. C. (1996, October). Dimensions of the impact of cancer pain in a four country sample: new information from multidimensional scaling. *Pain*, *67*(2), 267–273.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, 385–396.

Cohen, S., Kessler, R. C., & Gordon, L. U. (1995). Strategies for measuring stress in studies of psychiatric and physical disorders. *Measuring stress: A guide for health and social scientists*, 3–26.

Cohen, S., Underwood, L. G., & Gottlieb, B. (2000). *Social support measurement and intervention: a guide for health and social scientists*. New York, NY: Oxford University Press.

Cohen, S. & Williamson, G. (1988). Perceived stress in a probability sample of the united states. In S. Spacapam & S. Oskamp (Eds.), *The social psychology of health: claremont symposium on applied social psychology*. Newbury Park, CA: Sage.

Cohen, S. & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological bulletin*, *98*(2), 310.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009, May 1). Experience sampling methods: a modern idiographic approach to personality research. *Social and personality psychology compass*, *3*(3), 292–313.

Consolvo, S., Everitt, K., Smith, I., & Landay, J. A. (2006). Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 457–466). ACM.

Cook, A. J., Roberts, D. A., Henderson, M. D., Van Winkle, L. C., Chastain, D. C., & Hamill-Ruth, R. J. (2004). Electronic pain questionnaires: a randomized, crossover comparison with paper questionnaires for chronic pain assessment. *Pain*, *110*(1), 310–317.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 585–592). ACM.

Crawford, J. R. & Henry, J. D. (2004). 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.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity*. New Jersey: Lawrence Erlbaum Associates.

Cruise, C. E., Broderick, J., Porter, L., Kaell, A., & Stone, A. A. (1996). Reactive effects of diary self-assessment in chronic pain patients. *Pain*, *67*(2), 253–258.

Csikszentmihalyi, M. [Mihaly] & Larson, R. [Reed]. (1987). Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease*, *175*(9), 526–536.

Dai, P., Rzeszotarski, J. M., Paritosh, P., & Chi, E. H. (2015). And now for something completely different: improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 628–638). CSCW '15. New York, NY, USA: ACM.

de la Vega, R., Roset, R., Castarlenas, E., Sánchez-Rodríguez, E., Solé, E., & Miró, J. (2014, October). Development and testing of painometer: a smartphone app to assess pain intensity. *The Journal of Pain*, *15*(10), 1001–1007.

DeLongis, A., Folkman, S., & Lazarus, R. S. (1988). The impact of daily stress on health and mood: psychological and social resources as mediators. *Journal of personality and social psychology*, *54*(3), 486.

Dey, A. K., Wac, K., Ferreira, D., Tassini, K., Hong, J.-H., & Ramos, J. (2011). Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on ubiquitous computing* (pp. 163–172). ACM.

Dietz, P. H., Eidelson, B., Westhues, J., & Bathiche, S. (2009). A practical pressure sensitive computer keyboard. In *Proceedings of the 22nd annual ACM symposium on user interface software and technology* (pp. 55–58). ACM.

Downie, W. W., Leatham, P. A., Rhind, V. M., Wright, V., Branco, J. A., & Anderson, J. A. (1978). Studies with pain rating scales. *Annals of the Rheumatic Diseases*, *37*(4), 378–381.

Egleston, B. L., Miller, S. M., & Meropol, N. J. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. *Statistics in medicine*, *30*(30), 3560–3572.

Eich, E., Reeves, J. L., Jaeger, B., & Graff-Radford, S. B. (1985). Memory for pain: relation between past and present pain intensity. *Pain*, *23*(4), 375–380.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3), 169–200.

Epstein, D. A., Borning, A., & Fogarty, J. (2013). Fine-grained sharing of sensed physical activity: a value sensitive approach. In *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing* (pp. 489–498). ACM.

Epstein, D. H., Willner-Reid, J., Vahabzadeh, M., Mezghanni, M., Lin, J.-L., & Preston, K. L. (2009). Real-time electronic diary reports of cue exposure and mood in the hours before cocaine and heroin craving and use. *Archives of General Psychiatry*, *66*(1), 88–94.

Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., & Shah, S. (2011). AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM conference on embedded networked sensor systems* (pp. 274–287). ACM.

Evans, B. (2014). Mobile is eating the world. Retrieved April 28, 2015, from http://a16z.com/2014/10/28/mobile-is-eating-the-world/

Evers, V. & Day, D. (1997). The role of culture in interface acceptance. In S. Howard, J. Hammond, & G. Lindgaard (Eds.), *Human-computer interaction INTERACT '97* (pp. 260–267). IFIP — The International Federation for Information Processing. Springer US.

Fabrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: techniques for measuring the unobservable. *Persuasion: Psychological insights and perspectives*, *2*, 17–40.

Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., & Estrin, D. (2010). Diversity in smartphone usage. In *Proceedings of the 8th international conference on mobile systems, applications, and services* (pp. 179–194). ACM.

Farrar, J. T., Young Jr, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*, *94*(2), 149–158.

Fechner, G. T. (1860). Elements of psychophysics, 1860. In *Readings in the history of psychology* (pp. 206–213). Century psychology series. East Norwalk, CT, US: Appleton-Century-Crofts.

Ferraz, M. B., Quaresma, M. R., Aquino, L. R., Atra, E., Tugwell, P., & Goldsmith, C. H. (1990). Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *The Journal of rheumatology*, *17*(8), 1022–1024.

Ferreira, P., Sanches, P., Höök, K., & Jaensson, T. (2008). License to chill!: how to empower users to cope with stress. In *Proceedings of the 5th nordic conference on human-computer interaction: building bridges* (pp. 123–132). ACM.

Ferrell, B. A., Ferrell, B. R., & Rivera, L. (1995). Pain in cognitively impaired nursing home patients. *Journal of pain and symptom management*, *10*(8), 591–598.

Fogg, B. J. & Eckles, D. (2007). *Mobile persuasion: 20 perspectives on the future of behavior change*. Stanford Captology Media Standford, CA.

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes. *Journal of personality and social psychology*, *50*(5), 992.

Frayling, C. (1993). Research in art and design. *Royal College of Art Research Papers*, *1*(1), 1–5.

Free, M. M. (2002, April). Cross-cultural conceptions of pain and pain control. *Proceedings (Baylor University. Medical Center)*, *15*(2), 143–145.

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on mobile systems, applications and services* (pp. 57–70). ACM.

Gaertner, J., Elsner, F., Pollmann-Dahmen, K., Radbruch, L., & Sabatowski, R. (2004, September). Electronic pain diary: a randomized crossover study. *Journal of Pain and Symptom Management*, *28*(3), 259–267.

Gaggioli, A., Pioggia, G., Tartarisco, G., Baldus, G., Corda, D., Cipresso, P., & Riva, G. (2013). A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing*, *17*(2), 241–251.

Gaston-Johansson, F. (1996, September). Measurement of pain: the psychometric properties of the pain-o-meter, a simple, inexpensive pain assessment tool that could change health care practices. *Journal of Pain and Symptom Management*, *12*(3), 172–181.

Gay, G. (2004). *Activity-centered design: an ecological approach to designing smart tools and usable systems*. Mit Press.

Gay, G., Pollak, J. P., Adams, P., & Leonard, J. P. (2011, March 1). Pilot study of aurora, a social, mobile-phone-based emotion sharing and recording system. *Journal of Diabetes Science and Technology*, *5*(2), 325–332.

Glasgow, R. E., Fisher, L., Strycker, L. A., Hessler, D., Toobert, D. J., King, D. K., & Jacobs, T. (2014). Minimal intervention needed for change: definition, use, and value for improving health and health research. *Translational behavioral medicine*, *4*(1), 26–33.

Glass, T. A. & McAtee, M. J. (2006). Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Social science & medicine*, *62*(7), 1650–1671.

Goodenough, B., Piira, T., von Baeyer, C. L., Chua, K., Wu, E., Trieu, J. D. H., & Champion, G. D. (2005). Comparing six self-report measures of pain intensity in children. *The Suffering Child*, *8*, 1–25.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Guillory, J., Chang, P., Henderson, C. R., Shengelia, R., Lama, S., Warmington, M., … Reid, M. C. (2015, January 6). Piloting a text message-based social support intervention for patients with chronic pain: establishing feasibility and preliminary efficacy. *The Clinical Journal of Pain*.

Harmon-Jones, E., Harmon-Jones, C., Abramson, L., & Peterson, C. K. (2009). PANAS positive activation is associated with anger. *Emotion*, *9*(2), 183.

Harper, F. M., Li, X., Chen, Y., & Konstan, J. A. (2005). An economic model of user rating in an online recommender system. In *Proceedings of the 10th international conference on user modeling* (pp. 207–216). Edinburgh, UK.

Hawker, G. A., Mian, S., Kendzerska, T., & French, M. (2011). Measures of adult pain: visual analog scale for pain (vas pain), numeric rating scale for pain (nrs pain), mcgill pain questionnaire (mpq), short-form mcgill pain questionnaire (sf-mpq), chronic pain grade scale (cpgs), short form-36 bod-

ily pain scale (sf-36 bps), and measure of intermittent and constant osteoarthritis pain (icoap). *Arthritis care & research*, *63*, S240–S252.

Haythornthwaite, J. A., Menefee, L. A., Heinberg, L. J., & Clark, M. R. (1998). Pain coping strategies predict perceived control over pain. *Pain*, *77*(1), 33–39.

Healey, J. A. & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, *6*(2), 156–166.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. [Mihaly]. (2007). *Experience sampling method: measuring the quality of everyday life* (1 edition). Thousand Oaks, Calif: SAGE Publications, Inc.

Helme, R. D. & Gibson, S. J. (2001). The epidemiology of pain in elderly people. *Clinics in geriatric medicine*, *17*(3), 417–431.

Heo, J., Ham, D.-H., Park, S., Song, C., & Yoon, W. C. (2009). A framework for evaluating the usability of mobile phones based on multi-level, hierarchical model of usability factors. *Interacting with Computers*, *21*(4), 263–275.

Heo, S. & Lee, G. (2011). Force gestures: augmenting touch screen gestures with normal and tangential forces. In *Proceedings of the 24th annual ACM symposium on user interface software and technology* (pp. 621–626). UIST '11. New York, NY, USA: ACM.

Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: sensing stress of computer users. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems* (pp. 51–60). ACM.

Hicks, C. L., von Baeyer, C. L., Spafford, P. A., van Korlaar, I., & Goodenough, B. (2001). The faces pain scale–revised: toward a common metric in pediatric pain measurement. *Pain*, *93*(2), 173–183.

Hinrichs, J. R. (1964). Communications activity of industrial research personnel. *Personnel Psychology*, *17*(2), 193–206.

Hochschild, A. R. & Machung, A. (1989). *The second shift: working parents and the revolution at home*. Viking New York.

Hoggan, E., Trendafilov, D., Ahmaniemi, T., & Raisamo, R. (2011). Squeeze vs. tilt: a comparative study using continuous tactile feedback. In *CHI '11 extended abstracts on human factors in computing systems* (pp. 1309–1314). CHI EA '11. New York, NY, USA: ACM.

Humphreys, L., Gill, P., Krishnamurthy, B., & Newbury, E. (2013). Historicizing new media: a content analysis of twitter. *Journal of Communication*, *63*(3), 413–431.

Hurlburt, R. T. (1979). Random sampling of cognitions and behavior. *Journal of Research in Personality*, *13*(1), 103–111.

Huskisson, E. C. [E. C]. (1974). Measurement of pain. *The Lancet*. Originally published as Volume 2, Issue 7889, *304*(7889), 1127–1131.

Inc, A. (2014, November 1). *iOS human interface guidelines*.

Inc, G. (2014, November 1). *Android design*.

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., & Bao, L. (2003). A context-aware experience sampling tool. In *CHI'03 extended abstracts on human factors in computing systems* (pp. 972–973). ACM.

Isomursu, M., Tähti, M., Väinämö, S., & Kuutti, K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies*, *65*(4), 404–418.

Jamison, R. N., Gracely, R. H., Raymond, S. A., Levine, J. G., Marino, B., Herrmann, T. J., … Katz, N. P. (2002). Comparative study of electronic vs. pa-

per VAS ratings: a randomized, crossover trial using healthy volunteers. *Pain*, *99*(1), 341–347.

Jibb, L. A. [Lindsay A], Stevens, B. J., Nathan, P. C., Seto, E., Cafazzo, J. A., & Stinson, J. N. (2014). A smartphone-based pain management app for adolescents with cancer: establishing system requirements and a pain care algorithm based on literature review, interviews, and consensus. *JMIR Research Protocols*, *3*(1), e15.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 154–161). ACM.

Johannes, C. B., Le, T. K., Zhou, X., Johnston, J. A., & Dworkin, R. H. (2010). The prevalence of chronic pain in united states adults: results of an internet-based survey. *The Journal of Pain*, *11*(11), 1230–1239.

Johnson, C. (2005). Measuring pain. visual analog scale versus numeric pain scale: what is the difference? *Journal of Chiropractic Medicine*, *4*(1), 43–44.

Jonas, W. (2006). Research through DESIGN through research: a problem statement and a conceptual sketch. In *Design research society international conference, k. friedman, t. love, e. côrte-real, and c. rust eds., lisbon*.

Jonas, W. (2007). Research through DESIGN through research: a cybernetic model of designing design foundations. *Kybernetes*, *36*(9), 1362–1380.

Joyce, C. R. B., Zutshi, D. W., Hrubes, V., & Mason, R. M. (1975). Comparison of fixed interval and visual analogue scales for rating chronic pain. *European Journal of Clinical Pharmacology*, *8*(6), 415–420.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, ...M. Papka (Eds.), *Advances in visual computing* (7432, pp. 368–377). Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Kang, B., Lee, S., Oh, A., Kang, S., Hwang, I., & Song, J. (2015). Towards understanding relational orientation: attachment theory and facebook activities. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 1404–1415). CSCW '15. New York, NY, USA: ACM.

Kay, M., Choe, E. K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., & Kientz, J. A. (2012). Lullaby: a capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 226–234). ACM.

Keele, K. D. (1948, July 3). The pain chart. *The Lancet*. Originally published as Volume 2, Issue 6514, *252*(6514), 6–8.

Kessler, R. & Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. *American Journal of Preventive Medicine*, *40*(6), 637–644.

Kim, Y., Gay, G., Reynolds, L., & Hong, H. (2015). Mood.cloud: data as art. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems* (pp. 347–350). CHI EA '15. New York, NY, USA: ACM.

Klasnja, P., Harrison, B. L., LeGrand, L., LaMarca, A., Froehlich, J., & Hudson, S. E. (2008). Using wearable sensors and real time inference to understand

human recall of routine activities. In *Proceedings of the 10th international conference on ubiquitous computing* (pp. 154–163). ACM.

Klasnja, P. & Pratt, W. (2012). Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of biomedical informatics*, *45*(1), 184–198.

Klinger, E. (1978). Dimensions of thought and imagery in normal waking states. *Journal of Altered States of Consciousness*.

Kohl, A., Rief, W., & Glombiewski, J. A. (2013). Acceptance, cognitive restructuring, and distraction as coping strategies for acute pain. *The journal of pain*, *14*(3), 305–315.

Korotitsch, W. J. & Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment*, *11*(4), 415.

Kristensson, P. O. & Vertanen, K. (2014). The inviscid text entry rate and its application as a grand goal for mobile text entry. *Mobile HCI*.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*(3), 213–236.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*(1), 537–567.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In *Handbook of attitudes and attitude change* (pp. 21–76). Mahwah, NJ: Erlbaum.

Krosnick, J. A. & Presser, S. (2010). Question and questionnaire design. *Handbook of survey research*, *2*, 263–314.

Lane, N. D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., … Campbell, A. (2011). Bewell: a smartphone application to monitor, model and pro-

mote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare* (pp. 23–26).

Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, *50*(5), 372.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological psychiatry*, *44*(12), 1248–1263.

Larson, R. [R.] & Csikszentmihalyi, M. [M.]. (1983). The experience sampling method. In H. Reis (Ed.), *New directions for naturalistic methods in the behavioral sciences* (pp. 41–56). San Francisco: Jossey-Bass.

Laurans, G., Desmet, P., & Hekkert, P. (2009). The emotion slider: a self-report device for the continuous measurement of emotion. In *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on* (pp. 1–6). IEEE.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 719–728). ACM.

Lazarus, R. S. & Lazarus, B. N. (1994). *Passion and reason: making sense of our emotions*. Oxford University Press New York.

Lewinsohn, P. M. & Graf, M. (1973). Pleasant activities and depression. *Journal of consulting and clinical psychology*, *41*(2), 261.

Li, I., Dey, A. K., & Forlizzi, J. (2011). Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on ubiquitous computing* (pp. 405–414). ACM.

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 557–566). ACM.

Lindsley, O. R. (1968). A reliable wrist counter for recording behavior rates. *Journal of Applied Behavior Analysis*, *1*(1), 77–78.

Lookout. (2012). *Mobile mindset study*.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., … Choudhury, T. (2012). StressSense: detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 351–360). ACM.

Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., & Campbell, A. T. (2010). The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM conference on embedded networked sensor systems* (pp. 71–84). ACM.

Lupton, D. (2014). Self-tracking cultures: towards a sociology of personal informatics.

Machajdik, J. & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on multimedia* (pp. 83–92). ACM.

Mack, R. L. & Nielsen, J. (1994). *Usability inspection methods*. Wiley & Sons.

Mayer, J. D., DiPaolo, M., & Salovey, P. (1990). Perceiving affective content in ambiguous visual stimuli: a component of emotional intelligence. *Journal of personality assessment*, *54*(3), 772–781.

McCaffery, M. (1979). *Nursing management of the patient with pain*. Philadelphia, USA: JB Lippincot Company.

McDowell, I. (2006). *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press.

McDuff, D., Karlson, A., Kapoor, A., Roseway, A., & Czerwinski, M. (2012). AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 849–858). ACM.

McHorney, C. A., WARE JOHNE, J., & ANASTASIAE, R. (1993). The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical care*, *31*(3), 247–263.

Mehrabian, A. & Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Melzack, R. (1975). The McGill pain questionnaire: major properties and scoring methods. *Pain*, *1*(3), 277–299.

Melzack, R. (1987). The short-form McGill pain questionnaire. *Pain*, *30*(2), 191–197.

Melzack, R. & Casey, K. L. (1968). Sensory, motivational and central control determinants of pain: a new conceptual model. In *The skin senses* (pp. 423–443).

Melzack, R. & Torgerson, W. S. (1971). On the language of pain. *Anesthesiology*, *34*(1), 50–59.

Menon, G. (1997). Are the parts better than the whole? the effects of decompositional questions on judgments of frequent behaviors. *Journal of Marketing Research*, 335–346.

Meschtscherjakov, A., Reitberger, W., & Tscheligi, M. (2010). MAESTRO: orchestrating user behavior driven and context triggered experience sampling.

In *Proceedings of the 7th international conference on methods and techniques in behavioral research* (p. 29). ACM.

Meschtscherjakov, A., Weiss, A. [Astrid], & Scherndl, T. (2009). Utilizing emoticons on mobile devices within ESM studies to measure emotions in the field. *Proc. MME in conjunction with MobileHCI*, *9*.

Messick, S. (1995). *Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. Educational Testing Service. Princeton NJ USA.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237.

Miró, J., Huguet, A., Nieto, R., Paredes, S., & Baos, J. (2005, November). Evaluation of reliability, validity, and preference for a pain intensity scale for use with the elderly. *The Journal of Pain*, *6*(11), 727–735.

Mizobuchi, S., Terasaki, S., Keski-Jaskari, T., Nousiainen, J., Ryynanen, M., & Silfverberg, M. (2005). Making an impression: force-controlled pen input for handheld devices. In *CHI'05 extended abstracts on human factors in computing systems* (pp. 1661–1664). ACM.

Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., & Deleeuw, W. (2010). Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. *Journal of Medical Internet Research*, *12*(2).

Muller, M. J. & Kuhn, S. (1993, June). Participatory design. *Commun. ACM*, *36*(6), 24–28.

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological medicine*, *39*(9), 1533.

Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on computer supported cooperative work* (pp. 189–192). ACM.

Nainis, N., Paice, J. A., Ratner, J., Wirth, J. H., Lai, J., & Shott, S. (2006, February). Relieving symptoms in cancer: innovative use of art therapy. *Journal of Pain and Symptom Management*, *31*(2), 162–169.

Naz, K. & Helen, H. (2004). Color-emotion associations: past experience and personal preference. In *AIC 2004 color and paints, interim meeting of the international color association, proceedings* (Vol. 5, p. 31). Jose Luis Caivano.

Newman, M. W., Lauterbach, D., Munson, S. A., Resnick, P., & Morris, M. E. (2011). It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on computer supported cooperative work* (pp. 341–350). ACM.

Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J., & Stenild, S. (2006). It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In *Proceedings of the 4th nordic conference on human-computer interaction: changing roles* (pp. 272–280). ACM.

Niven, C. A. & Brodie, E. E. (1996). Memory for labor pain: context and quality. *Pain*, *64*(2), 387–392.

Norman, D. (2013, November 5). *The design of everyday things: revised and expanded edition* (Revised Edition edition). Basic Books.

Norman, D. & Nielsen, J. (2010). Gestural interfaces: a step backward in usability. *interactions*, *17*(5), 46–49.

Oulasvirta, A., Kurvinen, E., & Kankainen, T. (2003). Understanding contexts by being there: case studies in bodystorming. *Personal and Ubiquitous Computing*, *7*(2), 125–134.

Oulasvirta, A., Rattenbury, T., Ma, L., & Raita, E. (2012). Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, *16*(1), 105–114.

Oulasvirta, A., Wahlström, M., & Anders Ericsson, K. (2011, March). What does it mean to be good at using a mobile device? an investigation of three levels of experience and skill. *International Journal of Human-Computer Studies*, *69*(3), 155–169.

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. *PLoS ONE*, *8*(6), e68126.

Pearlin, L. I., Menaghan, E. G., Lieberman, M. A., & Mullan, J. T. (1981). The stress process. *Journal of health and social behavior*.

Peesapati, S. T., Wang, H.-C., & Cosley, D. (2010). Intercultural human-photo encounters: how cultural similarity affects perceiving and tagging photographs. In *Proceedings of the 3rd international conference on intercultural collaboration* (pp. 203–206). ACM.

Pels, T., Kao, C., & Goel, S. (2014). FatBelt: motivating behavior change through isomorphic feedback. In *Proceedings of the adjunct publication of the 27th annual ACM symposium on user interface software and technology* (pp. 123–124). ACM.

Phillips-Caesar, E. G., Winston, G., Peterson, J. C., Wansink, B., Devine, C. M., Kanna, B., … Charlson, M. E. (2015). Small changes and lasting effects

(SCALE) trial: the formation of a weight loss behavioral intervention using EVOLVE. *Contemporary Clinical Trials*, *41*, 118–128.

Picard, R. W. (2000, July 31). *Affective computing* (1st edition). Cambridge, Mass.: The MIT Press.

Picard, R. W. & Healey, J. (1997). Affective wearables. *Personal Technologies*, *1*(4), 231–240.

Picard, R. W. & Liu, K. K. (2007). Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress. *International Journal of Human-Computer Studies*, *65*(4), 361–375.

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'absi, M., ... Scott, M. (2011). Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Information processing in sensor networks (IPSN), 2011 10th international conference on* (pp. 97–108). IEEE.

Ploderer, B., Smith, W., Howard, S., Pearce, J., & Borland, R. (2012). Things you don't want to know about yourself: ambivalence about tracking and sharing personal information for behaviour change. In *Proceedings of the 24th australian computer-human interaction conference* (pp. 489–492). OzCHI '12. New York, NY, USA: ACM.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, *1*.

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on*, *57*(5), 1243–1252.

Pollak, J. P., Adams, P., & Gay, G. (2011). PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 725–734). ACM.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715–734.

Pressman, S. D. & Cohen, S. (2005). Does positive affect influence health? *Psychological bulletin*, *131*(6), 925.

Price, D. D., Bush, F. M., Long, S., & Harkins, S. W. (1994, February). A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. *Pain*, *56*(2), 217–226.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983, September). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45–56.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, *41*(1), 203–212.

Rammstedt, B. & Rammsayer, T. H. (2002). Gender differences in self-estimated intelligence and their relation to gender-role orientation. *European Journal of Personality*, *16*(5), 369–382.

Ramos, G., Boulos, M., & Balakrishnan, R. (2004). Pressure widgets. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 487–494). ACM.

Reid, S. C., Kauer, S. D., Dudgeon, P., Sanci, L. A., Shrier, L. A., & Patton, G. C. (2008, November 14). A mobile phone program to track young people's

experiences of mood, stress and coping. *Social Psychiatry and Psychiatric Epidemiology*, *44*(6), 501–507.

Reips, U.-D. & Funke, F. (2008). Interval-level measurement with visual analogue scales in internet-based research: VAS generator. *Behavior Research Methods*, *40*(3), 699–704.

Reis, H. T. & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222).

Reis, H. T. & Wheeler, L. (1991). Studying social interaction with the rochester interaction record. *Advances in experimental social psychology*, *24*, 269–318.

Richardson, J. E. & Reid, M. C. (2013). The promises and pitfalls of leveraging mobile health technology for pain care. *Pain Medicine*, *14*(11), 1621–1626.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: construct validation of a single-item measure and the rosenberg self-esteem scale. *Personality and social psychology bulletin*, *27*(2), 151–161.

Robinson, J. P. (1977). *How americans use time: a social-psychological analysis of everyday behavior*. Praeger New York.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A., Weiss, A. [Anna], & Mendelsohn, G. A. (1989). Affect grid: a single-item scale of pleasure and arousal. *Journal of Personality and Social psychology*, *57*(3), 493.

Ryan, T. & Xenos, S. (2011). Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, *27*(5), 1658–1664.

Saarni, C. (1999). *The development of emotional competence*. Guilford Press.

Saffer, D. (2013). *Microinteractions: designing with details*. " O'Reilly Media, Inc."

Sanches, P., Höök, K., Vaara, E., Weymann, C., Bylund, M., Ferreira, P., ... Sjölinder, M. (2010). Mind the body!: designing a mobile stress management application encouraging personal reflection. In *Proceedings of the 8th ACM conference on designing interactive systems* (pp. 47–56). ACM.

Sanders, E. B.-N. & Stappers, P. J. (2008, March 1). Co-creation and the new landscapes of design. *CoDesign*, *4*(1), 5–18.

Schaeffer, N. C. & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 65–88.

Scherer, K. R. (1986). Voice, stress, and emotion. In M. Appley & R. Trumbull (Eds.), *Dynamics of stress: physiological, psychological, and social perspectives* (pp. 157–179). Springer.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, *44*(4), 695–729.

Scherer, K. R. & Ekman, P. (2014). *Approaches to emotion*. Psychology Press.

Schlundt, D. G., Johnson, W. G., & Jarrell, M. P. (1985). A naturalistic functional analysis of eating behavior in bulimia and obesity. *Advances in Behaviour Research and Therapy*, *7*(3), 149–162.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, *54*(2), 93.

Scollon, C. N., Prieto, C.-K., & Diener, E. (2009). Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being* (pp. 157–180). Springer.

Scott, J. & Huskisson, E. C. [E. C.]. (1976, June). Graphic representation of pain. *Pain*, *2*(2), 175–184.

Scott, J. & Huskisson, E. C. [E. C.]. (1979). Vertical or horizontal visual analogue scales. *Annals of the rheumatic diseases*, *38*(6), 560.

Sengers, P., Boehner, K., Mateas, M., & Gay, G. (2008). The disenchantment of affect. *Personal and Ubiquitous Computing*, *12*(5), 347–358.

Serlin, R. C., Mendoza, T. R., Nakamura, Y., Edwards, K. R., & Cleeland, C. S. (1995, May). When is cancer pain mild, moderate or severe? grading pain severity by its interference with function. *Pain*, *61*(2), 277–284.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, *52*(6), 1061.

Shi, K., Irani, P., Gustafson, S., & Subramanian, S. (2008). PressureFish: a method to improve control of discrete pressure-based input. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1295–1298). CHI '08. New York, NY, USA: ACM.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol. 4*, 1–32.

Sparling, E. I. & Sen, S. (2011). Rating: how difficult is it? In *Proceedings of the fifth ACM conference on recommender systems* (pp. 149–156). RecSys '11. New York, NY, USA: ACM.

Srinivasan, M. A. & Chen, J.-s. (1993). Human performance in controlling normal forces of contact with rigid objects. *ASME DYN SYST CONTROL DIV PUBL DSC, ASME, NEW YORK, NY,(USA), 1993, 49*, 119–125.

Stephenson, N. L. & Herman, J. (2000, August). Pain measurement: a comparison using horizontal and vertical visual analogue scales. *Applied Nursing Research*, *13*(3), 157–158.

Stewart, C., Rohs, M., Kratz, S., & Essl, G. (2010). Characteristics of pressure-based input for mobile devices. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 801–810). ACM.

Stinson, J. N., Jibb, L. A. [Lindsay A.], Nguyen, C., Nathan, P. C., Maloney, A. M., Dupuis, L. L., … Strahlendorf, C., et al. (2013). Development and testing of a multidimensional iPhone pain assessment application for adolescents with cancer. *Journal of medical Internet research*, *15*(3).

Stinson, J. N., Kavanagh, T., Yamada, J., Gill, N., & Stevens, B. (2006). Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents. *Pain*, *125*(1), 143–157.

Stinson, J. N., Petroz, G. C., Tait, G., Feldman, B. M., Streiner, D., McGrath, P. J., & Stevens, B. J. (2006). E-ouch: usability testing of an electronic chronic pain diary for adolescents with arthritis. *The Clinical journal of pain*, *22*(3), 295–305.

Stone, A. A. & Broderick, J. E. (2007). Real-time data collection for pain: appraisal and current status. *Pain Medicine*, *8*, S85–S93.

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. *Pain*, *104*(1), 343–351.

Stone, A. A. & Shiffman, S. (2002). Capturing momentary, self-report data: a proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236–243.

Stuppy, D. J. (1998, May). The faces pain scale: reliability and validity with mature adults. *Applied Nursing Research*, *11*(2), 84–89.

Sundstrom, P., Stahl, A., & Hook, K. (2007). In situ informants exploring an emotional mobile messaging system in their everyday practice. *International journal of human-computer studies*, *65*(4), 388–403.

Suppes, P. & Zinnes, J. L. (1963). Basic measurement theory. *Handbook of mathematical psychology*, *1*(1).

Swearingen, K. & Sinha, R. (2002). Interaction design for recommender systems. In *Designing interactive systems* (Vol. 6, pp. 312–334). Citeseer.

Taylor, F. W. (1911). *The principles of scientific management*. New York, USA: Harper and Brothers.

Taylor, S. E., Welch, W. T., Kim, H. S., & Sherman, D. K. (2007). Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological Science*, *18*(9), 831–837.

Thoits, P. A. (1995). Stress, coping, and social support processes: where are we? what next? *Journal of health and social behavior*, 53–79.

Tian, Y.-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In *Handbook of face recognition* (pp. 247–275). Springer New York.

Tomkins, S. S. & Carter, R. (1964). What and where are the primary affects? some evidence for a theory. *Perceptual and motor skills*, *18*(1), 119–158.

Torsi, S., Nasr, N., Wright, P. C., Mawson, S. J., & Mountain, G. A. (2009). User-centered design for supporting the self-management of chronic illnesses: an interdisciplinary approach. In *Proceedings of the 2nd international conference on PErvasive technologies related to assistive environments* (p. 43). ACM.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Truong, K. N., Shihipar, T., & Wigdor, D. J. (2014). Slide to x: unlocking the potential of smartphone unlocking. In *Proceedings of the 32nd annual ACM*

conference on human factors in computing systems (pp. 3635–3644). CHI '14. New York, NY, USA: ACM.

Turk, D. C. & Melzack, R. (2011). *Handbook of pain assessment, third edition*. Guilford Press.

Vaish, R., Wyngarden, K., Chen, J., Cheung, B., & Bernstein, M. S. (2014). Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of the 32nd annual ACM conference on human factors in computing systems* (pp. 3645–3654). ACM.

Vannette, D. L. & Krosnick, J. A. (2014). Answering questions: a comparison of survey satisficing and mindlessness. *The Wiley Blackwell Handbook of Mindfulness*, *1*, 312.

Vastenburg, M., Romero Herrera, N., Van Bel, D., & Desmet, P. (2011). PMRI: development of a pictorial mood reporting instrument. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 2155–2160). ACM.

Venta, L., Isomursu, M., Ahtinen, A., & Ramiah, S. (2008). "my phone is a part of my soul"–how people bond with their mobile phones. In *Mobile ubiquitous computing, systems, services and technologies, 2008. UBICOMM'08. the second international conference on* (pp. 311–317). IEEE.

Verbrugge, L. M. (1980). Health diaries. *Medical care*, *18*(1), 73–95.

Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., ... Wang, X. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on*, *12*(7), 682–691.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063.

Wethington, E. (2005, May). An overview of the life course perspective: implications for health and nutrition. *Journal of Nutrition Education and Behavior*, *37*(3), 115–120.

Wethington, E. & Kessler, R. C. (1986). Perceived support, received support, and adjustment to stressful life events. *Journal of Health and Social behavior*, 78–89.

Wheeler, L. & Reis, H. T. (1991). Self-recording of everyday life events: origins, types, and uses. *Journal of personality*, *59*(3), 339–354.

Williams, A. C. d. C., Davies, H. T. O., & Chadury, Y. (2000). Simple pain rating scales hide complex idiosyncratic meanings. *Pain*, *85*(3), 457–463.

Wilson, G., Brewster, S. A., Halvey, M., Crossan, A., & Stewart, C. (2011). The effects of walking, feedback and control method on pressure-based interaction. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 147–156). MobileHCI '11. New York, NY, USA: ACM.

Wilson, T. D. & Dunn, E. W. (2004). Self-knowledge: its limits, value, and potential for improvement. *Psychology*, *55*.

Wood, C. & von Baeyer, C. (2011). Electronic and paper versions of a faces pain intensity scale: concordance and preference in hospitalized children. *BMC pediatrics*, *11*, 87.

Woodforde, J. M. & Merskey, H. (1971). Correlation between verbal scale and visual analogue scale and pressure algometer. *Journal of Psychosomatic Research*, *16*, 173.

Woolf, C. J. (2004). Pain: moving from symptom control toward mechanism-specific pharmacologic management. *Annals of internal medicine*, *140*(6), 441–451.

Wundt, W. (1874). Fundamentals of physiological psychology.

Younger, J., McCue, R., & Mackey, S. (2009). Pain outcomes: a brief review of instruments and techniques. *Current pain and headache reports*, *13*(1), 39–43.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 493–502). ACM.

Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., … Steinfeld, A. (2011). Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1677–1686). CHI '11. New York, NY, USA: ACM.